INVERSE PROBLEMS IN ENGINEERING

Theory and Practice

Edited by: Helcio R. B. Orlande

Rio de Janeiro, 2002

e-papers

4th INTERNATIONAL CONFERENCE ON INVERSE PROBLEMS IN ENGINEERING: THEORY AND PRACTICE

May 26 –31, 2002

Angra dos Reis, Brazil

Conference Chair:
Helcio R. B. Orlande
Federal University of Rio de Janeiro, Brazil

Conference Co-chairs:
Fernando Manuel Ramos
National Institute of Space Research, Brazil

Ashley F. Emery
University of Washington at Seattle, USA

Martin Raynaud
Institute National de Sciences Apliquées de Lyon, France

Shiro Kubo
Osaka University, Japan

United Engineering Foundation, Inc.

ORGANIZING COMMITTEE

Chair:

H. R. B. Orlande *(Brazil)*

Co-chairs:

F. M. Ramos *(Brazil)*
A. F. Emery *(USA)*
M. Raynaud *(France)*
S. Kubo *(Japan)*

F. Landis *(ex officio)*
K. A. Woodbury *(USA)*
J-E. Nordtvedt *(Norway)*
P. P. B. de Oliveira *(Brazil)*
G. Guimarães *(Brazil)*
N. J. Ruperti Jr. *(Brazil)*

SCIENTIFIC COMMITTEE

Honorary Members:

J. V. Beck *(USA)*
O. M. Alifanov *(Russia)*
M. N. Ozisik *(USA)*

A. Yagola *(Russia)*
A. J. Kassab *(USA)*
A. Denisov *(Russia)*
A. T. Watson *(USA)*
C. J. S. Alves *(Portugal)*
C. LeNiliot *(France)*
D. Lesnic *(UK)*
D. Delaunay *(France)*
E. Massoni *(France)*
G. Chavent *(France)*
H. F. C. Velho *(Brazil)*
H. D. Bui *(France)*
H. Sobieczky *(Germany)*
I. Egorov *(Russia)*
J. Howell *(USA)*
K. Dowding *(USA)*
L. C. Santos *(Brazil)*
M. Bonnet *(France)*
N-Z. Sun *(USA)*
N. Roberty *(Brazil)*
P. Husbands *(UK)*
W. S. Kim *(South Korea)*
A. El Badia *(France)*
A. Nenarokomov *(Russia)*
A. J. Silva Neto *(Brazil)*
B. Blackwell *(USA)*
C-H. Huang *(Taiwan)*
D. Maillet *(France)*
D. Petit *(France)*
D. Murio *(USA)*
E. A. Artioukhine *(France)*
G. S. Dulikravich *(USA)*
H. Busby *(USA)*
H. Engl *(Austria)*
H. Reinhardt *(Germany)*
J. C. Batsale *(France)*
K. Onishi *(Japan)*
L. Barichello *(Brazil)*
M. Bertero *(Italy)*
M. Tanaka *(Japan)*
N. Zabaras *(USA)*
N. McCormick *(USA)*
T. Burczynski *(Poland)*
Y. Jarny *(France)*

FOREWORD

This book contains the papers presented in the **4th International Conference on Inverse Problems in Engineering**: **Theory and Practice.** This conference is organized under the auspices of United Engineering Foundation and is held in a three-year cycle. Previous versions took place in Palm Coast, Florida, in 1993; in Le Croisic, France, in 1996; and in Port Ludlow, Washington State, in 1999. The series of **International Conferences on Inverse Problems in Engineering**: **Theory and Practice** finds its roots in the informal seminars organized by Prof. James V. Beck at Michigan State University, which were initiated in 80's.

The **4th International Conference on Inverse Problems in Engineering**: **Theory and Practice** was held during May 26 – 31, 2002, in the beautiful Hotel Portobello Resort & Safari, located near the city of Rio de Janeiro. The resort provided a unique atmosphere for 99 conference participants, from 21 different countries, to present their most recent research results and for the technical discussion of their findings. The **4th International Conference on Inverse Problems in Engineering**: **Theory and Practice** was co-promoted by the Brazilian Society of Mechanical Sciences (ABCM), the Brazilian Society of Computational and Applied Mathematics (SBMAC), and by COPPE, which is the graduate school in engineering of the Federal University of Rio de Janeiro (UFRJ). It was co-sponsored by the following agencies of the Brazilian Government: CNPq, from the Ministry of Science and Technology; CAPES, from the Ministry of Education and Culture; and the National Oil Agency (ANP), from the Ministry of Mines and Energy.

The **4th International Conference on Inverse Problems in Engineering**: **Theory and Practice** counted with 159 submitted abstracts, resulting on 104 accepted papers. A total of 98 papers were scheduled for presentation in the conference, distributed in 25 oral sessions and in 1 poster session. Invited keynote lecturers were presented by Prof. A. Yagola (Russia), Prof. G. Chavent (France), Prof. O. Alifanov (Russia), Prof. Y. Jarny (France) and Prof. N-Z. Sun (USA). Prof. K. Woodbury (USA) and Prof. B. Blackwell (USA) were invited to give tutorial sessions. I would like to express my gratitude to the members of the organizing and scientific committees for playing a fundamental role towards the success of the conference, as well as to the invited speakers for kindly accepting my invitation to share with the participants their knowledge on important subjects on the inverse problems field. Because of the large number of papers submitted, several other reviewers were invited to give their contributions to the conference by evaluating papers, in addition to the members of the organizing and scientific committees. They include Prof. J. P. Kaipio (Finland), Prof. L. Olson (USA), Prof. A. Haji-Sheikh (USA), Prof. J. G. Berryman (USA), Prof. R. Y. Qassim (Brazil), Prof. B. Dennis (Japan), Prof. G. R. Liu (Taiwan), Prof. H. Telega (Poland), Prof. V. Steffen Jr. (Brazil), Prof. F. Rochinha (Brazil) and Prof. M. D. Mikhailov (Brazil).

It was a great honor for Brazil to host the **4th International Conference on Inverse Problems in Engineering**: **Theory and Practice** and, personally for

myself, to be its chairman. The next conference shall take place in the United Kingdom in 2005 and Prof. Daniel Lesnic has agreed to lead the organizing committee for that event.

Helcio R. B. Orlande
Rio de Janeiro, Brazil

KEYNOTE LECTURERS

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

MATHEMATICAL AND EXPERIMENTAL SIMULATION IN DESIGNING AND TESTING HEAT-LOADED ENGINEERING OBJECTS

Oleg M. Alifanov

Aerospace School
Moscow Aviation Institute, Moscow, Russia
alf@cosmos.rcnet.ru

ABSTRACT

The paper deals with the inverse methodology in mathematical modeling and experimental studies of heat transfer processes while designing and testing thermally loaded structures. Among the problems under consideration the main are the investigation of thermophysical characteristics of materials, the transient heat measurements and identification of thermal processes. The identification of mathematical models of physical processes should be performed in such a way as to provide a correct consideration of physical laws and general rules in combination with the inverse methods for parameter estimation, test of hypothesis and model validation for adequacy. In particular, the results of studies on the construction and verification of models are presented to describe the process of heat propagation in high-porous fibrous materials and thermal protection structures made from them. The investigation of heat transfer in a heterogeneous gas flow is citied as another useful application of this technique. The experimental-and-design methods based on solving the inverse problems are widely used in the full-scale tests of different engineering systems. One such example given in the paper is a broad spectrum of thermophysical investigations that have been carried out in the course of flight tests of the reusable aerospace vehicle heat protection.

INTRODUCTION

A correct technology of scientific research and an engineering design assume the use of a system approach. A necessary aspect of the system approach is the modeling (simulation) of the physical processes and technical objects under study. The modeling can be experimental and mathematical. The role of mathematical modeling in different researches and developments is constantly growing. At the same time, the experiments and tests will always present a basis to validate the mathematical models and methods being used for their adequacy and verify the design decision correctness. Speaking here about the mutual relations of experimental and mathematical modeling we see that they become more ordered and substantiated from the viewpoint of final goal – to provide higher quality and efficiency of investigations and developments. Among the more important trends in achieving the above goal is an advanced methodology of mathematical model identification and physical process diagnostics based on inverse problem solving [1-6].

This methodology has received wide acceptance in different areas of science and technology. Rather high interest to solution of these problems is induced by practical needs of including of nonstationary, nonlinear and multifactor effects in the physical processes and the operational conditions of engineering systems under study. These effects restrict essentially the application of other methods and necessitate the development of new approaches, among which there are inverse methods. Their main advantage is that they allow experiments to be conducted in conditions maximally close to real ones, or directly during operation of real objects. Besides, such approach increases the informativeness, saves experimentation time compared with conventional methods.

In the most complete form the inverse methodology can be in potential realized it various areas of design and testing of engineering objects [5]. A well-organized process of development of some engineering structure, in the general case, should include the constructing of an ordered system of interconnected mathematical models of this structure and its components, as

well as the conditions of their operation. In this connection we point out that methods based on the inverse problem solution can be successfully used not only for solving the particular problems but they can form a base for development, structuralizution and test for adequacy of the desired mathematical models, providing them with the proper numerical information.

Such methodology is used for the construction of adequate enough mathematical models of physical processes as applied to thermally loaded structures and thermal protection materials. The methodology includes the following three general stages:

- construction of a model structure;
- parametric identification – parameter estimation of the structural models;
- validation of the models for adequacy.

The given process can be presented as an extended flow-chart shown in Fig.1

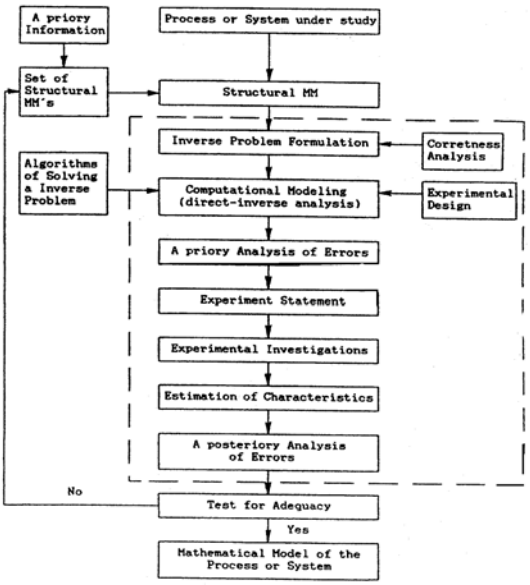


Figure 1: Flow-chart of the identification process

EXAMPLES OF INVERSE METHODOLOGY APPLICATION

One of the very broad field of inverse method application is the analysis of thermophysical properties of composite heat-protective materials acting in high temperature surroundings as, for example, when an orbiter is flying in the Earth's atmosphere. Thermophysical measurements based on classical approach methods for many materials could only be made at temperatures and heating rate changes much lower than at those realized in reality. To avoid this discrepancy it is possible to simulate the required conditions for model heating on the test stands with a further processing of temperature measurements through the methods of inverse heat transfer problems solving. The thermophysical properties thus obtained correspond to the heating conditions brought near to natural conditions in which a thermal protection should operate. In a number of cases, the inverse methods are unique ways for obtaining reliable experimental data on thermophysical characteristics of thermal protection, the insulation materials having complex compositions and structures.

The methodology based on solving similar inverse heat transfer problems poses a new field of thermophysics, the unsteady-state thermophysics of materials and media. In particular, it consists of a mathematical modeling of the heat transfer processes of advanced materials, working out recommendations for creating new materials with prescribed properties.

One more and a very wide application of inverse problem methods, which is directly concerned with thermal investigations of an aerospace vehicle, is the unsteady-state heat measurements. The point is that in heat testing of such engineering systems, or in studying heat transfer processes on the experimental facilities in thermal probing of hot gas flow and in other cases there appears a problem of determining the temperatures, heat fluxes and heat transfer coefficients at the surfaces of the bodies (various structural components, thermal shields, external protective coatings, etc.). Since the intensity of heat transfer to a body usually changes with time because of changes in the heating (cooling) rates, and the non-stationariness of experimental installations parameters, etc. it is especially important to be able to determine the unsteady-state parameters of heat transfer.

As a rule, it is impossible to actually measure the time-changing heat fluxes and heat transfer coefficients. The surface temperature of the objects often remains unaccessible for direct measurements. At the same time, there exists a possibility to measure temperatures at separate points within a body or on some surface part. Thus it becomes necessary to solve the corresponding inverse heat transfer problems, i.e.

to determine the desired thermal boundary conditions by calculation based on temperature measurements.

Such problems of heat measuring are often encountered in the simulation of thermal conditions on the test gasdynamic facilities, in the course of flight simulation, in full-scale tests of flying vehicles, and so on.

At present, similar inverse methods lie at the basis of a new efficient direction of heat measurements, the unsteady-state heat measurements.

THERMAL MODELING OF REUSABLE HEAT PROTECTION

Using the above mentioned procedure of identification the investigations have been carried out of fibrous ceramics and graphite materials for reusable thermal protection of aerospace vehicles, like Russian Buran and American Space Shuttle. The thermal properties and heat transfer characteristics were obtained for real high-temperature and transient conditions of heating including the simulation experiments, facility and full-scale flight tests. These investigations used a system of models of the unsteady-state heat fluxes at the surface and in the intertile clearances of thermal protection as well as the estimation of action of different catalytic properties on the external heat transfer in real high temperature flow of a non-equilibrium gas, the check for adequacy of developed mathematical models of heat transfer both on the surface and inside of thermal protection structures.

Let us dwell on the mathematical modeling and experimental testing of high-porous ceramic composits [7 - 9]. A general thermal mathematical model for the tiled heat protection shield developed in the Moscow Aviation Institute consists of the following components:

- an orbiter motion model in the atmosphere;
- a heat loading model to determine the surface heat fluxes in either points of the vehicle;
- a model of thermal protection structures;
- a thermal model of the high-porous material which itself includes the material structure model, the models of thermal, optic-radiative and hydrolic properties of the material, the conductive, convective and and radiative heat transfer models in the material;
- a heat transfer model in the thermal protection structure (depending on the problem being solved one-,two-, or three-dimensional models are used);
- a heat state model of the thermal protection structure.

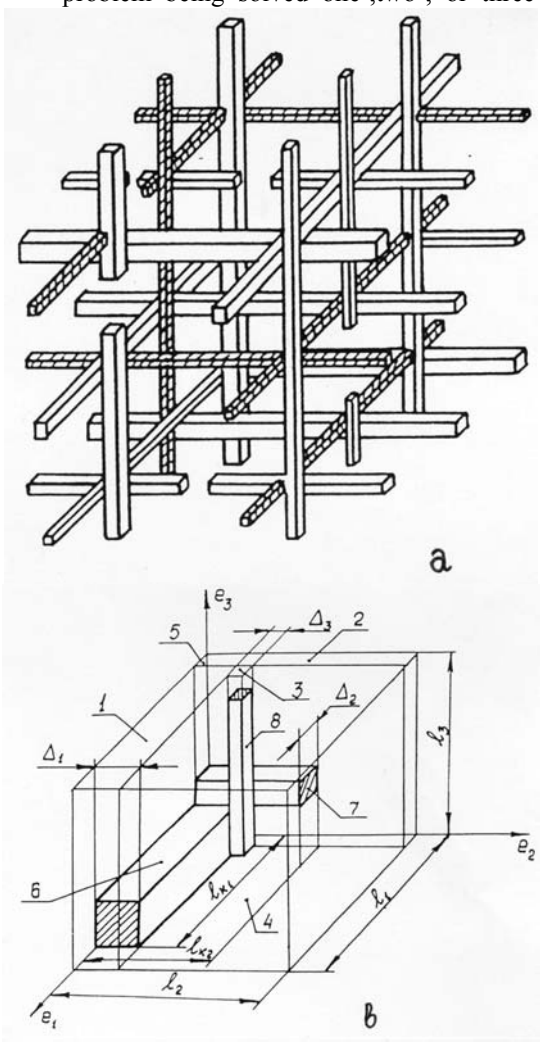


Figure 2: Structure model: a – general configuration; b – elementary volume

This model lets us encompass a significant number of problems arising at different stages of research and development of both the materials and structures.

The thermal model of a high-porous fibrous material.

A thorough analysis of the existing and advanced thermal protection fibrous ceramics showed that they have to be classified among random-and-inhomogeneous media in which the fibers are stochastically distributed by lengths,

diameters, orientatiions and physical properties. So, it is necessary to apply the probability theory methods for describing the material structures.

A model structure is the base for a thermal model of such a material. In the thermal model here considered a regular orthogonal anisotropic structure of fibers (Fig. 2, a) was proposed as a structure model. The analisis showed that a transfer from the original nonorthogonal structure to the orthogonal model is possible on conditions that a model has the thermophisical characteristics as well as the probability distributions of fiber lengths and diameters identical to those of the original material. Such a structure model is essentially an elementary volumes system (Fig. 2, b). Each of them is characterized by some random vector x. Its stochastic characteristics depend on the stochastic ones of fibers. The studies indicated that the mean value of any physical property F of a fibrouse material can be changed by the mean value of the same property but determined for the elementary volume. Thus, it is possible to calculate any physical property of a fibrouse material phovided that we have a mathematical model of this property for the elementary volume.

Based on this theory in combination with the known theories and models (for example, the theory Mi for calculation of optic-radiative charactrerictics of the fibrous media or the Prosolov's model of the gas thermal conductivity in a material) the analytic formulas were obtained for calculation of all required properties of a fibrous material, such as the apparent density, thermal conductivity at a given direction, volumetric heat capacity, spectral and integral optic-radiative coefficients. The developed model has been tested for adequacy through utilization of the experimental data on the basis of solving coefficient IHCPs and optimum experimental design. It was made not only for TZMK ceramics but also for Rigid, Fibrous Ceramic (RFC) composite materials based on the Lockheed HTP technology and flown on US Shuttle Orbiters. High Thermal Performance (HTP) technology uses various combinations of silica and alumina fibers.

The corresponding calculated and experimental values of the effective thermal conductivity for HTP materials (the silica and alumina fiber fractions by mass are 78% and 22% respectively, the mean diameter and variance for silica fibers are 4.3 mkm and 1.69 mkm, for alumina fibres are 3.68 mkm and 2.56 mkm respectively, structure anisotropy factor A is about 2) are presented in Fig. 3 for different pressures of ambient air. The similar data for TZMK materials are given in Fig. 4 for three values of a anisotropy factor A which is defined as an averaged ratio between the number of fibers situated along the longitudinal axis and the number of fibers oriented normally to the surface. The results of such comparative estimations allowed us to make the conclusion that the method developed may be used successfully to predict the structure and composition effect of multicomponent, random-and-inhomogeneous,

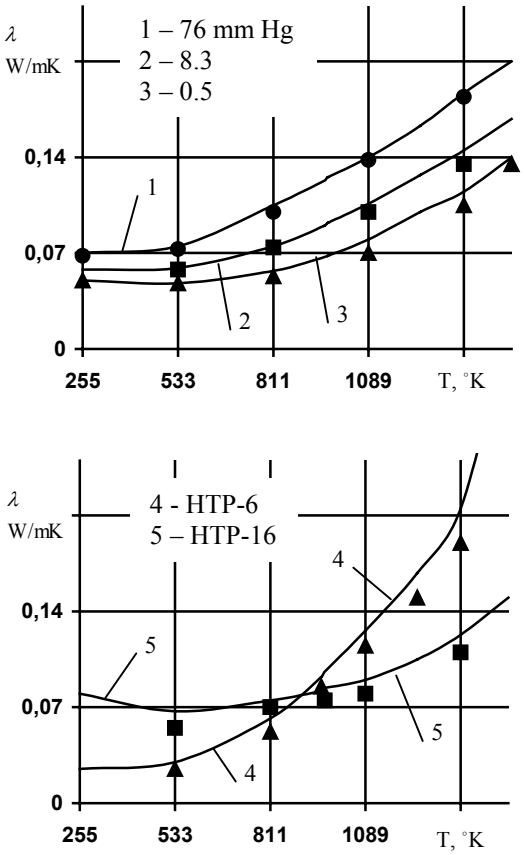


Figure 3: Calculated (solid lines) and experimental (symbols) values of the effective thermal conductivity, W/mK, of RFC materials (a – HTP-12-22 material, the density is 193 kg/m^3; b - HTP-6-22 and HTP-16-22 materials, the density are 96 and 256 kg/m^3, respectively): 1,2,3 – p = 76.0, 8.4, 0.5 mm Hg, respectively; 4,5 – the data for HTP-6-22 and HTP-16-22, respectively, at p = 760 mm Hg

fibrous materials on their heat insulation properties.

The models of heat transfer processes in fibrous ceramic layer. A computational analysis of the combined (conductive-convective-radiative) heat transfer is usually a very difficult problem. That is why, the effective thermal conductivity method has gained acceptance in the engineering practice. The so called "effective thermal conductivity" combines conventionally all effects of complex heat transfer. The values of this characteristic are related directly to its determination procedure and in a number of cases when the natural working conditions differ essentially from the experimental determination conditions of this magnitude such a method may give rise to big errors in estimating the heat state of a system studied. In this connection other and more accurate mathematical models are utilized at developing the thermal protection. The preliminary analysis indicated that the conduction and radiation contribute mainly in the heat transfer studied. So, most attention has been concentrated on the problem of investigating the conduction-and-radiation heat transfer.

The characteristic sizes of thermal protection components are many times the sizes of the material structure nonhomogeneities. This allows the high-porous fibrous materials to be considered as homogeneous media. Then the conduction-and-radiation heat transfer in these materials can be governed by the energy conservation and radiation transfer equations.

However, a direct solution of the radiation transfer equation in a sufficiently general statement is fraught with enormously bulky computations and this is not acceptable for practical implementations associated with repeated calculations of transient heat transfer

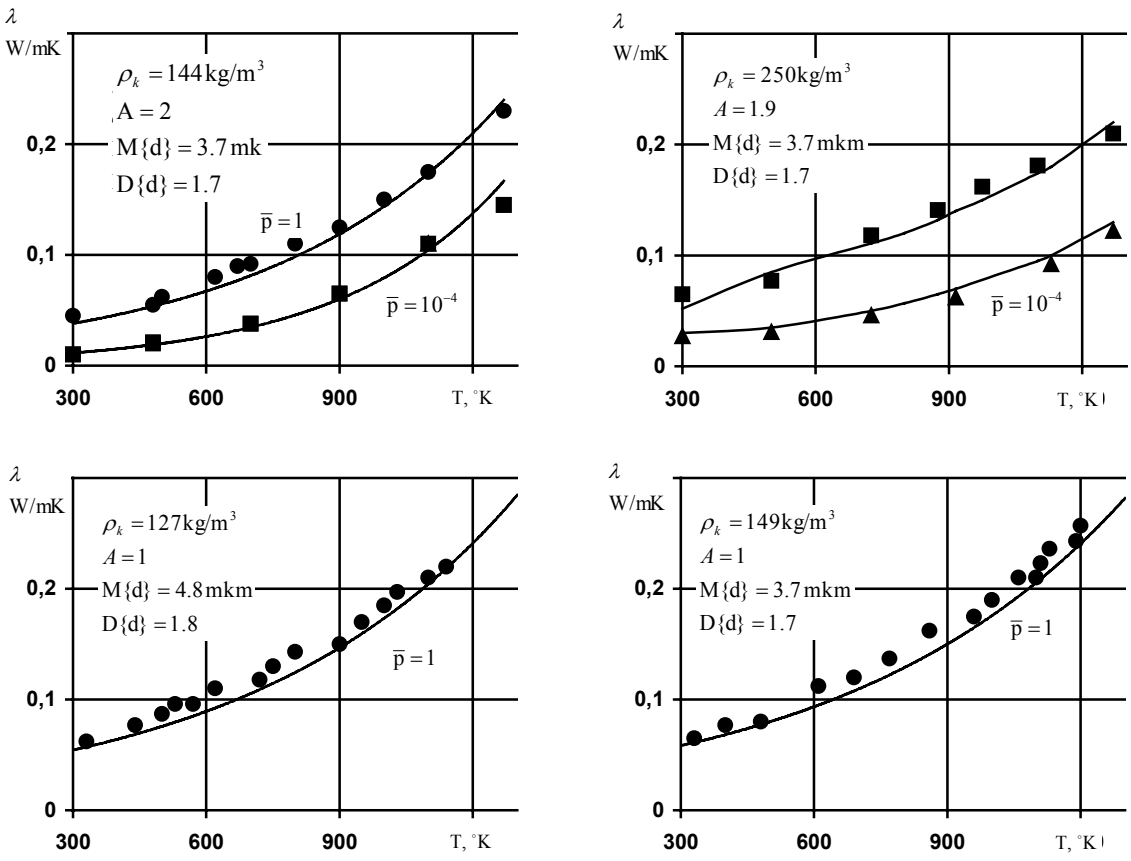


Figure 4: Temperature dependences of the effective thermal conductivity, W/mK, of TZMK materials; $\bar{p} = p/p_0$, $p_0 = 760$ mm Hg: symbols – experiment, solid lines - calculation

processes. Amongst the approximate mathematical models of radiative transfer, a diffusion approximation, as our studies showed, is of prime interest for computational investigations of thermal regimes both of the present-day and future aerospace vehicles. For a diffusive radiation model the radiance distribution in any point M is equiprobable for all directions coming from it. Then the monochromatic radiance $\mathrm{I}_\nu(\mathrm{M},l)$ corresponding to frequency ν is represented by formula

$$I_i(M,I) = A_{vo} + \sum_{i=1} A_{vi}(M)\cos(I,I_i),$$

where I is a unit-equal vector of radiation propagation direction at point M;
I_i is the basis vector;
$A_i, i = 0,2,3$ are some coefficients.

The governing differential equations for the considered problem are the following:

$$F_\nu(M) = -(3b_\nu(M))^{-1} gradU_\nu(M)$$
$$divF_\nu(M) = \alpha_\nu(M)\left[4\pi n^2{}_\nu I_{p\nu}(M) - U_\nu(M)\right],$$

where $F_\nu(M)$ is the monochromatic radiation flux of frequency ν; $U_\nu(M) = \int_{\Omega=4\pi} I_\nu(M,I)d\Omega$ is the monochromatic radiance moment of zero order;
α_ν, n_ν are the monochromatic absorption and refraction coefficients, respectively;
b_ν is the reduced coefficient of monochromatic scattering;
$I_{p\nu}(M)$ is the Planck's function.

The integral radiation flux is determined from formula

$$q_r = \int_0^\infty F_\nu(M)d\nu.$$

To find the temperature field $T(x,y,z,t)$ in a partially transparent scattering material it is necessary to solve the energy conservation equation
with one or other initial and boundary conditions.

$$C\frac{\partial T}{\partial \tau} = div(\lambda gradT) - divq_r.$$

Here C, λ are volumetric heat capacity and thermal conductivity, respectively.

The diffusion approximation gives a good accuracy of the results and it is used extensively in investigating .the reusable ceramic thermal protection. As an example, in Fig. 5 a comparison is presented between the temperatures calculated by means of this mathematical model and the experimental ones measured during a flight test.

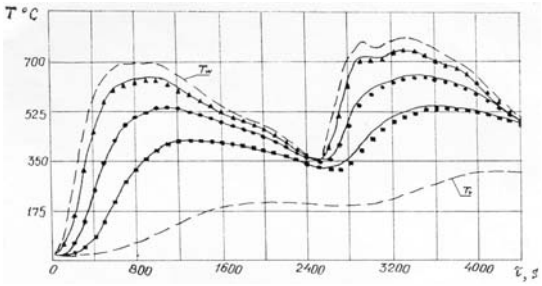


Figure 5: Temperature dependences on time at different points x_i away from the heated external surface, respectively; 1, 2, 3 – the experimental (symbols) and calculated (solid lines) data obtained at x =4.5, 11.9, 21.5 mm, respectively

Software. The MAI has developed an operational medium EXPRESS intended for carrying out studies of heat transfer in thermal shields. This operational medium consists of:
- tile research modules meant for solving different applied problems (prediction of composite materials properties, analysis of heat transfer processes in different thermal protection structures including both indestructible and ablative, in porous cooling structures, etc., thermophysical parameter estimation, diagnostics of heat loading conditions, and so on;
- the data bases on thermophysical properties of gases, homogeneous and composite materials, on structures of composite materials, on heat loadings acting in flight or in tests, on thermal protection structure parameters;
- the interface which makes it possible to operate efficiently the data bases and research modules, to prepare the input data, to analyze quickly the results.

In essence, the software EXPRESS development implies a swing to a new, more effective computer methodology of

thermophysical and thermoengineering problem analysis.

INVESTIGATION OF HEAT TRANSFER IN HETEROGENEOUS GAS FLOWS

Among problems concerning the development of reliable and efficient thermal protection for different types of re-entry vehicles, orbiters, solid propellant engines, etc. is the problem of investigation of heat and forced interaction between dusted gas flows and structural elements. It is well known that the presence of solid or liquid particles in gas flow can significantly increase the heat transfer rate and also may result in erosion of the body material. A characteristic example of such a heterogeneous medium for a re-entry vehicle is a cloud where the cloud particles can exist in the liquid (rain) or solid (hail, snow) phase.

To investigate the multifactor process of heat transfer while interacting of a heterogeneous supersonic gas flow with a solid body we have used a number of methods but the main was a method based on solving inverse heat conduction problems. This approach enables to make the investigations more systematically and obtain new results, in particular, at erosing the material under study.

It is very important that the inverse method allows one to define not only the general influence of the solid particle sizes and concentration on heat transfer but also to study the contribution of different factors of heat transfer (such as the external convective heat flux, the heat flux resulting from additional turbulence caused by solid particles, the additional heat flux resulting from an increase in the surface roughness and the heat flux generated due to particle kinetic energy accommodation) to the total balance of energy at the body surface. The corresponding results are presented, for example, in [10 - 11].

A FLIGHT TESTS OF REUSABLE THERMAL PROTECTION

The unsteady-state inverse methods both in thermophysics and in heat measurements considerably helped us in research and development of the orbiter reusable thermal protection system, in particular, in the parameter estimation and diagnostics of heat transfer in thermal protection / insulation materials and structures in the course of flight tests by Bor-4 automatic vehicles.

Dwell on the flight test application of these methods connected with the study of thermal modes of the tiled heat protection. In this case, heat diagnostics in flight tests were carried out in the following way:

- estimation of heat fluxes on the surface of the tiled thermal shield;
- quality analysis of the effects of physical-chemical reactions on the thermal shield surface with different catalytic properties;
- evaluation of the heat state of the thermal shield surface in the tile gaps;
- estimation of the inner heat state of the tile ceramic material under in-flight heating conditions;
- measurement of the surface pressure.

For these purposes, special measuring devices were developed in MAI and mounted in modified thermal protection tiles (Fig. 6). Outwardly, these tiles did not differ in any way from the standard ones. At the same time, they performed some measuring capabilities in addition to thermal shielding functions. In all of these experimental investigations the methods based on solving inverse heat transfer problems enable us to obtain the unique and reliable results which corroborate the validity of design treatments.

In this complex of investigations the unique data were obtained which have been of much interest both from the standpoint of scientific results and experimental development of the reusable thermal protection in the real conditions

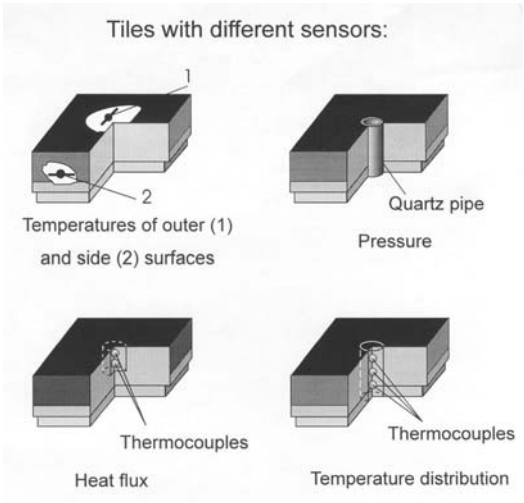


Figure 6: Modified tiles for studying the thermal modes during flight tests of “Bor-4” re-entry vehicle

of re-entry into the atmosphere. In particular, the surface heat fluxes and temperatures histories were determined using as the input data, the temperatures measured inside the tiles. As an example the results of data processing corresponding the measurements in one of the tiles during one of the test flights are shown in Fig.7.

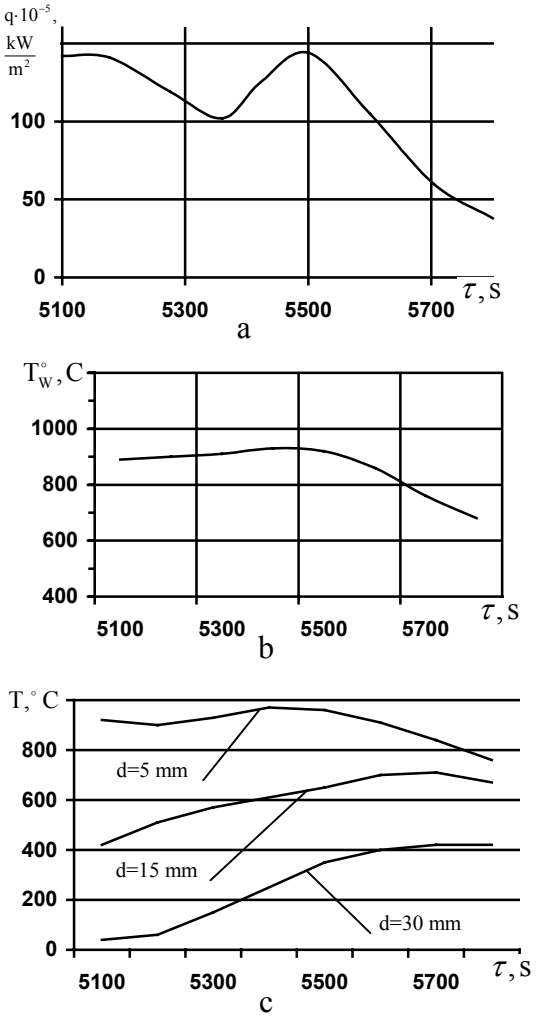


Figure 7: Results of the external heat flux (a) and surface temperature (b) reconstruction by the experimental data (c) : d is the distance between the external surface and a thermocouple

CONCLUSION

It may be said that the methodology based on inverse problems solution makes possible not only to successfully solve each specific problem, for example, in the list of the above – mentioned problems but also helps to set up a judicious combination of mathematical and experimental simulation and full-scale flight testing.

The inverse problems under consideration are ill-posed and in order to solve them we have used the methods based on different regularization procedures [3, 4, 12, 13]. In most cases, the iterative regularization method [3, 4] has given the best results, in particular, from the standpoint of the accuracy of determining desired characteristics. This method is advantageously distinguished by the simplicity and universality of algorithmic constructions in the solution of both the linear and nonlinear, one-dimensional and multi-dimensional inverse problems, including the inverse problems for different mathematical models, in particular, those described by the ordinary and partial differential equations, and by the integral and integro-differential equations. The method enables to take into account a priori information, both qualitative and quantitative, in solving the ill-posed problems. The iterative regularization method is validated rigorously. Its efficiency has been stregthened many times when solving the diversified ill-posed inverse problems arising in practice. The mathematical and applied theory of this method continues to evolve rapidly and the fields of its application are extended steadily [6, 14 -17].

The complex of theoretical and experimental investigations which is briefly considered in this paper has been carried out by a team of researchers in Moscow Aviation Institute in the intimate cooperation with a number of Russian industrial enterprises and research establishments, in particular, NPO Molniya, TsAGI, TsNIIMASH, VIO, NPO “Tekhnologia”.

REFERENCES

1. O.M.Alifanov, Identification of Heat Transfer Processes of Flight Vehicles (Introduction to Inverse Problem Theory), Mashinostroenie, Moscow, 1979 (in Russian).
2. J.V.Beck, B.Blackwell and C.R.Jr.Clair, Inverse Heat Conduction, Ill-posed Problems, A Wiley-Interscience Publication, New York, 1985.

3. O.M.Alifanov, Inverse Heat Transfer Problems, Springer-Verlag, Berlin-Heidelberg-New York, 1994.
4. O.M.Alifanov, E.A.Artyukhin and S.V.Rumyantsev, Extreme Methods of Solving Ill-Posed Problems with Applications to Inverse Heat Conduction Problems, Begell House Publisher, New York, Wallinford (UK) 1994.
5. O.M.Alifanov, Inverse Problem Methodology in Mathematical Modeling and Experimental Simulation, *Inverse Problems in Engineering: Theory and Practice*, ASME 1998, pp.393-409.
6. O.M.Alifanov, E.A.Artyukhin, A.V.Nenarokomov, Identification of Mathematical Models of Compex Heat Transfer, MAI Press, Moscow, 1999.
7. O.M.Alifanov, A Thermal Protection System for Buran Orbiter: Mathematical Modeling, Experimental Simulation and Testing, *Space Course 1995*, Univers: tat Stuttgart, 20, Februar – 3. Mars 1995, SFB 259, Band 2, pp.875-894.
8. Oleg Alifanov and Nikita Bojkov, Les Methodes des Previsions Informatiques des Materiaux Composites de Naute Porosite et de L'Analise Systemes de la Protection Thermique a Leur Base, *Proc. Conference on Spacecraft Structures, Materials & Mechanical Testing*, Grand Hotel Huis ter Duin, Noordwijk, The Netherlands, 27-29 March 1996 (ESA SP-386, June 1996).
9. N.A.Bojkov, S.N.Obroutch and V.K.Zantsev, Experimental and Theoretical Studies of Heat Transfer Complex in Fibrous Insulators, *J.Ing.Phys.*, Minsk, 1990, № 4, p.554-561.
10. O.M.Alifanov, E.A.Artyukhin, A.V.Nenarokomov and I.V.Repin, The Evaluation of Parameters Determining the Heat Interaction of Materials with Two-Phase Flows by the Method of Inverse Problems, *High Temperature*, Vol.31, № 3, 1993, pp.407-411.
11. O.M.Alifanov and I.V.Repin, The Investigation of Heat Transfer in Heterogeneous Flows Using the Method of Inverse Problems of Thermal Conductivity, *High Temperature*, Vol.31, № 1, 1993, pp.71-75.
12. A.N.Tikhonov and V.Y.Arsenin, Solutions of Ill-Posed Problems, V.H.Winston & Sons, Washington D.C., 1977.
13. V.A.Morozov, Regularization Method of solving Ill-Posed Problems, CRC Press, New York, 1993.
14. V.V.Vasin and A.L.Ageev, Ill-Posed Problems with A Priori Information. Vsp, Utrecht, 1995.
15. A.N.Tikhonov, A.V.Goncharskii, V.V.Stepanov and A.G.Yagola, Numerical Methods for the Solution of Ill-Posed Problems, Kluwer Academic Publishing, Dordrecht, 1995.
16. S.F.Gilyazov and N.L.Gol'dman, Regularization of Ill-Posed Problems by Iteration Methods, Kluwer Academic Publishing, Dordrecht, 1999.
17. Dinh Nho Hao, Methods for Inverse Heat Conduction Problems, Europaisher Verlag Der Weissenshaften, Frankfurt am Main, 1998.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

INVERSE HEAT TRANSFER PROBLEMS AND THERMAL CHARACTERIZATION OF MATERIALS

Yvon C. Jarny
Laboratoire de Thermocinetique UMR CNRS6607
Ecole polytechnique de l'université de Nantes
Nantes, France
yvon.jarny@polytech-nantes.univ.fr

ABSTRACT

Combined experimental and mathematical studies to determine thermophysical properties of materials are presented and applied to the characterization of metallic alloys, thermoplastic, thermoset polymers, composite and phase-change materials. The methodology is based on the solution of different inverse heat transfer problems. This approach is well adapted to characterize materials under experimental conditions which reproduce as close as possible some processing conditions which are difficult or even impossible to investigate with conventional techniques. Experimental results illustrate this approach. They focus on the characterization of materials during physical or chemical transformations which are temperature dependent.

INTRODUCTION

The use of advanced and new materials has been growing rapidly in a wide variety of fields (aerospace, aeronautic, automotive, tooling and sporting goods to name a few). In these high technology applications, it is important that the thermal properties of such materials be known for design purposes. Knowledge of the thermal properties is needed to model and to control heat transfer during the manufacturing processes as well as to predict thermal stresses developed when the materials are subjected to non-isothermal environments. The control of thermal phenomena can be a crucial aspect for the improvement in productivity and quality of components. Such control requires the ability to simultaneously predict both the temperature and the rate of the internal heat sources generated by chemical or physical transformations (if any) within the material. Moreover the thermal loads applied on the materials, for example in aerospace structures and vehicles, can induce large temperature gradients, which in turn result in the development of thermal stresses and thus possible structural failure. To prevent this, thermal stress analysis is essential in the design of such structures, which obviously necessitates an accurate knowledge of the thermal properties over large ranges of temperature.

A very large amount of works combining experimental and mathematical activities has been devoted to the determination of two relevant properties for modeling the heat conducting process, the density-specific heat and the thermal conductivity. To develop more accurate mathematical modeling, and to improve the experimental results, the trend is to combine the design of experiments and the solution of heat conduction inverse problems. This combination was defined as a "new research paradigm", J.V. Beck [1]. More degrees of freedom can then be accounted for modeling variable properties and anisotropic media. Moreover, optimal design of experiments allows to minimize the confidence region of the estimates. Further references can be found in [2-9].

In this paper some recent developments performed at Polytech'Nantes, which combine both experimental and computational techniques to determine thermophysical properties of materials, are presented and applied to the characterization of metallic alloys, thermoplastic, thermoset polymers, composites and phase-change materials. Most of the experimental apparatus and protocols, as well as the computational data processing procedures are specific, but they have been developed under a unique methodology based on the resolution of inverse heat transfer problems. The presentation will focus on the interests of this approach. One of them consists in the possibility to characterize materials under experimental conditions which reproduce as close

as possible some material processing conditions, when conventional testing techniques do not offer practical solutions.

For varying thermal properties, the resolution of the inverse problems aims to the determination of unknown functions, then regularization techniques have to be used to account for the numerical unstabilities which occur while solving the ill-conditionned problems and to compute stable solutions. It is well known that the accuracy of the estimated properties is directly related to the sensitivity of the measurements with respect to the unknown variable. When the unknown is considered as a function, the concept of sensitivity coefficient has to be extended. A lagrangian approach is preferred to compute the gradient of the least squares criterion to be minimized, and the standard conjugate gradient algorithm can be used for the minimization.

In the present paper we will report some developed methods and results of characterization for different materials. Varied approaches are considered depending on the modeling equations used to determine the unknown properties of the material. The experimental set up are briefly described.

SEMI-INFINITE MEDIUM - ESTIMATION OF CONSTANT THERMAL PROPERTIES

Isotropic medium

Consider the heat conduction process within an isotropic semi-infinite medium, with constant thermal properties, initially at zero temperature. For times $t > 0$, the material is heated by a line source at a constant rate *q*. In the normal plane *Oxy* to the line source direction, the resulting temperature rise $T(t)$ at the distance *r* to the line source, is solution of the linear modeling heat conduction equation, and is given by

$$T(t) = q\,.Exp\,\mathrm{int}\left[\frac{t}{t}\right], t > 0 \qquad (1a)$$

$$Exp\,\mathrm{int}(u) = \int_u^{\infty} \frac{e^{-x}}{x} dx \qquad (1b)$$

with the parameters (q, t) defined by

$$q = \frac{q}{4p\,l}, \quad t = \frac{r^2}{4a} \qquad (1c)$$

l, a are respectively the thermal conductivity and the thermal diffusivity of the material.

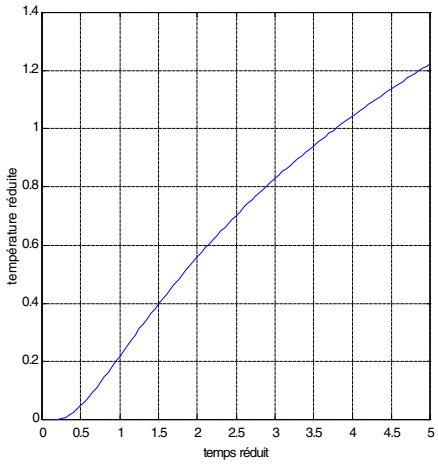


Fig. 1 *Temperature rise* $T(t)$ *resulting of a constant heat flux within a semi-infinite heat conducting medium*

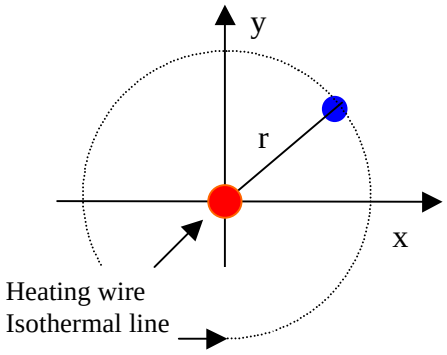


Fig. 2 *Isothermal line within an isotropic semi-infinite medium heated by a line source*

The parameter vector $b = (q, t)^t$ characterizes the thermal properties of the medium. A simple method to determine b consists in mimizing the least square criterion

$$S(b) = \left\| T(b) - Y \right\|^2 \qquad (2)$$

where $T(b)$ is the solution of eqs. (1) computed with the parameter b, *Y(t)* is the temperature measured in the medium

$$Y(t) = T(t; b) + d(t), \quad 0 < t < t_f \qquad (3)$$

d is an uncorrelated zero mean gaussian error with a constant variance s^2, and t_f the duration of the experiment.

The minimization can be performed according to the basic iterative Gauss-Newton algorithm

$$b^{k+1} = b^k + \left[X^{t(k)} X^{(k)}\right]^{-1} X^{t(k)} \left[Y - T^{(k)}\right] \qquad (4)$$

The notation $X^{t(k)} = \left[\nabla_b T^t (b^{(k)}\right]^t$ is used for the sensitivity matrix.

When the modelling eqs. (1)-(2) are exact, the last iteration k^* of the iterative process, is taken depending on the level s of the measurement noise, in order to satisfy the final condition $S(b^*) \le s^2$. Then the approximate variance-covariance of the parameter estimates is

$$\mathrm{cov}(b^*) \approx \left[X^{t(k^*)} X^{(k^*)}\right]^{-1} s^2 \qquad (5)$$

From eqs. (1), it is easy to check that the magnitude of the sensitivity coefficients are monotonously increasing with time, and that the error estimates decreases by increasing the duration of the experiment t_f.

In practice, the assumption of semi-infinite heat conducting medium is valid only for times $t < t_{\max}$. Consider for example the experimental apparatus fig. 3. It involves four flat plates of thermoplastic material arranged in a stack. An electrical heating wire is placed in the middle $(f = O.5mm)$ and thermocouples $(f = O.08mm)$ at the interfaces between the plates. At the outside surfaces, aluminium blocks are used to provide isothermal boundary conditions. Heating at a constant heat flux produces circular isothermal lines. Then the modeling equations (1) are still valid while the temperature rise at the interfaces with the Al-blocks remains less than $e_{\max} = 2s$, and the maximal duration of the experiment $t_{\max}$ is

$$t_{\max} = \frac{r_{\max}^2}{4a} \frac{1}{Exp\,\mathrm{int}^{-1}\left[\frac{e_{\max}}{q}\right]} \qquad (6)$$

where $r_{\max}$, the shortest distance between the heat source and the Al-blocks, depends on the thickness of the plates.

To improve the experiment design, the location(s) r of the sensor(s) and the heating power q can be optimized, but as usual for non linear estimation problems, the solution depends on the unknown parameters to be determined. Knowing the estimates $b^* = (q^*, t^*)$, the thermal parameters (l, a) are obtained from eqs. (1c). The error analysis leads to

$$\begin{cases} \dfrac{\Delta l}{l} \le \dfrac{\Delta q}{q^*} + \dfrac{\Delta q}{q} & (7a) \\ \dfrac{\Delta a}{a} \le \dfrac{\Delta t}{t^*} + 2\dfrac{\Delta r}{r} & (7b) \end{cases}$$

The heating power q has to be chosen a) to avoid a too high temperature rise in the sample which would be incompatible with the assumption of constant properties, and b) to maximize the signal-noise ratio. The sensor location r is chosen to account for the thermocouple location error Δr. When the distance r increases, both the relative error $\Delta r / r$ and the signal-noise ratio decreases, then an optimal value r_{opt} can be predicted [10]

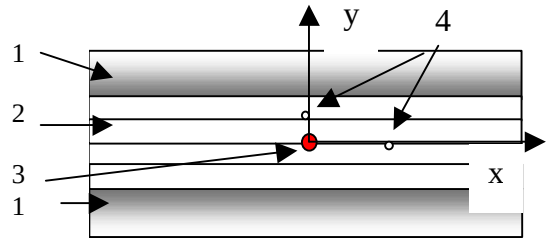


Fig.3 *Experimental set up (not scaled) for measuring the thermal conductivity and the thermal diffusivity . (1) Aluminium blocks, (2) sample plates, (3) heating wire, (4) thermocouples*

Orthotropic medium

Heat transfer within composite materials made up of thermoset matrix and reinforcing fibers (glass, carbon…) are usually modeled by considering these materials as orthotropic media. Conventional testing techniques (calorimeter, guarded hot plates, flash method,…) can be used to determine separately rC_p the specific heat and l_{xx}, l_{yy}, l_{zz} the three components of the thermal conductivity tensor But different experiments are required and substantial time must be devoted to characterize the material.

The method described above for isotropic medium is available for the simultaneous determination of three parameters when the orthotropic directions of the material are known. The heating wire is placed like in the previous experiment in the middle of a stack, the wire direction *Oz* is assumed to be one of the orthotropic directions. Thermocouples are placed parallel to the wire at the interfaces between the plates.

In the normal plane *Oxy* to the direction *Oz* of the line source, the resulting temperature rise $T(t)$ at the distance $r = \sqrt{x_s^2 + y_s^2}$ of the line source, is solution of the linear orthotropic modeling heat conduction equation, and is given by

$$T(t) = q.Exp\,\text{int}\left[\frac{t_x + t_y}{t}\right], t > 0 \qquad (8a)$$

with the parameters (q, t_x, t_y) defined by

$$q = \frac{q}{4p\sqrt{l_{xx}\, l_{yy}}}, \quad t_x = \frac{x_s^2}{4a_x}, \quad t_y = \frac{y_s^2}{4a_y} \qquad (8b)$$

$$a_x = \frac{l_{xx}}{rC_p}, a_y = \frac{l_{yy}}{rC_p} \qquad (8c)$$

For composite materials, the transverse component value of the thermal diffusivity is usually less than the plane component value ($a_y < a_x$), then a constant heat flux in the heating wire produces elliptic isothermal lines and the maximal duration of the experiment, eq. (6), becomes

$$t_{\max} = \min\left(\frac{x_{\max}^2}{4a_x}, \frac{y_{\max}^2}{4a_y}\right) \frac{1}{Exp\,\text{int}^{-1}\left[\frac{e_{\max}}{q}\right]} \qquad (9)$$

At least two sensors are required for the simultaneous determination of the three constant parameters l_{xx}, l_{yy} and rC_p. With two sensors placed at the location coordinates $(x_s, 0)$ and $(0, y_s)$, the set of three parameters is identifiable only if $l_{yy} x_s^2 \neq l_{xx} y_s^2$ [10].

In practice it is easy to place more than two thermocouples, some of them are not used to estimate the parameters but to validate the assumption that the orthotropic directions are correctly known.

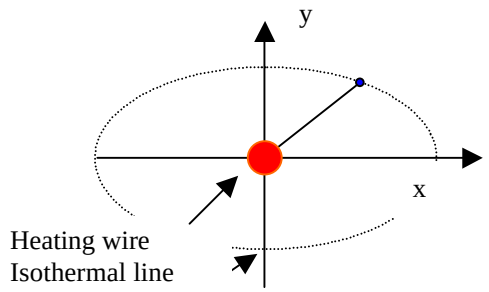


Fig 4 *Isothermal line within an orthotropic semi-infinite medium heated by a line source*

Experimental results (1)

Five thermocouples were placed within a stack of four squared plates of a composite material *(»*3.5x64x64mm^3*)* at the locations (x_s, y_s) given in table 1. The temperature rises shown in figure 5 were obtained with a heating flux *q = 11.35W/m*.

	1	**2**	**3**	**4**	**5**
x_s	0.0	3.80	7.67	11.73	31.84
y_s	6.86	6.86	0.0	0.0	0.0

Table 1 :sensor locations within the stack *(mm)*

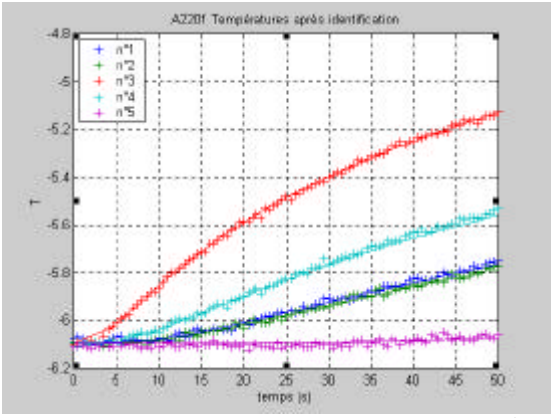


Fig.5 Temperature measurements resulting of a constant heat flux within an orthotropic medium, used to identify three parameters l_{xx}, l_{yy} *and* rC_p

The estimates values of the thermal parameters are $l_{xx} = 3.18 Wm^{-1}K^{-1}$, $l_{yy} = 0.66 Wm^{-1}K^{-1}$ and $rC_p = 1.483e+6 Jm^{-3}K^{-1}$. The maximal duration for this experiment, eq.(9), is $t_{\max} = 45s$. To account for the heat conduction process within the specimen for times $t > t_{\max}$, the assumption of a

semi-infinite medium is not valid. An inverse heat conduction algorithm [10] was developed for estimating the three constant parameters l_{xx}, l_{yy} and rC_p in the finite orthotropic case and was applied to the total duration of the experiment ($t_f = 180s$). It gave the same parameter estimates, but the computational time was much more longer. The numerical solution of the linear orthotropic heat equation has to be computed at each iteration.

Of course when simple thermal characterization methods like the heating wire method (HWM) are practicable, they have to be preferred. For estimating constant thermal parameters of composite materials, the HWM is efficient (3 parameters are estimated with one short experiment), experimentally it is easy to implement and the inverse heat conduction algorithm with the semi-infinite medium assumption, is among the simplest.

TEMPERATURE VARYING THERMAL PROPERTIES

Modeling equations

Modeling the heat conduction process over large temperature ranges leads to consider that thermal properties are not constant because some physical or chemical transformations have to be taken into account within the material over the investigated temperature range. In practice to determine the thermal properties in such conditions, two different approaches have been developed, depending on the ability of the non linear heat conduction model to model the heat transfer process during the transformation.

For example, to characterize the thermal properties of amorphous thermoplastics during solidification, it is sufficient to consider that the density-specific heat $rC_p(T)$ and the thermal conductivity $l(T)$ (isotropic case) are temperature varying, then the inverse heat conduction analysis can be based on the non linear equation

$$rC_p(T)\frac{\partial T}{\partial t} = \nabla\left[l(T)\nabla T\right] \qquad (10)$$

But to model the heat conducting process during the solidification of semi-crystalline thermoplastic materials, or the cooling process of some metallic alloys, or the curing of thermoset resins, eq.(10) is not sufficient. A coupling between the heat transfer and the kinetic of the exothermal transformation(s) has to be taken into account in the modeling equations. For thermal characterization purpose, the following coupled set of equations has been considered with success

$$rC_p(a,T)\frac{\partial T}{\partial t} = \nabla\left[l(a,T)\nabla T\right] + r\Delta H\frac{\partial a}{\partial t} \qquad (11a)$$

$$\frac{\partial a}{\partial t} = F(a,T) \qquad (11b)$$

where ΔH is the total energy (per unit of mass) which is liberated during the transformation(s). The scalar field a is introduced to describe the degree of transformation within the material. For a complete transformation, a varies from *0* to *1*.

The thermal characterization of such materials becomes much more complex, three varying parameters $rC_p(a,T)$, $l(a,T)$ and $F(a,T)$ have to be estimated. The main difficulty consists probably in the determination of the thermal conductivity $l(a,T)$, whose values are required for accurate modeling of the heat conducting process within thick parts of material. Of course characterization strategies which aim to eliminate the coupling between a and *T*, have to be preferred. Experimentally two of them are advisable.

First by selecting the temperature ranges where the transformation does not occur, it is possible to identify separetely the thermal properties $rC_p(a,T)$ and $l(a,T)$. Without transformation, $F(a,T) \approx 0$, that is before it starts ($a = 0$), and after it is completed ($a = 1$), the modeling eqs. (10) are valid.

Secondly when the thermal analysis of the material can be performed on "thin" enough parts, the gradient of temperature $\nabla T \approx 0$ is neglected in the part, and the modeling eqs.(11) integrated over the volume of the part reduce to the simple forms

$$mC_p(a,T)\frac{dT}{dt} = \Phi + m\Delta H\frac{da}{dt} \qquad (12a)$$

$$\frac{da}{dt} = F(a,T) \qquad (12b)$$

where Φ is the total heat flux entering at the outer surface of the part, *m* is the mass of the part.

Estimation of $C_p(T)$ and $l(T)$

The isotropic case

The experimental determination of the temperature varying parameter $C_p(T)$ is usually carried out with "thin" samples of material by using a scanning calorimeter (DSC). The apparatus measures directly the heat flux Φ, (eq. (12a) with $a = 0$), for a preset heating (or cooling) rate.

To determine the thermal conductivity $l(T)$, the heat conduction process is analyzed within "thick" parts of material. Standard inverse analysis is based on the one-dimensional heat equation (10). Different experimental set up have been used. For low conductivity materials, a variant of the set up shown on fig. 3, which works as a "scanning thermal conductimeter" is well adapted. No heating wire is needed in the middle of the stack, but both the outside surfaces of the sample are submitted to a temperature varying (heating or cooling) condition. Temperature histories are recorded at the interfaces of the stack.

The variations of the parameter over the temperature range $[T_{\min}, T_{\max}]$ are approximated by the sum

$$l(T) = \sum_{i=1,..,p} l_i w_i(T) \qquad (13)$$

where the set of basis functions $\{w_i, i = 1,, p\}$ is a prior given. It is convenient in practice to grid the temperature interval into *(p-1)* subintervals $[q_1 = T_{\min} < q_2 < .. < q_p = T_{\max}]$, and to take continuous piecewise linear functions such as $w_i(q_j) = d_{ij}, i, j = 1;;, p$. Then the p-components vector $b = [l_i]_{i=1}^{p}$ is estimated from the additional temperatures *Y(t)* measured within the sample during the experiment, by minimizing the output least square criterion, like in eq.(3).

The iterative Gauss-Newton algorithm, eq. (4), has to be adapted to account for the numerical ill-conditionness of the matrix $[X^{t(k)} X^{(k)}]$

$$b^{k+1} = b^{k} + e^{(k)} [P^{(k)}]^{-1} X^{t(k)} [Y - T^{(k)}] \qquad (14a)$$

$$P^{(k)} = [X^{t(k)} X^{(k)} + U^{(k)}] \qquad (14b)$$

where the diagonal terms of the matrices $U^{(k)}$ are chosen to ensure the stability and to improve the rate of convergence of the iterative process [11].

The numerical solutions of the modeling equation (10) and the *p* sensitivity equations are approximated by using standard finite differences. The vector size *p* is a useful degree of freedom, but it has to be chosen with care.

Experimental results (2). An experimental set up, figure 6a, was used to identify $l(T)$ of a thermoset material after polymerization. It involves a stack of two cured thermoset plates (1) pressed between two heating/cooling blocks (2)

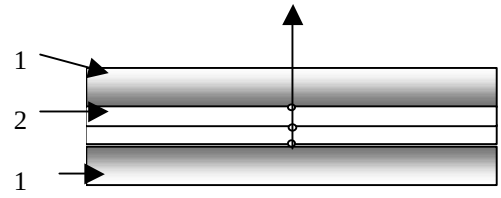


Fig6a- Experimental set up used to identify $l(T)$ for a thermoset material. (1) heating/cooling blocks, (2) specimen plate thickness=5mm.

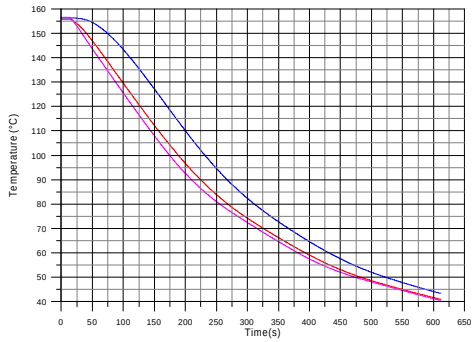


Fig.6b Temperature measurements during the cooling of a thermoset material used to identify $l(T)$

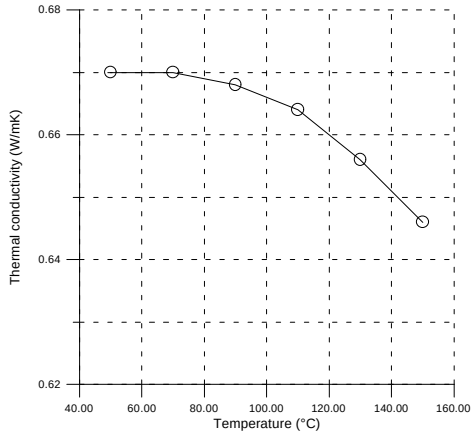


Fig.6c Estimated values of the thermal conductivity $l(T)$ of a thermoset material

The three temperature measurements shown in figure 6b were recorded during the cooling of the stack from $T = 160°C$ down to the ambient temperature in one experiment.The parameter values of $l(T)$, figure 6c, are estimated knowing the temperature varying specific heat determined with a DSC apparatus

$$C_p(T) = 2.0e+6 + 0.75(T-50)e+4 \quad (Jm^{-3}K^{-1})$$

In this simple experiment, no heat flux is measured, the simultaneous identification of $l(T)$ and $C_p(T)$ is not possible. By placing a thin electrical heater in the middle of the stack, and measuring the heat flux, the 1-D inverse heat conduction algorithm is available for both parameters [14].

Experimental results (3). The same 1-D inverse approach was used to determine the thermal conductivity of thermoplastic materials under molding conditions [11-12]. A specific experimental set up, figure 7a, was designed. It involves two molding cavities filled with a molten polymer under high pressure (250-300°C, up to 8 10^7 Pa), and two air coolers which drive the solidification of the polymer to the ambient temperature in less than five minutes. The knowledge of $C_p(T)$, and the temperature histories *Y(t)* recorded at the surface of the coolers, and in the thin central metallic plate located between the cavities are sufficient to estimate the unknown parameter $l(T)$.

High pressure are required to maintain filled up the molding cavities during solidification, and to compensate the tendency of the polymer to contract during solidification. However under the solidification temperature, the pressure drops and an air gap (few microns) modifies the thermal contact at the surface of the cavities. This phenomenon is taken into account in the modeling equations by estimating simultaneously the variations of the thermal contact resistance at the boundary of the molding cavities.

This approach is not valid for the thermal characterization of semi-crystalline polymers. The modeling eqs. (10) are not sufficient, a kinetic model has to be introduced as in eqs. (11) to describe accurately the heat transfer process during solidification and to account for the variation of the temperature solidification with respect to the cooling rate. However outside the phase change temperature interval, the procedure is still correct.

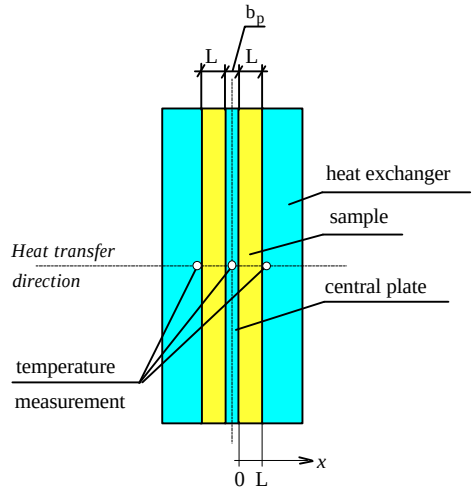


Fig 7a Experimental set up used to identify $l(T)$ of a molten thermoplastic material. Cavity width L=3mm

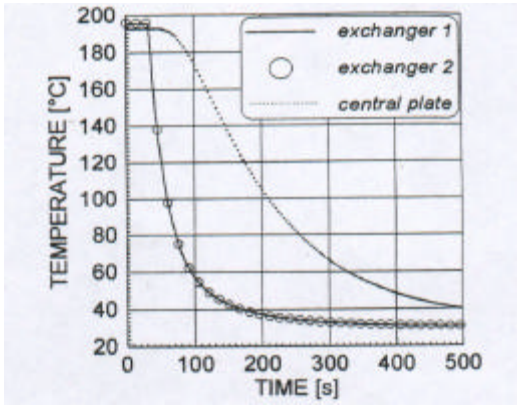


Fig.7b Temperature measurements during the cooling of a thermoplastic material(ABS) used to identify $l(T)$

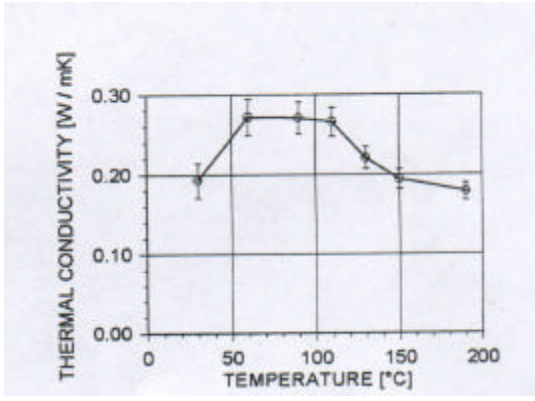


Fig.7c Estimated values of the thermal conductivity $l(T)$ of a thermoplsatic material (ABS)

Experimental results (4). Thermal characterization of metallic alloys was achieved according the same inverse approach [13]. The specific experimental set up, figure 8, involves a cylindrical $(f = 45mm, H = 60,80 or 200mm)$ sample (1) placed between two electrical heaters (3) located at the bases of the cylinder. The sample length *H* is chosen depending on the thermal conductivity of the alloy. The heaters are used to create a thermal gradient in the axis direction. The lateral sample surface is insulated (4) in order to neglect the radial heat losses. The sample is instrumented with thermocouples (2) $(f = 50mm)$ and it is placed in a temperature controlled oven (up to 1200°C).

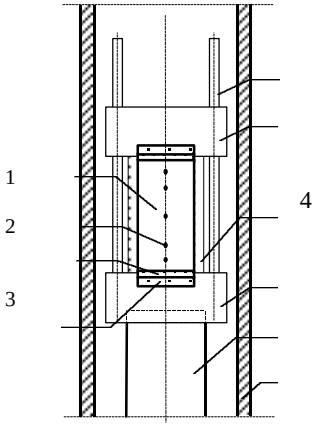


Fig 8 *Experimental set up used to identify* $l(T)$ *of a metallic alloy. Sample length H=80, 60 or200mm*

Estimation of $C_p(T)$ and $l(T)$

The orthotropic case

The inverse approach presented in the first section to estimate simultaneously three constant parameters was extended to the case of temperature varying parameters and orthotropic materials [15]. It is based on the same principle of the heating wire set up, fig. 3, and the approximation of the unknown functions already considered in eq. (13) was still adopted

$$l_{xx}(T) = \sum_{i=1,..,p1} l^i_{xx} w_i(T) \qquad (15a)$$

$$l_{yy}(T) = \sum_{i=1,..,p2} l^i_{yy} w_i(T) \qquad (15b)$$

$$rC_p(T) = \sum_{i=1,..,p3} C^i w_i(T) \qquad (15c)$$

The total size of the unknown vector b to be estimated is then $p = p1 + p2 + p3$, hence the computation of the sensitivity matrix $X^{t(k)} = \left[\nabla_b T^t(b^{(k)}\right]^t$ based on the derivation of the 2-D non linear orthotropic heat conduction equation (10) with respect to each component $b_j, j = 1,..,p$, becomes time prohibitive. The conjugate gradient algorithm combined with the adjoint method is then an advantageous alternative to minimize the output least square criterion, eq. (3). This approach avoids the computation of the sensitivity matrix $\left[\nabla_b T^t(b^{(k)}\right]^t$. It consists in introducing the adjoint variable Ψ solution of the linear 2-D equation

$$-rC_p(T)\frac{\partial \Psi}{\partial t} = \sum_{i=1,2} l_{ii}(T)\frac{\partial^2 \Psi}{\partial x_i^2} + \sum_n e_n \qquad (16a)$$

$$e_n = \left[T(b) - Y_n(t)\right] \otimes d(x_1 - x_1^n) d(x_2 - x_2^n) \qquad (16b)$$

where e_n is the deviation between the computed and the measured temperatures at the location of the sensor *n*.

The components of the gradient ∇S_b of the least square criterion is then computed according to the following equations

$$S'_{l^i_{xx}} = \int_0^{t_f}\int_\Omega \frac{\partial T}{\partial x}\frac{\partial \Psi}{\partial x} w_i(T) d\Omega dt \qquad (17a)$$
i = 1,..,p1

$$S'_{l^i_{yy}} = \int_0^{t_f}\int_\Omega \frac{\partial T}{\partial y}\frac{\partial \Psi}{\partial y} w_i(T) d\Omega dt \qquad (17b)$$
i = 1,..,p2

$$S'_{C^i} = \int_0^{t_f}\int_\Omega \frac{\partial T}{\partial t} \Psi w_i(T) d\Omega dt \qquad (17c)$$
i = 1,..,p3

Experimental results (5). The procedure was applied to the thermal characterization of a composite material made up of epoxy resin and carbon fibers (fiber volumetric ratio = 0.47) in the temperature interval *(10°C , 120°C).* During the experiment, the heat flux generated by the heating wire was time varying in the range 0-250 W/m. The resulting temperature histories of 9 sensors

within the stack of four plates, figure 9a, were used for the simultaneous estimation of the p = 6 unknown components of the parameter vector b .

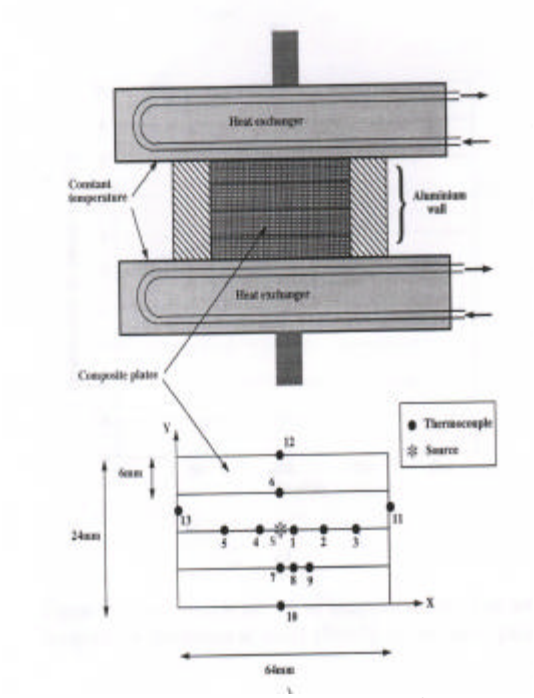


Fig 9a- Experimental set up used to identify the varying thermal parameters l_{xx}, l_{yy} and rC_p *of a composite material*

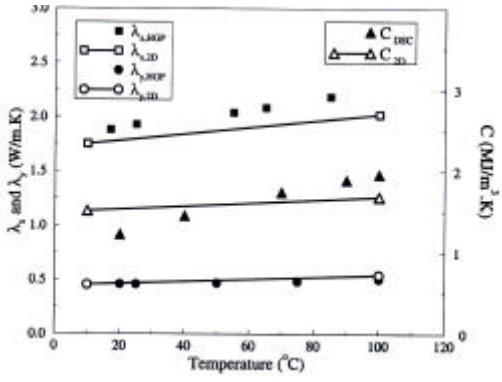


Fig.9b- Estimated values of the parameters l_{xx}, l_{yy} and rC_p *of a composite material*

The estimated parameters, fig. 9a, compared with the parameter values measured with conventional techniques (DSC, Guarded hot plates) are in good agreement. Only one experiment is sufficient However the inverse algorithm is more heavy to implement than the version developed for constant parameters.

Estimation of the kinetic parameter $F(a,T)$

To determine the kinetic parameter values $F(a,T)$ introduced in eqs. (11-12), the simplest experimental approach would consist in submitting a "thin" part of the material in a calorimeter to constant isothermal conditions T_{iso}, and to analyze the resulting signal heat flux $\Phi(t)$. Then eqs.(12) directly give

$$a(t) = \frac{-1}{m\Delta H}\int_{t_0}^{t} \Phi t)dt \qquad (18a)$$

$$F(a(t), T_{iso}) = \frac{-\Phi(t)}{m\Delta H} \qquad (18b)$$

By repeating the experiment for different temperature values T_{iso},, a table of the kinetic parameter values $F(a,T)$ would result. In practice, isothermal experiments are difficult, even impossible, to control accurately. This is true for "high" temperatures and for "fast" kinetics. A scanning approach of the temperature interval is more advisable.

Preliminary measurements of the specific heat $C_p(a,T)$ are required outside the transformation domain, that is for $a = 0$ and for $a = 1$. This is possible by selecting appropriate temperature intervals. Then a mixture law is used to extrapolate the parameter values in the temperature range of the transformation

$$C_p(a,T) = (1-a)C_p(a=0,T) + aC_p(a=1,T) \qquad (19)$$

Non isothermal experiments are carried out with "thin" parts of material, by scanning the temperature interval at different constant heating (or cooling) rate. With the measured heat flux signal $\Phi(t)$, the kinetic parameter $F(a,T)$ can be reconstructed from Eqs. (12). However, much care is required in the analysis of $\Phi(t)$ because only the temperature of the pan is controlled in the calorimeter. For high scanning rates, the imperfect thermal contact between the sample and the pan induces important temperature bias. This can be easily shown by recording the temperature within the sample, using micro-thermocouple.

Some variants of the method are possible depending on the transformation under study. Two examples are briefly presented.

Kinetic of curing. The curing of thick parts of composite materials is challenging because of the low thermal conductivity of the composite and the high heat of reaction during the cross-linking polymerization. This combination can lead to large thermal gradients, generation of residual stresses and polymer degradation. In order to improve the quality of thick parts, the processing temperature needs to be controlled so that the thermal gradients remain small.

The cure rate of thermoset materials is usually described according to the empirical autocatalytic model used by Kamal and Sourour [14]. In practice the chemical process of transformation cannot be reduced to cross-linking, it involves some inhibition period which depends on the thermal history of the material. To account for this induction time, the following curing kinetic model equations were adopted and applied with success to different resins and rubbers

$$\frac{da}{dt} = F(T,a) = \begin{cases} 0, t < t_{ind}(T) \\ k(T)g(a), t \geq t_{ind}(T) \end{cases} \qquad (20)$$

The unknown functions $g(a), t_{ind}(T), k(T)$, to be determined are taken in the following forms

$$g(a) = \sum_{i=1,..,p} g_i w_i(a), 0 < a < 1; g(1) = 0 \qquad (21a)$$

$$k(T) = k_{ref} \exp(-A\left[\frac{T_{ref}}{T} - 1\right]) \qquad (21b)$$

$$\int_0^{t_{ind}(T)} \exp(-A_{ind}\left[\frac{T_{ref}}{T} - 1\right])dt - t_{ref} = 0 \qquad (21c)$$

where the reference temperature T_{ref} is chosen in the temperature range of transformation, $t_{ind}(T), k(T)$ are in the form of Arhenius laws. The function $g(a)$ is approximated as in eq. (13). Typical values of the cure rate determined for an epoxy resin [17-18] are shown on fig. 10a, the parameters of the curing model are $T_{ref} = 423K, A = 17.6, k_{ref} = 0.007s^{-1}$.

The modeling equations (20)-(21) can be used to predict the cure of a "thin" part of resin. The influence of the heating cycle is illustrated on fig. 10b. The temperature of the part is rised from $T_0 = 300K$ up to $T_{\max} = 420K$ at different heating rates $(0.1°C/s,..,0.5°C/s)$, and hold at $T_{\max}$. The kinetic model together with the heat conduction eqs.(11), was validated by curing "thick" parts (thickness = 15mm) of composite material made up of epoxy resin and glass fibers. But to compare measured and computed temperatures, the thermal conductivity $l(a,T)$ has to be known. See the next section.

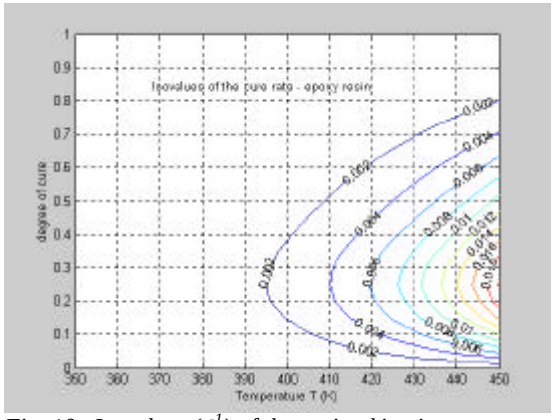


Fig. 10a Isovalues (s^{-1}) of the curing kinetic parameter $F(a,T)$ determined for an epoxy resin

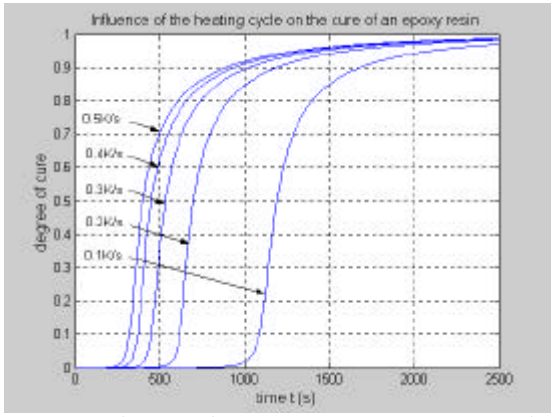


Fig. 10b Influence of the heating cycle on the cure of a "thin" part of epoxy resin.

Other approaches were explored to estimate the kinetic parameters of curing processes. One was based on genetic algorithms [19]. Others used the solution of an inverse heat transfer problems in thick parts of rubber (isotropic case) [20], or composite material (orthotropic case) [21].

Kinetic of solidification. Modeling the heat conducting process within "thick" parts of semi-crystalline thermoplastic materials, during solidification is also challenging. Because of the low thermal conductivity of the material, high cooling rates induce high thermal gradient in the part which in turn generate cristallinity gradient

and affect the mechanical properties of the material. In practice, in the injection molding conditions of such material, the cooling rate is greater than 15K/s, so in order to improve the quality of the part, it is important to well predict the coupling phenomena between cristallization and heat conduction during solidification.

The cristallization rate can be described according to the non isothermal kinetic model of Nakamura [24]

$$\frac{da}{dt} = F(T,a) = k(T)g(a), T_\infty < T < T_f \qquad (22a)$$

$$k(T) = k_0 \exp\left(-\frac{A}{T-T_\infty}\right)\exp\left(-\frac{B}{T(T_f-T)}\right) \qquad (22b)$$

$$g(a) = n(1-a)\left[\ln\frac{1}{1-a}\right]^{1-\frac{1}{n}}, 0 < a < 1 \qquad (22c)$$

Typical values of the crystallization rate for polypropylene are shown on fig. 11a.

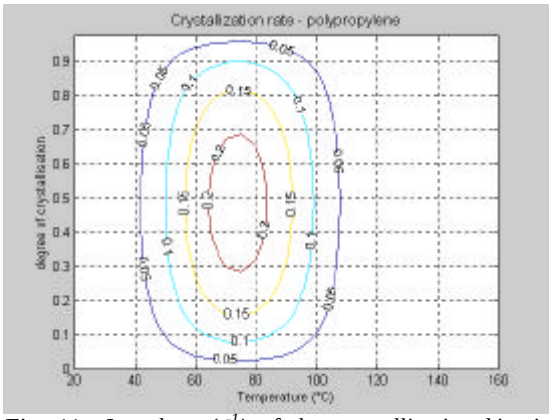


Fig. 11a Isovalues (s^{-1}) of the crystallization kinetic parameter $F(a,T)$ determined for a propylene

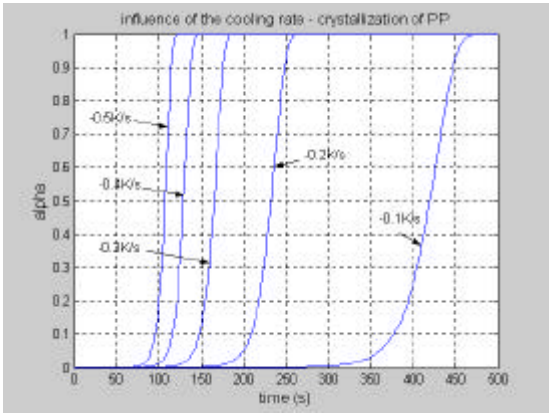


Fig. 11b Influence of the cooling conditions on the crystallization of a "thin" part of polypropylene.

The modeling equations (22) can be used to predict the crystallization of a "thin" part of polymer. The influence of the cooling conditions is illustrated on fig. 11b. The temperature of the part goes from $T_0 = 160°C$ down to $T_{\min} = 25°C$ at different cooling rates ,and is hold down at $T_{\min}$.

In fact with the calorimeter it is impossible to control sufficiently high cooling rates ($> 5K/s$). So the inverse approach described above for "thin" parts is not practicable for investigating the cooling conditions of the injection molding process. An experimental set up has been designed [25] in order to analyze the heat transfer process during the solidification of "thick" parts of such materials. The analysis is based on the kinetic model, eqs.(22), coupled to the heat conduction eqs.(11). Thermal conductivity $l(a,T)$ has to be known.

Estimation of the thermal conductivity $l(a,T)$

Modeling the heat conduction process within "thick" parts of materials characterized by low thermal conductivity, while occurs some chemical or physical transformation with kinetic highly sensitive to the temperature, and which generates internal heat sources, is not an easy task. The two previous examples (curing and solidification processes) explain why it is challenging and illustrate the different steps of our approach for the thermal characterization. The last step consists in checking the ability of the model, eqs. (11), to predict heat flow within "thick" parts. Then the parameter values of $l(a,T)$ are required. In practice, "on-line" experimental measurements of the temperature within the part are available during the transformation, and can be compared to the predicted values. This is not the case for the variable a , hence the complete validation remains difficult. However experiments have been done for the curing and the solidification processes described above. Experimental set up are based on the same principle than on figure 6. In both cases, the following linear approximation was adopted to model the variations of the parameter $l(a,T)$

$$l(a,T) = (1-a)\,l(a=0,T) + a\,l(a=1,T) \qquad (23)$$

During the curing process of "thick" parts of composite material at $T = 140°C$, an overheating is observed in the middle of the part, fig.12. It is well predicted by the solution of the modeling equations.

.

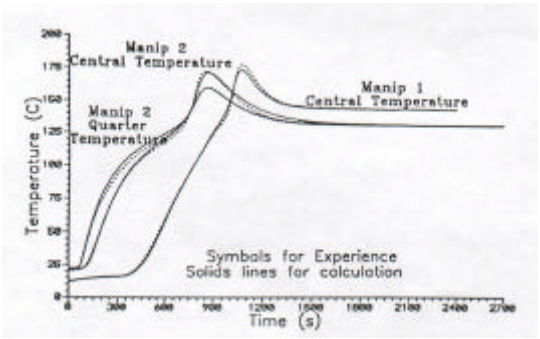


Fig. 12 Curing of thick parts of composite material (epoxy resin/ glass fibers) - Temperatures measured and computed in the middle of the part.

Moreover to complete the validation of the kinetic model for "thick" parts, a strategy based on "partial curing" was developed with success [22-23]. It consists in the determination of optimal heating conditions to apply at the boundary of the part in order to hold a spatially uniform degree of cure at a preset value. The heating cycles are determined by solving an inverse optimal control problem.

During the solidification process of "thick" parts of semi-crystalline polymer, it is observed that high cooling rates at the boundary of the part shift down the solidification temperature. This observation is predicted by the modeling eqs. (22). But more experimental investigation is needed to confirm the adequacy of the model for high cooling rates. An inverse approach is under study to estimate the kinetic function $k(T)$, eq. (22b), from temperature measurements recorded during solidification. The influence of the thermal contact resistance between the part and the mold has also to be taken into account at the boundary conditions.

CONCLUSION

The interests in using methods based on the resolution of inverse heat transfer problems for the thermal characterization of materials were illustrated. This approach usually involves several main steps: a) Choice of the mathematical model of the heat transfer process , b) Development of the inverse problems (IP) algorithms and validation by numerical experiments, c) Design of experiments and experimental data gathering, d) use of the IP algorithms and analysis of the results, to verify finally the adequacy of the process description. The presented results were developed according to a fruitful combination of all these steps. More often, conventional testing techniques offer limited practical solutions to characterize thermal properties under conditions close to processing conditions. It was shown how the use of specific experimental set up together with adapted inverse algorithms enables us to overcome these limits. Most of the examples were related to the thermal characterization of polymers because the control of heat transfer in the manufacturing processes of these materials (like curing or injection molding) is challenging for the improvement in productivity and quality of polymer components.

REFERENCES

1.- J.V. Beck (1990) Inverse problems in Heat Transfer. *Proc. of SFT Conf.*, Nantes, **1,** 47-76- ISBN No2950447104

2.- E.A. Artyukhin, A.S. Okhapin, (1984), Parametric analysis of the accuracy of solution of a non linear inverse problem of recovering the thermal conductivity of a composite material. *J. Eng. Phys.*, **45**, n°5, 1281-1286.

3.- Y.Jarny, D Delaunay and J Bransier, (1986), *8th Proc. Int. Heat Transfer Conf., San Francisco,* **4,** 1811-1816

4.- R. Taktak, J.V.Beck, E.P. Scott (1993), Optimal experiment design for estimating thermal properties of composite materials. *Int. J. Heat Mass Transfer*, **36**,n°12, 2977-2986

5.- K.Dowding, J.V.Beck, B.Blackwell, (1996), Estimation of directional-dependent thermal properties in a carbon-carbon Composite, *Int.J. Heat Mass Transfer*, **39,** 3157-3164

6.- D.Lesnic, L. Elliott and D.B. Ingham , (1996)– Identification of the thermal conductivity and heat capacity in unsteady nonlinear heat conduction problems using the boundary element method. *J. of Comp. Physics*, **126**, 410-420

7.- M.M. Mejias,H.R.B. Orlande, M.N.Ozisik (1999) – A comparison of different parameter estimation techniques for the identification of thermal conductivity components of orthotropic solids. *3rd Proc. of ICIPE, Port-Ludlow WA*, 325-332

8.- G. Carvalho, A.J. Silva Neto-(1999) An inverse analysis for polymers thermal properties estimation. *3rd Proc. of ICIPE, Port-Ludlow WA*, 495-500

9.- C. Aviles-Ramos, A. Haji-Shikh, (2001), Estimation of thermophysical properties of

composites using multi-parameter estimation and zetoth-order regularization, *Inverse Problems in Eng.*,**9**, 507-536

10.- Y Jarny, P Guillemet (2001)Estimation simultanée de la conductivité thermique et de la chaleur spécifique de matériaux orthotropes *Proc. of the SFT Conf., Nantes*, 609-614, Elsevier

11.- T. Jurkowski, D. Delaunay, Y.Jarny (1997)– Estimation of thermal conductivity of thermoplastics under molding conditions- *Int. J. Heat Mass Transfer***, 40**, n°17,4169-4181

12.- A Sommier, T Jurkowski, D Delaunay, Y Jarny, (1998). Characterization of thermophysical properties of thermoplastic materials. *High Pressure-High Temperature*, **30**, 159-164

13.-T Jurkowski, S Heas, A Sarda, Y Jarny, (1998) Solidification d'un alliage d'aluminium-Détermination des propriétés thermophysiques au cours du changement d'état" *Proc. of the SFT Conf. , Marseille*, 320-325, Elsevier.

14.- B. Garnier, D. Delaunay, J.V. Beck, (1992)- Estimation of thermal properties of composite materials without instrumentation inside the samples, *Int. J. of Thermophysics*, **13,** n°6, 1097-1111

15.- R. Aboukachfe, J.L. Bailleul, Y. Jarny, (2000) – The simultaneous determination of thermal conductivity and heat capacity within an orthotropic medium by using conjugate gradient algorithm. *Proc. of the 16th IMACS World Congress*, Lausanne

16.- M.R. Kamal, S Sourour**,** (1973) – Kinetics and thermal characterization of thermoset cure- *Poymer Eng. Sc.*, **13,** 59-64

17.- J.L. Bailleul, D. Delaunay and Y.Jarny, (1996) Determination of temperature variable properties of composite materials – Methodology and experimental results. *J. of Reinforced Plastics and Composites*, **15**, 479-495

18.- J.L. Bailleul, G. Guyonvarch, B. Garnier, D. Delaunay and Y.Jarny , (1996) Identification des propriétés thermiques de composites fibres de verre/resins thermodurcissables. *Rev. Gen. Therm.*, **35**, 65-77

19.- S Garcia, B. Garnier, Y Jarny, (1999) - Simultaneous estimation of kinetic parameters using genetic algorithms- *3rd Proc. ICIPE, Port-Ludlow WA*, 309-316

20- J.S. LeBrizaut, D Delaunay, B. Garnier, Y. Jarny , (1993)– Implementation of an inverse method for identification of reiculation kinetics from temperature measurements on a thick sample. *Int. J. Heat Mass Transfer*, **36**, n°16, 4039-4047

21.- R. Aboukachfe, Y. Jarny , (1999).– Résolution numérique d'un problème d'estimation de source thermodépendante dans un domaine bidimensionnel. *Proc. of the SFT Conf., Arcachon*, 3-8, Elsevier

22.- J.L. Bailleul, D. Delaunay, Y.Jarny , T. Jurkowski , (2001)– Thermal conductivity of undirectional reinforced composite materials- Experimental measurement as function of state of cure. *J. of Reinforced Plastics and Composites*, **20**,n°1,52-64

23.- J.L. Bailleul, D. Delaunay, Y.Jarny, (1998). – Optimal thermal processing of composite materials- An inverse algorithm and its experimental validation. *Proc. 11th IHTC*, **5**, n°5, 87-92, Kyongju, Korea

24.- K. Nakamura (1972), Relationship between crystallization, temperature, crystallinity and cooling conditions, *J. Applied Polym. Sc.* **16,** 1077-1091

25.- G Poutot, P Le Bot, D Delaunay, Y Jarny (2001), Analyse des phénomènes thermiques lors de la cristallisation d'un thermoplastique. *Proc. of the SFT Conf., Nantes*, 615-620, Elsevier

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

PARAMETER STRUCTURE IDENTIFIABILITY AND EXPERIMENTAL DESIGN IN GROUNDWATER MODELING

Ne-Zheng Sun
Department of Civil & Environmental Engineering, University of California at Los Angeles, Los Angeles, CA 90095
nezheng@ucla.edu

Frank T.-C. Tsai
Department of Civil & Environmental Engineering, University of California at Los Angeles, Los Angeles, CA 90095
ftsai@seas.ucla.edu

William W.-G. Yeh
Department of Civil & Environmental Engineering, University of California at Los Angeles, Los Angeles, CA 90095
williamy@seas.ucla.edu

ABSTRACT

The paper presents a new methodology for identifying a distributed parameter with complex and unknown structure, such as the hydraulic conductivity of a heterogeneous aquifer. The basic idea of the methodology is to find a simplest structure from all equivalent structures with respect to the given model applications. A series of new concepts, such as the identifiability of parameter structure, the reliability of model application and the sufficiency of observation data are rigorously defined. Some quantitative relationships between them are derived. Based on these theoretical results, the paper presents an algorithm that can judge the sufficiency and robustness of an experimental design before it is actually conducted in the field. A numerical example is given that shows how a robust experimental design is found by a heuristic procedure.

NOMENCLATURE

AE	Structure error measured in the prediction space.
D	An experimental design
RE	Minimum fitting residual.
SE	Structure error.
$\mathbf{g}_E$	Objectives of prediction alternative E.
$\mathbf{u}_D$	Designed measurements without observation error.
$\tilde{\mathbf{u}}_D$	Designed measurements with observation error.
η	The norm of observation error.
ε	Accuracy requirement of application.
$\grave{e}$	The unknown parameter.
μ	Weighting coefficient.
Θ	Admissible region of the unknown parameter.
$(S,\boldsymbol{q})$	A parameterization representation (PR).

INTRODUCTION

The identification of hydraulic conductivity of a heterogeneous aquifer is a very challenging problem. During the past four decades, this problem was studied by many hydrogeologists and petroleum engineers (Jacquard and Jain 1965, Neuman 1973, Chavent et al, 1975, Yeh and Yoon, 1981, Kitanidis and Vomvoris 1983, Sun and Yeh 1985, Carrera and Neuman 1986, Woodbury and Smith 1988, Sun 1994, MaLaughlin and Townley 1996, among others). From the point of view of mathematics, hydraulic conductivity is the coefficient of the second-order terms of a parabolic or an elliptic PDE. The coefficient identification problem of these types of equations has been studied extensively in mathematics and many engineering fields (Beck et al. 1985, Chavent et al. 1995, Engl et al. 1996, Isakov 1998, Grimstad and Mannseth 2000**,** among others).

A major difficulty of identifying the hydraulic conductivity of an aquifer is the determination of its structure. This difficulty is caused by the complexity and high variability in the structure of natural formations. In most of previous studies, it is assumed that the structure of the unknown parameter is known *a priori* and only the values associated with the structure need to be identified. The parameter identification problem is thus transferred into an optimization problem of best

fitting the existing observed data. Unfortunately, the so identified parameter is often unreliable when it is used for prediction or management purposes even though the fitting residual is small. Table 1 shows how the error in model prediction (or model application) is impacted by the error in parameter structure and the error in parameter values associated with the structure. In Case 1, both structure and value errors are small, and accurate result in model prediction can be expected (small + small = small). This case is ideal but difficult to achieve in practice because of the limitations in prior information and observation data. Case 2 is often seen when the unknown parameter is over-parameterized (attempting to estimate a complex parameter structure with limited data). In this case, the result of model prediction may become very unreliable (small + large = large). Case 3 is seen when the parameter structure is roughly estimated but the parameter values are conditioned by the directly measurements at some locations. In this case, the result of model prediction is again unreliable (large + small = large). When both structure and value errors are large, two cases are possible. Besides Case 4 (large + large = large), we may have Case 5 (large + large = small), in which, the two types of errors cancel each other.

Table 1. The combinations of errors

Case	Error in structure	Error in values	Error in prediction
1	Small	Small	Small
2	Small	Large	Large
3	Large	Small	Large
4	Large	Large	Large
5	Large	Large	Small

It seems that only Case 5 is feasible and practical in the field of groundwater modeling. In this case, the identified parameter structure and values are not their true physical counterparts and thus can only be called as the "*representative structure*" and "*representative values*." The methodology described in the paper attempts to lead us to this case by identifying the best "representative structure" and its associated best "representative values."

The classical theory of inverse problems aims at finding conditions and methods to make the inverse solution to be unique and stable. When the parameter structure is not exactly known, however, to require the *uniqueness* of the inverse solution becomes meaningless because different parameter values may be identified when different structures are used to represent the unknown distributed parameter. In this paper, we introduce a generalized inverse problem that circumvents the uniqueness of the identified parameter in both its structure and its values, instead, it requires finding the simplest "representative structure" to assure the reliability when the model is used for prediction or management purposes. With the concept of "*structure identifiability*" defined in this paper, a complex parameter structure can be identified in a reduced level of complexity provided that the observation data can overcome the impacts of both observation and structure errors. For a given structure, the *worst- case parameter* (WCP) is such a parameter that is the most difficult one to be identified than all other parameters in the admissible region. One can prove that if the WCP is identifiable then all other parameters with the same structure or simplified structures must be identifiable too.

To successfully solve an inverse problem, we must have sufficient information, including the prior information and the information extracted from observed data. When the existing data are insufficient, we must conduct experiments to collect more data. A successful experimental design should ensure that sufficient information would be provided when the designed experiment is actually conducted in the field. The *optimal design* seeks either to maximize the information provided by the experiment or to minimize the cost for conducting the experiment. Several criteria of optimal experimental design have been used in the field of groundwater modeling and other fields of engineering (Qureshi et al. 1980, Rafajlowicz 1986, Sun 1994, Wouwer et al. 2000, Ucinski 2000, among others). Most of these criteria were borrowed from the theory of linear systems. When they are used for nonlinear systems, the optimal design will depend on the unknown parameters and a sequential experiment-design process is needed. For groundwater modeling, however, to conduct such a process is often impractical. When the structure of the unknown distributed parameter is also unknown, the optimal design problem becomes extremely difficult because a more complex structure needs more information to identify. If we cannot judge the sufficiency of a design, the optimal design problem becomes meaningless. In this paper, the optimal design is chosen only from such designs that are sufficient for identifying the simplest

"representative structure" and the WCP. We can prove that if a design is sufficient for identifying the WCP, it must be sufficient for identifying all other parameters in the admissible region. In other words, it is a *robust design*.

In the following sections, a systematic methodology is introduced that allows us to find a reliable model for prediction or management purposes when parameter structures of the model are complicated and unknown. Quantitative relationships between the reliability of model prediction, the identifiability of parameter structure, and the sufficiency of data are established. Algorithms for finding the WCP, solving the GIP and designing the best experiment for parameter structure identification are given. A numerical example shows how this methodology is used for identifying the hydraulic conductivity of a heterogeneous aquifer.

A GENERALIZED INVERSE PROBLEM

It is impossible to identify a distributed parameter $q(\mathbf{x})$ with limited data when the dimension (or the degree of freedom) of the unknown parameter is very high or infinite. *Parameterization* is a way to approximate a distributed parameter by a function with lower degree of freedom. The general form of parameterization can be represented by

$$q(\mathbf{x}) \approx \sum_{j=1}^{m} q_j f_j(\mathbf{x}, \mathbf{v}) \tag{1}$$

where the integer m is called the dimension of parameterization, $\{q_j\}$ ($j=1,2, \ldots,m$) is a set of coefficients, $\{ f_j(\mathbf{x},\mathbf{v}) \}$ is a set of basis functions with a set of shape parameters $\mathbf{v}$ (vector). We will use the combined notation $(S,\boldsymbol{q})$ to denote a *parameterization representation* (PR) of a distributed parameter $q(\mathbf{x})$, where S represents a parameter structure determined by m basis functions, and $\boldsymbol{q} = (q_1, q_2, \ldots, q_m)^T$ is a vector representing the parameter values associated with the structure. The same distributed parameter may have different PRs when it is approximated by different structures.

In Sun and Sun (2002), three types of inverse problems are identified: *the classical inverse problem* (CIP), *the extended inverse problem* (EIP), and *the generalized inverse problem* (GIP). In CIP, it is assumed that structure S is given and only the parameter values $\boldsymbol{q}$ need to be identified. With certain assumptions on the probability distribution of observation error, the estimated values $\hat{\boldsymbol{q}}$ of the unknown parameter is obtained by solving the following optimization problem:

$$\hat{\boldsymbol{q}} = \arg\min_{\boldsymbol{q}} \left\{ \left\| \mathbf{u}^{obs} - \mathbf{u}^{M}(\boldsymbol{q}; \mathbf{x}^{obs}) \right\|_D + \lambda \left\| \boldsymbol{q} - \boldsymbol{q}_0 \right\| \right\} \tag{2}$$

In (2), $\|\cdot\|_D$ and $\|\cdot\|_P$ are norms defined in the observation and parameter spaces, respectively, $\mathbf{u}^{obs}$ is the observed system state at a set of observation locations and times $\mathbf{x}^{obs}$, $\mathbf{u}^{M}$ is a set of corresponding model outputs, λ is a regularization coefficient, and $\boldsymbol{q}_0$ the initial guess of the unknown parameter based on prior information. If there exists no such kind of prior information, but we know the range of the unknown parameter: $\underline{\boldsymbol{q}} < \boldsymbol{q} < \bar{\boldsymbol{q}}$, where $\underline{\boldsymbol{q}}$ and $\bar{\boldsymbol{q}}$ are the lower and upper bounds of the unknown parameter vector, we can take $\lambda=0$ in (2) and add constraint $\boldsymbol{q} \in \Theta$ to the optimization problem, where $\Theta = (\underline{q}, \bar{q})$ is a m-dimensional box and called the *admissible region* of $\boldsymbol{q}$. Usually, it is difficult to find the global solution of problem (2) because of its non-convex nature. When a gradient-based approach is used, only a local minimum can be found. In fact, the most difficult problem of solving CIP is how to determine the complexity of parameter structure. When m is too small (under-parameterized), the fitting residual

$$R = \left\| \mathbf{u}^{obs} - \mathbf{u}^{M}(\hat{\boldsymbol{q}}; \mathbf{x}^{obs}) \right\|_D + \lambda \left\| \hat{\boldsymbol{q}} - \boldsymbol{q}_0 \right\|_P \tag{3}$$

may have a large value. On the other hand, when m is too large (over-parameterized), the model prediction becomes unreliable. Moreover, even we can find an appropriate m, the identified parameter may be still very different from the real one if the structure pattern of the unknown parameter is not correctly assumed.

In EIP, structure S and parameter values $\boldsymbol{q}$ are identified simultaneously by solving the following optimization problem

$$(\hat{S}, \hat{\boldsymbol{q}}) = \arg\min_{(S,\boldsymbol{q})} \{ \left\| \mathbf{u}^{obs} - \mathbf{u}^{M}(S, \boldsymbol{q}, \mathbf{x}^{obs}) \right\|_D + \lambda \left\| (S, \boldsymbol{q}) - (S_0, \boldsymbol{q}_0) \right\| \} \tag{4}$$

The second term on the right-hand side can be deleted when the lower-upper bound constraint $\boldsymbol{q} \in \Theta(S)$ is used instead. Equation (4) is a combinatorial optimization problem and is very difficult to solve because the dimension of the shape vector **v** in (1) may be high. Sun and Sun (2002) presented a tree regression procedure to solve EIP that can find a nearly optimal solution with less computation effort. Tsai et al. (2002) used the genetic algorithm (GA) in combination with a local search to solve the EIP, in which the unknown parameter is represented by the natural neighbor parameterization. By sequentially increasing the number of basis points and optimizing their locations, the fitting residuals can be effectively decreased and the over-parameterization problem can be avoided.

Note that the parameter structure obtained by solving the EIP is only based on existing observations without considering the reliability of model applications. After the EIP is solved, we cannot answer whether or not the complexity of the identified parameter structure is appropriate, and whether or not the observation data are sufficient.

The GIP aims at finding an appropriate parameter structure to satisfy the accuracy requirement of model applications. Let $\mathbf{g}_E$ be a set of predictions or management decisions, the reliability requirement may be stated as

$$\left\| \mathbf{g}_E(S,\boldsymbol{q}) - \mathbf{g}_E(\boldsymbol{q}^t) \right\|_E < e \qquad (5)$$

where $\|\cdot\|_E$ is a norm defined in the objective (prediction or management) space, $\boldsymbol{q}^t$ is the true parameter, which , of course, is unknown. Condition (5) can be satisfied by different PRs with different parameter structures and different parameter values. The GIP requires finding the simplest parameter structure and its associated parameter values from all PRs that satisfy the accuracy requirement (5).

The so defined GIP has the following advantages. First, the reliability of model application is incorporated into the identification procedure. Second, the uniqueness requirement of the inverse solution is avoided and replaced by a weak requirement (5). This condition may be satisfied by such parameters that are not close to the true parameter in the parameter space. Third, the data requirement is minimized because the GIP attempts to find the simplest parameter structure. Once the complexity of parameter structure is determined, the sufficiency of existing data can be judged.

The stepwise regression method presented by Sun et al. (1998) can be used to solve the GIP, in which a max-min problem must be solved in each iteration step. With the theorem developed in the next section, however, we can find a more effective method to solve the GIP.

STRUCTURE ERROR AND STRUCTURE REDUCTION

Letting $(S_A,\boldsymbol{q}_A)$ and $(S_B,\boldsymbol{q}_B)$ be two different PRs of a distributed parameter $\boldsymbol{q}(\mathbf{x})$, the distance between them can be measured in parameter, observation and prediction (or management) spaces, respectively, by

$$d_P(S_A,\boldsymbol{q}_A;S_B,\boldsymbol{q}_B) = \left\| \bar{\boldsymbol{q}}_A - \bar{\boldsymbol{q}}_B \right\|_P$$
$$d_D(S_A,\boldsymbol{q}_B;S_B,\boldsymbol{q}_B) = \left\| \mathbf{u}_D(S_A,\boldsymbol{q}_A) - \mathbf{u}_D(S_B,\boldsymbol{q}_B) \right\|_D$$
$$d_E(S_A,\boldsymbol{q}_B;S_B,\boldsymbol{q}_B) = \left\| \mathbf{g}_E(S_A,\boldsymbol{q}_A) - \mathbf{g}_E(S_B,\boldsymbol{q}_B) \right\|_E$$

where $\bar{\boldsymbol{q}}_A$ and $\bar{\boldsymbol{q}}_B$ are spans of $\boldsymbol{q}_A$ and $\boldsymbol{q}_B$ to the parameter space P, $\mathbf{u}_D$ is the model outputs corresponding to an observation design D, $\mathbf{g}_E$ is a vector of model applications corresponding to a set of objectives E, $\|\cdot\|$ means a norm defined in a space as denoted by its subscript.

In Sun et al. (1998), the *distance d* between two PRs $(S_A,\boldsymbol{q}_A)$ and $(S_B,\boldsymbol{q}_B)$ is defined by

$$d(S_A,\boldsymbol{q}_A;S_B,\boldsymbol{q}_B) = d_E + \mu d_D + l d_P \qquad (6)$$

were μ and l are weighting coefficients. In this paper, we take l=0, i.e, we do not consider the difference between the two PR's in the parameter space. Let $(S_A,\boldsymbol{q}_A)$ be a PR and S_B be a structure different from S_A. A PR $(S_B,\boldsymbol{q}_{AB})$ is called a *projection* of $(S_A,\boldsymbol{q}_A)$ onto the structure S_B, when

$$\boldsymbol{q}_{AB} = \arg \min_{\boldsymbol{q}_B} d(S_A,\boldsymbol{q}_A;S_B,\boldsymbol{q}_B)$$
$$\text{s.t. } \boldsymbol{q}_B \in \Theta(S_B) \qquad (7)$$

To find $\boldsymbol{q}_{AB}$ from (7) is equivalent to solving a classical inverse problem, i.e. using a fixed

parameter structure S_B but changing the parameter values $\boldsymbol{q}_B$ to best fit both the model output $\mathbf{u}_D(S_A,\boldsymbol{q}_A)$ and model application $\mathbf{g}_E(S_A,\boldsymbol{q}_A)$.

Definition 1. The *structure error* $SE(S_A,S_B)$ of using parameter structure S_B to replace parameter structure S_A is defined by the following max-min problem:

$$SE(S_A,S_B)=\max_{\boldsymbol{q}_A}\min_{\boldsymbol{q}_B} d(S_A,\boldsymbol{q}_A;S_B,\boldsymbol{q}_B)$$
$$\text{s.t. } \boldsymbol{q}_A\in\Theta(S_A) \text{ and } \boldsymbol{q}_B\in\Theta(S_B) \qquad (8)$$

Generally, $SE(S_A,S_B)\neq SE(S_B,S_A)$. When S_B is a simplification of S_A, we have $SE(S_B,S_A)=0$. If $SE(S_A,S_B)$ and $SE(S_B,S_A)$ are both less than a given error bound, the two structures are said to be *equivalent*. When S_B is a simplification of S_A, to judge their equivalence we only need to calculate $SE(S_A,S_B)$. If we take λ=0 and μ=0 in (6), the equivalence of two parameter structures means that we can use one structure to replace another one for specified model applications. The max-min problem (8) is very difficult to solve.

Definition 2. A PR (S_A, $\tilde{\boldsymbol{q}}_A$) is called the *worst-case parameter* (WCP) for simplifying a structure S_A to a structure S_B, when it satisfies

$$SE(S_A,S_B)=\min_{\boldsymbol{q}_B} d(S_A,\tilde{\boldsymbol{q}}_A;S_B,\boldsymbol{q}_B),$$
$$\text{s.t. } \boldsymbol{q}_B\in\Theta(S_B) \qquad (9)$$

If we know the WCP, the structure error can be obtained by solving the min problem (9) rather than the max-min problem (8).

Theorem 1. When a *k*-zone structure S_A is simplified into a one-zone structure S_B, the WCP must be located at such vertices of the admissible region Θ_A where the differences between the *k* parameter values reach either their upper bounds or their lower bounds.

The proof of Theorem 1 and its more general form can be found in Sun (2002). For the case of *k* = 2, we have $SE(S_A,S_B)=cL^2$, where *c* is a coefficient and *L* is the maximum difference between the parameter values of the two zones in the admissible region. For example, if the ranges of the parameter values associated with the two zones are $10\le q_{A,1}\le 20$ and $8\le q_{A,2}\le 30$, respectively, then we have $L=30-10=20$. For the case of $k>2$, the WCP may depend on the flow conditions (boundary conditions and/or since/source terms). Although Theorem 1 cannot give us a unique solution of the WCP, it limits the search of the WCP to a small set. From physics, WCP is the most unlikely one to be simplified to a homogeneous one, and thus, it can often be guessed from the available prior information of the physical problem under study.

SOLUTION OF GIP

To solve the GIP by the stepwise regression method presented by Sun et al. (1998), we construct the following structure sequence

$$S_1\subset S_2\subset S_3\subset\cdots\subset S_m\subset S_{m+1}\subset\cdots \qquad (10)$$

where S_1 is a homogeneous structure, S_2 is a structure with two zones, and so forth. Generally, S_{m+1} is obtained from S_m by dividing one zone of S_m into two sub-zones. In this case, the shape vector is described by the location of a linear boundary dividing one zone of S_m into two zones of S_{m+1} and thus the shape parameter **v** in the representation of parameterization has a low dimension. For each complexity level *m*, we solve the EIP to find the optimal PR by minimizing the following fitting residual:

$$RE_m=\min_{S_m,\boldsymbol{q}_m}\left\|\mathbf{u}_D^{obs}-\mathbf{u}_D(S_m,\boldsymbol{q}_m)\right\|_D$$
$$\text{s.t. } \boldsymbol{q}_m\in\Theta(S_m) \qquad (11)$$

where $\mathbf{u}_D^{obs}$ are the observed values of **u**, and $\mathbf{u}_D(S,\boldsymbol{q})=\mathbf{u}^M(S,\boldsymbol{q};\mathbf{x}^{obs})$ are the corresponding model outputs. At the same time, we calculate the maximum model application error of using S_{m-1} to replace S_m, which is defined by

$$AE_m=\max_{\boldsymbol{q}_m}\min_{\boldsymbol{q}_{m-1}}\left\|\mathbf{g}_E(S_m,\boldsymbol{q}_m)-\mathbf{g}_E(S_{m-1},\boldsymbol{q}_{m-1})\right\|_E$$
$$\text{s.t. } \boldsymbol{q}_{m-1}\in\Theta(S_{m-1}),\quad \boldsymbol{q}_m\in\Theta(S_m) \qquad (12)$$

Let us consider three cases. (1) If $AE_m > e$ and $RE_m > 2h$, where e is the given accuracy requirement of model applications and h is the norm of observation error, we increase m to m+1. S_{m+1} is obtained by dividing such a zone of S_m into two zones that is the most sensitive one to the specified model applications. The boundary between the two zones is determined by minimizing the fitting residual $RE_{m+1.}$ (2) If $AE_m < e$, stop and use $(\hat{S}_m, \hat{\boldsymbol{q}}_m)$ as the identified parameter. (3) If $AE_m > e$ but $RE_m < 2h$, new data need to be collected.

Because S_{m+1} is obtained from S_m by dividing one zone of S_m into two sub-zones, the WCP is very easy to determined and AE_m can be calculated by only solving a min problem rather than a max-min problem. As a result, the above stepwise regression procedure becomes very effective.

STRUCTURE IDENTIFIABILITY

The classical identifiability requires that the mapping between the observation space and the parameter space be a one to one mapping. This requirement can never be satisfied when the observation error exits. The output least square identifiability (Chavent 1987) requires that the solution of (2) be unique and continuously depend on observation data. The extended identifiability defined in Sun and Yeh (1990) uses the reliability of model application to replace the requirement on the uniqueness of the identified parameter. In the statistic framework, a parameter is identifiable if a change in parameter is always accompanied by a change in the probability distribution of the observed data (Stark 2000). In all of the previous definitions on identifiability, it assumes that the structure of the unknown parameter is known. In practice, however, this assumption is often unsatisfied and, instead, the structure of the unknown parameter must be identified together with its unknown values. In this section, we will define a new kind of identifiability that does not require knowing the parameter structure. In fact, it allows the non-uniqueness in both parameter structure and parameter values.

Definition 3. A parameter $\boldsymbol{q}_A$ with structure S_A, i.e., a PR $(S_A, \boldsymbol{q}_A)$, is said to be *d-e identifiable* at a simplified structure level S_B (or *d-e-*S_B identifiable) if there is an observation design D that

$$\left\|\mathbf{g}_E(S_A,\boldsymbol{q}_A) - \mathbf{g}_E(S_B,\boldsymbol{q}_B)\right\|_E < e \tag{13}$$

is satisfied for any PR $(S_B, \boldsymbol{q}_B)$, provided

$$\left\|\mathbf{u}_D(S_A,\boldsymbol{q}_A) - \mathbf{u}_D(S_B,\boldsymbol{q}_B)\right\|_D < d \tag{14}$$

The values $\{\mathbf{u}_D(S_A,\boldsymbol{q}_A)\}$ in (14) can be considered as the observations under design *D* without observation error. Condition (14) means that we can fit these observations to a certain extent by a parameter $\boldsymbol{q}_B$ with a simplified structure S_B. Equation (13) means that when $(S_B, \boldsymbol{q}_B)$ is used to replace $(S_A, \boldsymbol{q}_A)$ as the model parameter, reliable results of model application can be obtained. Therefore, if a distributed parameter is *d-e-*S_B identifiable, we can identify it at the simplified structure level S_B. The following theorem gives a sufficient condition for the *d-e-*S_B identifiability.

Theorem 2. If the projection $(S_B, \boldsymbol{q}_{AB})$ of $(S_A, \boldsymbol{q}_A)$ onto S_B is d_1-e_1 identifiable, i.e.,

$$\begin{aligned}&\left\|\mathbf{u}_D(S_B,\boldsymbol{q}_B) - \mathbf{u}_D(S_B,\boldsymbol{q}_{AB})\right\|_D < d_1 \text{ implies}\\ &\left\|\mathbf{g}_E(S_B,\boldsymbol{q}_B) - \mathbf{g}_E(S_B,\boldsymbol{q}_{AB})\right\|_E < e_1\end{aligned} \tag{15}$$

for any parameter $\boldsymbol{q}_B$ with the structure S_B, then $(S_A, \boldsymbol{q}_A)$ is *d-e-*S_B identifiable, where

$$\begin{aligned}d &= d_1 - d(S_A, \boldsymbol{q}_A; S_B)/m\\ e &= e_1 + d(S_A, \boldsymbol{q}_A; S_B)\end{aligned} \tag{16}$$

The Proof of Theorem 2 can be found in Sun (2002). When observation error is involved, Equation (14) is replaced by

$$\left\|\tilde{\mathbf{u}}_D(S_A,\boldsymbol{q}_A) - \mathbf{u}_D(S_B,\boldsymbol{q}_B)\right\|_D < d \tag{17}$$

where $\tilde{\mathbf{u}}_D(S_A,\boldsymbol{q}_A) = \mathbf{u}_D(S_A,\boldsymbol{q}_A) + \boldsymbol{\eta}$ and $\boldsymbol{\eta}$ is the observation error (vector), In this case, we have to change (16) to

$$\begin{aligned}d &= d_1 - d(S_A, \boldsymbol{q}_A; S_B)/m - h\\ e &= e_1 + d(S_A, \boldsymbol{q}_A; S_B)\end{aligned} \tag{18}$$

In the above equation, h is the upper bound of $\|\boldsymbol{h}\|_D$.

Definition 4. A parameter structure S_A is said to be *d-e-*S_B identifiable, if all PRs (S_A, $\boldsymbol{q}_A$) within the admissible region Θ_A are *d-e-*S_B identifiable.

Theorem 3. If all PRs (S_B,$\boldsymbol{q}_B$) are d_1-e_1 identifiable (Sun, 1994), i.e.,

$$\begin{aligned}&\left\|\mathbf{u}_D(S_B,\boldsymbol{q}_{B,1})-\mathbf{u}_D(S_B,\boldsymbol{q}_{B,2})\right\|_D<d_1 \text{ implies}\\&\left\|\mathbf{g}_E(S_B,\boldsymbol{q}_{B,1})-\mathbf{g}_E(S_B,\boldsymbol{q}_{B,2})\right\|_E<e_1\end{aligned}\qquad(19)$$

for any two PRs (S_B,$\boldsymbol{q}_{B,1}$) and (S_B,$\boldsymbol{q}_{B,2}$) within Θ_B, then structure S_A is *d-e-*S_B identifiable, where

$$\begin{aligned}&d=d_1-SE(S_A,S_B)/m-h \quad\text{and}\\&e=e_1+SE(S_A,S_B)\end{aligned}\qquad(20)$$

This theorem can be proved similarly as Theorem 2 (Sun, 2002). Equation (20) clearly shows that to make a distributed parameter to be identifiable, the information provided by observation data must be able to overcome the effects of both structure error and observation error. The effect of structure error is deterministic rather than random, and, in most cases, it dominates the effect of observation error. If the structure error between the true and simplified structures is too large, no observation design can make the unknown parameter to be identifiable at the simplified structure level.

ROBUST EXPERIEMENTAL DESIGN

When the structure of a distributed parameter is unknown, the problem of experimental design for parameter identification becomes extremely difficult because a more complicated parameter structure needs more data to identify. If we cannot judge the sufficiency of a design, both the concepts of optimal design and robust design become meaningless. In this section, we will define the sufficiency of a design when the parameter structure is unknown and then present an effective approach for finding a robust design. A design D for identifying a distributed parameter system consists of a set of decisions on where, when and how the system is excited, and where and when the states of the system are observed (Sun, 1994).

Definition 5. A design *D* is said to be sufficient for identifying a PR (S_A,$\boldsymbol{q}_A$), if the PR is $e_0-d_0-S_B$ identifiable, i.e.,

$$\begin{aligned}&\left\|\tilde{\mathbf{u}}_D(S_A,\boldsymbol{q}_A)-\mathbf{u}_D(S_B,\boldsymbol{q}_B)\right\|_D<d_0 \text{ implies}\\&\left\|\mathbf{g}_E(S_A,\boldsymbol{q}_A)-\mathbf{g}_E(S_B,\boldsymbol{q}_B)\right\|_E<e_0\end{aligned}\qquad(21)$$

for any PR (S_B,$\boldsymbol{q}_B$) in $\Theta(S_B)$. Here, d_0 must be at least larger than $2h$, and e_0 should not exceed the given accuracy requirement ε of model application. From this definition, when a sufficient design is actually conducted and the data are collected, we can assure that an equivalent parameter for model application can be identified from the data.

Theorem 4. If a design D satisfies the following condition at a complexity level S_B:

$$\begin{aligned}&\left\|\mathbf{u}_D(S_B,\boldsymbol{q}_{B,1})-\mathbf{u}_D(S_B,\boldsymbol{q}_{B,2})\right\|_D<\\&2\left[SE(S_A,S_B)/m+h\right] \quad\text{implies}\\&\left\|\mathbf{g}_E(S_B,\boldsymbol{q}_{B,1})-\mathbf{g}_E(S_B,\boldsymbol{q}_{B,2})\right\|_E<\\&e-SE(S_A,S_B)\end{aligned}\qquad(22)$$

where S_A is the structure of the unknown parameter, then the design should be sufficient for identifying a parameter at level S_B to satisfy the following accuracy requirement of model application:

$$\left\|\mathbf{g}_E(S_B,\hat{\boldsymbol{q}}_B)-\mathbf{g}_E(S_A,\boldsymbol{q}_A^t)\right\|<\varepsilon\qquad(23)$$

where $\boldsymbol{q}_A^t$ is the true parameter but unknown and $\hat{\boldsymbol{q}}_B$ is obtained by solving a CIP with the fixed structure S_B, i.e.,

$$\hat{\boldsymbol{q}}_B=\arg\min_{\boldsymbol{q}_B\in\Theta_B}\left\|\mathbf{u}_D^{obs}-\mathbf{u}_D(S_B,\boldsymbol{q}_B)\right\|_D$$

The proof of Theorem 4 can be found in Sun (2002). Theorem 4 tells us that the inverse

solution obtained by solving the CIP can be reliable for model application if the observation data can provide sufficient information to overcome the effects of both structure and observation errors. The following theorem provides a basis of finding a sufficient and robust design.

Theorem 5. Under the conditions of Theorem 3, if a design D is sufficient for identifying the WCP $(S_A,\tilde{\boldsymbol{q}}_A)$ of an admissible region Θ_A, then it should be sufficient for identifying all parameters with structure S_A or with a structure that is a simplification of S_A in the admissible region.

This theorem is a deduction of Theorem 3. When the structure error in (20) is decreased, δ will be increased and ε will be decreased. This means that the data obtained from the same design are sufficient for satisfying the requirement of identifiability with $d > d_0$ and $e < e_0$.

ALGORITHMS FOR SUFFICIENT AND ROBUST DESIGN

Experimental design depends heavily on how much prior information that is available. Different physical parameters in different fields may have different prior information. For example, if the unknown parameter is the hydraulic conductivity of an aquifer, prior information may be obtained from well logs, well tests, local pumping tests, tracer tests, and various geophysical measurements. If we can use more prior information, less information is needed from the designed experiment. A good design approach should demonstrate how prior information could be effectively and quantitatively incorporated into the design procedure. The design method presented in this paper requires that after transformation, analysis and judgment, all prior information can be integrated into the following form: The definition region can be divided into L zones $\{\Omega_i \mid i = 1,2,..L\}$ and the values of the unknown parameter are relatively homogeneous within each zone. Moreover, the upper and lower bounds of the unknown parameter $\boldsymbol{q}(\mathbf{x})$ at these zones can be estimated, i.e., we have two sets of numbers: $\{\underline{q}(\Omega_i)\}$ and $\{\bar{q}(\Omega_i)\}$, such that $\underline{q}(\Omega_i) \le q(\mathbf{x} \in \Omega_i) \le \bar{q}(\Omega_i)$ for all $i = 1,2,...,L$. With this information, we can use the following algorithm to find a candidate of the WCP:

Step 1. If the parameter value associated with zone (Ω_i) has not been assigned, then let it be its upper bound $\bar{q}(\Omega_i)$.

Step 2. Consider all neighboring zones (Ω_j) of (Ω_i). If the parameter value associated with zone (Ω_j) has not been assigned, then assign $\underline{q}(\Omega_j)$ to (Ω_j) when $\bar{q}(\Omega_i)$ is assigned to (Ω_i), or assign $\bar{q}(\Omega_j)$ to (Ω_j) when $\underline{q}(\Omega_i)$ is assigned to (Ω_i).

Obviously, the so determined parameter satisfies the condition of Theorem 1, i.e. the parameter values between neighboring zones have the maximum difference. To start the algorithm with different *i*, we can move from one candidate to another candidate of the WCP. Note that the WCP is dependent on sink/source terms and boundary conditions. Usually, we start from such zones where sink/source or inflow/outflow are involved.

Based on the concepts and theorems developed above, we present the following algorithm for judge the sufficiency and robustness of a design *D*:

Step 1. Compile all available prior information.

Step 2. Set a most possibly complex structure S_A and guess its WCP $\tilde{\boldsymbol{q}}_A$.

Step 3. Run the simulation model to generate a set of "observation data" $\mathbf{u}_D(S_A,\tilde{\boldsymbol{q}}_A)$ according to the designed excitation strengths, observation locations and times.

Step 4. Run the model for given model applications to generate a set of "prediction data" $\mathbf{g}_E(S_A,\tilde{\boldsymbol{q}}_A)$.

Step 5. Form the structure series (10) as in the solution of the GIP. From S_{m-1} to S_m, the most sensitive zone to the given model applications is selected to divide into two sub-zones, and the boundary between the two zones is determined by minimizing the fitting residual

$$RE_m = \min_{S_m \boldsymbol{q}_m} \left\| \mathbf{u}_D(S_A,\tilde{\boldsymbol{q}}_A) - \mathbf{u}_D(S_m,\boldsymbol{q}_m) \right\|_D \quad \text{s.t.} \quad \boldsymbol{q}_m \in \Theta(S_m) \tag{24}$$

Then, calculate the model application error $AE_m = \left\| \mathbf{g}_{\mathbf{E}}(S_A, \tilde{\boldsymbol{q}}_A) - \mathbf{g}_{\mathbf{E}}(S_m, \hat{\boldsymbol{q}}_m) \right\|_E$, where $(S_m, \hat{\boldsymbol{q}}_m)$ is the solution of (24).

Step 6. If $AE_m > \varepsilon$ and $RE_m > 2h$, increase *m* by 1 and repeat the above procedure to find S_{m+1}.

Step 7. When *m* increases, the value of AE_m decreases to zero and the value of RE_m decreases to until less than 2η. Thus, finally we must have the following cases: (1) $AE_m < \varepsilon$ but $RE_m \geq 2h$, (2) $AE_m < \varepsilon$ for all $RE_m < 2h$, and (3) $AE_m \geq \varepsilon$ but $RE_m < 2h$. According to Definition 5, if we only have the cases (1) and (2) during the optimization procedure, we can conclude that the design is sufficient. Otherwise, when Case (3) appears, the design is insufficient.

Since the so obtained design is sufficient for the WCP, according to Theorem 5, it should be a robust design.

THE OPTIMAL DESIGN

The principle of optimal design is either to minimize the experimental cost while the information provided by the experiment is sufficient, or to maximize the information with a certain budget. After we know how to judge the sufficiency of a design, we can define the optimal design problem as follows.

Definition 6. Let $f(D)$ be the cost of an experiment design *D*. The optimal design D^* is the solution of the following optimization problem:

$$D^* = \min_D f(D), \text{ s.t. } D \in \{D\} \tag{25}$$

where $\{D\}$ contains such designs that must be feasible and sufficient.

Problem (25) is a mixed integer-programming problem with very complex constraints, and thus it is very difficult to solve. In practice we often search a sub-optimal solution with less computation effort instead of solving (25). The following is a proposed heuristic procedure that might be useful for many environmental and geophysical systems.

Step 1. Collect all existing records on excitation locations and strengths, observation locations and frequencies.

Step 2. Use the procedure given above to test if the existing data are sufficient and robust. If the answer is "no", then go to the next step.

Step 3. Perform sensitivity analysis for existing observations. We calculate the sensitivities of $\mathbf{g}_E$ to the parameters for all *m* zones of S_m. The zone with the maximum sensitivity is selected to divide into two sub-zones. Now we calculate the sensitivity of each observation $u_{D,i}$ to the parameter of each zone. Locations where the observations only make contributions to the identification of those parameters that are not sensitive to model applications can be deleted from the further observation design.

Step 4. Perform sensitivity analysis for planned observations. Either increase the strength of excitation or increase the number of observation locations and frequencies depending on which one is more effective and feasible. New observation locations should be so selected that they make the maximum contribution to the identification of the most sensitive parameters to the model applications. This can be done by the adjoint state method (Sun, 1994).

Step 5. After new observations are planned in the last step, test the sufficiency of the new design and calculate its cost. Repeat Step 4 and Step 5 several times, a nearly cost-effective and feasible design may be found.

NUMERICAL EXAMPLE

In this section, the identification of hydraulic conductivity of an aquifer is used as an example to explain the presented methodology. Figure 1 shows a two-dimensional confined aquifer. It is assumed that the head is fixed to be 100 *m* at the boundary sections AB and CD and there is no flow through other boundary sections. The initial head is 100 *m* everywhere. The purpose of this study is to predict the steady state head values in three pumping wells: W_1, W_2 and W_3 when their pumping rates are 2000, 10000 and 4000 (m^3/day), respectively. The prior information available for the hydraulic conductivity includes: (1) It consists of 24 homogeneous zones, and (2)

the upper and lower bounds of each zone can be estimated. Their ranges vary from 5 to 50 $mday^{-1}$.

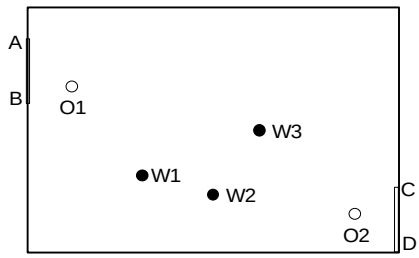


Figure 1. The flow field

One candidate of the WCP is shown in Figure 2, which is obtained by the procedure described in the last section with consideration of the flow conditions. With this WCP, the values of steady state head in the three pumping wells are $W_1 = 78.60m$, $W_2 = 66.81m$ and $W_3 = 74.02m$, respectively. These values, of course, will change significantly when the distribution of hydraulic conductivity changes.

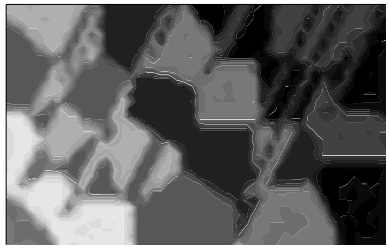


Figure 2. The worst-case parameter

Now we want to design a pumping test to identify the hydraulic conductivity so that the accuracy requirement of model prediction, $e = 1.0\,m$, can be reached, i.e., the norm of the differences between the model predicted heads and the real heads in the three wells must be less than 1 meter. Note that to compare with the large drawdown in these wells (more than 20*m*), this accuracy requirement is very high. The upper bound of observation error is assumed to be $h = 0.1\,m$.

Using the adjoint state method for sensitivity analysis, we can find that the local hydraulic conductivity around Well 2 plays the most important role for the short-term pumping test. For the long-term steady state, on the other hand, the values of hydraulic conductivity along the inflow boundaries AB and CD play the most important role. Using the sensitivity equation method, we can find that O_1 and O_2 in Figure 1 are the best observation locations for identifying these values. The sensitivity analysis methods based on the forward or reverse modes of auto-differentiation are important tools for the presented design process.

The first pumping test design D_1 consists of (1) pumping $1{,}000\,m^3 day^{-1}$ from W_2, (2) three observation locations at W_1, W_2 and W_3, (3) five observation times at *t* = 0.01, 0.05, 0.1, 0.5 and 1.0 (*day*). Following the steps described in the last section, we find $RE_1 = 0.09\,m$, which is less than $2h = 0.2\,m$, when $K_{1,1} = 27.29\ mday^{-1}$, but $AE_1 = 4.1m$, which is larger than the accuracy requirement $e = 1.0\,m$. Therefore, we can conclude that the design D_1 is insufficient. The second pumping test design D_2 is the same as D_1 but increasing the pumping rate in W_2 to $2{,}000\,m^3/day$. Repeating the steps described in the last section, we find $RE_1 = 0.15\,m < 2h$ when $K_{1,1} = 28.02\ m/day$, but $AE_1 = 4.60m > e$. Therefore, the design D_2 is insufficient too. The design D_3 is formed by adding two observation wells O_1 and O_2 to the design D_2. Unfortunately, it is still insufficient. Design D_4 is formed from the design D_3 by pumping 500 from W_1, 2000 from W_2 and 1000 (m^3/day) from W_3. With D_4, we find $RE_1 = 0.31 > 2h$, when $K_{1,1} = 25.28$, and $AE_1 = 3.01m > \varepsilon$. This means that the design allows us to identify a two-zone structure. Under the optimized two-zone structure, we find $RE_2 = 0.18 < 2h$, when $K_{2,1} = 33.13$, and $K_{2,2} = 11.39$ (m/day), but $AE_2 = 1.87m > \varepsilon$. Therefore the design D_4 is still insufficient. A sufficient design, D_5, has been found which is based on D_4 but increasing the period of pumping test to 3 days. The five observation times are *t* = 0.05, 0.1, 0.5, 1.0 and 3.0 (*days*). In this case, during the search of the best fitting two-zone parameter (including both pattern and values), we always have either ($AE_2 < e$ and $RE_2 < 2h$) or ($AE_2 > e$ and $RE_2 > 2h$). The best fitting parameter values are $K_{2,1} = 17.99\,m/day$ and $K_{2,2} = 34.61\,m/day$, for which $RE_2 = 0.18$ and

$AE_2 = 0.52$. When the Zone 1 of the two-zone structure is further divided into two zones, we always have $AE_2 < e$ and $RE_2 < 2h$ during the procedure of pattern optimization. The best three-zone pattern is shown in Figure 3 and the best fitting parameter values are $K_{3,1} = 20.27$ *m/day*, $K_{3,2} = 33.27$ *m/day*, and $K_{3,3} = 7.02$ *m/day*. With this parameter, the steady state heads in the three wells predicted by the model are $h(W_1) = 77.82m$, $h(W_2) = 67.24m$, and $h(W_3) = 74.17m$. The identified parameter is thus equivalent to the WCP. Note that the boundaries of the three zones in Figure 3 do not represent any real physical boundaries of discontinuity. What we found is that the real hydraulic conductivity distribution can be replaced by such a three-zone structure for model prediction.

Figure 3. An equivalent structure to the WCP

To verify the robustness of the design D_5, two randomly generated hydraulic conductivity distributions are tested. Their values in each zone are randomly specified within the range of the given upper and lower bounds. First, we run the simulation model to obtain the heads at the specified times and locations according to the design D_5, then a set of observation errors with norm $\eta = 0.1m$ is added to them to obtain a set of "observation data". For the first test case, we found that the condition $RE < 2h$ is satisfied when the number of zones is increased to 3 during the solution of GIP. For the second test case, the condition $RE < 2h$ is satisfied when the number of zones is increased just to 2. According to our theory, the reliability requirement of prediction $e = 1.0\,m$ must be satisfied by these equivalent parameters. From Table 2 we can find that the heads in all three wells can be predicted with very high accuracy indeed. Note that these results can be expected at the design stage based on the new methodology.

Table 2. Results of the two test cases

	Case 1		Case 2	
Head	Real	Model predicted	Real	Model predicted
W_1	79.84	79.67	80.43	80.57
W_2	70.48	70.51	72.13	72.05
W_3	75.05	75.04	75.89	76.18

CONCLUSIONS

In this paper, we have introduced a new methodology for identifying a distributed parameter when its dimension is high and its structure is unknown, such as the hydraulic conductivity in groundwater modeling. The basic idea is to solve a weak inverse problem, the GIP, to find an equivalent parameter that can give almost the same results for model application as what the true parameter does. The weak solution has the minimum complexity in structure and thus it needs minimum data to identify.

We have found that the error of structure reduction can be calculated effectively by solving a CIP if we know the worst-case parameter (WCP). We have proved that the WCP is always located at one vertex of the admissible region of the unknown parameter. A set of sufficient conditions for structure identifiability is presented that requires the information provided by the observation data be able to overcome the impacts of both structure and observation errors. We have proved that if an experimental design is sufficient for identifying the WCP then it must be sufficient for identifying all other parameters in the admissible region, and thus, it is a robust design. Based on these new concepts and theorems, we have presented algorithms for determining the WCP from prior information, for judging the sufficiency of a design, and finally for finding a cost-effective, sufficient and robust design.

A numerical example is given, in which the unknown hydraulic conductivity with complex structure is replaced by a very simple but equivalent structure for the given model application. In this example, we have shown how the WCP can be found from prior information and flow conditions and how a sufficient and robust design can be found through a heuristic process.

Acknowledgement: This material is based upon work supported by NSF under award EAR-0001082 and the U. S. Army Research Office under grant DAAD19-01-1-0510.

REFERENCES

Beck, J. V., B. Blackwell, and St. Jr. C. R. Clair, *Inverse heat conduction: ill-posed problems*, Wiley, New York, 1985.

Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 2. Uniqueness, stability, and solution algorithms, *Water Resour. Res.*, 22(2), 211-227, 1986.

Chavent, G., M. Dupuy, and P. Lemonnier, History matching by use of optimal theory, *Soc. Pet. Eng. J.*, 15(1), 74-86, 1975.

Chavent, G., Identifiability of parameters in the output least square formulation. In *Identifiability of parametric models*, edited by E. Walter, Pergamon, New York, 1987.

Chavent, G., G. Papanicolaou, P. Sacks, and W. W. Symes, *Inverse Problems in Wave Propagation*, Springer-Verlag, New York, 1995.

Engl, H. W., M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Kluwer Academic Publishers, 1996.

Grimstad, A. A., and T. Mannseth, Nonlinearity, scale, and sensitivity for parameter estimation problems. *SIAM J. Sci. Comput.*, 21, 2096-2113, 2000.

Isakov, V., *Inverse problems for partial differential equations*, Springer-Verlag, New York, 1998.

Jacquard, P., and C. Jain, Permeability distribution from fiel pressure data, Soc. Pet. Eng. J., 5(4), 281-294, 1965.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resources Research*, 19(3), 677-690, 1983.

McLaughlin, D., and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resour. Res.*, 32(5), 1131-1162, 1996.

Neuman, S. P., Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res.*, 9(4), 1006-1021, 1973.

Qureshi, Z. H., t. s. Ng, and G. C. Goodwin, Optimum experimental design for identification of distributed parameter systems, *Int. J. of Control*, 31, 21-29, 1980.

Rafajlowicz, E., Optimum choice of moving sensor trajectories for distributed parameter system identification. *Int. J. of Control*, 43, 1441-1451, 1986.

Stark, P. B., Inverse problems as statistics. In *Surveys on Solution Methods for Inverse Problems*, edited by D. Colton, H. W. Engl, A. K. Louis, J. R. McLaughlin, and W. Rundell, Springer Wien, New York, 2000.

Sun, N.-Z., *Inverse problems in groundwater modeling*, Kluwer Academic Publishers, 1994.

Sun, N.-Z., 2002, Parameter structure identification and robust experimental design for distributed parameter systems: a new methodology, submitted to *Int. Journal of Control.*

Sun, N.-Z., and W. W.-G. Yeh, Identification of parameter Structure in Groundwater Inverse Problems, *Water Resour. Res.*, 21(6): 869-883, 1985.

Sun, N.-Z., and W. W.-G. Yeh, Coupled inverse problems in groundwater modeling, 2, identifiability and experimental design. *Water Resour. Res.*, 26(10), 2527-2540, 1990.

Sun, N.-Z., S. Yang, and W.W-G. Yeh, A proposed stepwise regression method for model structure identification. *Water Resour. Res.*, 34(10), 2561-2572, 1998.

Sun, N.-Z., and A. Y. Sun, Parameter identification of environmental systems, Chapter 8, in *Environmental Fluid Methods--Theories and Applications*, ASCE PUBLICAION, 2002.

Tsai, F. T-C., N.-Z. Sun, and W. W.-G. Yeh, Groundwater parameter structure identification and parameterization with natural neighbor method, submitted to *Water Resour. Res.*, 2002.

Ucinski, D., Optimal sensor location for parameter estimation of distributed processes, *Int. J. of Control*, 73, 1235-1248, 2000.

Woodbury, A. D., and L. Smith, Simultaneous inversion of hydrogeologic and thermal data, 2, Incorporation of thermal data, Water Resour. Res., 24(3), 356-372, 1988.

Wouwer, A. V., N. Point, S. Porteman, and M. Remy, An approach to the selection of optimal sensor locations in distributed parameter systems. *J. of Process Control*, 10, 291-300, 2000.

Yeh, W. W.-G., and Y. S. Yoon, Parameter identification with optimum dimension in parameterization, Water Resour. Res., 17(3), 664-672, 1981.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

NUMERICAL METHODS AND REGULARIZATION TECHNIQUES FOR THE SOLUTION OF ILL-POSED PROBLEMS

Anatoly G. Yagola
Department of Mathematics, Faculty of Physics, Moscow State University, Moscow 119899 Russia
e-mail: yagola@inverse.phys.msu.su

Valery N. Titarenko
e-mail: ill-posed@mail.ru

ABSTRACT

Consider linear ill-posed problems with *a priori* information about their unknown solutions. We discuss what one should know about the given data to construct a regularizing algorithm. It is shown how properties of regularizing algorithms depend on known properties of the solutions. The uniform and *a posteriori* errors of an approximate solution, rates of convergence for regularizing algorithms are considered for problems on sourcewise represented and compact sets of exact solutions. For various *a priori* restrictions to the exact solution several numerical methods are offered to construct regularizing algorithms. These methods are applied to inverse problems in astrophysics, electronic microscopy and vibrational spectroscopy.

INTRODUCTION

The majority of all problems investigated in a modern science are inverse problems. When a researcher has enough information about properties of an unknown solution of a problem, then he or she almost always may £nd a solution that are stable in relation to perturbations of input data. It is well known that a century ago many scientists thought that only such stable problems are in nature and all other problems are only model mathematical ones. Therefore, to study these "real" problems J. Hadamard offered a notion of a well-posed problem in [1].

Let us consider a linear inverse problem written in the form of the operator equation

$$A\bar{z} = \bar{u} \qquad \bar{z} \in Z, \bar{u} \in U \tag{1}$$

where Z, U are normed spaces. The problem (1) is called well-posed on the class of its "admissible" data if for any pair $\{A, \bar{u}\}$ from the set of "admissible" data the solution of (1):

1. exists,
2. is unique,
3. continuously depends on errors in A and $\bar{u}$ (is stable).

Stability means that if instead of $\{A, \bar{u}\}$ we are given "admissible" $\{A_h, u_\delta\}$ such that $\|A_h - A\| \leqslant h$, $\|u_\delta - \bar{u}\| \leqslant \delta$, the approximate solution converges to the exact one as $h, \delta \to 0$. The numbers h and δ are error estimates for the approximate data $\{A_h, u_\delta\}$ of the problem (1) with the exact data $\{A, \bar{u}\}$. Denote $\eta = (h, \delta)$. If at least one of the mentioned requirements is not met, then the problem (1) is called ill-posed. Remark that the most important requirement is the third one, since the others may often be made just. For example, the £rst requirement is ful£lled if instead of the solution of (1) we introduce some generalized solution. If one makes additional restrictions for the considered problem, then the problem (1) often becomes a problem with a unique solution. Regretfully, stability of the problem (1) depends on properties of the given spaces Z and U, which can not be changed by other spaces in practice.

As a generalized solution, it is often taken the so-called normal pseudosolution (a solution in the sense of the least-squares method with a minimum norm or sometimes with a minimum distance from a given £xed element). This solution $\tilde{z}$ exists and is unique for any exact data of the problem (1) if $A \in L(Z, U)$, $\bar{u} \in R(A) \oplus R^{\perp}(A)$, $\tilde{z} = A^{+}\bar{u}$. Here $R(A)$ and $R^{\perp}(A)$ denote the ranges of the operator A and its orthogonal complement in U, and A^{+} stands for the operator pseudoinverse to A. See, e.g., [2] for details. In the paper we £nd $\bar{z}$ as a normal pseudosolution, i.e. $\bar{z} = \tilde{z}$.

As opposed to well-posed problems ill-posed ones are in some sense underdetermined problems. This means that a researcher has not enough information to solve an ill-posed problem. Therefore, he or she should assume that the solution has additional properties. Some assumptions make the considered problem well-posed as it will be shown for the compact sets. Unfortunately, for many problems these assumptions often help to construct only so-called regularizing algorithms.

In spite of the existence and uniqueness of a normal pseudosolution $\tilde{z}$ for any admissible data, the problem of its £nding, as well as the problem (1) itself, may be unstable with respect to perturbations of A and $\bar{u}$. Thus, it is important to answer the question, what means to "solve" such an unstable problem. Tikhonov answered this question in his famous de£nition of a regularizing algorithm [3, 4]. To solve an ill-posed problem means to produce a map (regularizing algorithm) $R(A_h, u_\delta, \eta)$ such that

1. brings an element $z_\eta = R(A_h, u_\delta, \eta)$ into correspondence with any data $\{A_h, u_\delta, \eta\}$, $A_h \in L(Z,U)$, $u_\delta \in U$, of the problem (1);

2. has the convergence property $z_\eta \to \bar{z} = A^+\bar{u}$ as $\eta \to 0$, $\bar{u} \in R(A) \oplus R^\perp(A)$.

A mathematical problem is (Tikhonov) regularizable if there exists a regularizing algorithm. It is evident that a well-posed problem is regularizable. Unfortunately, regularizing algorithms do not exist for all mathematical problems. Therefore, we may divide all inverse problems into three groups:

1. well-posed problems,

2. ill-posed regularizable problems,

3. ill-posed nonregularizable problems.

More than 40 years ago the question arose whether it is possible to construct a regularizing algorithm that would not depend exactly on the estimates of errors h and δ. Regretfully, such an approach can solve only well-posed problems.

Theorem 1 [5]: Let $R(A_h, u_\delta)$ be a map of the set $L(Z,U) \otimes U$ into Z. If $R(A_h, u_\delta)$ is a regularizing algorithm (not depending explicitly on η), then the map $P(A, \bar{u}) = A^+\bar{u}$ is continuous on its domain $L(Z,U) \otimes (R(A) \oplus R^\perp(A))$.

Proof The second condition in the de£nition of regularizing algorithm implies in the equality $R(A, \bar{u}) = A^+\bar{u} = P(A, \bar{u})$ valid for each $(A, \bar{u}) \in L(Z,U) \otimes (R(A) \oplus R^\perp(A))$ and the convergence $P(A_h, u_\delta) = R(A_h, u_\delta) \to A^+\bar{u} = P(A, \bar{u})$ when $h, \delta \to 0$ valid for any $(A, \bar{u}) \in L(Z,U) \otimes (R(A) \oplus R^\perp(A))$, $(A_h, u_\delta) \in L(Z,U) \otimes (R(A) \oplus R^\perp(A))$. Therefore, the map $P(A, u)$ is continuous on $L(Z,U) \otimes (R(A) \oplus R^\perp(A)) \subset L(Z,U) \otimes U$. The theorem is proved.

It is clear from Theorem 1 that a regularizing algorithm not using h and δ explicitly can only exist for problems (1) well-posed on the set of the data $L(Z,U) \otimes (R(A) \oplus R^\perp(A)) \subset L(Z,U) \otimes U$. The theorem generalized the assertion proved by Bakushinskii in [6].

Let us discuss the very principal question: is it possible to estimate an error of an approximate solution of an ill-posed problem? Regretfully, the answer is negative. The main and very important result was obtained by Bakushinskii (see, [7] or [8]). For simplicity we assume $h = 0$, i.e., $A_h = A$. Let $R(u_\delta, \delta)$ be a regularizing algorithm that depends on δ, u_δ only. Denote by $\triangle(R, \delta, \bar{z}) = \sup\{\|R(u_\delta, \delta) - \bar{z}\| : \forall u_\delta \in U, \|A\bar{z} - u_\delta\| \leqslant \delta\}$ the error of a solution of the ill-posed problem (1) at the point $\bar{z}$ using the algorithm R. If the problem (1) is regularizable by a continuous map R and there is an error estimate, which is uniform on D,

$$\sup\{\triangle(R, \delta, \bar{z}) : \bar{z} \in D\} \leqslant \varepsilon(\delta) \to 0$$

then the restriction of A^{-1} to $AD \subset U$ is continuous on AD.

Usually the accuracy of the approximate solution $z_\delta = R(u_\delta, \delta)$ of the problem (1) could be estimated as

$$\|z_\delta - \bar{z}\| \leqslant K\varphi(\delta) \tag{2}$$

where K does not depend on δ and the function $\varphi(\delta)$ de£nes the convergence rate of z_δ to $\bar{z}$.

Note that pointwise and uniform error estimates (2) should be distinguished. For pointwise estimates the exact solution $\bar{z}$ is £xed, the constant K and the function $\varphi(\delta)$ depend on $\bar{z}$. For the case of uniform estimates the inequality (2) is just for some set M of exact solutions $\bar{z}$. Then K and $\varphi(\delta)$ depend on properties of the set M. Since the exact solution $\bar{z}$ is unknown, pointwise error estimates have no signi£cant sense.

We consider the results obtained by Vinokurov in [9]. Let A be a linear continuous injective operator acting in Banach space Z and the inverse operator A^{-1} is unbounded on its domain $D(A^{-1})$. Suppose that $\varphi(\delta)$ is an arbitrary positive function such that $\varphi(\delta) \to 0$ as $\delta \to 0$, and R is an arbitrary method to solve the problem. Then the following equality holds for elements $\bar{z}$ except maybe for a £rst category set in Z:

$$\limsup_{\delta \to 0} \left\{ \frac{\triangle(R, \delta, \bar{z})}{\varphi(\delta)} \right\} = \infty$$

This means that a uniform error estimate can only exist on a £rst category subset in Z.

In uniform estimates the rate of accuracy of an approximate solution $\varphi(\delta)$ does not depend on an exact solution. Therefore uniform accuracy estimates are widely spread in the theory of ill-posed problems. However, uniform accuracy estimates do not exist on any set M.

A compact set is a typical example of the £rst category set in a normed space Z. For this set special regularizing algorithms may be used [8, 2] and a uniform error estimation may be constructed.

Clearly, a uniform error estimate exists only for well-posed problems. For general ill-posed problems we can't £nd an error of an approximate solution z_η and estimate the convergence rate of z_η to the exact solution $\bar{z}$. Fortunately, for some ill-posed problems it is possible to £nd a so-called *a posteriori* error estimation. Following [10], for the case, when there is an exact injective operator A with closed graph and Z is a σ-compact space, we introduce a function $\kappa(u_\delta, \delta)$ such that $\forall \bar{z} \in Z$ $\exists \delta(\bar{z}) > 0, \forall \delta \in (0, \delta(\bar{z})], \forall u_\delta \in U \ \|u_\delta - \bar{u}\| \leqslant \delta$: $\|\bar{z} - R(u_\delta, \delta)\| \leqslant \kappa(u_\delta, \delta)$. The function $\kappa(u_\delta, \delta)$ is *a posteriori* error estimation for the problem (1), if $\kappa(u_\delta, \delta) \to 0$ as $\delta \to 0$.

THE GENERALIZED DISCREPANCY METHOD

Tikhonov in his papers [3, 4] not only clearly de£ne the meaning of solving an ill-posed problem (1), but also give a practical regularizing algorithm to solve (1).

We follow [8]. Let Z, U be Hilbert spaces, $D \subset Z$ be a closed convex set of *a priori* constraints such that $0 \in D$, A, A_h be linear operators. Given a set of the data $\{A_h, u_\delta, \eta\}$ we introduce the Tikhonov's functional:

$$M^\alpha[z] = \|A_h z - u_\delta\|^2 + \alpha \|z\|^2 \qquad (3)$$

where $\alpha > 0$ is a regularization parameter.

Consider the following extreme problem:

$$\inf\{M^\alpha[z] : z \in D\} \qquad (4)$$

For any $\alpha > 0$, $u_\delta \in U$ and bounded linear operator A_h the problem (4) is solvable and has a unique solution $z_\eta^\alpha \in D$.

The idea to construct a regularizing algorithm using the extreme problem (4) for $M^\alpha[z]$ consists of constructing of a function $\alpha = \alpha(\eta)$ such that $z_\eta^{\alpha(\eta)} \to \bar{z}$ as $\eta \to 0$. We may £nd a regularization parameter *a priori* or *a posteriori*. In [8] it is shown that if A is a injective operator, $\bar{z} \in D$ and $\alpha(\eta) \to 0$, $(h + \delta)^2/\alpha(\eta) \to 0$ as $\eta \to 0$, then $z_\eta^{\alpha(\eta)} \to \bar{z}$ as $\eta \to 0$, i.e., there is the *a priori* choice of α.

De£ne the incompatibility measure of (1) with the approximate data on D as

$$\mu_\eta(u_\delta, A_h) = \inf\{\|A_h z - u_\delta\| : z \in D\}$$

Assume that the incompatibility measure can be computed with an error $\kappa > 0$, i.e., instead of $\mu_\eta(u_\delta, A_h)$ there is $\mu_\eta^\kappa(u_\delta, A_h)$ such that

$$\mu_\eta(u_\delta, A_h) \leqslant \mu_\eta^\kappa(u_\delta, A_h) \leqslant \mu_\eta(u_\delta, A_h) + \kappa$$

Let us introduce the so-called generalized discrepancy:

$$\rho_\eta^\kappa(\alpha) = \|A_h z_\eta^\alpha - u_\delta\|^2 - \big(\delta + h\|z_\eta^\alpha\|\big)^2 - \big(\mu_\eta^\kappa(u_\delta, A_h)\big)^2$$

The generalized discrepancy $\rho_\eta^\kappa(\alpha)$ is continuous and monotonically non-decreasing for $\alpha > 0$.

Now we state the generalized discrepancy principle to choose the regularization parameter:

1. If the condition $\|u_\delta\|^2 > \delta^2 + \big(\mu_\eta^\kappa(u_\delta, A_h)\big)^2$ is not ful£lled, then we take $z_\eta = 0$ as an approximate solution of (1);

2. If the condition $\|u_\delta\|^2 > \delta^2 + \big(\mu_\eta^\kappa(u_\delta, A_h)\big)^2$ is ful£lled, then the generalized discrepancy has a positive zero α^* and $z_\eta = z_\eta^{\alpha^*}$.

If A is a injective operator, then $\lim_{\eta \to 0} z_\eta = \bar{z}$. Otherwise, $\lim_{\eta \to 0} z_\eta = z^*$, where z^* is the normal solution of (1), i.e., $\|z^*\| = \inf\{z \in D : Az = \bar{u}\}$.

It is known that we can put $\mu_\eta^\kappa(u_\delta, A_h) = 0$ even if $u_\delta \notin A_h D$. However, we should change the generalized discrepancy principle as follows.

1. If $\|u_\delta\| > \delta$ is not ful£lled, then $z_\eta = 0$;

2. If $\|u_\delta\| > \delta$ is ful£lled, then:

   (a) if there is an $\alpha^* > 0$, which is a zero of the function $\rho_\eta(\alpha)$, then $z_\eta = z_\eta^{\alpha^*}$;

   (b) if $\rho_\eta(\alpha) > 0$ for all $\alpha > 0$, then $z_\eta = \lim_{\alpha \to 0} z_\eta^\alpha$.

For the case, when A, A_h are bounded linear operators, D is a closed convex set containing the

point 0, $\bar{z} \in D$, it is proved in [8] that the generalized discrepancy principle are equivalent to the generalized discrepancy method: £nd

$$\inf\Big\{\|z\| : z \in D, \quad \|A_h z - u_\delta\|^2 \leqslant (\delta + h\|z\|)^2 + \big(\mu_\eta^\kappa(u_\delta, A_h)\big)^2\Big\}$$

You may £nd the generalized principles of discrepancy, quasisolutions and smoothing functional for linear incompatible and non-linear general ill-posed problems in [8, 2].

For simplicity suppose $A_h = A$. Consider the sets represented in the form $M_r = \{z : z = Bv, v \in V, \|v\| \leqslant r\}$, where V is an auxiliary Hilbert space, $B : V \to Z$ is a linear, injective and compact operator, r is a £xed parameter. For a method R of solving (1) we de£ne

$$\triangle(R, \delta, r) = \sup\{\|R(u_\delta, \delta) - \bar{z}\| : \bar{z} \in M_r, \|u_\delta - \bar{u}\| \leqslant \delta\} \quad (5)$$

Then, for a class $\mathcal{R}$ of all possible methods R for solving (1) the optimal accuracy is

$$\triangle_{\text{opt}}(\delta, r) = \inf\{\triangle(R, \delta, r) : R \in \mathcal{R}\}$$

A method R is said to be optimal in order on sets M_r if the following inequality holds for its accuracy (5):

$$\frac{\triangle(R, \delta, r)}{\triangle_{\text{opt}}(\delta, r)} \leqslant k = \text{const}$$

as $\delta \to 0$ and k does not depend on δ, r.

The generalized principles of discrepancy, quasisolutions and smoothing functional are optimal in order on sets M_r with $k = 2$ [2].

Let us apply the generalized discrepancy principle to solve a model example of an inverse problem for the heat conduction equation

$$\begin{cases} w_t = a^2 w_{xx} \quad x \times t \in (0, l) \times (0, T) \\ w(0, t) = 0 \\ w(l, t) = 0 \end{cases} \quad (6)$$

There is a function $u_\delta(\xi) \equiv w(\xi, T) \in L^2[0, l]$, we want to £nd $z(x) \equiv w(x, 0) \in W_1^2[0, l]$ such that $z(x) \to \bar{z}(x)$ as $\eta \to 0$. We may write that

$$\|u(\xi)\|^2 = \int_0^l |u(\xi)|^2 d\xi,$$

$$\|z(x)\|^2 = \int_0^l \left(|u(x)|^2 + \left|\frac{\partial u(x)}{\partial x}\right|^2\right) dx$$

The problem may be written in the form of integral equation

$$u(\xi) = \int_0^l G(\xi, x, T) z(x) dx$$

where $G(\xi, x, t)$ is the Green function:

$$G(\xi, x, t) = \frac{2}{l} \sum_{n=1}^{+\infty} \sin\left(\frac{\pi n \xi}{l}\right) \sin\left(\frac{\pi n x}{l}\right) \times \exp\left(-\left(\frac{\pi n a}{l}\right)^2 t\right)$$

The problem is solved for the parameters $a = 1.0$, $T = 0.1$, $l = 1.0$, the function $u_\delta(\xi)$ is taken such that $\delta = 0.05 \cdot \|\bar{u}\|$. In Figure 1 there are the exact function $\bar{z}(x)$ and the found solution $z_\eta(x)$.

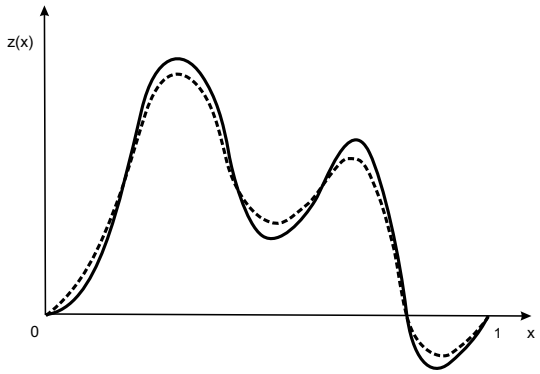


Figure 1. The exact solution $\bar{z}(x)$ (——) and the approximate solution $z_\eta(x)$ ($- - - -$) for the generalized discrepancy method.

NUMERICAL METHODS

Consider a Tikhonov's functional $M^\alpha[z]$ written as (3), which is a strongly convex functional in a Hilbert space. We recall that a necessary and suf£cient condition for z_η^α to be a minimum point of $M^\alpha[z]$ on a set D of *a priori* constraints is that

$$\big((M^\alpha[z_\eta^\alpha])', z - z_\eta^\alpha\big) \geqslant 0 \qquad \forall z \in D$$

If z_η^α is an interior point of D, then this condition takes the form $(M^\alpha[z_\eta^\alpha])' = 0$, or

$$A_h^* A_h z_\eta^\alpha + \alpha z_\eta^\alpha = A_h^* u_\delta \quad (7)$$

Thus, in this case we may solve the Euler equation (7) instead of minimizing $M^\alpha[z]$.

To solve ill-posed problems it is usually necessary to approximate the initial, often in£nite dimensional, problem by a £nite dimensional one, for which numerical algorithms and computer programs have been devised.

Consider the Fredholm integral equation of the £rst kind

$$Az = \int_a^b K(x,s)z(s)\,ds = u(x) \quad c \leqslant x \leqslant d$$

We take $U = L^2[c,d]$, $Z = W_1^2[a,b]$. Assume that $K(x,s)$ is a real-valued function de£ned and continuous on $\Pi = [a,b] \times [c,d]$. Suppose that instead of $K(x,s)$ we know a function $K_h(x,s)$ such that $\|K_h - K\|_{L^2(\Pi)} \leqslant h$. Then $\|A_h - A\|_{W_1^2 \to L^2} \leqslant h$, where A_h is the integral operator with kernel $K_h(x,s)$.

Let us choose grids $\{s_j\}_1^m \subset [a,b]$, $\{x_i\}_1^m \subset [c,d]$. The £nite dimensional operator is a linear operator with matrix $\hat{A} = \{a_{ij}\}$. The simplest version of the approximation is given by the formulas

$$a_{ij} = \begin{cases} K_h(x_i,s_j), & j = \overline{2,n-1} \\ K_h(x_i,s_j)/2, & j = 1,n \end{cases}$$

for $i = \overline{1,m}$.

For simplicity we use uniform grids with steps h_s and h_x. We put $z(s_j) = z_j$, $u_\delta(x_i) = u_i$, $\hat{z} = (z_1,\ldots,z_n)$, $\hat{u} = (u_1,\ldots,u_m)$. Using the rectangle formula to approximate the integrals, we obtain

$$\hat{M}^\alpha[\hat{z}] = \sum_{i=1}^m \left[\sum_{j=1}^n a_{ij} z_j h_s - u_i\right]^2 h_x + \alpha \sum_{j=1}^n z_j^2 h_s + \alpha \sum_{j=2}^n \frac{(z_j - z_{j-1})^2}{h_s}$$

Set $b_{jk} = h_x \sum_{i=1}^n a_{ik}a_{ij}$, $f_j = h_x \sum_{i=1}^m a_{ij}u_i$. Thus, we arrive at the problem of solving the system of equations

$$\hat{B}^\alpha \hat{z} = \hat{B}\hat{z} + \alpha \hat{C}\hat{z} = \hat{f} \tag{8}$$

where $\hat{B} = \{b_{jk}\}$, $\hat{f} = (f_1,\ldots,f_n)$ and $\hat{C}$ you may £nd in [8].

We can use various numerical methods to solve the system of linear equations (8). Moreover, we should take into account that the matrix $\hat{B}^\alpha$ is symmetric and positive de£nite. Therefore it is possible to very ef£cient methods to solve (8).

The square-root method is one such method. We may write $\hat{B}^\alpha = (\hat{T}^\alpha)^*\hat{T}^\alpha$, where $\hat{T}^\alpha$ is an upper triangular matrix. The system (8) takes the form

$$(\hat{T}^\alpha)^*\hat{T}^\alpha\hat{z}^\alpha = \hat{f}$$

Introducing the notation $\hat{y}^\alpha = \hat{T}^\alpha\hat{z}^\alpha$, we obtain two equations

$$(\hat{T}^\alpha)^*\hat{y}^\alpha = \hat{f}, \qquad \hat{T}^\alpha\hat{z}^\alpha = \hat{y}^\alpha$$

Each of these equations can be elementary solved, since each involves a triangular matrix.

Let write the equation (8) in the form of Euler equation

$$(\hat{A}_h^*\hat{A}_h + \alpha\hat{C})\hat{z}^\alpha = \hat{A}_h^*\hat{u}$$

Using the square-root method, the tridiagonal matrix $\hat{C}$ can be written as $\hat{C} = \hat{S}^*\hat{S}$, where $\hat{S}$ is bidiagonal. Changing to $\hat{y}^\alpha = \hat{S}\hat{z}^\alpha$, we obtain

$$(\hat{A}_h^*\hat{A}_h + \alpha\hat{C})\hat{S}^{-1}\hat{y}^\alpha = \hat{A}_h^*\hat{u}$$

Multiplying the lefthand side by $(\hat{S}^{-1})^*$, we obtain

$$(\hat{D}^*\hat{D} + \alpha\hat{E})\hat{y}^\alpha = \hat{D}^*\hat{u}, \qquad \hat{D} = \hat{A}_h\hat{S}^{-1}$$

where $\hat{E}$ is the identity matrix. The matrix $\hat{D}$ may be written as $\hat{D} = \hat{Q}\hat{P}\hat{R}$, where $\hat{Q}$ is an orthogonal $(m \times m)$-matrix, $\hat{R}$ is an orthogonal $(n \times n)$-matrix, and $\hat{P}$ is a right bidiagonal $(m \times n)$-matrix.

Now we make change of variables $\hat{x}^\alpha = \hat{R}\hat{y}^\alpha$ and obtain $(\hat{R}^*\hat{P}^*\hat{Q}^*\hat{Q}\hat{P}\hat{R} + \alpha\hat{E})\hat{R}^{-1}\hat{x}^\alpha = \hat{D}^*\hat{u}$, or $(\hat{P}^*\hat{P} + \alpha\hat{E})\hat{x}^\alpha = \hat{R}\hat{D}^*\hat{u} = \hat{f}$. The matrix $\hat{P}^*\hat{P}$ is tridiagonal, and the latter equation can be solved by the sweep method. The operator $\hat{S}^{-1}\hat{R}^{-1}$ realizes the inverse transition form $\hat{x}^\alpha$ to $\hat{z}^\alpha$.

Of course, to minimize $\hat{M}^\alpha[\hat{z}]$ one may use the method of conjugate gradients.

In [8] you may £nd programs implementing the considered algorithms.

SOURCEWISE REPRESENTED SETS

Consider the operator equation (1), where $A : Z \to U$ is a linear bounded injective operator, Z and U are normed spaces. Assume the next *a priori* information: the exact solution $\bar{z}$ is sourcewise represented with a linear compact operator B acting from a re¤exive Banach space V into Z:

$$\bar{z} = B\bar{v} \qquad \bar{z} \in Z, \bar{v} \in V \tag{9}$$

For reasons of simplicity we suppose that the operator B is injective, the operator A is known exactly and instead of $\bar{u}$ there is u_δ such that $\|u_\delta - \bar{u}\| \leqslant \delta$.

We want to construct a regularizing algorithm to solve (1) with *a priori* information (9) using the data $\{u_\delta, \delta\}$. Set $n = 1$ and de£ne the set Z_n:

$$Z_n = \{z \in Z : z = Bv, v \in V, \|v\| \leqslant n\}$$

Then we minimize the discrepancy $F(z) = \|Az - u_\delta\|$ on the set Z_n. If $\min\{\|Az - u_\delta\| : z \in Z_n\} \leqslant \delta$, then the solution is found. We denote $n(\delta) = n$. Otherwise, we change n to $n + 1$ and reiterate the process. If $n(\delta)$ is found, then we de£ne the approximate solution $z_{n(\delta)}$ of (1) as an arbitrary solution of the inequality

$$\|Az - u_\delta\| \leqslant \delta \quad z \in Z_{n(\delta)}$$

Theorem 2 [11]: The process described above converges: $n(\delta) < +\infty$. There exists $\delta_0 > 0$ (generally speaking, depending on $\bar{z}$) such that $n(\delta) = n(\delta_0)$ for any $0 < \delta < \delta_0$. Approximate solutions $z_{n(\delta)}$ strongly converge to $\bar{z}$ as $\delta \to 0$.

Proof The ball $V_n = \{v \in V : \|v\| \leqslant n\}$ is a bounded closed set in V. The set Z_n is a compact in Z for any n, since B is a compact operator. Due to Weierstrass theorem the continuous functional $F(z)$ attains its exact lower bound on Z_n.

Clearly, $\bar{z} = B\bar{v} \in Z_N$, where

$$N = \begin{cases} \|\bar{v}\| & \|\bar{v}\| \text{ is a positive integer} \\ [\|\bar{v}\|] + 1 & \text{otherwise} \end{cases}$$

$[\cdot]$ is the integer part of a number. Therefore $n(\delta)$ is a £nite number and there is δ_0 such that $n(\delta) = n(\delta_0)$ for any $\delta \in (0, \delta_0]$. The inequality $n(\delta) \geqslant N$ for any $\delta > 0$ is evident. Thus, for all $\delta \in (0, \delta_0]$ the approximate solutions $z_{n(\delta)}$ belong to the compact set $Z_{n(\delta_0)}$, and the method coincides with the quasisolutions method [12] for all suf£ciently small positive δ. The convergence $z_{n(\delta)} \to \bar{z}$ follows from the general theory of ill-posed problems [8].

Remark 1: The method is a variant of the method of extending compacts proposed in [13].

Theorem 3 [11]: For the method described above there exists an *a posteriori* error estimate. It means that a functional $\kappa(u_\delta, \delta)$ exists such that $\kappa(u_\delta, \delta) \to 0$ as $\delta \to 0$ and $\|z_{n(\delta)} - \bar{z}\| \leqslant \kappa(u_\delta, \delta)$ at least for all suf£ciently small positive δ.

Proof De£ne the function $\kappa(u_\delta, \delta)$ as

$$\kappa(u_\delta, \delta) = \max\{\|z_{n(\delta)} - z\| : z \in Z_{n(\delta)}, \\ \|Az - u_\delta\| \leqslant \delta\}$$

Since the operator A is bounded and $Z_{n(\delta)}$ is a compact set, then $\{z \in Z_{n(\delta)} : \|Az - u_\delta\| \leqslant \delta\}$ is a compact set too. Therefore, $\kappa(u_\delta, \delta) < +\infty$. Note that $\bar{z} \in Z_{n(\delta)}$. Then the inequality $\|z_{n(\delta)} - \bar{z}\| \leqslant \kappa(u_\delta, \delta)$ is just for all $\delta \leqslant \delta_0$. Since the method coincides with the quasisolutions method, then $\kappa(u_\delta, \delta) \to 0$ as $\delta \to 0$.

Remark 2: The existence of the *a posteriori* error estimation follows from [10]. If by $\bar{Z} \subset Z$ we denote the space of sourcewise represented with the operator B solutions of (1), then $\bar{Z} = \bigcup_{n=1}^{\infty} Z_n$. Since Z_n is a compact set, then $\bar{Z}$ is a σ-compact space.

An *a posteriori* error estimate is not an error estimate in general meaning that is impossible in principle for ill-posed problems [8, 2, 7]. But it becomes an upper error estimate of the approximate solution for "small" errors $\delta < \delta_0$, where δ_0 depends on the exact solution $\bar{z}$.

Let A be a linear injective compact operator, Z and U are Hilbert spaces. Consider the case, when $\bar{z} = (A^*A)^{p/2}\bar{v}$, $\bar{v} \in Z$, $p = \text{const} > 0$.

Lemma 1: The operator $(A^*A)^{p/2}$ is a compact injective operator from Z to Z for any $p > 0$.

Proof The operator A^*A is compact and self-adjoint. The compactness of $(A^*A)^{p/2}$ follows from the properties of eigenvalues of linear compact selfadjoint operators [14]. The injectiveness is obvious.

Consider the extending compacts method in the case: Z and U are Hilbert spaces, $V = Z$, $A : Z \to U$ is a linear compact injective operator, $B = (A^*A)^{p/2}$, $p = \text{const} > 0$.

Theorem 4 [11]: For this case the method of extending compacts is an optimal in order of accuracy regularizing algorithm.

Proof Both Theorems 2 and 3 are valid, therefore the method of extending compacts is a regularizing algorithm with an *a posteriori* error estimation. For all $\delta \in (0, \delta_0]$, where δ_0 is de£ned in Theorem 1, the method coincides with the quasisolutions method on the convex balanced compact set $BV_{n(\delta_0)}$. Thus, the method is optimal in order of accuracy [7]. From [15] it follows that the accuracy of the method is at least $O(\delta^{p/(p+1)})$ for all $p > 0$.

Clearly, in the method of extending compacts instead of integer numbers $n = 1, 2, \ldots$ one may use another increasing sequence $r_1, r_2, \ldots$ of positive numbers such that $\lim_{n\to\infty} r_n = +\infty$.

The case, when the operators A and B are known with errors, is considered in [11]. In the paper we write only the results obtained there. Let there be linear operators A_{h_A}, B_{h_B} and errors h_A, h_B such that $\|A_{h_A} - A\| \leqslant h_A$, $\|B_{h_B} - B\| \leqslant h_B$. Denote the vector of errors by $\eta \equiv (\delta, h_A, h_B)$. For any integer n de£ne a compact set

$$Z_{n,h_B} \equiv \{z \in Z : z = B_{h_B} v, v \in V, \|v\| \leqslant n\}$$

Set the problem: £nd a minimal positive integer number $n = n(\eta)$ such that the inequality

$$\|A_{h_A} z - u_\delta\| \leqslant \delta + (h_A \|B_{h_B}\| + h_B \|A_{h_A}\| + h_A h_B)\, n(\eta)$$

has a nonempty set of solutions. Then the *a posteriori* error estimation is

$$\kappa(u_\delta, A_{h_A}, B_{h_B}, \eta) \equiv h_B n(\eta) + \max\{\|z - z_{n(\eta)}\| : z \in Z_{n(\eta),h_B}, \|A_{h_A} z - u_\delta\| \leqslant \delta + (h_A \|B_{h_B}\| + h_B \|A_{h_A}\| + h_A h_B)\, n(\eta)\}$$

We now apply the method of extending compacts for the solution of the inverse problem for the heat conduction equation (6). It is evident that for any moment of time $-t_\varepsilon < 0$ there is

$$z(\xi) = Bv(x) = \int_0^l G(\xi, x, t_\varepsilon) v(x) dx$$

where $v(x) = w(x, -t_\varepsilon)$. Suppose that $V = Z = U = L^2[0, l]$.

Clearly, we have obtained the problem (1) with *a priori* constraints (9). Therefore, we may solve the problem using the method of extending compacts. Let $a = 1.0$, $l = 1.0$, $t_\varepsilon = 0.02$, $T = 0.1$, $\delta = 0.03 \cdot \|\bar{u}\|$. As a function $\bar{v}(x)$

$$\bar{v}(x) = \begin{cases} 10 & 0.3 < x < 0.5 \\ -4 & 0.5 < x < 0.8 \\ 0 & \text{otherwise} \end{cases}$$

is taken. Solving the problem we go to the Fourier coef£cients of the function $v(x)$ and estimate their ranges. After that in any point of the interval $[0, l]$ we £nd the maximal and the minimal values of a function that has Fourier coef£cients in the found intervals. In Figure 2 there are the found $z_\eta(x)$ solution and the area, which is the *a posteriori* error estimation for $z_\eta(x)$. We obtain $n(\delta) = 5$.

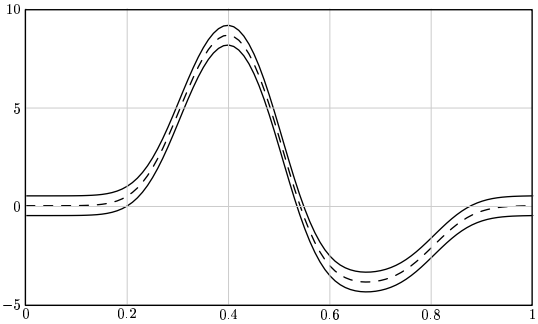


Figure 2. The approximate solution $z_\eta(x)$ $(- - -)$ and its *a posteriori* error estimation.

Other regularizing algorithms for ill-posed problems with sourcewise represented solutions may be found in [16].

COMPACT SETS

Suppose that there is the additional *a priori* information that the exact solution $\bar{z}$ of (1) belongs to a compact set M and A is a linear continuous injective operator. It is shown in [8] that as a set of approximate solutions of (1) it is possible to accept

$$Z_M^\eta \equiv \{z \in M : \|A_h z - u_\delta\|_U \leqslant h\|z\|_Z + \delta\}$$

Then $z_\eta \to \bar{z}$ as $\eta \to 0$ in Z for any $z_\eta \in Z_M^\eta$.

In practical problems there is often *a priori* information that the exact function of (1) is a monotone or convex bounded function or a function with a given Lipschitz constant. These functions are given on line segments $[a, b]$. Then the sets may be considered as compact sets. In [8] it is proved that on some subsets of $[a, b]$ any approximate function $z_\eta(x)$ in $L^p[a, b]$ converges to $\bar{z}(x)$ uniformly. Moreover, there is the algorithm to £nd the error of the approximate solution for the sets [17].

Assume that z is a function $z(x)$ on $[a, b]$, u is a function $u(\xi)$ on $[c, d]$, $Z = L^2[a, b]$, $U = L^2[c, d]$. For many problems instead of the function $u_\delta(\xi)$ there are only its grid values. For the function $u_\delta(\xi)$ we assume that $\|u_\delta - \bar{u}\|_{C[c,d]} \leqslant g(h)$, where $g(h) \to 0$ as $h \to 0$. Therefore the grid values for the function $u_\delta(\xi)$ are close to the grid values for the function $\bar{u}(\xi)$. For the compact sets M it is convenient to use grid values of the function $z(x)$, since the conditions of compactness for M may be easily written for the grid values. For example, these conditions for monotone non-decreasing function are

$$z_{i+1} - z_i \leqslant 0 \quad i = \overline{1, n-1},$$
$$C_1 \leqslant z_i \leqslant C_2 \quad i = \overline{1, n}$$

where z_i are grid values of $z(x)$ on a grid $\{x_i\}_1^n \subset [a,b]$. Thus to £nd an approximate solution of (1) we go to £nite dimensional Euclidean spaces. Denote a grid for $u(\xi)$ by $\{\xi_j\}_1^m \subset [c,d]$ and grid values by $\{u_j\}_1^m$. For any function $z(x) \in M$ we may introduce the piecewise linear function $z_n(x)$ such that

$$z_n(x) = z_i + \frac{z_{i+1} - z_i}{x_{i+1} - x_i}(x - x_i) \quad \forall x \in [x_i, x_{i+1}], i \in \overline{1, n-1} \tag{10}$$

By $u_n(\xi)$ denote the function $Az_n(x)$. Suppose the operator A_h is known exactly. De£ne an approximate operator A_n as an operator such that $\forall z \in Z$: $A_n z(x) \equiv u_n(\xi)$. Since for the considered functions $z(x)$ the inequalities $C_1 \leqslant z(x) \leqslant C_2$ are valid, then $\|z_n - z\| \to 0$ as $n \to +\infty$. The norm of the operator A is bounded, so $\|A_n z - Az\| = \|A(z_n - z)\| \leqslant \|A\| \cdot \|z_n - z\| \to 0$ as $n \to +\infty$. As the set of approximate solutions we take

$$Z_M^\eta \equiv \{z \in M : \|A_n z - u_\delta\|_U \leqslant \Delta\}$$

where $\Delta = H + \delta$, $H \equiv \sup\{\|A_n z - Az\| : z \in M\}$.

To distinguish variables for the in£nite dimensional problem (1) from variables for the appropriate £nite dimensional one we use the symbol $\hat{}$ ("hat"). Thus, to £nd the in£nite dimensional set Z_M^η we should £nd the appropriate £nite dimensional set

$$\hat{Z}_M^\eta \equiv \{\hat{z} \in \hat{M} \subset Z^n : \|\hat{A}\hat{z} - \hat{u}_\delta\|_{U^m} \leqslant \hat{\Delta}\} \tag{11}$$

where $\hat{z} = (z_1, \ldots, z_n)$, $\hat{u} = (u_1, \ldots, u_m)$, Z^n and U^m are £nite dimensional Euclidean spaces, $\hat{A}$ is an $m \times n$ matrix, $\hat{\Delta}$ is an error of the £nite dimensional problem. The sets $\hat{M}$ of *a priori* constraints are convex polyhedrons in the paper. Instead of the exact grid values $\bar{u}_j$ of the function $\bar{u}(\xi)$ there are vectors $\hat{u}_\delta = (u_1, \ldots, u_m)$ and $\hat{\delta} = (\delta_1, \ldots, \delta_m)$ such that $|u_j - \bar{u}_j| \leqslant \delta_j$, $j = \overline{1, m}$. For the vectors $\hat{u}_\delta$ and $\hat{\delta}$ we may construct the linear piecewise functions $u_\delta^l(\xi)$ and $u_\delta^u(\xi)$ similarly to (10) using the grid values $u_j - \delta_j$ and $u_j + \delta_j$, $j = \overline{1, m}$, accordingly. As above, we assume that $\forall \xi \in [c,d]$: $u_\delta^l(\xi) \leqslant \bar{u}(\xi) \leqslant u_\delta^u(\xi)$.

Let $\hat{A}$, $\hat{u}_\delta$, $\hat{M}$, $\hat{\Delta}$ be known. To £nd a £nite dimensional solution of the problem (1) an approximate solution $\hat{z}_\eta \in \hat{Z}_M^\eta$ of (11) should be found. We obtain a problem to minimize the discrepancy $\hat{\Phi}[\hat{z}] = \|\hat{A}\hat{z} - \hat{u}_\delta\|^2$, which is convex and differentiable, on the convex polyhedron $\hat{M}$. Clearly, the Fréchet derivative $\hat{\Phi}'[\hat{z}] = 2(\hat{A}^*\hat{A}\hat{z} - \hat{A}^*\hat{u}_\delta)$. One may £nd all the vertex of $\hat{M}$ using the method to cut convex polyhedrons [17]. Therefore it is possible to use the method of conditional gradient. Note, for the sets M of monotone, convex functions these vertex are found analytically in [8]. However, it is better to use the method of projection of conjugate gradients. The programs to solve these problems are in [8]. After the vector $\hat{z}_\eta$ has been found we construct the function z_η using the formula (10). There are several applications of this approach in astrophysics [18, 19] and in electronic microscopy [20].

To £nd the error of the found solution we should construct the set Z_M^η or a set approximated it. For this purpose we do the following. First, we £nd the minimum and the maximum values for each coordinate of $\hat{Z}_M^\eta$. Denote them by z_i^l, z_i^u, $i = \overline{1, n}$. Secondly, using the found grid values we construct functions $z^l(x)$ and $z^u(x)$ closed to Z_M^η such that $\forall z \in Z_M^\eta$: $z^l(x) \leqslant z(x) \leqslant z^u(x)$ for each $x \in [a,b]$. Clearly, on the segment $[a,b]$ it our aim are the functions $z^l(x) = \inf\{z(x) : z \in Z_M^\eta\}$ and $z^u(x) = \sup\{z(x) : z \in Z_M^\eta\}$.

The set $\hat{Z}_M^\eta = \hat{Z}^\eta \cap \hat{M}$, where $\hat{M}$ is a convex polyhedron and $\hat{Z}^\eta = \{\hat{z} : \|\hat{A}\hat{z} - \hat{u}_\delta\| \leqslant \hat{\Delta}\}$ is an ellipsoid. Thus to £nd $\hat{z}^l$, $\hat{z}^u$ we should minimize linear functions on the convex bounded set $\hat{Z}_M^\eta$. One may circumscribe a convex polyhedron near $\hat{Z}_M^\eta$. The problem to minimize the linear function will be reduced then to a linear programming problem, which can be solved with the usage of the simplex-method. Since it is necessary to solve $2n$ linear programming problems and the minimum of a linear function are in its vertex, it is better to £nd all vertexes of the polyhedron. For these purpose it is possible to use the method to cut convex polyhedrons [17].

Now we consider the problem to £nd the functions $z^u(x)$, $z^l(x)$ for the set Z_M^η and M is a set of non-decreasing functions. Clearly, $\forall x \in [a,b]$: $z^l(x) = \inf\{z_n^l(x) : \hat{z} \in \hat{Z}_M^\eta\}$, $z^u(x) = \sup\{z_n^u(x) : \hat{z} \in \hat{Z}_M^\eta\}$. Let the vectors $\hat{z}^l = (\hat{z}_1^l, \ldots, \hat{z}_n^l)$, $\hat{z}^u = (\hat{z}_1^u, \ldots, \hat{z}_n^u)$ be known exactly. Then,

$$z^l(x) = \begin{cases} z_i^l & x \in [x_i, x_{i+1}) \\ z_n^l & x = b \end{cases}$$

$$z^u(x) = \begin{cases} z_1^u & x = a \\ z_{i+1}^u & x \in (x_i, x_{i+1}] \end{cases}$$

When we solve the problem numerically, the vectors $\hat{z}^l$, $\hat{z}^u$ are found approximately. Therefore suppose that instead of $\hat{z}^l$, $\hat{z}^u$ vectors $\hat{z}^{l*} = (z_1^{l*}, \ldots, z_n^{l*})$, $\hat{z}^{u*} = (z_1^{u*}, \ldots, z_n^{u*})$ are known such that $\forall i \in \overline{1,n}$: $z_i^{l*} \leqslant z_i^l$, $z_i^{u*} \geqslant z_i^u$. Using the vectors $\hat{z}^{l*}$, $\hat{z}^{u*}$ we should construct the least vector $\hat{z}^l$ and the most vector $\hat{z}^u$ that satisfy to the last inequalities. The least vector $\hat{z}^l$ is considered as the vector each coordinate of it is equal to the minimum value of the appropriate coordinate for the set $\hat{Z}_M^\eta$. The most vector $\hat{z}^u$ is considered as the vector each coordinate of it is equal to the maximum value. From the de£nition of a non-decreasing function we obtain $\forall i \in \overline{1,n-1}$: $z_i \leqslant z_{i+1}$ and $z_i^u \leqslant z_{i+1}^u$, $z_i^l \leqslant z_{i+1}^l$. Therefore we set $z_1^l = z_1^{l*}$ and $\forall i \in \overline{2,n}$: $z_i^l = \max(z_{i-1}^l, z_i^{l*})$. Similarly, for $\hat{z}^u$ we set $z_n^u = z_n^{u*}$ and $\forall i \in \overline{n-1,1}$: $z_i^u = \min(z_{i+1}^u, z_i^{u*})$. Evidently, if $\exists i \in \overline{1,n}$: $z_i^l > z_i^u$, then $Z_M^\eta = \varnothing$, i.e., the problem has no solution.

Note that for the considered approach the main data is the function $u_\delta(\xi)$. Its grid values are used only to approximate this function by the piecewise linear one. This way is in some sense arti£cial, since we really have only the vector $\hat{u}_\delta$ of grid values. The inequality $\|u_\delta - \bar{u}\| \leqslant \delta$ determines a wider set of functions in U than the inequalities $|u_j - \bar{u}_j| \leqslant \delta_j$, $j = \overline{1,m}$. Therefore if the problem to construct the set Z_M^η is solved by minimization of $\|\hat{A}\hat{z} - \hat{u}_\delta\|$ exactly, then the found set is wider than the "real" set of approximate solutions. To be de£nite, for this "real" set we use symbol $\mathcal{Z}_M^\eta$.

We use the problem to minimize the considered quadratic function for the following reasons. First, it is necessary to associate the in£nite dimensional spaces Z, U with the appropriate £nite dimensional ones Z^n, U^m more closely. Secondly, if only an approximate solution should be found and we do not want to construct the functions $z^l(x)$ and $z^u(x)$, then this problem is reduced to the problem to minimize a quadratic function on the convex polyhedron, i.e., it is solved very fast. Thirdly, the considered problem is changed very easy to £nd an approximate solution of a real ill-posed problem, i.e., when there is no information that the exact solution $\bar{z}(x)$ belongs to a compact set M. Then we should minimize a Tikhonov's functional.

When we want to £nd not an error estimation of an approximate solution but to construct the functions $z^l(x)$ and $\hat{z}^u(x)$, then we may use another approach. Instead of the function $u_\delta(\xi)$ we consider the vector $\hat{u}_\delta$ of the grid values as the given data. The choice of the norm for the space U is not important, since the function $u_\delta(\xi)$ is not approximated by a function of the vector $\hat{u}_\delta$.

Therefore, we may obtain the following inequalities

$$-H_j^- - \delta_j \leqslant A^j\hat{z} - u_j \leqslant H_j^+ + \delta_j \qquad j = \overline{1,m} \quad (12)$$

where A^j are n vectors. The set of all the points $\hat{z} \in Z^n$ satisfying these inequalities is denoted by $\hat{\mathcal{Z}}_M^\eta$. For the reasons written above the inclusion $\hat{\mathcal{Z}}_M^\eta \subset Z_M^\eta$ is valid. After the functions $z^l(x)$, $z^u(x)$ have been constructed for the set $\hat{\mathcal{Z}}_M^\eta$ we obtain the "real" set $\mathcal{Z}_M^\eta$ of the approximate solutions of the problem (1). The construction of the functions $z^l(x)$, $z^u(x)$ for the set $\hat{\mathcal{Z}}_M^\eta$ is the same as for the set $\hat{Z}_M^\eta$. Since the inequalities (12) are the equations of half-spaces in Z^n, i.e., the set $\hat{\mathcal{Z}}_M^\eta$ is a polyhedral set, then to construct the vectors $\hat{z}^l$, $\hat{z}^u$ some linear programming problems should be solved.

There is a difference between the approaches considered in this section. In the £rst approach to £nd the vectors $\hat{z}^l$, $\hat{z}^u$ it is necessary to £nd an approximate solution $\hat{z} \in \hat{Z}_M^\eta$, since we approximate the convex set $\hat{Z}_M^\eta$ by a convex polyhedron when solving the problems to minimize linear functions. In the second approach an approximate solution is not found, it may always be constructed after the vectors $\hat{z}^l$, $\hat{z}^u$ have been found.

Let us solve the problem (6) on the set of convex upward functions $z(x)$ such that $0 \leqslant z(x) \leqslant C$. Assume that $a = 1.0$, $l = 1.0$, $T = 1.0$, $C = 1.2$, $n = 20$. In Figure 3 there are the exact solution $\bar{z}(x)$ and the functions $z^l(x)$, $z^u(x)$ found for the set $\mathcal{Z}_M^\eta$.

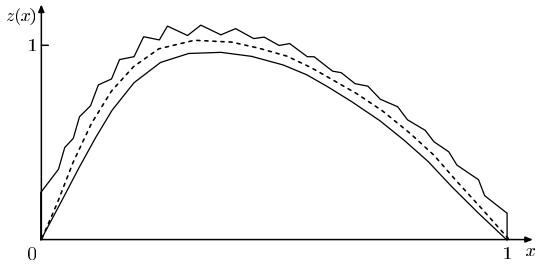


Figure 3. The exact solution $\bar{z}(x)$ (– – ––), the functions $z^l(x)$, $z^u(x)$ for the sets $\mathcal{Z}_M^\eta$ (——).

CONCLUSIONS

In the paper we have shown how to construct regularizing algorithms and provide an er-

ror estimation using *a priori* information about the exact solution. We discuss the algorithms for a general ill-posed problem (minimization on Tikhonov's smoothing functional), for a problem with a sourcewise represented set of solutions (the method of extending compacts) and for ill-posed problems on compact sets. If there is more information about the exact solution, then there is more information about properties of approximate solutions, i.e., the convergence, the *a posteriori* error estimation, the uniform error estimation.

The regularizing algorithms considered in the paper are applied in astrophysics [18, 19], electronic microscopy [20], vibrational spectroscopy [21].

The authors thank the Russian Foundation for Basic Research for the £nancial support (grant 02-01-00044).

REFERENCES

1. J. Hadamard, *Lectures on Cauchy's problem in linear partial differential equations*, Yale Univ. Press, New Haven, 1923

2. A. N. Tikhonov, A. S. Leonov, and A. G. Yagola, *Nonlinear ill-posed problems*, Chapman and Hall, London, 1998

3. A. N. Tikhonov, Solution of incorrectly formulated problems and the regularization method, *Sov. Math., Dokl.*, **5**, 1035–1038, (1963)

4. A. N. Tikhonov, Regularization of incorrectly posed problems, *Sov. Math., Dokl.*, **4**, 1624–1627, (1963)

5. A. S. Leonov and A. G. Yagola, Can an ill-posed problems be solved if the data error is unknown? *Moscow Univ. Physics Bull.*, **50**(4), 25–28, (1995)

6. A. B. Bakushinskii, Remark about the choice of the regularization parameter by the quasioptimality criterion and the propertion criterion, *Comput. Math. Math. Phys.*, **24**(8), 1258–1259, (1984)

7. A. B. Bakushinskii and A. V. Goncharskii, *Ill-posed problems: theory and applications*, Kluwer Academic Publishers, Dordrecht, 1994

8. A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical methods for the solution of ill-posed problems*, Kluwer Academic Publishers, Dordrecht, 1995

9. V. A. Vinokurov, Regularizable functions in topological spaces and inverse problems, *Sov. Math., Dokl.*, **20**, 569–573, (1979)

10. V. A. Vinokurov and Yu. L. Gaponenko, A posteriori estimates of the solutions of ill-posed inverse problems, *Sov. Math., Dokl.*, **25**, 325–328, (1982)

11. A. G. Yagola and K. Yu. Dorofeev, Sourcewise representation and a posteriori error estimates for ill-posed problems, In: Ramm, A. G. *et al.*, eds., *Fields Institute Communications: Operator Theory and Its Applications*, **25**, 543–550, AMS, Providence, RI, (2000)

12. V. K. Ivanov, V. V. Vasin, and V. P. Tanana, *The theory of linear ill-posed problems and its applications*, Nauka, Moscow, 1978 (in Russian)

13. I. N. Dombrovskaya and V. K. Ivanov, Some questions to the theory of linear equations in abstract spaces, *Sibirskii Mat. Zhurnal*, **16**, 499–508, (1965) (in Russian)

14. F. Riesz and B. Sz.-Nagy, *Functional analysis*, Dover Publications Inc., New York, 1990

15. G. M. Vainikko, *Methods for the solution of linear incorrectly formulated problems in Hilbert spaces. Textbook*, Tartu University Press, Tartu, 1982

16. A. S. Leonov and A. G. Yagola, Special regularizing methods for ill-posed problems with sourcewise represented solutions, *Inverse Problems*, **14**(6), 1539–1550, (1998)

17. V. N. Titarenko and A. G. Yagola, A method to cut convex polyhedrons and its applications to ill-posed problems, *Numerical Methods and Programming*, **1**, 8–13, (2000) (http://num-meth.srcc.msu.su)

18. A. V. Goncharsky, A. M. Cherepashchuk, and A. G. Yagola, *Ill-posed problems of astrophysics*, Nauka, Moscow, 1985 (in Russian)

19. A. V. Goncharsky, A. M. Cherepashchuk, and A. G. Yagola, *Numerical methods for the solution of inverse problems in astrophysics*, Nauka, Moscow, 1978 (in Russian)

20. V. D. Rusov, Yu. F. Babikova, and A. G. Yagola, (1991). *Image restoration in electronic microscopy autoradiography of surfaces*, Energoatomizdat, Moscow, 1991 (in Russian)

21. A. G. Yagola, I. V. Kochikov, G. M. Kuramshina, and Yu. A. Pentin, *Inverse problems of vibrational spectroscopy*, VSP, Zeist, 1999

TUTORIAL SESSIONS

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

Sensitivity and Uncertainty Analysis for Thermal Problems

Bennie F. Blackwell
Sandia National Laboratories
PO Box 5800
MS 0828
Albuquerque, NM 87185 USA
bfblack@sandia.gov

Kevin J. Dowding
Sandia National Laboratories
PO Box 5800
MS 0828
Albuquerque, NM 87185 USA
kjdowdi@sandia.gov

ABSTRACT

A discussion of various methods used to compute sensitivity information is presented. The methods include differentiation of analytical solutions, finite difference, complex step, software differentiation, sensitivity equation method, and adjoint methods. Example calculations are presented for several of these methods. It is emphasized that sensitivity information is important in its own right as opposed to simply being one of the many ingredients necessary to perform parameter estimation and/or optimization calculations. Strengths and weaknesses of the various sensitivity methods are discussed.

NOMENCLATURE

A	area, m^2
C	$= \rho c_p$, volumetric heat capacity, J/m^3-K
c_p	specific heat, J/kg-K
h	convective heat transfer coefficient, W/m^2-K
k	thermal conductivity, W/m-K
[K]	global conduction matrix
L	thickness of slab, m
n_p	number sensitivity coefficients (parameters)
n_s	number of sensors
$\{p\}$	parameter vector
p_i	element of $\{p\}$
q	heat flux, W/m^2
T	temperature, K
T_p	scaled sensitivity coefficient, $= p\partial T/\partial p$
t	time, s
α	thermal diffusivity, m^2/s
ε	emittance
σ	Stefan-Boltzmann constant

INTRODUCTION

When analyzing the response of a thermal system, a large number of parameters must be specified to characterize the system. These parameters include material properties (density, specific heat, thermal conductivity, emittance, etc.), geometry, and boundary conditions (heat flux, convective heat transfer coefficient, etc.). During the design phase of a project, some of the parameters may change as the design evolves. In many instances, these parameters are not known with a high degree of precision. Also, a designer may be free to choose among many different competing materials. As an example, a design might call for 304 stainless steel; alternative stainless steels such as 316 might work equally well and could be used interchangeably (depending on availability). Even if we consistently use 304 stainless, there may be manufacturer-to-manufacturer variability as well as lot-to-lot variability from a single manufacturer. Consequently, we need to play "what if" scenarios with regard to the material properties in order to assure ourselves that the lot-to-lot or manufacturer-to-manufacturer variability does not produce undesirable consequences.

Historically, these "what if" scenarios have been performed on an ad hoc basis; selected parameter values would be changed and the analysis repeated. Without the aid of a computer, only a limited number of parameter studies could be performed. However, today's computers now make it possible to perform a wide range of parameter studies. Even with the computer, these parameter studies still tend to be performed on an ad hoc basis. Based on a designer's intuition, the most important parameters would be varied over some range. Since intuition is only as good as prior experiences, it is possible for even an experienced designer or analyst to miss an important parameter. It is also time consuming to study parameters in an ad-hoc manner. Consequently, a more formal procedure needs to be developed. This is where *sensitivity analysis* plays an important role.

A desirable goal of the design process is to produce *robust designs* that can reliably function

over a wide range of operating conditions. This is particularly true for mission critical components. For example, the on-board flight controller for a rocket system must be capable of functioning when the ambient temperatures range from high altitude space to desert launch pad. Is it possible for the hardware to operate if the system parameters differ from the nominal values assumed in the analysis phase? The robustness of a design can be investigated by means of a *sensitivity analysis*.

Sensitivity analysis is defined as the study of how variations in input parameters of a computational model cause variations in output. Input parameters would include material properties, boundary/initial conditions, and geometry; output variables might be displacement, stress and/or temperature. A measure of this sensitivity is termed the sensitivity coefficient and is (mathematically) defined as the partial derivative of the output variable with respect to the parameter of interest. For a general discussion of sensitivity coefficients, see Beck and Arnold [1] and Beck, et al. [2]. Since the focus of this work is on thermal problems, let us define the first order sensitivity coefficient of the temperature field with respect to the thermal conductivity k as

$$\text{thermal conductivity sensitivity coefficient} = \frac{\partial}{\partial k} T(\hat{x}, t; k) \qquad (1)$$

where it is understood that all parameters other than thermal conductivity are held fixed during the differentiation. The sensitivity coefficient is also a field variable in that it depends on position and time just like temperature. In order to understand how one might use the sensitivity coefficient to predict how a system responds if you perturb the thermal conductivity, let us expand the temperature field in a Taylor series about the mean value of the thermal conductivity

$$T(\hat{x}, t; k) = T(\hat{x}, t; \bar{k}) + \left.\frac{\partial T}{\partial k}\right|_{\bar{k}} (k - \bar{k}) + \left.\frac{1}{2}\frac{\partial^2 T}{\partial k^2}\right|_{\bar{k}} (k - \bar{k})^2 + \ldots \qquad (2)$$

From this expansion, one can see how the first order sensitivity coefficient is needed for a first order analysis and higher order sensitivity coefficients are required for higher order analysis. This work will focus on first order sensitivity analysis as the computational load scales approximately linearly with the number of parameters. For a second order analysis, the computational load scales as the square of the number of parameters; this may become prohibitive for problems with hundreds of parameters. If the system response and first order sensitivity coefficients are known for the nominal parameter values, Eq. (2) (with higher order terms neglected) can be used to compute the response at a neighboring point in parameter space. If higher order sensitivity coefficients are required, an initial approach might be to compute second order sensitivity coefficients only for those selected parameters that have large first order sensitivity coefficients.

In some cases, only the sign of the sensitivity coefficient is important. If for example, the length of a system is increased, does the critical temperature go up or down?

In many instances, the sensitivity coefficient is often required as an intermediate step in the solution of parameter estimation, function estimation, uncertainty propagation, and optimization problems. The emphasis of this work is on calculating sensitivity coefficients because they have importance themselves as opposed to just numbers that are fed into a parameter estimation or optimization process.

SUMMARY OF METHODS FOR COMPUTING SENSITIVITY COEFFICIENTS

Sensitivity coefficients can be calculated by many methods. These methods include the following:

- differentiation of analytical solutions
- finite difference
- complex step
- software differentiation (e.g. ADIFOR/ ADIC)
- sensitivity equation method
- adjoint methods

Each of these methods has it place. In many instances, multiple methods will be used on the same problem to verify that the primary method produces the desired results. Example calculations will be presented for most of these methods.

DIFFERENTIATION OF ANALYTICAL SOLUTIONS

Differentiation of analytical solutions is probably the simplest method. It involves differentiating an analytical solution with respect to the parameter(s) of interest. If the analytical expressions are very complex, then symbolic algebra programs such as Mathematica®, Maple®, Macsyma®, etc. will prove invaluable. Obviously, this approach is limited to problems in which analytical solutions are available; this severely limits the problems which can be addressed by this method. Recent work by McMasters, et al. [3] using Green's functions has produced software

that will provide and evaluate analytical solutions for a wide variety of 3-D time dependent heat conduction problems in rectangular geometries. It has been the author's experience that one of the most significant uses of this method is to serve as a tool to verify other approximate numerical methods.

As an example of this method, consider a 1-d

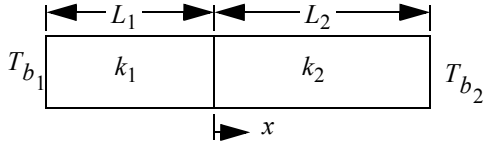


Figure 1. Problem definition for 1-D two layer slab problem.

configuration that might be used in the comparative method for the determination of thermal conductivity. This geometry is shown in Fig. 1. Fixed boundary temperatures are applied on the two ends and the steady state temperature profile is measured. The conductivity of one specimen is known and the other is to be determined. The analytical expressions for the temperature profile are

$$T_1(x) = \frac{k_1L_2T_{b_1} + k_2L_1T_{b_2}}{k_1L_2 + k_2L_1} - \frac{k_2L_1(T_{b_1} - T_{b_2})}{k_1L_2 + k_2L_1}\frac{x}{L_1} \quad (3)$$

$$T_2(x) = \frac{k_1L_2T_{b_1} + k_2L_1T_{b_2}}{k_1L_2 + k_2L_1} - \frac{k_1L_2(T_{b_1} - T_{b_2})}{k_1L_2 + k_2L_1}\frac{x}{L_2}. \quad (4)$$

In the design of this experiment, one would like to have a large sensitivity to the unknown thermal conductivity and a small sensitivity to the known thermal conductivity. The two thermal conductivity sensitivity coefficients can be computed analytically and are given by

$$k_1\frac{\partial T_1}{\partial k_1} = T_{k_1} = \frac{k_1L_2k_2L_1}{(k_1L_2 + k_2L_1)^2}(T_{b_1} - T_{b_2})\left(1 + \frac{x}{L_1}\right), \quad (5)$$

$$T_{k_1} = -T_{k_2}, \quad -L_1 \le x \le 0$$

$$k_1\frac{\partial T_1}{\partial k_1} = T_{k_1} = \frac{k_1L_2k_2L_1}{(k_1L_2 + k_2L_1)^2}(T_{b_1} - T_{b_2})\left(1 - \frac{x}{L_2}\right), \quad (6)$$

$$T_{k_1} = -T_{k_2}, \quad 0 \le x \le L_2.$$

The temperature and conductivity sensitivity coefficient profiles are shown in Fig. 2. Each profile consists of two straight line segments. The sensitivity coefficient T_{k_1} is positive throughout, which indicates that increasing k_1 increases the

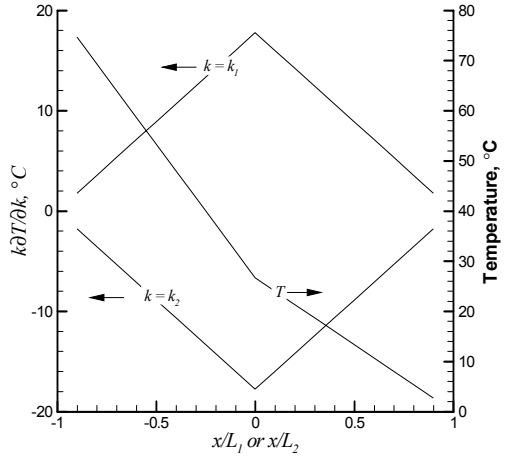


Figure 2. Temperature profile and conductivity sensitivity coefficients for two layer slab problem.

temperature. The sensitivity coefficient T_{k_2} is negative throughout, which indicates increasing k_2 decreases the temperature. The location of maximum thermal conductivity sensitivity T_{k_1} is at the interface between the two regions. This might seem strange at first, but remember that the conductivity sensitivity is identically zero on the boundaries and positive in between. Consequently, this forces the maximum sensitivity to be at an interior point. Note that the two conductivity sensitivity coefficients are correlated. This means that from a parameter estimation perspective, you can not estimate both thermal conductivities from the same experiment. Hence, the name comparator is appropriate.

The above sensitivity coefficients can be used to choose the specimen lengths as well as sensor locations.

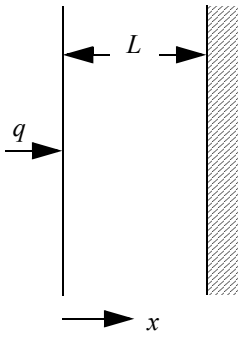
FINITE DIFFERENCE DETERMINATION OF SENSITIVITY COEFFICIENTS

If a sensitivity analysis requires an analytical solution, then we would be severely limited in the problems that we can address. Fortunately, general purpose software is available to numerically solve complex three dimensional, time dependent thermal problems. The discretization schemes include finite difference, finite volume, control volume finite element and finite element methods. In industry, commercially available software is often used and source code is generally not available. For this case, sensitivity coefficients are often computed by running the software for two different values of a parameter and using a first order forward difference. Scaled sensitivity coefficients are then determined from

$$T_{p_i} = p_i[T(p_1, p_2, \ldots, p_i + \Delta p_i, \ldots, p_n) - T(p_1, p_2, \ldots, p_i, \ldots, p_n)]/\Delta p_i + O(\Delta p_i). \quad (7)$$

This approach requires n+1 solutions for the temperature field and will be first order accurate in Δp_i. If a second order accurate central difference is used instead, then 2n+1 solutions of the temperature field will be required. There are many examples in the literature where this approach has been successfully used. An advantage of this approach is that you numerically solve the problem with different inputs. Since no source code modification is required, the software development costs for this method will be minimal. Commercial software can be used to accomplish this, provided the computational results are available with sufficient precision.

Some numerical experimentation is strongly recommended to determine an acceptable value for the finite difference step size Δp. If it is too large, the truncation errors will be excessive; if it is too small, machine round off may become a problem. In order to emphasize the importance of this issue, consider the one dimensional, constant flux problem given in Fig. 3. This example was



$q = 4\text{x}10^5$ W/m^2
$k = 10$ W/m-K
$L = 0.01$ m
$T_o = 300$ K
$\rho = 8000$ kg/m^3
$c_p = 500$ J/kg-K
$\alpha = 2.5\text{x}10^{-6}$ m^2/s
$qL/k = 400$ K
adiabatic back wall

Figure 3. Schematic of constant heat flux problem for 1-d planar slab.

solved numerically using a control volume finite element code with a lumped capacitance matrix and a fully implicit time integrator; double precision on a 32-bit machine (nominally 15 significant digits) was utilized. The final problem time was 20 s, which corresponded to a dimensionless time $\alpha t/L^2 = 0.5$. The dimensionless thermal conductivity sensitivity coefficient at $x = 0$ and $t = 20$ s was computed, utilizing a range of values for Δk for a forward difference approximation. The relative error in each computation was computed from

$$\%\ \text{Error} = 100\frac{T_{k_n} - T_{k_a}}{T_{k_a}} \quad (8)$$

where the subscripts n and a represent numerical and analytical, respectively; the error results are presented in Fig. 4. Uniform grids of 10 and 20

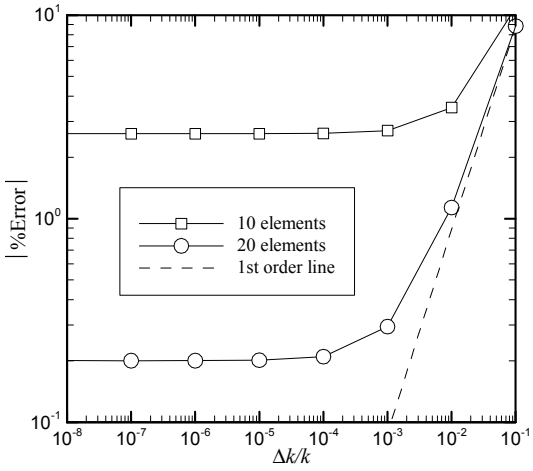


Figure 4. Thermal conductivity sensitivity coefficient errors at $x = 0$ and $t = 20$ s as a function of $\Delta k/k$ for two grid refinements.

elements were used along with a fixed grid scale Fourier number ($\alpha\Delta t/\Delta x^2 = 5$). Focusing on the upper right hand corner of this figure, as you decrease the relative finite difference step size ($\Delta k/k$), the error decreases initially and then reaches a plateau. From theoretical considerations of the Taylor series truncation error, the errors in the finite difference approximation to the sensitivity coefficient (forward difference) would decay linearly with decreasing finite difference step size. However, there are additional discretization errors in the numerically generated temperature field. The curve in Fig. 4 labeled "first order" is such a linear relationship and is shown for reference. Due to the errors in the numerical solution for the temperature field, it is obvious that the sensitivity coefficient errors do not decay linearly as the finite difference step size is decreased. This is particularly noticeable for the 10 element case. As the mesh is refined from 10 to 20 elements, the errors are closer to the linear relationship in the upper right hand portion of the figure. If the finite difference step size is made even smaller, it is possible that the errors will become even larger. An example of this behavior is shown in the next section.

COMPLEX STEP METHOD

The main criticism of the *finite difference method* for computing sensitivity coefficients is that the computational results exhibit a step size dependence. In practice, this means that for each (class of) problem(s), one needs to perform a step size parameter study for each sensitivity coefficient. Unfortunately, this could be a time consuming process for practical problems with

ten's of parameters. The *complex step method* offers a practical method of eliminating the step size dependence of finite difference methods, but at the expense of software modification and increased run time.

The derivation of the complex step method follows from a Taylor series expansion of the real valued function f about the complex (imaginary) parameter value $p + i\Delta p$ with $i = \sqrt{-1}$. This series is

$$f(p + i\Delta p) = f(p) + \left.\frac{\partial f}{\partial p}\right|_p i\Delta p - \frac{1}{2!}\frac{\partial^2 f}{\partial p^2}\Delta p^2 - \frac{1}{3!}\frac{\partial^3 f}{\partial p^3} i\Delta p^3 + O(\Delta p^4). \quad (9)$$

Taking the imaginary part of Eq. (9), one obtains

$$Im[f(p + i\Delta p)] = \left.\frac{\partial f}{\partial p}\right|_p \Delta p - \frac{1}{3!}\frac{\partial^3 f}{\partial p^3}\Delta p^3 + \ldots . \quad (10)$$

Solving for the first derivative yields

$$\left.\frac{\partial f}{\partial p}\right|_p = \frac{Im[f(p + i\Delta p)]}{\Delta p} + O(\Delta p^2) . \quad (11)$$

This simple result says that the derivative of the function is obtained by taking the *imaginary* part of the complex function $f(p + i\Delta p)$ divided by the real parameter step size Δp. Note that this result is second order accurate while Eq. (7) is first order accurate. Computational results for the complex step method have been presented by Martins, et al. [4]; they have applied the method to both analytical expressions as well as multidimensional structural and fluid dynamic codes. They demonstrated that the step size can be made arbitrarily small without suffering the loss of accuracy associated with the finite difference method. However, this result comes at the expense of code modification and increased run times

The computational procedure is as follows: 1) in the source code, declare all parameter values and the function f to be complex variables; 2) for the parameter p of interest, replace p by $p + i\Delta p$ in the input; 3) execute the software to compute f as a complex variable; 4) evaluate Eq. (11) for the sensitivity coefficient for parameter p. The process outlined above must be repeated for each parameter. Hence, the process of generating multiple sensitivity coefficients is similar to that for the *finite difference method* in that multiple runs of the same software are required. However, the big difference is that the step size Δp in Eq. (11) can be set to roughly machine zero and the results will be independent of step size. The *complex step method* will eliminate the (generally annoying) step of a parametric study in step size.

The complex step method has been applied to the same example considered in DIFFERENTIATION OF ANALYTICAL SOLUTIONS and the results are given in Fig. 5.

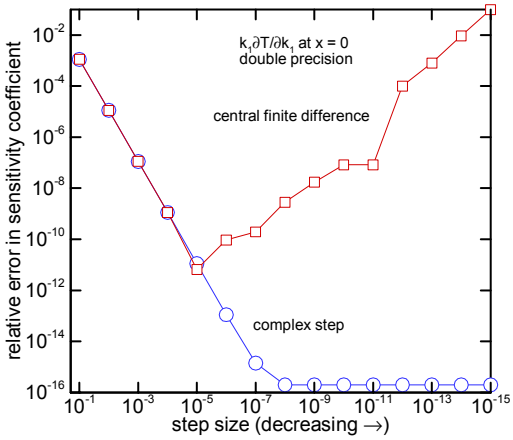


Figure 5. Relative error in conductivity sensitivity coefficient as a function of finite difference step size.

For comparison purposes, the central finite difference results are also shown. Both methods are second order accurate in the finite difference step size. Therefore, one would expect similar accuracy provided there are no machine precision problems. Both methods display second order behavior for step sizes down to 10^{-5}, as evidenced by the straight line behavior with a slope of -2. Further reductions in step size below 10^{-5} cause the central difference method errors to increase. Contrast this with the complex step method which is capable of driving the errors to machine zero. Although there may be a wide range of step sizes for which the central difference method produces acceptable accuracy, one is never sure what this range might be for a different problem. The rule of thumb on step size that is given in Nocedal and Wright [5] suggested that $2\text{x}10^{-11}$ would be appropriate for a central difference. For this problem, it appears that this rule of thumb is optimistic.

The authors have also applied the *complex step method* to a finite volume heat conduction code and computed sensitivity to multiple thermal conductivities and a contact conductance. References [6]-[9] apply the *complex step method* to compute sensitivity coefficients for a variety of aerodynamic problems. Some computational aids for converting a code from real to complex can be found in Reference [10].

SOFTWARE DIFFERENTIATION

The software differentiation method is a recent

development. An existing source code (FORTRAN 77 or C) is input into a special pre-processor (ADIFOR or ADIC) that performs line-by-line differentiation of the original source code while producing a new source code for the sensitivity coefficient. Examples of this technology are presented by Bischof, et al. [11], where they have successfully applied it to large codes. If there are multiple parameters for which sensitivity is desired, multiple runs of the pre-processor ADIFOR or ADIC are required. If the original source code is modified (enhancements, bug fixes, etc.), then the pre-processor must be run again. Since our work has been focused primarily on techniques that can be readily applied to software under development as well as many parameter problems, we have not personally exercised the software differentiation methods.

SENSITIVITY EQUATION METHOD

A sensitivity coefficient is a field variable just like temperature and will have its own describing equation. In this section, we will demonstrate how to derive the field equation(s) for sensitivity coefficients; this method is termed the *Sensitivity Equation Method* (SEM). This process involves the differentiation of the describing equation, along with associated initial/boundary conditions, with respect to the parameters of interest. These sensitivity equations are then solved numerically, using the same kind of algorithm as is used to solve the energy equation. To demonstrate this process, consider a 1-D planar slab with a radiation boundary condition on one face and convection boundary condition on the other face; this problem is shown schematically in Fig. 6. Due

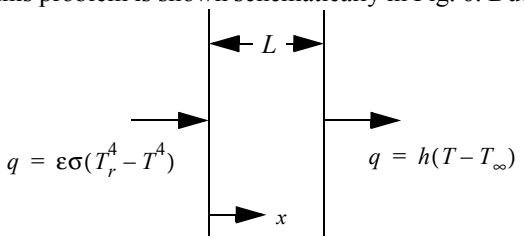


Figure 6. Schematic of 1-D problem with radiation and convection boundary conditions.

to the nonlinear radiation boundary condition, this problem is unlikely to have an exact analytical solution. Hence, a numerical solution will be explored. The energy equation and boundary conditions can be written as follows:

$$C\frac{\partial T}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{12}$$

$$q = -k\frac{\partial T}{\partial x} \tag{13}$$

$$q\big|_{x=0} = -k\frac{\partial T}{\partial x}\bigg|_{x=0} = \varepsilon\sigma(T_r^4 - T^4)\big|_{x=0} \tag{14}$$

$$q\big|_{x=L} = -k\frac{\partial T}{\partial x}\bigg|_{x=L} = h(T - T_\infty)\big|_{x=L} \tag{15}$$

$$T(x, 0) = T_i\,. \tag{16}$$

The parameters for this problem are given by the vector

$$\{p\}^T = \{C \;\; k \;\; \varepsilon \;\; T_r \;\; h \;\; T_\infty\}^T \tag{17}$$

Now, we differentiate Eq. (12)-Eq. (16) with respect to each element in the parameter vector, Eq. (17). Starting with the volumetric heat capacity sensitivity coefficient, we will differentiate Eq. (12) with respect to C, resulting in

$$\frac{\partial}{\partial C}\left(C\frac{\partial T}{\partial t}\right) + \frac{\partial}{\partial C}\left(\frac{\partial q}{\partial x}\right) = C\frac{\partial}{\partial t}\left(\frac{\partial T}{\partial C}\right) + \frac{\partial T}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\partial q}{\partial C}\right) = 0 \tag{18}$$

where it has been assumed that the order of differentiation can be interchanged. Again, we will introduce the scaled sensitivity coefficient by multiplying Eq. (18) through by C to obtain

$$C\frac{\partial}{\partial t}\left(C\frac{\partial T}{\partial C}\right) + C\frac{\partial T}{\partial t} + \frac{\partial}{\partial x}\left(C\frac{\partial q}{\partial C}\right) = 0\,. \tag{19}$$

The volumetric heat capacity sensitivity coefficient is readily identified in Eq. (19). The sensitivity of the heat flux to changes in the volumetric heat capacity can be determined by differentiating Fourier's Law with respect to C, resulting in

$$q_C = C\frac{\partial q}{\partial C} = -k\frac{\partial}{\partial x}\left(C\frac{\partial T}{\partial C}\right) = -k\frac{\partial T_C}{\partial x}\,. \tag{20}$$

Eq. (19) can now be written as

$$C\frac{\partial T_C}{\partial t} - \frac{\partial}{\partial x}\left(k\frac{\partial T_C}{\partial x}\right) = -C\frac{\partial T}{\partial t}\,. \tag{21}$$

Eq. (19) is the partial differential equation that describes the field variable T_C. Note that the left hand side of this equation is identical in form to that of the original energy equation. However, there is an apparent source term on the right hand side that was not present in the energy equation. If the temperature field is known, then this source term is a known function of position and time.

We will continue the development of the equations governing the behavior of T_C by differentiating the initial/boundary conditions with

respect to C. Evaluating Eq. (20) at the $x = 0$ boundary and differentiating Eq. (14) with respect to C, this boundary condition becomes

$$C\frac{\partial q}{\partial C}\bigg|_{x=0} = -k\frac{\partial T_C}{\partial x}\bigg|_{x=0} = -4\varepsilon\sigma T^3 T_C\big|_{x=0}. \quad (22)$$

While the left hand side boundary condition for the energy equation was nonlinear and inhomogeneous, the corresponding T_C boundary condition is *linear* and *homogeneous*. This assumes again that the temperature field is known prior to the computation of the sensitivity field. Through a similar procedure, the boundary condition for the right face is given by

$$C\frac{\partial q}{\partial C}\bigg|_{x=L} = -k\frac{\partial T_C}{\partial x}\bigg|_{x=L} = hT_C\big|_{x=L}. \quad (23)$$

This boundary condition is linear and homogeneous. Since the initial condition is independent of the volumetric heat capacity, the corresponding sensitivity initial condition is the zero condition

$$C\frac{\partial T}{\partial C}\bigg|_{x,0} = T_C\big|_{x,0} = 0. \quad (24)$$

The formulation of the field equation and associated boundary/initial conditions for T_C is complete and is given by Eq. (20)-Eq. (24). Due to the similarities in form of the energy equation and the volumetric heat capacity sensitivity equation, the same technique can be used to numerically solve these equations. It does not matter if the discretization algorithm is finite difference, finite volume, control volume finite element or finite element. In fact, the existing software coding used to include the effects of capacitance, diffusion and source terms for the energy equation can be used to form the analogous terms for the sensitivity equation. The computational procedure is to first time march the energy equation one time step and then solve the sensitivity equation. The source term for T_C in Eq. (21) is known from the temperature solution. Even though the original energy equation was nonlinear because of the radiation boundary condition, the corresponding sensitivity equation is a *linear* equation. This linearity may afford computational savings, depending on the algorithm used to solve the nonlinear algebraic equations resulting from the discretization of the energy equation.

From Eq. (21), *the time rate of change of temperature drives the volumetric heat capacity sensitivity field*. If the temperature field is not changing with time, the volumetric heat capacity sensitivity will tend toward zero. For a problem with a positive temperature rise rate, this source term is negative, suggesting a negative sensitivity to the volumetric heat capacity. Similarly, for a body that is cooling, the source term is positive which suggests a positive sensitivity to the volumetric heat capacity. As one can see, insight into the thermal response can be gained by simply studying the describing equation for the sensitivity coefficients. In some cases, trends may be determined without actually solving the sensitivity equations. However, it is best to continue the process and numerically solve the sensitivity equations in order to gain maximum insight.

Next, we will derive the equation for the thermal conductivity sensitivity. The thermal conductivity sensitivity equation will be more complicated because thermal conductivity appears in both the differential equation and boundary conditions. Following the same procedure as above, the differential equation for T_k can be written as

$$C\frac{\partial}{\partial t}\left(k\frac{\partial T}{\partial k}\right) + \frac{\partial}{\partial x}\left(k\frac{\partial q}{\partial k}\right) = C\frac{\partial T_k}{\partial t} + \frac{\partial q_k}{\partial x} = 0. \quad (25)$$

Differentiating Fourier's Law with respect to thermal conductivity k, we obtain

$$q_k = k\frac{\partial q}{\partial k} = -k\frac{\partial T_k}{\partial x} - k\frac{\partial T}{\partial x}. \quad (26)$$

While Fourier's Law involves a single term, the sensitivity of Fourier's Law with respect to the thermal conductivity involves two terms. The first term involves what can be thought of as a flux of sensitivity information plus a second term that is the heat flux itself. Combining Eq. (25) and Eq. (26), the T_k equation becomes

$$C\frac{\partial T_k}{\partial t} - \frac{\partial}{\partial x}\left(k\frac{\partial T_k}{\partial x}\right) = \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = -\frac{\partial q}{\partial x}. \quad (27)$$

Again, the left hand side of the T_k equation is identical in form to the energy equation; the right hand side has a fictitious source term that is equal to the negative of the gradient of the heat flux. *Gradients in local heat flux drive the thermal conductivity sensitivity field.*

Care must be exercised in deriving the boundary conditions for the T_k equation. Intuitively, one might be inclined to derive a boundary condition on $k\partial T_k/\partial x$. However, we need a condition on q_k which can simply be derived by differentiation of the right hand sides of Eq. (14) and Eq. (15) with respect to the thermal conductivity. The results are

$$q_k\big|_{x=0} = k\frac{\partial q}{\partial k}\bigg|_{x=0} = -4\varepsilon\sigma T^3 T_k\big|_{x=0} \quad (28)$$

$$q_k\big|_{x=L} = k\frac{\partial q}{\partial k}\bigg|_{x=L} = hT_k\big|_{x=L}. \qquad (29)$$

Again, the nonlinear, inhomogeneous boundary condition for the energy equation has become a linear, homogeneous boundary condition for the thermal conductivity sensitivity equation, provided the temperature field is known. The initial condition simply becomes

$$k\frac{\partial T}{\partial k}\bigg|_{x,0} = T_k\big|_{x,0} = 0. \qquad (30)$$

The only inhomogeneous term in the thermal conductivity sensitivity equation is the gradient of heat flux term on the right hand side of the T_k equation in Eq. (27).

We have addressed two of the three gradient type boundary conditions that commonly occur. The third type of gradient boundary condition is a specified heat flux. This kind of boundary condition can occur, for example, when there is an electric heater present. Since the magnitude of a specified flux is independent of either the thermal conductivity or the volumetric heat capacity, this boundary condition becomes "adiabatic like"

$$\frac{\partial T_C}{\partial x}\bigg|_b = q_k\big|_b = 0 \qquad (31)$$

where the subscript b designates the generic boundary along which this boundary condition is applied. Note that for the thermal conductivity sensitivity coefficient, $\partial T_k/\partial x \neq 0$ along a specified flux boundary. This is a subtle point that requires careful thought.

The last boundary condition type that we will address is the specified temperature boundary condition. Again, since this is an imposed boundary condition that is independent of volumetric heat capacity or thermal conductivity, specified temperature boundary conditions become specified sensitivity coefficient boundary conditions with a value of zero.

$$T_C\big|_b = T_k\big|_b = 0, \text{ along specified T boundaries} \qquad (32)$$

Of the parameters listed in Eq. (17), the thermal conductivity k and volumetric heat capacity C are special in that they both appear in the describing differential equation. For their respective sensitivity describing equation, inhomogeneous terms are present. For all other parameters that do not appear in the energy differential equation, their sensitivity describing equation can be written as

$$C\frac{\partial T_p}{\partial t} - \frac{\partial}{\partial x}\left(k\frac{\partial T_p}{\partial x}\right) = 0, p \neq k, C \qquad (33)$$

where T_p is the sensitivity coefficient for parameter p; note that this is a homogeneous equation.

The boundary conditions for the four remaining parameters in Eq. (17) will now be addressed. Differentiating the $x = 0$ boundary condition given by Eq. (14) with respect to these parameters results in

$$\varepsilon\frac{\partial q}{\partial \varepsilon}\bigg|_{x=0} = \varepsilon\left[\sigma(T_r^4 - T^4) - 4\varepsilon\sigma T^3\frac{\partial T}{\partial \varepsilon}\right]\bigg|_{x=0} \qquad (34)$$
$$= [\varepsilon\sigma(T_r^4 - T^4) - 4\varepsilon\sigma T^3 T_\varepsilon]\big|_{x=0}$$

$$\Delta T_r\frac{\partial q}{\partial T_r}\bigg|_{x=0} = \Delta T_r\left[4\varepsilon\sigma T_r^3 - 4\varepsilon\sigma T^3\frac{\partial T}{\partial T_r}\right]\bigg|_{x=0} \qquad (35)$$
$$= 4\varepsilon\sigma(T_r^3\Delta T_r - T^3 T_{T_r})\big|_{x=0}$$

$$h\frac{\partial q}{\partial h}\bigg|_{x=0} = h\left(-4\varepsilon\sigma T^3\frac{\partial T}{\partial h}\right)\bigg|_{x=0} = -4\varepsilon\sigma T^3 T_h\big|_{x=0} \qquad (36)$$

$$\Delta T_\infty\frac{\partial q}{\partial T_\infty}\bigg|_{x=0} = \Delta T_\infty\left(-4\varepsilon\sigma T^3\frac{\partial T}{\partial T_\infty}\right)\bigg|_{x=0} \qquad (37)$$
$$= -4\varepsilon\sigma T^3 T_{T_\infty}\big|_{x=0}$$

Rather than using T_r and T_∞ to scale their respective sensitivity coefficients, a temperature change ΔT_r or ΔT_∞ is used. This eliminates problems with zero temperature when absolute temperature units are not used. These reference temperature changes represent a characteristic temperature change for the problem. As an example, one might choose the maximum temperature rise of the system, T_{max} - T_i. The same reference temperature rise could be used for both T_r and T_∞ sensitivities, although this is not necessary. Since the describing equation is homogeneous for those parameters that do not appear in the energy equation, the inhomogeneities in the boundary or initial conditions will drive the remaining sensitivities. For example, the emissivity sensitivity is driven in Eq. (34) by the radiative heat flux term $\varepsilon\sigma(T_r^4 - T^4)$.

By now, one should see a pattern developing in the sensitivity equations. With this in mind, the remaining results will be given as

$$\varepsilon\frac{\partial q}{\partial \varepsilon}\bigg|_{x=L} = hT_\varepsilon\big|_{x=L} \qquad (38)$$

$$\Delta T_r \left.\frac{\partial q}{\partial T_r}\right|_{x=L} = hT_{T_r}\Big|_{x=L} \tag{39}$$

$$h\left.\frac{\partial q}{\partial h}\right|_{x=L} = [h(T-T_\infty)+hT_h]\Big|_{x=L} \tag{40}$$

$$\Delta T_\infty \left.\frac{\partial q}{\partial T_\infty}\right|_{x=L} = h(T_{T_\infty}-\Delta T_\infty)\Big|_{x=L}. \tag{41}$$

The inhomogeneities in the h and T_∞ sensitivity boundary conditions at $x = L$ are the convective heat fluxes $h(T-T_\infty)$ and $h\Delta T_\infty$ respectively.

We have discussed the initial conditions for both T_C and T_k, Eq. (24) and Eq. (30) respectively; they are both zero. It is easy to see that if the parameter of interest is anything other than the initial temperature itself, the initial condition for T_p will be zero. The initial conditions can be summarized as follows:

$$p_i \left.\frac{\partial T}{\partial p_i}\right|_{x,\,0} = T_{p_i}\Big|_{x,\,0} = \begin{pmatrix} 0, p_i \neq T_i \\ \Delta T_i, p_i = T_i \end{pmatrix}. \tag{42}$$

As with other sensitivity coefficients related to temperature, we have used a temperature change as a scale factor.

After the implementation of the sensitivity equations, the first step is to perform verification calculations to insure that the equations are being solved correctly. Since we already have evaluated the analytical solution for the problem described in Fig. 3, we will repeat the solution to this problem using a control volume finite element method with a lumped capacitance and fully implicit time integration scheme. We will compute a percent error as a function of a grid metric using Eq. (8). This will allow one to verify that the order of convergence of the scheme as the grid is refined. All calculations were performed with a fixed grid scale Fourier number ($\alpha\Delta t/\Delta x^2 = 5.0$). We focus on the spatial location $x = 0$ for times of 4 s ($\alpha t/L^2 = 0.1$) and 20 s ($\alpha t/L^2 = 0.5$). The grid refinement results are shown in Fig. 7. The 4 s results are all less accurate than the 20 s results. The errors for T_k are highest for a given time; however, the ordering of the errors for the other two sensitivity coefficients are not consistent as time increases. Although not shown, the error in temperature rise will be the same as the error in the heat flux sensitivity. The line labeled "2nd order" is a reference line indicating a slope of -2; the algorithm used to numerically solve the equations is theoretically second order accurate for a spatially uniform mesh. These results confirm the approximate second order behavior of the numerical algorithm for sensitivity

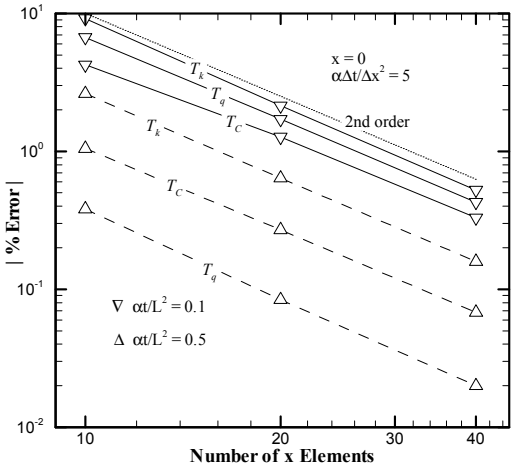


Figure 7. Grid refinement error in sensitivity coefficient calculation for 1-D planar slab with constant flux.

coefficients. The results of Fig. 7 also point out that if a certain level of accuracy is required for all sensitivity coefficients, then a different mesh may be required for each sensitivity coefficient. For example, if an error of approximately 0.4% is required, then 10, 20, and 40 elements would be required for T_q, T_C and T_k respectively for $\alpha t/L^2 = 0.5$.

Additional details on the sensitivity equation method can be found in references [12]-[30].

COMPARISON OF SEM AND DISCRETE ADJOINT METHODS FOR STEADY STATE PROBLEMS

The discrete form of the steady state energy equation can be written in matrix-vector form as

$$[K]\{T\} = \{S\} \tag{43}$$

where $[K]$ is the global conduction matrix, $\{T\}$ is the vector of unknown temperatures and $\{S\}$ is the source/right hand side vector. Sensitivity coefficients can be computed by differentiating the discrete energy equation with respect to p_i, an arbitrary element of the parameter vector $\{p\}$, to obtain

$$[K]\frac{\partial\{T\}}{\partial p_i} + \frac{\partial[K]}{\partial p_i}\{T\} = \frac{\partial\{S\}}{\partial p_i}, \quad i = 1, \ldots, n_p. \tag{44}$$

Experience indicates that scaled sensitivity coefficients, which are defined by

$$T_{p_i} = p_i \frac{\partial T}{\partial p_i} \tag{45}$$

are useful concepts. Multiplying Eq. (44) by the

nominal parameter value p_i and rearranging, the linear system of equations that determines the scaled sensitivity coefficient becomes

$$[K]\{T_{p_i}\} = p_i\frac{\partial\{S\}}{\partial p_i} - p_i\frac{\partial[K]}{\partial p_i}\{T\},\ i = 1, \ldots, n_p \quad (46)$$

For each parameter value p_i, an additional system of linear equations must be solved. This solution will give the scaled sensitivity coefficient at each nodal point in the computational domain. For parameter sensitivity studies, it may be desirable to have the sensitivity coefficient at every point in the computational domain. However, in parameter estimation work, the sensitivity coefficients may be desired only at selected locations. For example, in the estimation of thermal properties from temperature measurements, a finite number of sensors are used and the sensitivity coefficients are desired only at the temperature sensor locations.

Adjoint methods offer some potential computational savings when the number of sensor locations are few and the number of parameters are large. Following Kirsch [31], the adjoint method can be developed by multiplying the sensitivity coefficient equation, Eq. (46), by the inverse of the global conduction matrix.

$$[K]^{-1}[K]\{T_{p_i}\} = [K]^{-1}\left[p_i\frac{\partial\{S\}}{\partial p_i} - p_i\frac{\partial[K]}{\partial p_i}\{T\}\right]. \quad (47)$$

The left hand side of Eq. (47) yields the vector of sensitivity coefficients for parameter p_i at all nodal locations; however, we are only concerned with the sensitivity coefficient at a few selected locations in the computational domain. To extract the sensitivity coefficient at a single location in the computational domain, define a row-vector that has zeros everywhere except for unity at the j-th nodal location

$$\{I_j\}^T = \{0, \ldots, 0, 1, 0, \ldots 0\} \quad (48)$$

and multiply Eq. (47) by this vector results in

$$\{T_{p_i}\}_j = \{I_j\}^T[K]^{-1}[K]\{T_{p_i}\} = \{I_j\}^T[K]^{-1}\left[p_i\frac{\partial\{S\}}{\partial p_i} - p_i\frac{\partial[K]}{\partial p_i}\{T\}\right]. \quad (49)$$

Eq. (49) gives the sensitivity coefficient for parameter p_i at nodal location j. It is computationally convenient to define the coefficient of the square brackets on the right hand side of Eq. (49) as the adjoint variable vector and is

$$\{\xi_j\}^T = \{I_j\}^T[K]^{-1}. \quad (50)$$

Taking the transpose of Eq. (50) yields

$$\{\xi_j\} = ([K]^{-1})^T\{I_j\} = ([K]^T)^{-1}\{I_j\} \quad (51)$$

which can be written as

$$[K]^T\{\xi_j\} = \{I_j\},\ j = 1, \ldots, n_s. \quad (52)$$

Although Eq. (52) is valid at all n-nodal locations, the adjoint variable approach is attractive only when the number of sensors n_s is a small subset of n. Note that Eq. (52) is independent of the particular sensitivity coefficient one is trying to compute; this means that the adjoint variable vector depends only on the spatial (sensor) location in the computational domain (and $[K]$). Once Eq. (52) has been solved for the adjoint variable vector $\{\xi_j\}$, the sensitivity coefficient for all parameters of interest (at this nodal location) can be computed from Eq. (49), which is written as

$$\{T_{p_i}\}_j = \{\xi_j\}^T\left[p_i\frac{\partial\{S\}}{\partial p_i} - p_i\frac{\partial[K]}{\partial p_i}\{T\}\right],\ i = 1, \ldots, n_p. \quad (53)$$

Eq. (46) defines the discrete form of the SEM while the Eq. (52) and Eq. (53) define the discrete adjoint equations. Both approaches have a single left hand side matrix but multiple right hand side vectors. The number of right hand side vectors can be used as a rule of thumb for when one method is computationally more efficient than the other.

use SEM when $n_p < n_s$

use adjoint when $n_s < n_p$

Obviously, when n_p and n_s are approximately equal, this rule of thumb will have to be inspected more closely.

FIRST ORDER PROPAGATION OF UNCERTAINTY IN COMPUTATIONAL MODELS

Sensitivity coefficients are used in the propagation of uncertainty through computational models. The process is very analogous to experimental uncertainty estimation. Following Coleman and Steele [32], the first order uncertainty propagation equation is

$$\sigma_T^2 = \left(T_{\beta_1}\frac{\sigma_{\beta_1}}{\beta_1}\right)^2 + \left(T_{\beta_2}\frac{\sigma_{\beta_2}}{\beta_2}\right)^2 + \ldots. \quad (54)$$

Note that the uncertainty propagation equation has been written in terms of scaled sensitivity coefficients. If the sensitivity coefficients are computed at every nodal location in a

computational domain, then the uncertainty estimation due to parameter uncertainty is just a post processing of all the field variables. An uncertainty estimation for a thermally activated battery is given in Blackwell, et al. [15]. Additional details on uncertainty propagation are contained in Fadale [33] and Fadale and Emery [34].

SUMMARY

Six methods for computing sensitivity coefficients have been discussed. Example calculations were presented for several of them. The methods discussed can be divided into two broad categories; code invasive and code non-invasive. The *finite difference method* is non-invasive and probably the most general; it can be applied when the source code is not available. This means it can be used in conjunction with commercially available software. An objection to the *finite difference method* is that for non-linear problems such as temperature dependent properties, each perturbed parameter solution is a non-linear solve. If the same problem is solved using the *sensitivity equation method* (very code invasive), the sensitivity coefficient equations are linear equations. If the source code is available and the sensitivity equation method is not a viable option, then the *complex step method* should be seriously considered since it eliminates the step size issue. No matter which method is chosen to be the primary method, differentiation of analytical solutions is an important part of the process of verifying that your equations have been implemented correctly.

ACKNOWLEDGMENTS

This work was funded by Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

REFERENCES

1. J. V. Beck and K. J. Arnold, *Parameter Estimation in Engineering and Science*, Wiley & Sons, New York, 1977.
2. J. V. Beck, B. F. Blackwell, and C. R. St. Clair, *Inverse Heat Conduction-Ill Posed Problems*, Wiley & Sons, New York, 1985.
3. R. J. McMasters, K. J. Dowding, J. V. Beck, and D. H. Y. Yen, Methodology to Generate Accurate Solutions for Verification in Transient Three-Dimensional Heat Conduction, accepted for publication in *Numerical Heat Transfer*, *Part B*, 2002.
4. J. Martins, I. Kroo, and J. Alonso, An Automated Method for Sensitivity Analysis using Complex Variables, AIAA Paper 2000-0689, Proceedings of the 38th Aerospace Sciences Meeting, Reno, NV, January 2000.
5. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
6. James C. Newman, III and David L. Whitfield, A Step-Size Independent Approach for Multidisciplinary Sensitivity Analysis and Design Optimization, AIAA-99-3102.
7. J. R. R. A. Martins, I. M. Kroo, and J. J. Alonso, An Automated Method for Sensitivity Analysis using Complex Variables, AIAA-2000-0689, Presented at 38th Aerospace Sciences Meeting and Exhibit, Reno, NV, January 10-13, 2000.
8. J. M. Janus and J. C. Newman III, Aerodynamic and Thermal Design Optimization for Turbine Airfoils, AIAA-2000-0840, Presented at 38th Aerospace Sciences Meeting & Exhibit, Reno, NV, January 10-13, 2000.
9. L. Massa and J. M. Janus, Aerodynamic Sensitivity Analysis of Unsteady Turbine Stages, AIAA-2001-2579, Presented at 15th AIAA Computational Fluid Dynamics Conference, Anaheim, CA, June 11-14, 2001.
10. http://aero-comlab.stanford.edu/jmartins
11. C. Bischof, P. Khademi, A. Mauer, and A. Carle, Adifor 2.0: Automatic Differentiation of Fortran 77 Programs, *IEEE Computational Science & Engineering*, pp. 18-32, Fall 1996.
12. B. F. Blackwell, R. J. Cochran, and K. J. Dowding, Development and Implementation of Sensitivity Coefficient Equations for Heat Conduction Problems, Proceedings of 7th AIAA/ASME Joint Thermophysics and Heat Transfer Conference, ASME/HTD Vol. 357-2, pp. 303-316, 1998.
13. K. J. Dowding, B. F. Blackwell, and R. J. Cochran, Application of Sensitivity Coefficients for Heat Conduction Problems, Proceedings of 7th AIAA/ASME Joint Thermophysics and Heat Transfer Conference, ASME/HTD Vol. 357-2, pp. 317-327, 1998.
14. K. J. Dowding, J. V. Beck, and B. F. Blackwell, Estimating Temperature Dependent Properties of Carbon-Carbon Composite, AIAA 98-2933, presented at 7th AIAA/ASME Joint Thermophysics and Heat Transfer Conference, Albuquerque, NM, 1998.
15. B. F. Blackwell, K. J. Dowding, R. J. Cochran, and D. Dobranich, Utilization of Sensitivity Coefficients to Guide the Design of a Thermal Battery, Proceedings of ASME Heat Transfer Division, HTD-Vol. 361-5, pp. 73-82, 1998.

16. Kevin J. Dowding, Ben F. Blackwell, and R. J. Cochran, Study of Heat Flux Gages Using Sensitivity Analysis, Proceedings of ASME Heat Transfer Division, HTD-Vol. 361-5, pp. 595-602, 1998.
17. Kevin Dowding, and Ben Blackwell, Design of Experiments to Estimate Temperature Dependent Thermal Properties, presented at Third International Conference on Inverse Problems in Engineering, Port Ludlow Washington, June 13-18, 1999.
18. Bennie F. Blackwell and Kevin J. Dowding, Sensitivity Analysis and Uncertainty Propagation in a General-Purpose Thermal Analysis Code, Presented at 3rd ASME/JSME Joint Fluids Engineering Conference & FED Annual Summer Meeting/Exposition, July 18-22, San Francisco, CA.
19. B. F. Blackwell, K. J. Dowding, and R. J. Cochran, Development and Implementation of Sensitivity Coefficient Equations for Heat Conduction Problems, *Numerical Heat Transfer, Part B*, 36:15-32, 1999.
20. K. J. Dowding, B. F. Blackwell, and R. J. Cochran, Application of Sensitivity Coefficients for Heat Conduction Problems, *Numerical Heat Transfer, Part B*, 36:33-55, 1999.
21. Kevin J. Dowding and Bennie F. Blackwell, Sensitivity Analysis for Nonlinear Heat Conduction, *ASME Journal of Heat Transfer*, Vol. 123, pp. 1-10, February 2001.
22. K. J. Dowding, J. Beck, A. Ulbrich, B. F. Blackwell, and J. Hayes, Estimation of Thermal Properties and Surface Heat Flux in Carbon-Carbon Composite, *Journal of Thermophysics and Heat Transfer*, Vol. 9, No. 2, pp. 345-351, 1995.
23. K. J. Dowding, J. V. Beck, and B. F. Blackwell, Estimation of Directional-Dependent Thermal Properties in a Carbon-Carbon Composite, *International Journal of Heat and Mass Transfer*, Vol. 39, No. 15, pp. 3157-3164, 1996.
24. E. Turgeon, D. Pelletier, and J. Borggaard, A Continuous Sensitivity Equation Approach to Optimal Design in Mixed Convection, AIAA 99-3625, Presented at 33rd Thermophysics Conference, Norfolk, VA, June 28-July 1, 1999.
25. E. Turgeon, D. Pelletier, and J. Borggaard, A General Continuous Sensitivity Equation Formulation for Complex Flows, AIAA 2000-4732, Presented at 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA, Sept. 6-8, 2000.
26. E. Turgeon, D. Pelletier, and J. Borggaard, A Continuous Sensitivity Equation Method for Flows with Temperature Dependent Properties, AIAA 2000-4821, Presented at 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA, Sept. 6-8, 2000.
27. E. Turgeon, D. Pelletier, and J. Borggaard, Sensitivity and Uncertainty Analysis for Variable Property Flows, AIAA 2001-0139, Presented at 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 8-11, 2001.
28. Andrew G. Godfrey and Eugene M. Cliff, Sensitivity Equations for Turbulent Flows, AIAA 2001-1060, Presented at 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 8-11, 2001.
29. E. Turgeon, D. Pelletier, and J. Borggaard, Application of a Sensitivity Equation Method to the k-ε Model of Turbulence, AIAA 2001-2534, Presented at 15th AIAA Computational Fluid Dynamics Conference and Exhibit, Anaheim, CA, June 11-14, 2001.
30. E. Turgeon, D. Pelletier, and J. Borggaard, A General Continuous Sensitivity Equation Formulation for the k-ε Model of Turbulence, AIAA 2001-3000, Presented at 15th AIAA Computational Fluid Dynamics Conference and Exhibit, Anaheim, CA, June 11-14, 2001.
31. U. Kirsch, *Structural Optimization*, Springer-Verlag, New York, 1993.
32. H. W. Coleman and W. G. Steele, *Experimentation and Uncertainty Analysis for Engineers*, 2nd ed., Wiley, New York, 1999.
33. T. D. Fadale, "Uncertainty Analysis using Finite Elements," Ph.D. Thesis, University of Washington, 1993.
34. T. D. Fadale, and A. F. Emery, "Transient Effects of Uncertainties on the Sensitivities of Temperatures and Heat Fluxes Using Stochastic Finite Elements," *ASME Journal of Heat Transfer,* Vol. 116, pp. 808-814, 1994.

Application of Genetic Algorithms and Neural Networks to the Solution of Inverse Heat Conduction Problems

A Tutorial

Keith A. Woodbury
Department of Mechanical Engineering
The University of Alabama
Tuscaloosa, Alabama, USA
woodbury@me.ua.edu

ABSTRACT

Genetic Algorithms and Neural Networks are relatively new techniques for optimization and estimation. These techniques can, of course, be applied to the solution of inverse problems. This paper presents a tutorial for application of these techniques to the solution of some simple inverse problems. A description of each of the techniques precedes presentation of the algorithms. MATLAB is used to solve these problems.

INTRODUCTION

Genetic Algorithms are a class of search methods, which are patterned after evolutionary processes. These algorithms have existed for perhaps 20 years, but were first popularized following the publication of David Goldberg's text on the subject (Goldberg, 1989). These algorithms search a solution space by manipulating populations of candidate solutions. These populations are evaluated to determine the best members of each generation, and each generation reproduces to create the next generation. Notions from evolution are borrowed to manipulate the population after reproduction: randomly triggered crossover and mutation enter in to widen the search region. At the end of a pre-specified number of generations, the results are examined.

A Genetic Algorithm is a type of search procedure. Simply put, it is a localized random search. It requires no evaluation of the derivative of the performance measure, and is therefore highly suited to nonlinear problems.

Neural Networks describe another type of algorithm that is borrowed from nature. Neural Networks are an attempt to model the massively parallel operation of a brain. A collection of simple *neurons* is interconnected with links. Each link will be assigned a *weight* during the *training* of the network. During operation of the network, the output of each neuron is the result of the weighted sum of all the inputs connected to it passed through an *activation function*. This activation function is some suitable non-linear mathematical function. It is through the sum total action of many connected neurons that the network is able to "learn" its behavior and produce intelligible results.

A Neural Network is an interpolative procedure. Through training, the network "learns" an association between a collection of inputs and their corresponding outputs. In operation, when the network is presented inputs, which were not present in the training set, the network will produce a result, which is consistent with the training data.

This paper has two parts, one devoted to Genetic Algorithms and one dedicated to Neural Networks. In each section a basic description of the method is given, followed by application to a simple optimization problem of parameter identification. Then each method is applied to a classic inverse boundary problem. The MATLAB programming language is used to illustrate the algorithms.

GENETIC ALGORITHMS

These algorithms mimic the evolutionary processes that have led to development of higher organisms in nature. An initial population of candidates reproduces to create a new generation of the population. In each generation there are random occurrences of mutation in the population. Above all, *survival of the fittest* ensures that the "best" members of the population are retained.

Randomness plays a central role in the genetic algorithm search process. A random number generator will be called thousands of times during the execution of a genetic algorithm.

A "genetic algorithm" is one that possesses the following characteristics:

1. An initial population of a fixed number of candidate solutions is selected.
2. The "fitness" of each of the members of the population is determined using the performance measure for the problem.
3. The members of the current generation of the population reproduce to form the next generation. The reproduction should favor the better members as "parents". During reproduction, crossover of the genes results in new members not originally in the previous generation but related to them.
4. Random mutation of some of the "children." These mutations introduce new characteristics not in the previous generation and not directly related to the previous generation but which may result in a "more fit" child.
5. The reproduction continues until a preset number of generations have been created.

At least two types of encoding are possible for the members of the population: binary strings and real number arrays. The latter are more useful for numerical problems, but the former are classic in the genetic algorithms and will be discussed first.

Binary Encoding

Early applications of genetic algorithms clung to the notion of an organism (member of the population) represented as a *gene* through a binary string of *chromosomes*. The binary string could be interpreted as a color code, and ASCII letter code, an integer code, etc., depending on the problem at hand. But the use of binary encoding facilitates application of the analogy to evolutionary processes.

Initial Population. The initial population is typically seeded randomly, but this need not be the case. In the case of binary strings, this could be done bit-by-bit with random selection of a zero or a one for each location.

Selection of Parents. Parents are chosen for reproduction based on their fitness. One complicated method for selection of parents is called roulette wheel selection (Davis (1991)), however any method that favors the fitter members of the population may be used.

Reproduction. Two Binary strings reproduce through crossover of their chromosomes. After two parents are selected, the child of the two parents is created by splitting the gene (binary string) at one or more points (randomly chosen) and splicing the pieces together.

Figure 1 illustrates reproduction with binary strings. Once the two parents are identified, a crossover point is randomly selected. The Child "AB" results from splicing the first portion of the genetic string "A" onto the second portion of the genetic string "B". Note that a scond child "BA" could easily be produced by splicing the remaining portions of the strings.

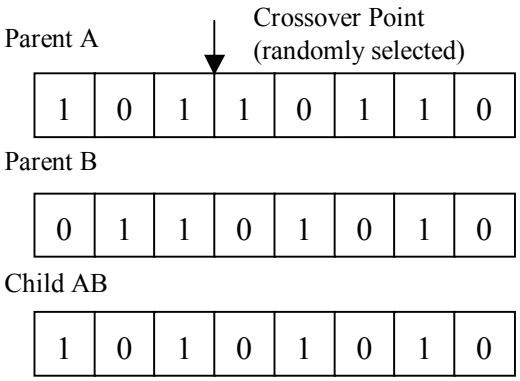


Figure 1. Reproduction through crossover with binary strings.

Mutations. After the new generation is created, mutations occur. For each member of the population a random number in the range (0,1) is generated. If the random number is below some prespecified mutation threshold, then the gene is allowed to mutate. Mutation in a binary string is accomplished by selecting one or more (but not all!) of the chromosomes in the string to be altered. Of course "alteration" of a binary digit can only mean changing its sense (flipping it's bit): if it's a zero, make it a one, and if it's a one, make it a zero.

Real Number Encoding

Binary encoding can literally be applied to any problem at hand. However, if the problem at hand happens to be numerical, and the unknowns of the problem happen to be continuous real numbers, then it is very convenient to represent the members of the population using real number

encoding. A simple example that will be presented below is identification of the slope and intercept of a line based on knowledge of a few of the points on the line. Although these two real numbers (the slope and intercept) most certainly can be represented as a string of binary digits (indeed, in the heart of our computers that is the only form in which they exist!), it is much easier to let the "gene" be an real array of length two. For such a representation we must discover suitable methods for initial population generation, reproduction, and mutation.

Initial Population. The initial real number arrays are of course generated randomly. There are at least two possibilities for this and either one could be used, but in any case *the domain of the individual variables* must be known. Remember that the genetic algorithm is only a search mechanism, and the limits of the search region must be known. One possible method of seeding the initial population is to generate a random number for each member of each of the arrays. A second possibility is to generate a random number for each array in the population and assign each array a constant value.

Selection of Parents. My technique for selection of parents for reproduction of the next generation is simple. After evaluating the fitness of all the members of the population, sort them in best-to-worst order, then use a specified number of the "best" as parents for the new generation. This assures that the next generation is reproduced using characteristics of the best of the previous generation.

Reproduction. The key to reproduction is some sort of crossover mechanism that combines characteristics from each of the parents. In arrays of real numbers both the magnitude of the individual members and the order of the array are important. At least two methods of crossover are possible that address these two characteristics.

To alter the magnitude, averaging of the two parent strings will achieve the desired effect (Davis, 1991). Rather than a simple arithmetic average, I have employed a weighted average with the weight being chosen randomly.

To modifiy the order of the numbers in each child array, I use a crossover technique identical to that in Fig. 1 for binary strings. Choose a random location in the array, and crossover the sub-arrays from each partner.

Mutation. The main purpose of mutation is to introduce new information into the population that can't be obtained directly from the parents. The method I use was suggested by (Davis, 1991) and is a simple replacement of the array with a randomly generated array within the search space. A mutation threshold must be passed first before the mutation is applied.

Creep. This is really a second type of mutation but proves very necessary to refine the search. Creep (Davis, 1991) refers to the drift of the members of a real array around their present value. For each member of the population, a creep threshold is applied, and if the member is eligible to creep, the magnitude of each number in the array is scaled by a random number in the range $(1 - C, 1 + C)$, where C is a fraction between 0 and 1.

Elitism. One final mechanism that often is introduced into genetic algorithms is *elitism*. When elitism is employed, the best (or N_{elite} best) members of each generation are retained in the next generation. This allows the characteristics of the "super-individual" to dominate over several generations.

A Simple Example

To illustrate the techniques for real number encoding, consider a simple parameter estimation problem. Suppose we have a number N_{data} of (x_i,y_i) data pairs and we want to know the equation of a straight line that passes thorugh these points (or close to them):

$$y = b + mx; \tag{1}$$

in other words, the constants b and m are to be determined. The classic solution to this problem involves the minimization of the sum of the squared errors between the model-predicted value and the corresponding data value:

$$S = \sum_{i=i}^{N_{data}} (y_i - y)^2 \tag{2}$$

This same methodology will be used to solve this problem using genetic algorithms.

MATLAB was used to code a genetic algorithm to solve this problem, and the main function is shown in Listing 1. Several parameters are passed to the routine: the *xvals* at which the known *ydata* are supplied, the domain of the search (*low, high*), which applies to both the slope m and intercept b. Other parameters must be specified for the search: *Npop* is the number of members of the population; *Nbest* is the number of the best members of the population to used for reproduction at each new generation; *Ngen* is the number of generations to produces before the program terminates; *mut_chance* is the

probability threshold for a gene mutation; *creep_chance* is the threshold for creep of the member and *creep_amount* is the maximum magnitude of the random creep (the parameter *C* introduced above).

The routine begins with initialization of some arrays and constants. The value *Nelite* is set to one, which means that only the best member of the population is retained from one generation to the next. The *population* is an array of real numbers with *Npop* rows and two columns (one for intercept *b* and one for slope *m*). This array is initialized using uniform random numbers in the range (*low, high*) for the intercept and slope.

The main loop of the routine performs the following steps. The model (Eq. (1)) is used to compute the values *ytest* using the function *straight_line* and all the current members of the population. These *ytest* values are compared to the *ydata* values supplied and a fitness index is computed for each member of the population using Eq. (2). These fitness values are used to sort the population from best to worst, and the *Nbest* members are used for reproduction of the next generation. The reproduction is performed by the routine *reproduce_by_weighted_avg* using the weighted averaging scheme described earlier. Only after new children are produced, the crossover mechanism is applied to all the new members, and the mutation and creep mechanisms are applied randomly based on the thresholds specified.

As a demonstration, the data for a straight line with intercept $b = 1$ slope $m = 2$ is used (*xvals* = [1 2 3 4 5]; and *ydata* = [3 5 7 9 11]. The parameter *mut_chance* was set to 0.1, meaning that there is a 10% chance that a child will mutate (have its values completely replaced by randomly generated numbers in the domain (*low*, *high*)). The parameters *creep_chance* and *creep_amount* were set to 0.90 and 0.25, respectively, meaning that there is a 90% chance that the child will have its value randomly scaled by +/- 25%.

As a first attempt, 50 generations are computed using a population of only 10 members and allowing only the best two for reproduction. At the end of the 50 generations, the best member of the population was $b = -0.0756$and $m = 2.2994$. The resulting "convergence history", which is the error of the best member of the population at each generation, is shown in Fig 2. The corresponding estimates for the points on the line are shown in Fig. 3. These results were obtained in 0.22 seconds of CPU time on a 1000 MHz Pentium 4 processor.

The values obtained are not very good. The size of the population is too small to allow the effects of mutation and creep to widen the search. Note that without mutation and creep, the search space will be constrained to that enclosed by the initial randomly generated population as the weigted averaging cannot create a candidate outside that domain.

For the next attempt, the population is increased to 50 and the number of the best to use for reproduction is increased to 10. The number of generations to compute is increased to 100. At the end of the 100 generations, the best member of the population was $b = 0.9851$ and $m = 2.0042$. The convergence history is shown in Fig. 4 and the computed *y* values are compared with the data

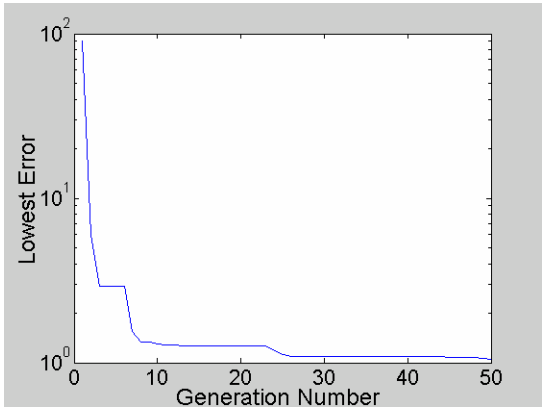


Figure 2. Convergence history for *Npop* = 10 and *Nbest* = 2.

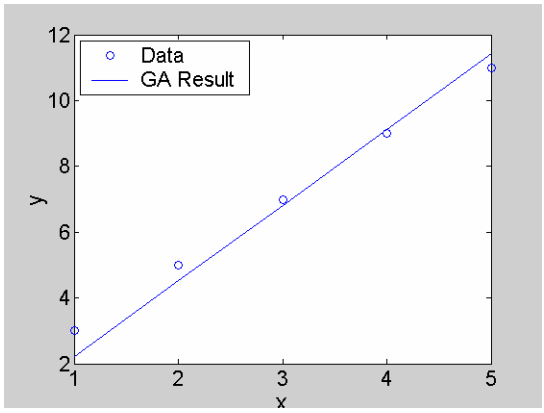


Figure 3. Model estimates (circles) and exact values (line) for *Npop* = 10 and *Nbest* = 2. (*Ngen* = 50)

in Fig. 5. These results were obtained in 10.22 CPU seconds on the 1000 MHz Pentium 4 processor.

Note that the results here are much better than those obtained previously, with the minimum sum squared error below 10^{-3}. From the convergence history (Fig. 4) we can see that, after the initial drop, there is little improvement in the lowest error. In fact, a good solution is obtained after 20 generations or so.

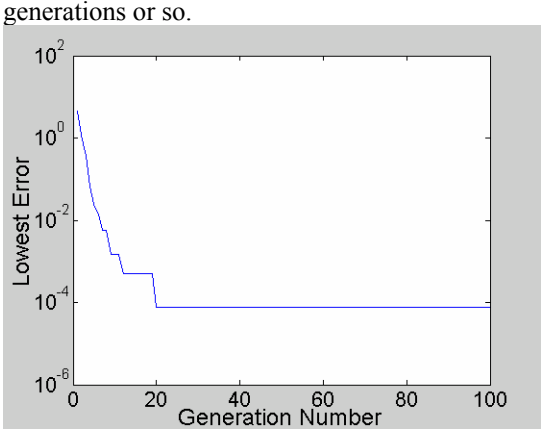


Figure 4. Convergence history for *Npop* = 50 and *Nbest* = 10.

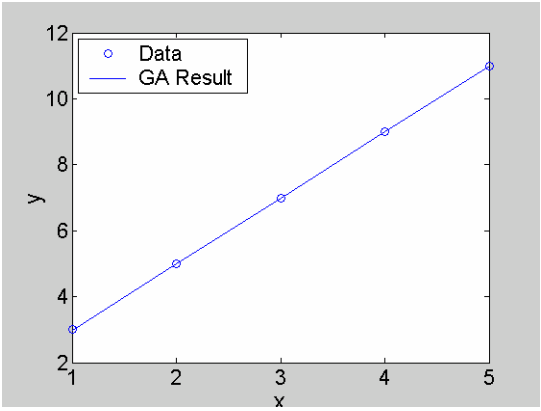


Figure 5. Model estimates (circles) and exact values (line) for *Npop* = 50 and *Nbest* = 10 (*Ngen* = 100)

Classic Function Estimation Example

As a much more challenging example, consider the classical problem of estimation of the surface heat flux history for a one dimensional slab insulated at $x = L$. This problem was considered previously by Raudensky, et al. (1995), but they worked with an unknown heat transfer coefficient rather than the heat flux. The triangular heat flux history popularized by Beck, et al. (1985) will be used as the "unknown function" to generate data for the estimation problem. For simplicity, the parameters are taken as $k = \rho c_p = L = 1$, which of course is the same as using non-dimensional data. Two cases of data are considered: data with an interval 0.18 s and another with interval 0.06 s. The heat flux history has the following character:

$$q(t) = \begin{cases} 0, & t < 0.24 \\ t - 0.24 & 0.24 \le t < 0.84 \\ 0.6 - (t - 0.84) & 0.84 \le t < 1.44 \\ 0 & t \ge 1.44 \end{cases} \quad (3)$$

Data were generated for a sensor located at $x = L$, and the data generated for the two data intervals are shown in Table 1 and Table 2.

Table 1. Artificial data for large time interval (0.18 s).

t, secs	*T, C*
0.00	0.00000
0.18	0.00000
0.36	0.00037
0.54	0.01338
0.72	0.05446
0.90	0.12720
1.08	0.21959
1.26	0.29501
1.44	0.34067
1.62	0.35655
1.80	0.35942
1.98	0.35990

Data Representation. To estimate the heat flux variation, a suitable parameterization of the heax flux function *q(t)* is necessary. We will choose a piecewise constant heat flux, and will estimate one component of heat flux between every temperature data point. So, in the case of the large time interval data of Table 1, there will be 12 unknown heat flux comonents, and the in case of the small data interval in Table 2, there will be 35 unknown components. (Our algorithm estimates a heat flux component for every data point, even the first one; this assumes a zero temperature initial condition). Again, each member of the population will be a real vector of the appropriate length.

Function Evaluation. The objective function for fitness will again be the sum of the squared errors between the model-computed values (objective function for fitness will again

be the sum of the squared errors between the model-computed values (*ytest*) and the data (*ydata*). The model-computed values will be produced using the Duhamel's summation as

Table 2. Artificial data for small time interval (0.06 *s*)

t, secs	*T, C*	*t, secs*	*T, C*
0.00	0.00000	1.08	0.21959
0.06	0.00000	1.14	0.24768
0.12	0.00000	1.20	0.27293
0.18	0.00000	1.26	0.29501
0.24	0.00000	1.32	0.31371
0.30	0.00001	1.38	0.32895
0.36	0.00037	1.44	0.34067
0.42	0.00217	1.50	0.34882
0.48	0.00632	1.56	0.35376
0.54	0.01338	1.62	0.35655
0.60	0.02366	1.68	0.35809
0.66	0.03732	1.74	0.35894
0.72	0.05446	1.80	0.35942
0.78	0.07515	1.86	0.35968
0.84	0.09939	1.92	0.35982
0.90	0.12720	1.98	0.35990
0.96	0.15788	2.04	0.35995
1.02	0.18929		

described in Chapter 3 of Beck, et al. Note that this computation will be approximate, especially for the larger time intervals.

Algorithm Description. The MATLAB main function for this genetic algorithm is shown in Listing 2. Many of the features are the same as in the SimpleGA function, but several enhancements have been made. Specifically, the parameters for the problem (*Ngen*, *mut_chance, creep_chance, creep_amount*) are all vector quantities to facilitate modification of these parameters during the simulation. Also, some regularization has been added via a Tikhonov term, and the coefficients for this are passed through the funciton call. These enhancements will be described in more detail below.

Large time interval data. The case of a larger time step in the data is easier from a classical solution point of view and thus this case will be take first.

As a starting point, use the same parameters that were used at the end of SimpleGA – *Npop* = 50, *Nbest* = 10, and *Ngen* = 100. After 100 generations, the sum squared error *S* is less than 10^{-3}, and the results for the computed temperatures *ytest* can be seen compared to the data in Fig. 6. The convergence history can be seen in Fig. 7. These results are obtained on a 1.6 GHz Pentium 4 in 5.7 CPU seconds.

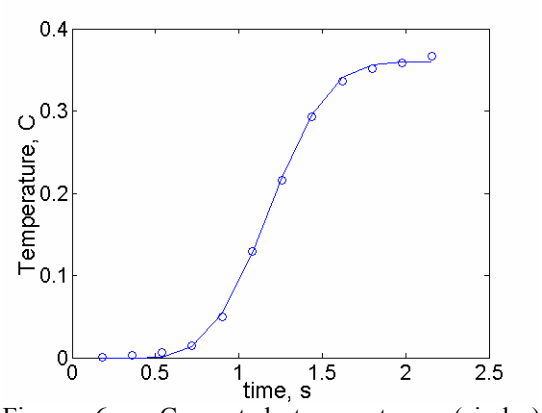


Figure 6. Computed temperatures (circles) compared to data (line) history for first large time interval run.

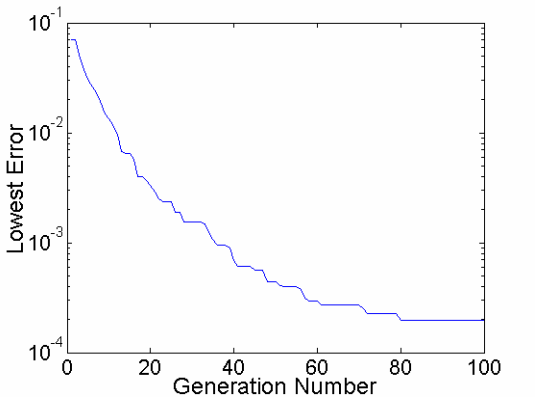


Figure 7, Convergence history for first large time interval run.

The sumsquared error is pretty low, and the computed values of *T(t)* are relatively close to the data values (as shown in Fig. 6). But the estimated heat flux components bear little resemblence to the actual input (see Fig. 8). Considering the results (Fig 8.) and the convergence history, it seems plausible that the problem is not converged well enough. Note that the convergence history decreases past 80 generations, but then levels out. Note also that the history exhibits a mix of large-scale changes (probably caused by mutations) and smaller scale decreases (perhaps brought about by creep). But after 80 generations, the large scale changes (in the domain (-1, 1) and small scale changes (on the order of 25%) are too large to bring any improvement. What is needed is a multiple parameter approach.

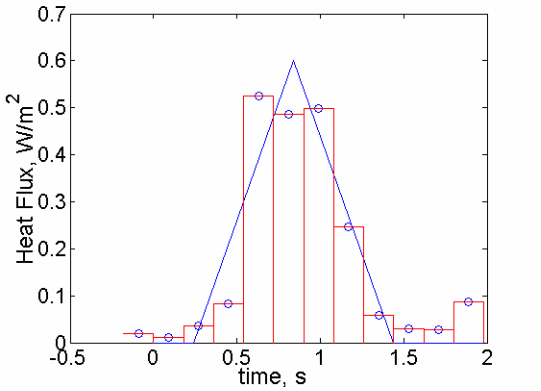


Figure 8. Computed (circles) and actual (line) heat flux input for first large time interval run.

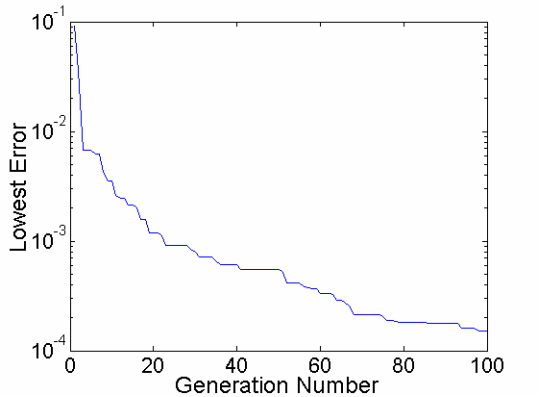


Figure 9. Convergence history for second large time interval run.

A modification to the algorithm allows for this. Next let's try the allowing larger more mutations for the first 50 genreations (*mut_chance* = 0.2), but then decrease the mutation chance to 0.1, and keep the creep chace the same, but decrease the creep amount: *mut_chance* = 0.1, *creep_chance* = 0.9, and *creep_amount* = 0.1). The idea is to let the search fine-tune the result after "getting close".

One other modification was made to the program. After each "break" in the generational loop (after the 50 generations, say), the domain for mutations is changed from the initial values to the (minimum, maximum) of the heat flux vector. The idea is to keep the mutation changes within the most reasonable range.

The convergence history, seen in Fig. 9, shows improvement in both the final value and the sustained decrease past 80 generations. The final *S* parameter is about 2×10^{-4}, which is pretty good. Considering Fig. 9, more generations may help reduce the error.

The convergence history for several subsequent runs are seen in Fig. 10. The final run corresponds to the lowest line, which achieved an *S* parameter of almost 8×10^{-4}. This last run corresponds to a parameter strategy of *Ngen* = [50 100 200 300], *mut_chance* = [0.2 0.1 0.05 0.02], *creep_chance* = [0.9 0.9 0.9 0.9], and *creep_amount* = [0.2 0.1 0.05 0.02]. The graphical comparison of the computed *ytest* and the given data *ydata* is seen in Fig. 11, and it can be seen that the comparison is quite good. The estimates for the heat flux history, seen in Fig. 12, are quite good, and do reproduce the input curve favorably.

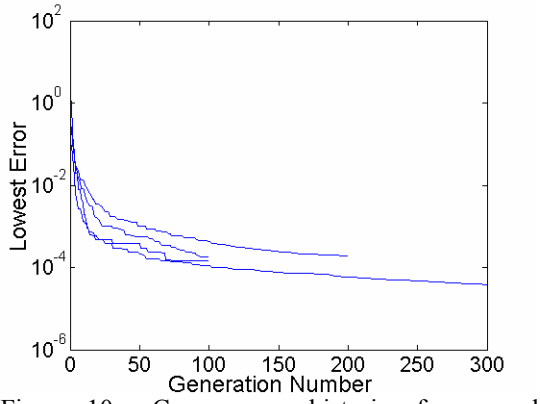


Figure 10. Convergence histories for several subsequent simulations.

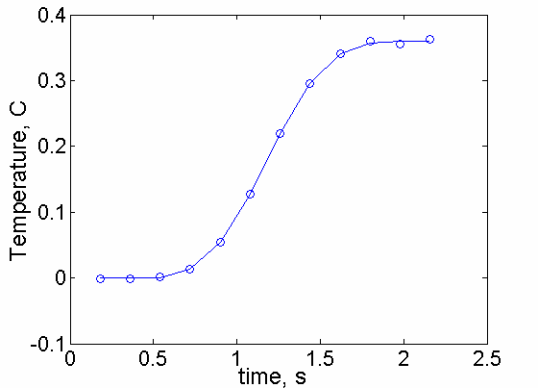


Figure 11. Computed (circles) values of temperature compared with the data (line)

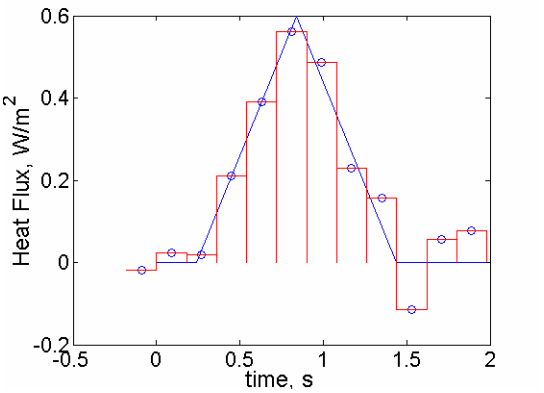


Figure 12. Heat flux estimates for last simulation.

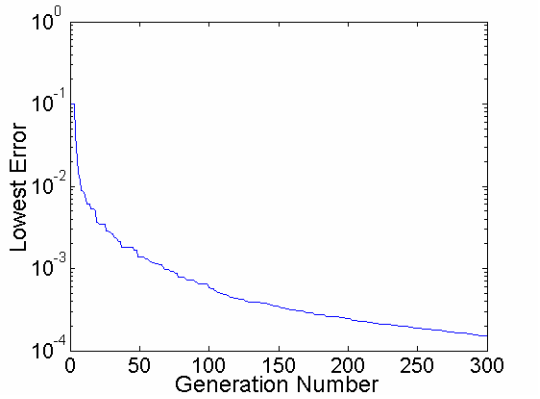


Figure 13. Convergence History for small time step data.

Small Time Interval Data. The data from Table 2, which has a (dimensionless) time step of 0.06, is known to cause estimation problems using unregularized methods (such as Stoltz data matching). We next apply the genetic algorithm search to this data.

The same parameters used were the same as in the last run with large time step in the data (*Ngen* = [50 100 200 300], *mut_chance* = [0.2 0.1 0.05 0.02], *creep_chance* = [0.9 0.9 0.9 0.9], and *creep_amount* = [0.2 0.1 0.05 0.02]). The convergence history (Fig. 13) suggests the results should be good, and the comparison between the computed and measured temperatures confirms this view (Fig. 14). However, the plot of the estimated heat fluxes shows that the underlying ill-posedness of the problem prevents a reasonable result from being obtained (Fig. 15.).

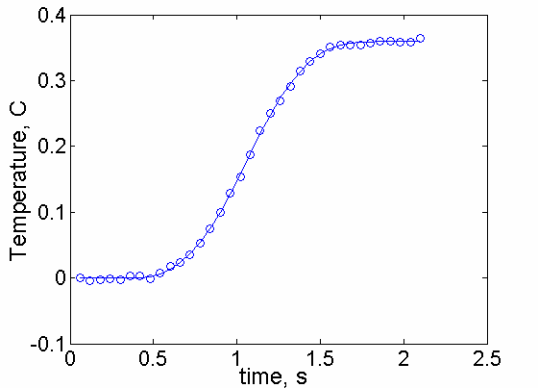


Figure 14. Computed (circles) values of temperature compared with the data (line) for first simulation with small time steps.

To get a better result in the face of ill-posedness, some regularization must be added to the problem. A familiar Tikhonov regularization term can be added to the objective function (fitness measure) to penalize changes in the first derivative of the heat flux. This is implemented in a discrete form (as described in Chapter 4 of Beck, et al, 1985) as:

$$S = \sum_{i=i}^{N_{data}} (y_i - y)^2 + \sum_{j=1}^{N_{data}-2} \alpha_1 (q_{j+1} - q_j)^2 \qquad (3)$$

The α_1 parameter is the regularizing parameter and must be specified.

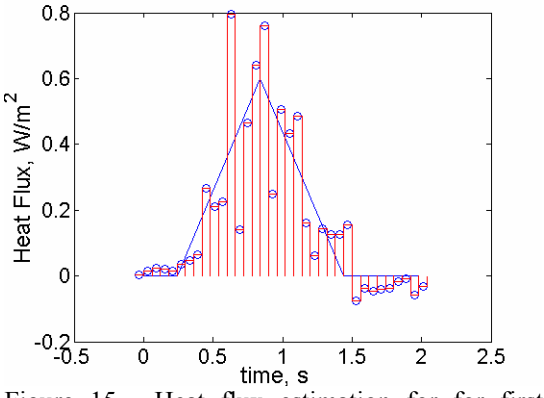


Figure 15. Heat flux estimation for for first simulation with small time steps.

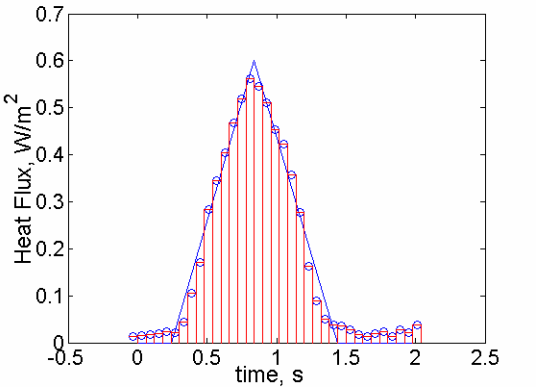


Figure 16. Heat flux estimation with same parameters as Fig. 15, but with Tikhonov regularization (α_1 = 1.e-3).

Figure 16 shows the results from a simulation using the same parameters as before, but with a First order Tikhonov regularizing term (α_1 = 1.e-3). The converged solution more clearly defines the input heat flux. Note that the algorithm has difficulty where the heat flux is zero (Figs. 8, 12, 15, and 16).

Genetic Algorithms Conclusions

Genetic algorithms are a random search procedure that search in a fixed domain without using function gradient information. They can be applied to linear or nonlinear problems and are by nature computationally intensive. Real number arrays can be used as "genes" in the population to represent engineering data. Genetic algorithms can be applied to ill-posed problems such as the inverse heat conduction problem, but this solution technique does not evade the inherent ill-posedness of the problem. Some regularization, such as Tikhonov regularization, must be applied in the objective (fitness) function to combat the ill-posedness.

NEURAL NETWORKS

Neural networks have been used for perhaps 50 years, dating from the early works of Frank Rosenblatt (Rosenblatt, 1961). The main feature of neural metworks is in pattern recognition. The network "learns" the relationship between given input and output, and then generalizes this "knowledge". The result is that when the network is given inputs that are not exactly the same as those from the training data, the output from the network will be something consistent with the training data. In this way the neural network can be considered an interpolative algorithm.

An early application of Neural Networks was in simple pattern recognition. A classic example is a network designed to "recognize" letters based on a set of optically encoded inputs. A network might be designed and "trained" to identify the letters of the alphabet based on the sense of a *x6 grid of inputs. But if the network was well trained using, say, a Times Roman font, we might expect the netowrk to yield resonable results if it was shown letters from another font family, such as Ariel.

Application of Neural Networks, then, have two distinct phases: training and simulation. In the training phase, many pairs of inputs and outputs are shown to the network and the weights within the network are adjusted until the network (hopefully) produces the desired output. In the simulation phase, the training algorithm is deactivated, and the network merely computes the output based on the given inputs.

There are many classifications of Neural Networks according to the construction and of the network. Two broad classes are *concurrent* and *recurrent* or *dynamic* networks. Concurrent networks have all their input given at once and the output of the network depends only on these inputs. In contrast, recurrent or dynamic networks receive their inputs sequentially, and the output of some of the neurons in the network are fed back into the input for subsequent computations. In this paper I consider only concurrent Neural Networks.

Neural Network Topology

A schematic of a Neural Network is shown in Fig. 17. A typical network is composed of one or more *hidden layers* of *neurons* which are interconnected by *weighted* links. The output of each neuron is typically passed through some linear or non-linear *filter* or *activation function*.

A neuron is shown schematically in Fig. 18. The neuron receives inputs from perhaps n neurons in the previous layer. The output of the summation node is the dot product of the weights w_i and the value of the inputs:

$$SUM_{out} = \sum_{i}^{n} w_i p_i \qquad (4)$$

where p_i is the value of the input 'i'. If an output filter is used on the neuron, then the output of the neuron is the result of the filter on the SUM_{out}. Several output filter are possible, including

linear, tangent sigmoid, or hyperbolic sigmoid. A sigmoid function has a mathematical character similar to:

$$OUT_j = \frac{1}{1+\exp(-SUM_{out})} \qquad (5)$$

which asymptotically approaches constant values as SUM_{out} becomes very large or very small.

During the training phase of the network, the network processes given inputs in an attempt to produce given target outputs. The weights of the interconnections are adjusted by an appropriate algorithm to produce the desired outputs. This is

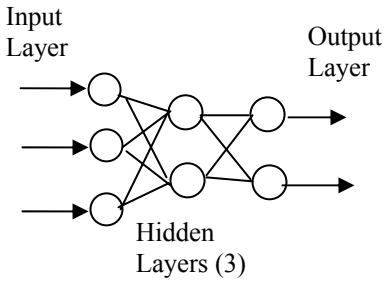


Figure 17. A Schematic of a Neural Network

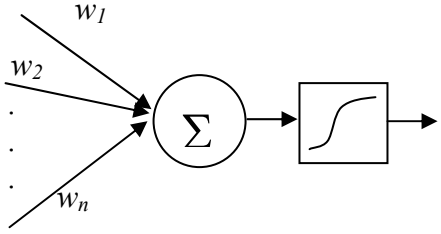


Figure 18. A Schematic illustrating the neuron.

an inherently iterative process, and the number of passes (called *epochs*) through the network during the training may be in the thousands. To train a Neural Network to solve an inverse problem, the mathematical model (forward solution) is used to generate training data.

MATLAB Toolbox

MATLAB has an excellent toolbox add-in for Neural Network analysis. This collection of programs and interfaces, written by Mark Demuth and Mark Beale, allow easy design and training of a wide range of networks: backpropagation networks, cascade feedforward networks, radial basis function networks, and many recurrent networks as well. The examples presented here make use of this toolbox add-in.

A Simple Example

As a simple example, consider the parameter identification problem considered earlier: the estimation of the slope and intercept of a line based on knowledge of several data points. We will design our network to estimate the slope m and intercept b of a line over $0 < x < 1$, and we restrict the range of b and m to the interval [0,1]. Furthermore, the values of y at specified x locations of 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0 will be given to the network to estimate b and m.

A backpropagation network with 6 inputs (for the 6 values of y) and two outputs (for the two values b and m) is created. One hidden layer with 12 neurons is employed, and a tangent sigmoid is chosen for the activation function on the hidden layer. The input layers have linear activation functions (filters).

A set of training data are generated which covers the solution space. I used the b and m pairs in Table 3 to generate y data on the specified intervals for x, resulting in 20 training vector pairs: input vectors of length 6 containing the values of y at the specified locations, and the corresponding output values of b and m in vectors of length 2.

Table 3. Intercept and slope used to generate data sets for training.

b	*m*	*b*	*m*
0.00	1.00	0.00	0.00
0.25	0.75	0.25	0.25
0.50	0.50	0.50	0.50
0.75	0.25	0.75	0.75
1.00	0.00	1.00	1.00
1.00	0.00	1.00	1.00
0.75	0.25	0.75	0.75
0.50	0.50	0.50	0.50
0.25	0.75	0.25	0.25
0.00	1.00	0.00	0.00

The network was set up and trained in the MATLAB toolbox. A scaled conjugate gradient training method (Demuth and Beale, 2001) was used to adjust the weights in the network. After 3000 epochs in the training, the sum squared error between the network output and the targets was 3.E-7.

After training, the network was tested with the eight vector inputs shown in Table 4. Note that

these vector inputs are not in the training set but do cover the range of inputs used in the training. The actual parameters corresponding to the rows of y values in Table 4 are shown in Table 5, along with the values computed using the trained Neural Network. As can be seen in Table 5, the values estimated from the Neural Network are reasonably good: the RMS errors are 0.0255 for b and 0.0069 for m.

Table 4. Test vectors (in rows) for the Line Identification Neural Network

y_1	y_2	y_3	y_4	y_5	y_6
0.90	1.04	1.18	1.32	1.46	1.60
0.10	0.28	0.46	0.64	0.82	1.00
0.90	0.92	0.94	0.96	0.98	1.00
0.30	0.40	0.50	0.60	0.70	0.80
0.50	0.56	0.62	0.68	0.74	0.80
0.30	0.44	0.58	0.72	0.86	1.00
0.70	0.76	0.82	0.88	0.94	1.00
0.70	0.88	1.06	1.24	1.42	1.60

Table 5. Actual and Computed Line Parameters for the rows in Table 4 using single hidden layer

ACTUAL		NN Output	
b	m	b	m
0.9	0.7	0.8533	0.6928
0.1	0.9	0.1001	0.8997
0.9	0.1	0.9003	0.0997
0.3	0.5	0.2766	0.4958
0.5	0.3	0.5140	0.3139
0.3	0.7	0.3001	0.6999
0.7	0.3	0.7000	0.2999
0.7	0.9	0.7476	0.9110

Another Neural Network was constructed and an additional hidden layer with 12 neurons was added. This network was trained using the same data (generated from Table 3) and after 3000 epochs the sum squared error was 2.3E-8. The test data from Table 4 was again shown to the network, and the results in Table 6 were obtained. The RMS errors for this case are 0.0008 for b and 0.0210 for m. Note the improvement in the estimates with an additional hidden layer in the network.

Table 6. Actual and Computed Line Parameters for the rows in Table 4 using two hidden layers

ACTUAL		NN Output	
b	m	b	m
0.9	0.7	0.9009	0.6601
0.1	0.9	0.0997	0.9003
0.9	0.1	0.9005	0.0997
0.3	0.5	0.3021	0.4871
0.5	0.3	0.4996	0.3170
0.3	0.7	0.3001	0.7000
0.7	0.3	0.6999	0.3001
0.7	0.9	0.6997	0.9384

Another approach to the line parameter identification problem is to try to train the network to learn the relationship between arbitrary groups of (x,y) data and the parameters b and m. We'll try this by adding an extra 6 inputs to the network, corresponding to the x locations. The network with two hidden layers was trained using the same data as before, but giving the x values as input also. The network learned this relationship very easily – in 6 epochs the SSE is less that 10^{-29}. Next, the network was tested by generating y data values for x values not in the training set: (0.1, 0.3, 0.45, 0.55, 0.7, 0.9). The six (x,y) data pairs for the eight test lines were input to the trained network, and the results are seen in Table 7. The results are not as good as those obtained previously with fixed x value inputs.

Table 7. Actual and Computed Line Parameters for the rows in Table 4 using one hidden layer and (x,y) inputs

ACTUAL		NN Output	
b	m	B	m
0.9	0.7	0.92865	0.68718
0.1	0.9	0.14523	0.87785
0.9	0.1	0.93050	0.092988
0.3	0.5	0.33520	0.48794
0.5	0.3	0.53171	0.29446
0.3	0.7	0.33841	0.68561
0.7	0.3	0.73234	0.29215
0.7	0.9	0.73396	0.88615

Application to Heat Flux Estimation

Time and space constraints do not allow the demonstration of the neural network to the solution of the inverse heat conduction problem.

However, this problem was considered by Krejsa, et. al (1999). In that work they considered two possible approaches: the whole domain estimation problem (as was considered in the genetic algorithm problem earlier, where all heat flux components are estimated simultaneously) and a sequential estimation scheme. Only concurrent neural networks were considered. Their conclusion was that the whole domain method offered the best possibility for solution of the inverse heat conduction problem using concurrent networks. However, I note here that recurrent networks may offer the possibility of sequential estimation.

The approach to solving the whole domain estimation of heat fluxes is similar to that taken in the line parameter estimation problem. Training data must be generated over the whole solution space of (t, q). This might be done by considering a range of different types of inputs: linear ramps of different slope, steps, parabolas, etc. The key is that the generated training data must cover the whole space of possible inputs for the neural network.

Neural Network Conclusions

Neural networks offer the possibility of solution of parameter estimation problems and also boundary inverse problems. Proper design of the network itself *and* the training data set is essential for successful application of this approach.

REFERENCES

1. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning,* Reading, Mass.: Addison Wesley (1989)

2. L. Davis (ed.), *Handbook of Genetic Algorithms*, Van Nostrand Reinhold (1991)

3. M. Raudensky, K. A. Woodbury, J. Kral, and T. Brezina,, "Genetic Algorithm in Solution of Inverse Heat Conduction Problems," Numerical Heat Transfer, Part B: Fundamentals, Vol 28, no 3, Oct.-Nov. 1995, pp. 293-306.

4. J. V. Beck, B. Blackwell, and C. St. Clair, *Inverse Heat Conduction: Ill-posed problems*, New York: John Wiley (1985).

5. F. Rosenblatt, *Principles of Neurodynamics*, Washington, DC: Spartan Books (1961).

6. H. Demuth and M Beale, *Neural Network Toolbox: For Use with MATLAB* (*User's Guide*), Natick, MA: The Mathworks (2001)

7. J. Krejsa, K. A. Woodbury, J. D. Ratliff., and M. Raudensky, "Assessment of Strategies and Potential for Neural Networks in the IHCP," Inverse Problems in Engineering, Vol 7, n 3, pp. 197-213. (1999)

Listing 1. Main function for the estimation of parameters of a straight line

```
% function simpleGA( xvals, ydata, Npop, low, high, Nbest, Ngen, mut_chance, creep_chance,
    creep_amount )
% function to perform simple GA search
%   find the slope and intercept of a line described the
%   data in 'xvals' and 'ydata'
%   1) randomly initialize 'Npop' candidate vectors in range
%      'low' to 'high'
%   2) evaluate the fitness using least squares criteria
%   3) sort to find best 'Nbest' candidates to use for reproducing
%   4) repopulate using random selection from the Nbest and random recombination
%      (use elitism - keep the top performer)
%   5) allow mutation of new generation at 'mut_chance' rate of magnitude
%       mut_amount
%   6) repeat for 'Ngen' new generations
function simpleGA( xvals, ydata, Npop, low, high, Nbest, Ngen, mut_chance, creep_chance,
    creep_amount )
Nunknown = 2;   % slope and intercept are the two unknowns
Nelite = 1;     % clone the super-individual
Ndata = length( ydata );

% generate the initial population
population = gen_rand_real( Npop, Nunknown, low, high );
% create matrix to keep best values and fitness for each generation
best = zeros( Ngen, Nunknown + 1 );
index = zeros( Ngen, 1);

% loop for the generations
for gen = 1 : Ngen
   % use model to compute values from population
   ytest = straight_line( population, xvals );
   % compute the fitness using least squares
   fitness = sum_square_fitness( ytest, ydata );
   % sort them
   sorted = sort_by_fitness( fitness, population );
   % copy generation champion into storage array
   best( gen, : ) = sorted( 1, : );
   index(gen)=gen;
   % copy best candidates into top of population
   population( 1:Nbest, : ) = sorted( 1:Nbest, 2:Nunknown+1 );
   % reproduce to fill the bottom of the population
   population = reproduce_by_weighted_avg( Nbest, population, Nelite );
   % crossover the children
   population = crossover( Nelite, population );
   % creep the children
   population = creep( Nelite, population, creep_chance, creep_amount );
   % mutate the children
   population = mutate( Nelite, population, mut_chance, low, high );
end
figure(1);
semilogy( index, best(:,1) );
b_best = best( Ngen, 2 );
```

```
m_best = best( Ngen, 3);
ybest = b_best * ones( 1, Ndata ) + m_best * xvals;
figure(2);
plot( xvals, ydata, '-' );
hold on;
plot( xvals, ybest, 'o' );
hold off;
best
sorted;
population;
```

Listing 2. Main Function for Estimating Heat Flux History

```
% function estimateQ_GA( xvals, ydata, dt, Nunknown, Npop, low, high, Nbest, Ngen, mut_chance,
    creep_chance, creep_amount, alpha )
% function to estimate a heat flux function of time using simple GA search
%   the location(s) of the sensors are contained in 'xvals', and the corresponding
%   measurements vectors are in 'ydata' (for more than one location xvals(1),
%   xvals(2), etc., the data are ydata = [ ydata(1), ydata(2), etc. ].
%   The time step in the data is 'dt'
%   The vector 'alpha' is length 2 and contains the tikhonov regularization
%   scalar weights.
%
%   1) randomly initialize 'Npop' candidate vectors in range
%      'low' to 'high'
%   2) evaluate the fitness using least squares criteria
%   3) sort to find best 'Nbest' candidates to use for reproducing
%   4) repopulate using random selection from the Nbest and random recombination
%      (use elitism - keep the top performer)
%   5) allow mutation of new generation at 'mut_chance' rate of magnitude
%       mut_amount
%   6) repeat for 'Ngen' new generations
function estimateQ_GA( xvals, ydata, dt, Nunknown, Npop, low, high, Nbest, Ngen, mut_chance,
    creep_chance, creep_amount, alpha )
nx = length(xvals);
Nloop = length( Ngen );
Ndata = round( length(ydata) / nx ); % number of values in the unknown vector
Nelite = 5;     % clone the super-individuals

% generate the initial population
% for random distribution
% population = gen_rand_real( Npop, Nunknown, low, high );
% for uniform distribution
values = gen_rand_real( Npop, 1, low, high );
population = ones( Npop, Nunknown);
for i = 1:Npop
   population( i, : ) = population( i, : )* values(i);
end

% create matrix to keep best values and fitness for each generation
best = zeros( Ngen(Nloop), Nunknown + 1 );
index = zeros( Ngen(Nloop), 1);

t_x = [ dt xvals ];  % special data vector for the evaluation function
```

```
% loop for the generations
gen = 1;
for loop = 1 : Nloop
   for gen = gen : Ngen(loop)
      % use model to compute values from population
      ytest = eval_qvec( population, t_x, Ndata );
      % compute the fitness using least squares
      fitness = sum_square_fitness( ytest, ydata );
      fitness = fitness + tikhonov_term( population , alpha );
      % sort them
      sorted = sort_by_fitness( fitness, population );
      % copy generation champion into storage array
      best( gen, : ) = sorted( 1, : );
      index(gen)=gen;
      % copy best candidates into top of population
      population( 1:Nbest, : ) = sorted( 1:Nbest, 2:Nunknown+1 );
      % reproduce to fill the bottom of the population
      population = reproduce_by_weighted_avg( Nbest, population, Nelite );
      % crossover the children
      population = crossover( Nelite, population );
      % creep the children - modify slightly each value (by chance)
      population = creep( Nelite, population, creep_chance(loop), creep_amount(loop) );
      % mutate the children - random replace of chromosome
      population = mutate( Nelite, population, mut_chance(loop), low, high );
   end
   range = minmax( best( gen, 2:Nunknown+1 ) )
   low = range(1);
   high = range(2);
end
figure(1);
semilogy( index, best(:,1) );
time_max = Ndata * dt;
dt_unk = time_max / Nunknown;
time = zeros( Ndata, 1);
time_half = zeros( Nunknown, 1);
time(1) = dt;
time_half(1) = dt_unk/2;
for i = 2:Nunknown
   time_half(i) = time_half(i-1) + dt_unk;
end
for i = 2:Ndata
   time(i) = time(i-1) + dt;
end
last = best( Ngen(Nloop), 2:Nunknown+1);
figure(2);
plot( time_half, last, 'o' );
t_exact = [ 0 .24 .84 1.44 1.8 ];
q_exact = [ 0   0  0.6  0   0  ];
hold on;
plot( t_exact, q_exact, '-' );
hold off;
qbest = best(Ngen(Nloop),2:Nunknown+1)
% use model to compute values from population
```

```
ybest = eval_qvec( qbest, t_x, Ndata );
figure(3);
plot( time, ybest, 'o' );
hold on;
plot( time, ydata, '-' );
hold off;
sum_sq_err = sum_square_fitness( ybest, ydata );
rms_error = sqrt(sum_sq_err/Nunknown)
sorted;
population;
```

ALGORITHMS,THEORETICAL AND MATHEMATICAL ASPECTS

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

A FEYNMAN-KAC METHOD FOR THE DETERMINATION OF THE STEFAN'S FREE BOUNDARY

Eduardo Souza de Cursi
Laboratoire de Mécanique de Rouen, UMR 6138 CNRS
Institut National des Sciences Appliquées de Rouen, INSA
Avenue de l'Université, BP 08
76801 Saint-Etienne du Rouvray CEDEX, France
souza@insa-rouen.fr

ABSTRACT

This work presents numerical methods for the determination of the free boundary in the two-phase Stefan problem. This paper does not directly concern the idenfication from measured data, which will be a futur development. The method is based on a Feyman-Kac representation of the solution: the position of the free boundary is the solution of an *algebraical equation* involving the means of random variables. This equation can be numerically solved by iterative methods and the free boundary can be determined by *algebraical calculations*.

The approach introduced is based on a Feynman-Kac representation involving the mean of a convenient random variable. The numerical methods have some interesting properties: they are meshless (i.e., they may be implemented without introducing a spatial discretization). Other features are the independence of the dimension and a natural parallelism. In addition, the methods are adapted to both the tracking of the front and the direct evaluation of the position at a few given moments without discretization in time.

The method has been tested in two and three spatial dimensions. The result of a numerical experiment is presented.

NOMENCLATURE

a_i ,b_i	lower and upper bounds for x_i
c	diffusivity; c_s on $\Omega_S(t)$, c_L on $\Omega_L(t)$
c_L	diffusivity in the liquid region
c_s	diffusivity in the solid region
c_ε	regularized function c
E(Y | X)	Mean of Y conditional to X
div(**u**)	Divergence of **u** = (u_1, u_2, u_3), div(**u**) = $\partial u_1/\partial x_1 + \partial u_2/\partial x_2 + \partial u_3/\partial x_3$
Id	Identity Matrix
n	$\nabla\phi / \lvert\nabla\phi\rvert$, unitary normal to $\Sigma(t)$
$N(0,\sigma)$	Normal distribution having mean zero, standard deviation σ
$N(0,\sigma\mathbf{Id})$	Multidimensional normal distribution having mean zero, Covariance matrix σ **Id**
Q	$\Omega \times T$
Q_L	$\{(x,t) \in Q \mid \theta(x,t) > \theta_c \}$
Q_S	$\{(x,t) \in Q \mid \theta(x,t) < \theta_c \}$
R	The set of real numbers
S	$\{(x,t) \in Q \mid \theta(x,t) = \theta_c \}$
t, T	time variable and maximum time
v	velocity of the free boundary
x, x_i	spatial variable, a component of **x**
W_t	Wiener process
$\partial\Omega$	boundary of Ω
ε	regularization parameter
χ	indicator of the solid region
χ_ε	regularized function χ
ϕ	equation of the free boundary
φ	numerical approximation of ϕ
λ	latent heat of the material
Σ (t)	free boundary: $\theta = \theta_c$, $\phi = 0$
θ	field of temperatures
θ_c	temperature of solidification
θ_S, θ_L, θ_s	restriction of θ to Q_S,Q_L or S
θ_0	initial field of temperatures
$\theta_{\partial\Omega}$	field of temperatures on $\partial\Omega$
∇u	Gradient of u, $\nabla u = (\partial u/\partial x_1, \partial u/\partial x_2, \partial u/\partial x_3)$
Ω	spatial domain
$\Omega_L(t)$	liquid region: $\theta > \theta_c$, $\phi > 0$
$\Omega_S(t)$	solid region: $\theta < \theta_c$, $\phi < 0$

INTRODUCTION

Multiphase and multiregion problems arise in several significant situations in Engineering. In those problems, the determination of the interfaces, i. e., of the surfaces or regions separating the different phases, is a crucial point, which leads to mathematical and numerical difficulties. In this framework, many works have considered a heat transfer problem frequently introduced as a simple model for melting or solidification phenomena: the two-phase Stefan problem.

The usual formulation of a Stefan problem leads to evolution equations describing the temperature of the material and the moving boundary. The major difficulty in a direct problem lies in the fact that the unknown boundary intervenes explicitly in the equations giving the thermal state of the system. The problem is frequently rewritten in order to eliminate the unknown boundary (see, for instance [1], [2]).

A different standpoint considers the temperature as an auxiliary variable (instead of the position of the moving boundary). In this case, the main variable is the position of the free boundary : the field of temperatures is determined by solving two linear heat equations on each region (liquid and solid), once the moving boundary has been found. Level set methods may be considered as included in this approach (see, for instance, [3],[4])

The numerical resolution of two-phase Stefan problem has been extensively treated in the literature since 30 years (see, for instance, [5], [6], [7]). In this paper, we introduce an original numerical approach based, on the one hand, on a formulation of the Stefan problem leading to a non-linear evolution equation verified on the whole domain (see, for instance, [8]) and, on the other hand, on Feynman-Kac representations of the solutions of linear parabolic equations (see, for instance, [9]).

An interesting feature of the numerical methods associated to Feynman-Kac representations is the *absence of discretization in space* (i. e., they are *meshless*). Usually, when we are looking for the position of the free boundary on the interval (0,T), we must introduce N subintervals of length Δt involving N+1 values t_i such that $0 = t_0 < t_1 < t_2 < \ldots < t_{N+1} = T$ and the solution is constructed at the N discrete times t_1 , …, t_N. The construction of the solution at each one of the considered times t_i involves a Finite Difference or Finite Element Method. *The proposed approach does not involve such a discretization in space*.

Other interesting properties of the Feynman-Kac approach are the independence of the number of dimensions and its natural parallelism (see, for instance [10], [11]).

In the next section, we recall the formulation of the Stefan problem which will be used. Then we recall some elements concerning the Feynman-Kac representation and, at last, we shall present the results of some numerical experiments show that the resulting numerical method is effective to calculate.

A MODEL FOR THE PROCESS

In this section, we shall present the notations and a model for the situation previously described. Main results and formulations mentioned are stated in [8] and will be not detailed here.

Description

The solidifying material is inside a rectangular cavity. In a first approximation, we consider that the temperature is known on the boundary: the method extends to given fluxes or Newton's conditions. We assume that the system is well described by a two dimensional problem. These simplifications allow us to point the essential difficulties and are not essential: on the one hand, as previously observed, the results and the method can be extended to upper dimension, and, on the other hand, cavities of arbitrary but regular enough shape can be considered. Thus, we consider the open bounded domain $\Omega \subset R^3$ defined by

$$\Omega = \{ \mathbf{x} = (x_1, x_2, x_3) \mid a_i < x_i < b_i \} \qquad (1)$$

The boundary of Ω is denoted by $\partial\Omega$. The time is denoted by $t \in T = (0, T)$, $T > 0$ and we set $Q = \Omega \times T$. The field of temperatures is a function $\theta : Q = \Omega \times T \rightarrow R$. At each time t, Ω is partitioned as follows:

$$\Omega = \Omega_S(t) \cup \Omega_L(t) \cup \Sigma(t) \qquad (2)$$

where

$$\Omega_S(t) = \{ x \in \Omega \mid \theta(\mathbf{x},t) < \theta_c \}, \qquad (3)$$

$$\Omega_L(t) = \{ x \in \Omega \mid \theta(\mathbf{x},t) > \theta_c \}, \qquad (4)$$

$$\Sigma(t) = \{ x \in \Omega \mid \theta(\mathbf{x},t) = \theta_c \}. \quad (5)$$

We assume that the solid/liquid interface $\Sigma(t)$ can be represented at each time t by a curve described by an equation $\phi(\mathbf{x},t) = 0$ and the sets $\Omega_S(t)$, $\Omega_L(t)$, $\Sigma(t)$ are characterized respectively by $\phi(\mathbf{x},t) < 0$, $\phi(\mathbf{x},t) > 0$, $\phi(\mathbf{x},t) = 0$. Thus, the different regions can be also characterized by

$$\Omega_S(t) = \{ \mathbf{x} \in \Omega \mid \phi(\mathbf{x},t) < 0 \}, \quad (6)$$

$$\Omega_L(t) = \{ \mathbf{x} \in \Omega \mid \phi(\mathbf{x},t) > 0 \}, \quad (7)$$

$$\Sigma(t) = \{ \mathbf{x} \in \Omega \mid \phi(\mathbf{x},t) = 0 \}. \quad (8)$$

We assume also that ϕ is regular enough in order that the unitary normal $\mathbf{n}$ to $\Sigma(t)$ is defined for any $\mathbf{x} \in \Sigma(t)$ and $t \in T$:

$$\mathbf{n} = \nabla\phi / |\nabla\phi| . \quad (9)$$

We notice that n points inwards $\Omega_L(t)$ (and outwards $\Omega_S(t)$).

The natural choice for ϕ is $\phi = \theta - \theta_c$, but ϕ can take infinitely many values: the previous equations hold for an arbitrary function such that $\text{sign}(\phi) = \text{sign}(\theta - \theta_c)$, where

$$\text{sign}(\alpha) = \begin{cases} 1, & \alpha > 0; \\ 0, & \alpha = 0; \\ -1, & \alpha < 0. \end{cases} \quad (10)$$

The Stefan condition

Let us introduce λ, the latent heat of the material; c_S and c_L, the diffusivity in the solid and liquid parts, respectively; θ_S, θ_L the temperatures in the solid and liquid parts, respectively:

$$\theta_S = \theta|_{Q_S} \; ; \; \theta_L = \theta|_{Q_L} \quad (11)$$

$$Q_S = \{(x,t) \in Q \mid x \in \Omega_S(t) \} \; ; \quad (12)$$

$$Q_L = \{(x,t) \in Q \mid x \in \Omega_L(t) \} \; . \quad (13)$$

As previously observed, two heat equations are verified on each region (solid and liquid):

$$\frac{\partial \theta_S}{\partial t} - \text{div}(c_S \nabla \theta_S) = 0 \quad \text{on} \quad Q_S \quad (14)$$

$$\frac{\partial \theta_L}{\partial t} - \text{div}(c_L \nabla \theta_L) = 0 \quad \text{on} \quad Q_L \quad (15)$$

On the free boundary, the Stefan condition is satisfied:

$$c_L \nabla\theta_L.\mathbf{n} - c_S \nabla\theta_s.\mathbf{n} = \lambda \mathbf{v}.\mathbf{n} \quad \text{on} \quad S \, , \quad (16)$$

where $\mathbf{v}$ is the velocity of the free boundary.

We denote by $c(\bullet)$ the function

$$c(\alpha) = \begin{cases} c_S, & \alpha < 0; \\ c_L, & \alpha > 0. \end{cases} \quad (18)$$

Let us introduce a function $\chi(\bullet)$ such that:

$$\chi(\alpha) = \begin{cases} 1, & \alpha < 0; \\ 0, & \alpha \geq 0. \end{cases} \quad (19)$$

With these notations, (11)-(16) is equivalent to (Cf. [8]):

$$\frac{\partial \theta}{\partial t} - \text{div}(c(\phi)\nabla\theta) = -\lambda \frac{\partial}{\partial t}(\chi(\phi)) \quad \text{on} \quad \Omega \quad (20)$$

$$\text{sign}(\phi) = \text{sign}(\theta - \theta_c) \qquad \text{on} \quad \Omega \quad (21)$$

We point that the equations (20)-(21) are verified on the whole Ω and, as above mentioned, infinitely many choices of ϕ are possible.

The evolution problem

We denote by θ_0, the initial field of temperatures:

$$\theta(x,0) = \theta_0(x) \quad \text{on} \quad \Omega \quad (22)$$

We assume that θ_0 has square summable partial derivatives: $\theta_0 \in H^1(\Omega)$. We assume also that the temperature on the boundary is known:

$$\theta(x,t) = \theta_{\partial\Omega}(x,t) \quad \text{on} \quad \partial\Omega \quad (23)$$

We assume also that $\theta_{\partial\Omega}$ is square summable: $\theta_{\partial\Omega} \in L^2(\partial\Omega)$. Moreover, the following compatibility condition is satisfied:

$$\theta_{\partial\Omega}(x,0) = \theta_0(x). \qquad (24)$$

The unknowns (θ, ϕ) satisfy the following boundary value problem:

Problem 1: Find (θ, ϕ) satisfying (20)-(24).

From [8], we have the following result:

Theorem 1: The field of temperatures θ and the regions $\Omega_S(t)$, $\Omega_L(t)$, $\Sigma(t)$ are uniquely determined.

As previously observed, ϕ is not uniquely determined.

Regularization

In order to obtain a more regular problem, we shall introduce continuously differentiable and lipschitzian approximations of c and χ, denoted c_ε and χ_ε, respectively. In order to construct such an approximation, let us consider

$$\eta(\alpha) = \begin{cases} 0, & \alpha \le 0; \\ 3\alpha^2 - 2\alpha^3, & 0 \le \alpha \le 1; \\ 1, & \alpha \ge 1. \end{cases} \qquad (25)$$

We observe that η is continuously differentiable on R. η is not uniquely determined. More regular approximations involving differentiability at an arbitrary order n can be introduced by convenient choices.

Let $\varepsilon > 0$ be a given parameter. We shall approximate the discontinuous functions $c(\bullet)$ and $\chi(\bullet)$ by

$$c_\varepsilon(\alpha) = c_S\eta\left(\frac{\varepsilon-\alpha}{2\varepsilon}\right) + c_L\eta\left(\frac{\alpha+\varepsilon}{2\varepsilon}\right) \qquad (26)$$

$$\chi_\varepsilon(\alpha) = \eta\left(\frac{-\alpha}{\varepsilon}\right) \quad . \qquad (27)$$

We have

$$c_\varepsilon(\phi) = c(\phi) \ , \ \text{if} \ |\phi| \ge \varepsilon \ ; \qquad (28)$$

$$\chi_\varepsilon(\phi) = \chi(\phi) \ , \ \text{if} \ \phi \le -\varepsilon \ \text{or} \ \phi \ge 0 \, . \qquad (29)$$

c_ε and χ_ε are used in order to obtain a regularized boundary value problem. The equations of Problem 1 are approximated as follows:

$$\frac{\partial\theta_\varepsilon}{\partial t} - \mathrm{div}(c_\varepsilon(\phi_\varepsilon)\nabla\theta_\varepsilon) = -\lambda\frac{\partial}{\partial t}(\chi_\varepsilon(\phi_\varepsilon)) \ \text{on} \ \Omega \qquad (30)$$

$$\mathrm{sign}(\phi_\varepsilon) = \mathrm{sign}(\theta_\varepsilon - \theta_c) \quad \text{on} \ \Omega \qquad (31)$$

$$\theta_\varepsilon(x,0) = \theta_0(x) \quad \text{on} \ \Omega \qquad (32)$$

$$\theta_\varepsilon(x,t) = \theta_{\partial\Omega}(x,t) \quad \text{on} \ \partial\Omega \qquad (33)$$

and we denote by $(\theta_\varepsilon, \phi_\varepsilon)$ the solution of the associated boundary value problem:

Problem 2: Find $(\theta_\varepsilon, \phi_\varepsilon)$ verifying (30) – (33)

From [8], we have the following result:

Theorem 2: The field of temperatures θ_ε and the associated approximated regions $\Omega_{S\varepsilon}(t)$, $\Omega_{L\varepsilon}(t)$, $\Sigma_\varepsilon(t)$ are uniquely determined. Moreover,

$$\theta_\varepsilon \xrightarrow[\varepsilon\to 0+]{} \theta \quad in \quad L^2(0,T;H^1(\Omega)) \qquad (34)$$

DISCRETIZATION OR ITERATION IN TIME FOR THE REGULARIZED PROBLEM

As previously observed, the numerical resolution of two phase Stefan problem has been extensively treated in the literature and many methods of discretization in time have been proposed. We present here only the particular methods used for the Feynman-Kac approximation.

In order to alleviate the notations, we shall drop the index ε.

Iterative solution of the regularized problem

The problem 2 may be solved by an iterative procedure. Let us set

$$\phi = \theta - \theta_c \quad . \qquad (35)$$

We have

$$\frac{\partial}{\partial t}(\chi(\phi)) = \chi'(\phi)\frac{\partial\theta}{\partial t} \qquad (36)$$

Thus, Equation (30) is equivalent to

$$(1+\lambda\chi'(\phi))\frac{\partial\theta}{\partial t}-\mathrm{div}(c(\phi)\nabla\theta)=0 \qquad (37)$$

Let be given an initial guess (θ^0, ϕ^0) (such as, for instance, $\theta^0 = \theta_0$, $\phi^0 = \theta_0 - \theta_c$). We define a sequence $\{(\theta^k , \phi^k)\}_{k\geq 0}$ by

$$(1+\lambda\chi'(\phi^k))\frac{\partial\theta^{k+1}}{\partial t}-\mathrm{div}(c(\phi^k)\nabla\theta^{k+1})=0 \text{ on } \Omega \qquad (38)$$

$$\phi^{k+1} = \theta^{k+1} - \theta_c \qquad (39)$$

$$\theta^{k+1}(x,0) = \theta_0(x) \quad \text{on } \Omega \qquad (40)$$

$$\theta^{k+1}(x,t) = \theta_{\partial\Omega}(x,t) \quad \text{on } \partial\Omega \qquad (41)$$

We have the following result:

Theorem 3: For each $k \geq 0$, the field of temperatures θ^{k+1} and ϕ^{k+1} are uniquely determined. Moreover,

$$\theta^k \xrightarrow[k\to+\infty]{} \theta_\varepsilon \quad in \quad L^2(0,T;H^1(\Omega)) \qquad (42)$$

Approximation of the free boundary by interpolation

In order to alleviate the notations, we shall drop the indexes k and k+1.

Let us introduce an integer $N > 0$ and, for $0 \leq i \leq N$,

$$\Delta t = T/N \; ; \; t_i = i\,\Delta t \qquad (43)$$

$$\theta_i(x) = \theta(x, t_i) \; ; \quad \phi_i(x) = \phi(x, t_i) \qquad (44)$$

The function ϕ is approximated by a function φ: $\phi \cong \varphi$. The approximation uses an interpolation procedure involving the the values $\phi_0, \ldots, \phi_N$. Many interpolation procedures can be considered. Here, we shall consider two kinds of approximation: a polynomial approximation of degree N and piecewise constant approximations.

Polynomial interpolation. In this approximation, we set

$$\varphi(x,t) = k_0(x) + k_1(x)\,t + \ldots + k_N(x)\,t^N \; , \qquad (45)$$

where

$$k_i = k_i(\phi_0, \phi_1, \ldots, \phi_N) \qquad (46)$$

is such that

$$\varphi(x, t_i) = \phi_i(x) \qquad (47)$$

For instance, for N=1, we have

$$k_0 = \phi_0 \; ; \quad k_1 = (\phi_1 - \phi_0)/T \qquad (48)$$

while for N = 2 we have

$$k_0=\phi_0;\; k_1=(4\phi_1-3\phi_0-\phi_2)/T;\; k_2=2(\phi_2-2\phi_1+\phi_0)/T^2 \qquad (49)$$

We observe that, when (48) is used, $\phi(x,T)$ is calculated *without discretization in time*. Into an analogous way, the use of (49) furnishes $\phi(x,T/2)$ and $\phi(x,T)$ *without the evaluation of the free boundary for other values of t.*

Piecewise constant approximation. In this case, we set

$$\varphi(x,t) = \phi_i(x) \quad \text{for} \quad t_i \leq t \leq t_{i+1} \qquad (50)$$

This approximation leads to an explicit method: on each each subinterval (t_i, t_{i+1}), ϕ_i is known and ϕ_{i+1} can be determined by solving

$$(1+\lambda\chi'(\phi_i))\frac{\partial\theta}{\partial t}-\mathrm{div}(c(\phi_i)\nabla\theta)=0 \text{ on } \Omega\times(t_i,t_{i+1}) \qquad (51)$$

$$\theta(x,t_i) = \theta_i(x) \quad \text{on} \quad \Omega \qquad (52)$$

$$\theta(x,t) = \theta_{\partial\Omega}(x,t) \quad \text{on } \partial\Omega \qquad (53)$$

$$\phi_{i+1} = \theta_{i+1} - \theta_c \qquad (54)$$

This approach leads to the tracking of the free boundary by its calculation at the times $t_1, \ldots t_N$, but can also be used analogously to (48) by performing a few steps in order to get $\phi(x,T)$.

A FEYNMAN-KAC METHOD FOR THE DETERMINATION OF THE FREE BOUNDARY AT A GIVEN TIME

As mentioned in the introduction, we shall introduce a method based on Feynman-Kac representations of the solutions. By reasons of limitation of the room, we give only some elements concerning the construction of representations. The reader interested in more

complete developments is invited to refer to [9], [10], [11].

Ito's Formula

The main tool for the construction of Feynman-Kac representations is the Ito's formula.

Wiener processes. Let $\{W_t\}_{t\geq 0}$ be a family of real random variables. We say that $\{W_t\}_{t\geq}0$ is a standard Wiener process if and only if

$$W_0 = 0 \; ; \tag{55}$$

$$W_t - W_s \text{ is } N(0, (t-s)^{1/2}) \quad (t \geq s) \tag{56}$$

$$W_t - W_s \text{ is independent of } W_z \; (t \geq s \geq z) \tag{57}$$

Multidimensional Wiener processes are defined into an analogous way, by considering $W_t - W_s$ as Gaussian vectors having the distribution $N(0, (t-s)^{1/2}\,\mathbf{Id})$.

Ito's stochastic Integrals. Let γ: $R \rightarrow R$ be a function and $\{Z_t\}_{t\geq 0}$ be a stochastic process. We define

$$I_1(\gamma, Z_t) = \int_0^\tau \gamma(Z_t)dt \tag{58}$$

$$I_{2(}\gamma, Z_t) = \int_0^\tau \gamma(Z_t)dW_t \tag{59}$$

as follows: let $p > 0$ be an integer and $h = \tau/p$. We note

$$Z_i = Z_{t_i} \; ; \; W_i = W_{t_i} \; ; \; t_i = i\,h \tag{60}$$

And we consider the finite sums:

$$I_1^p(\gamma, Z_t) = h\sum_{i=1}^{p} \gamma(Z_{i-1}) \tag{61}$$

$$I_2^p(\gamma, Z_t) = \sum_{i=1}^{p} \gamma(Z_{i-1})(W_i - W_{i-1}) \tag{62}$$

We define

$$I_i(\gamma, Z_t) = \lim_{p\to+\infty} I_i^p(\gamma, Z_t) \quad , \quad i = 1,2 \tag{63}$$

when such a limit exists.

Ito's formula. Let us denote by $\{W_t\}_{t\geq 0}$ a standard Wiener process and X_t the following Ito's diffusion

$$d\mathbf{X}_t = \alpha\, d\mathbf{W}_t \quad ; dS_t = -\beta(\mathbf{X}_t, t)\, dt \tag{64}$$

$$\mathbf{X}_0 = \mathbf{x} \quad ; \; S_0 = s \tag{65}$$

Then, the diffusion

$$Y_t = u(\mathbf{X}_t, S_t) \tag{66}$$

verifies the following Ito's formula:

$$\begin{aligned} dY_t = &\left(\operatorname{div}\left(\frac{\alpha^2}{2}\nabla u(\mathbf{X}_t, S_t)\right) - \beta(\mathbf{X}_t, S_t)\frac{\partial u}{\partial t}(\mathbf{X}_t, S_t)\right)dt \\ &+ \alpha(\mathbf{X}_t, S_t)\nabla u(\mathbf{X}_t, S_t)dW_t \end{aligned} \tag{67}$$

Numerical determination of the free boundary

Representation of the solutions of linear parabolic problems. Let u satisfy the equation

$$\beta(\mathbf{x}, t)\frac{\partial u}{\partial t} - \operatorname{div}\left(\frac{\alpha^2}{2}\nabla u\right) = f(\mathbf{x}, t) \text{ on } \Omega \times (0, T) \tag{68}$$

Then equation (67) becomes

$$dY_t = -f(\mathbf{X}_t, S_t)dt + \alpha(\mathbf{X}_t, S_t)\nabla u(\mathbf{X}_t, S_t)dW_t \tag{69}$$

So, we have, for any $\tau > 0$:

$$Y_\tau - Y_0 = -\int_0^\tau f(\mathbf{X}_t, S_t)dt + \int_0^\tau \alpha(\mathbf{X}_t, S_t)\nabla\theta(\mathbf{X}_t, S_t)dW_t$$

and, taking (57) into account :

$$E(Y_\tau - Y_0) = E\left(-\int_0^\tau f(\mathbf{X}_t, S_t)dt\right)$$

Thus,

$$u(\mathbf{x}, s) = Y_0 = E\left(Y_\tau + \int_0^\tau f(\mathbf{X}_t, S_t)dt\right) \tag{70}$$

Numerical solution of a parabolic equation using the representation. Equation (70)

furnishes a simple way for the numerical evaluation of u(x,s): let us assume that the following data is given:

$$u(\bullet,0) = u_0(\bullet) \text{ on } \Omega \quad (71)$$

$$u(\bullet,t) = u_{\partial\Omega}(\bullet,t) \text{ on } \partial\Omega \quad (72)$$

$$u(\bullet,t) = u_c(\bullet,t) \text{ if } \varphi(\bullet,t) = 0 \quad (73)$$

Then, we consider τ as the first time where u takes a known value, i. e. :

$$\tau = \inf \{ t \mid (\mathbf{X}_t, S_t) \notin Q \} ; \quad (74)$$

So,

$$Y = u(\mathbf{X}_\tau, S_\tau) + \int_0^\tau f(\mathbf{X}_t, S_t)dt \quad (75)$$

corresponds to the given data (71)-(73) and we have:

$$u(\mathbf{x},s) = E(Y) \quad (76)$$

Thus the value of u(x,s) may be approximated by using an empirical mean: we may generate NS values of Y, denoted by $Y_1, \ldots, Y_{NS}$ and we have

$$u(\mathbf{x},s) \cong (Y_1 + \ldots + Y_{NS}) / NS \quad (77)$$

The values of $Y_1, \ldots, Y_{NS}$ can be generated by simulation of (64)-(65). Methods of simulation can be found, for instance, in [10], [11]. A simple method of generation is the Euler's discretization involving a step $h > 0$ and a gaussian vector $\mathbf{Z} \sim N(0, \mathbf{Id})$ ($\mathbf{Z}_i$ denotes a value from $\mathbf{Z}$)

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \alpha h \mathbf{Z}_i \quad ; S_{i+1} = S_i - h\beta(\mathbf{X}_i,t) \quad (78)$$

Determination of the interface. As previously observed, the free boundary satisfies an algebraical equation such as, for instance,

$$\phi = \theta - \theta_c$$

This equation may be considered as being of algebraical type and solved by iterative methods such as fixed point or Newton's iterations. For instance, we can consider the iterations

$$\phi^{k+1} = \phi^k + \mu (\theta^{k+1} - \theta_c - \phi^k)$$

where $\mu \in R$ is a parameter to be conveniently chosen. These iterations imply the evaluation of ϕ on the whole Ω. In order to limit the evaluation to the single interface, we may introduce θ_s, the restriction of θ to S and observe that

$$\theta_s = \theta_c$$

Thus, we can also consider the iterations

$$\phi^{k+1} = \phi^k + \mu (\theta^{k+1,s^k} - \theta_c) \quad (79)$$

where θ^{k+1,s^k} is the restriction of θ^{k+1} to S^k. Equations (38) and (51) can be reduced to the form (68). Thus, θ^{k+1} or its restriction $\theta^{k+1}{}_{s^k}$ can be determined by using (76)-(78) and the iterations can be performed without use of a Finite Element of Finite Difference method for the evaluation of the temperatures.

If the free boundary verifies:

$$\phi(\mathbf{x},t) = x_3 - \rho(x_1, x_2, t)$$

the iterations reads as follows:

$$\rho^{k+1} = \rho^k - \mu (\theta^{k+1,s^k} - \theta_c) \quad (80)$$

A NUMERICAL EXAMPLE

By reasons of limitation of the room, we present here the results of a single experiment concerning the determination of the free boundary at T = 1.

Let us consider the situation where $\Omega = (-1,1)^3$, $c_S = 10$; $c_L = 11$; $\lambda = 4$; $\theta_c = 0$;

$$\theta(\mathbf{x},t) = (x_1+1)^2+(x_2+1)^2+(x_3+1)^2 - \exp(-t) \quad (81)$$

In order to numerically perform the iterations (81), we introduce a discrete set of np ***calculation*** (and not measurement) points distributed on the horizontal plane (Ox_1, Ox_2):

$$\mathbf{p}(n) = (x_1(i), x_2(j)) \ , \ 0 \le i \le n_1 \ , \ 0 \le j \le n_2 \quad (82)$$

$$n = i + (j-1)n_1; \quad np = (n_1+1)(n_2+1) \quad (83)$$

$$x_1(i) = -1+2i/n1; \ x_2(j) = -1+2j/n2 \quad (84)$$

The results presented concern $n_1 = n_2 = 40$. The error in the evaluation of the free boundary is measured by the mean quadratic deviation

$$e_k = \sqrt{\frac{1}{np}\sum_{n=1}^{np}\left(\rho^k(n)-\rho(n)\right)^2} \qquad (85)$$

$$\rho(n) = \rho(\mathbf{p}(n)) \; ; \; \rho^k(n) = \rho^k(\mathbf{p}(n)) \qquad (86)$$

For a generic point **x**, the value of ρ^k is obtained by linear interpolation using the values (86). The initial guess is the initial position of the free boundary: $\rho^0 = \rho(x_1,x_2,0)$. The results furnished by (80) with ns = 10^4; $\mu = 0.5$ and the interpolation (48) are shown in Figures 1 and 2. Results furnisheed by the Robbins-Monro procedure with $\mu_k = 0.5/(1+k/4)$; ns = 10^2 and the interpolation (48) are shown in Figure 3.

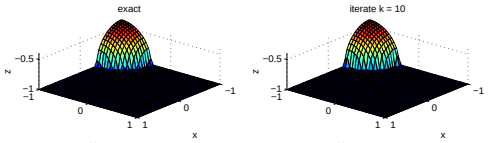


Figure 1– Free Boundary at t = 1.

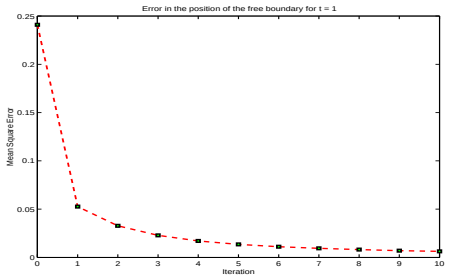


Figure 2– Evolution of the error (e_{10} = 6E-3)

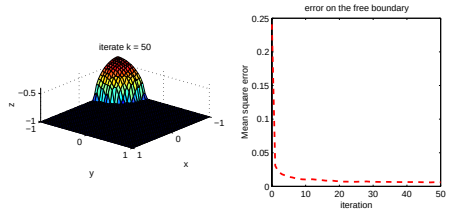


Figure 3–Results obtained by Robbins-Monro

CONCLUDING REMARKS

We have presented a numerical method for the determination of the Stefan free boundary, based on a formulation of the Stefan problem as a non-linear evolution equation verified on the whole domain and a Feynman-Kac representation of the solution of a linear parabolic equation. The method does not involve spatial discretization: the values of the temperatures are evaluated by simulating a stochastic diffusion. The approach is naturally adapted to parallel computation. It has been tested in two or three dimensional situations and has shown to be effective to calculate. Improvements may be obtained by using more sophisticated spatial and temporal interpolation of the free boundary ore other methods of simulation of the stochastic diffusion.

REFERENCES

1. E. Magenes, C. Verdi, A. Visintin, Theoretical and numerical results on the two-phase Stefan problem, SIAM *Journal of Numerical Analysis*, Vol. 26, 1425 - 1438(1989).

2. J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Ed. Dunod, Paris, France (1969).

3. S.Chen, B. Merriman, S. Osher, P. Smereka, *A simple level set method for solving Stefan problems*, Technical Report 96-21, http://www.math.ucla.edu/applied/cam/index.html (1996)

4. S. Osher, P.Fedkiw, *Level Set Methods: an overview and some recent results*, TR 00-08, http://www.math.ucla.edu/applied/cam/index.html (2000)

5. J. F. Ciavaldini, *Résolution Numérique d'un problème de Stefan à deux phases*, Ph. D. Thesis, Rennes, France (1972).

6. J. W. Jerome, M. Rose, Error estimates for the multidimensional two-phase Stefan problem in two space dimension, *Mathematics of Computation*, Vol. 39, 377-414 (1982).

7. R. H. Nochetto, M. Paolin, C. Verdi, An Adaptive Finite Element Method for Stefan problem in two space dimension part II, *SIAM J Sci Statist Comput*, Vol 12, 1207- 1244 (1991).

8. J. P. Humeau, J. E. Souza de Cursi, Regularization and numerical resolution of a bidimensional Stefan problem, *Journal of Mathematical Systems, Estimation and Control*, Vol. 3 no. 4, 473 - 497(1993).

9. R. Dautray, P. L. Lions, E. Pardoux et al., *Méthodes Probabilistes pour les équations de la physique*, Ed. Eyrolles, Paris, France (1989).

10. J. E. Souza de Cursi, Numerical methods for linear boundary value problems based on Feyman-Kac representations, *Mathematics and Computers in Simulation*, 36, 1-16 (1994).

11. J. P. Morillon, Numerical Solution of Linear Mixed Boundary Value Problems using stochastic representations, *IJNME*, 40, 387- 405 (1997).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

SOURCE TERMS IDENTIFICATION FOR THE DIFFUSION EQUATION

Zhuobiao Yi
Department of Mathematical Sciences
University of Cincinnati
Cincinnati, OH 45221-0025, USA
yizo@email.uc.edu

Diego A. Murio
Department of Mathematical Sciences
University of Cincinnati
Cincinnati, OH 45221-0025, USA
diego@dmurio.csm.uc.edu

ABSTRACT

We present an automated technique for the approximate reconstruction of arbitrary spatial and time varying source terms using the observed solutions to the forward problem on a discrete set of points. The numerical method is based on computations of the derivatives of filtered versions of the noisy data by discrete mollification and generalized cross validation.

The unknown forcing terms are identified in a compact subset of the domain where the solutions are measured, the compact subset being automatically determined by the amount of noise in the data. We restore continuous dependence on the data, estimate the rate of convergence when certain conditions are met, and provide several numerical examples of interest.

1. INTRODUCTION

For $0 < x < x_1, 0 < t < t_{\max}$, consider a linear partial differential equation of the form

$$u_t = (a(x,t)u_x)_x + f(x,t), \qquad (1)$$

together with the corresponding initial (IC) and boundary conditions (BC)

$$u(0,t) = u_0(t), 0 \le t \le t_{\max}, \qquad \text{BC}$$
$$u(x_1,t) = u_1(t), 0 \le t \le t_{\max}, \qquad \text{BC}$$
$$u(x,0) = u^0(x), 0 \le x \le x_1. \qquad \text{IC}$$

Ordinarily, $f(x,t)$ and $a(x,t)$ are known functions and we are asked to determine the solution functions $u(x,t)$ so as to satisfy equation (1) and (BC-IC). So posed, this is a direct problem.

There is, however, an interesting inverse problem that can be formulated. The objective of this new problem is to determine part of the structure of the system, in our case the forcing term $f(x,t)$, from experimental information given by the approximate knowledge of the function $u(x,t)$ at a discrete set of points in its domain. This question belongs to a general class of inverse problems, known as system identification problems, and, in particular it is an ill-posed problem because small errors in the function $u(x,t)$ might cause large errors in the computation of the partial derivatives $u_t(x,t)$, $u_x(x,t)$, and $u_{xx}(x,t)$ which are needed in order to estimate the forcing term function $f(x,t)$.

To better illustrate the poor stability properties of the mapping from the data function u to the solution function f, let's consider a slab in the (x,t) plane, where the temperature function u satisfies

$$u_t = u_{xx} + f(x,t), 0 < x < \pi, 0 < t < \infty,$$

with homogeneous (zero) boundary and initial conditions. We wish to reconstruct $f(x,t)$ from the **exact** transient temperature history $T(t) = u(x_0,t)$ given at some point x_0, $0 < x_0 < \pi$. Separation of variables leads to the integral representation

$$T(t) = \int_0^t \int_0^\pi k(x,t-s) f(x,s)\, dx\, ds,$$

where the kernel function is given by

$$k(x,\tau)=\frac{2}{\pi}\sum_{n=1}^{\infty}e^{-n^2\tau}\sin(nx_0)\sin(nx).$$

Introducing the sequence of data functions

$$T_n(t)=n^{-\frac{3}{2}}(2-e^{-n^2t})\sin(nx_0), n\geq 0,$$

the source terms f_n are independent of t and we obtain $f_n=2\sqrt{n}\sin(nx)$. From here, it is clear that when $n\to\infty$, $T_n\to 0$ and

$$\max_{0<x<\pi}|f_n(x)|=2\sqrt{n}\to\infty,$$

showing that the problem is greatly ill-posed with respect to perturbations in the data.

The identification of source terms in the one-dimensional inverse heat conduction problem (IHCP) has been extensively explored. However, the available results are based on the assumptions that the source term f depends only on one variable (Cannon and DuChateau [1]) or that it can be separated into spatial and temporal components (Ewin and Lin [3], Nanda and Das [5], Coles and Murio [2].) A historical and technical review of general inverse source problems can be found in the classical book of Isakov [4].

The basic idea of the method presented in this paper begins by attempting to reconstruct mollified versions of several partial derivative functions. The approximations are generated initially by filtering the noisy data by discrete convolution with an averaging kernel and then using finite differences to numerically solve the associated well-posed problems. Once the approximate derivative functions have been computed, the function f is evaluated providing an estimate for the unknown forcing term. The efficiency of this "direct" and simple approach is demonstrated in section 3 where several numerical examples of interest are presented. In section 2 the stabilized problem is introduced and the corresponding error bounds are derived.

2. STABILIZED PROBLEM

In what follows we consider, without loss of generality, the temperature function $u(x,t)$ measured in the unit square $I=I_x\times I_t=[0,1]\times[0,1]$ of the (x,t) plane, i.e., we set $x_1=1$ and $t_{\max}=1$ in equation (1). On the basis of this information we discuss the problem of estimating the forcing term function $f(x,t)$ in some suitable compact set $K\subseteq I$. We denote the x and t - projections of K by K_x and K_t, respectively.

If $C^0(I)$ represents the set of continuous real functions over I with norm

$$\|g\|_{\infty,I}=\max_{(x,t)\in I}|g(x,t)|,$$

we assume that the functions $u(x,t)$, $a(x,t)$, $a_x(x,t)$, $u_t(x,t)$, $u_x(x,t)$, $u_{xx}(x,t)$ and $f(x,t)\in C^0(I)$. We also assume that instead of the function $u(x,t)$, we know some data function $u^{\varepsilon}(x,t)$ such that $\|u-u^{\varepsilon}\|_{\infty,I}\leq\varepsilon$.

In order to stabilized the source problem, we introduce the function

$$\rho_{\delta,p}(x)=\begin{cases}A_p\dfrac{1}{\delta}\exp(-\dfrac{x^2}{\delta^2}), & |x|\leq p\delta,\\ 0, & |x|>p\delta,\end{cases}$$

with $\delta>0$, $p>0$, and $A_p=(\int_{-p\delta}^{p\delta}\exp(-s^2)ds)^{-1}$.

$\rho_{\delta,p}\in C^{\infty}(-p\delta,p\delta)$, is nonnegative, and satisfies $\int_{-p\delta}^{p\delta}\rho_{\delta,p}(x)dx=1$.

For $g(x)\in L^1(I_x)$, and for $x\in K_x$, we define the δ-mollification of g by

$$J_{\delta}g(x)=(\rho_{\delta}*g)(x)=\int_{I_x}\rho_{\delta}(x-s)g(s)ds$$

$$= \int_{x-p\delta}^{x+p\delta} \rho_\delta(x-s)g(s)ds\,,$$

with the p - dependency on the kernel dropped for simplicity of notation. We observe that $p\delta = dis\text{tance}(K_x, \partial I_x)$.

The following lemma and theorem are needed for the stability analysis. The proofs can be found, for example, in Murio, Mejía and Zhan [6].

In all cases, the discrete (sampled) functions $G, G^\varepsilon = \{g(x_j), g^\varepsilon(x_j) : j \in Z\}$ are defined on a uniform partition of I_x, with step size Δx, and satisfy $\| G - G^\varepsilon \|_{\infty, K_x} \le \varepsilon$. The symbol C represents a generic positive real parameter.

Lemma 1 If $g(x), \frac{d}{dx}g(x)$ and $\frac{d^2}{dx^2}g(x) \in C^0(I_x)$, there exist constants C and C_δ, independent of δ and Δx, respectively, such that

$$\| J_\delta G^\varepsilon - J_\delta G \|_{\infty, K_x} \le C(\varepsilon + \Delta x)\,,$$

$$\| J_\delta G^\varepsilon - g \|_{\infty, K_x} \le C(\varepsilon + \delta + \Delta x)\,,$$

$$\| \frac{d}{dx} J_\delta G^\varepsilon - \frac{d}{dx} g \|_{\infty, K_x} \le C(\delta + \frac{\varepsilon + \Delta x}{\delta})\,,$$

$$\| \frac{d^2}{dx^2} J_\delta G^\varepsilon - \frac{d^2}{dx^2} g \|_{\infty, K_x} \le C(\delta + \frac{\varepsilon + \Delta x}{\delta^2})\,,$$

$$\| D_0(J_\delta G^\varepsilon) - \frac{d}{dx} g \|_{\infty, K_x} \le C(\delta + \frac{\varepsilon + \Delta x}{\delta}) + C_\delta(\Delta x)^2$$

and

$$\| D_0^2(J_\delta G^\varepsilon) - \frac{d^2}{dx^2} g \|_{\infty, K_x} \le C(\delta + \frac{\varepsilon + \Delta x}{\delta^2}) + C_\delta(\Delta x)^2,$$

where D_0 and D_0^2 denote the centered and backward - forward finite differences approximations to the first and second derivatives, respectively.

In what follows we will use D_0^t, D_0^x and $D_0^{2,x}$ to indicate the corresponding finite differences approximations to the partial derivatives.

Lemma 1 shows that attempting to reconstruct derivatives of mollified noisy data functions is a stable problem with respect to perturbations in the data, in the maximum norm. This regained stability property is naturally inherited by the mollified reconstructed source term $J_\delta F^\varepsilon(x,t)$, which is obtained as a linear combination of partial derivatives of the measured temperature function. More precisely, we have

$$J_\delta F^\varepsilon(x,t) = a_x D_0^x(J_\delta u^\varepsilon) + aD_0^{2,x}(J_\delta u^\varepsilon) - D_0^t(J_\delta u^\varepsilon)\,. \qquad (2)$$

We can now state our main theoretical result.

Theorem 1 Under the conditions of Lemma 1, for fixed $\delta > 0$, the reconstructed mollified source term $J_\delta F^\varepsilon$, given by formula (2), satisfies

$$\| J_\delta F^\varepsilon - f \|_{\infty, K} \le C\{\delta + \frac{\varepsilon + \Delta x + \Delta t}{\delta^2}\} + C_\delta[(\Delta x)^2 + (\Delta t)^2]\,.$$

Proof Rearranging terms in equations (1) and (2), subtracting, and using maximum norms, we have

$$\| J_\delta F^\varepsilon - f \|_{\infty, K} \le \| D_0^t(J_\delta u^\varepsilon) - u_t \|_{\infty, K} + \| a_x \|_{\infty, K} \| D_0^x(J_\delta u^\varepsilon) - u_x \|_{\infty, K} + \| a_x \|_{\infty, K} \| D_0^2(J_\delta u^\varepsilon) - u_{xx} \|_{\infty, K}\,,$$

and by Lemma 1 we have

$$\| J_\delta F^\varepsilon - f \|_{\infty, K} \le C(\delta + \frac{\varepsilon + \Delta t}{\delta}) + C_\delta(\Delta t)^2 + \| a_x \|_{\infty, K} [C(\delta + \frac{\varepsilon + \Delta x}{\delta}) + C_\delta(\Delta x)^2]$$

$$+ \| a \|_{\infty,K} [C(\delta + \frac{\varepsilon + \Delta x}{\delta^2}) + C_\delta (\Delta x)^2].$$

Setting $M = \max(\| a \|_{\infty,K}, \| a_x \|_{\infty,K}, 2)$, we obtain the desired estimate

$$\| J_\delta F^\varepsilon - f \|_{\infty,K} \le CM \{\delta + \frac{\varepsilon + \Delta x + \Delta t}{\delta^2}\} + C_\delta [(\Delta x)^2 + (\Delta t)^2].$$

Corollary To get formal convergence, the ill-posedness of the problem requires to relate all the parameters involved.

The choice $\delta = (2(\varepsilon + \Delta x + \Delta t))^{1/3}$ shows that $\| J_\delta F^\varepsilon - f \|_{\infty,K} = O(\varepsilon + \Delta x + \Delta t)^{1/3}$, which implies formal convergence as ε, Δx, and $\Delta t \to 0$.

Remarks In practice, when modeling, the selection $\delta = \delta(\varepsilon)$ is performed automatically by combining the mollification method with the statistical procedure of generalized cross validation, as described in Murio, Mejía and Zhan [6].

We also note that the choice of δ automatically defines the compact subset $K \subseteq I$ where we seek to reconstruct the unknown forcing term $f(x,t)$.

3. NUMERICAL PROCEDURE

Let $h = \Delta x = 1/M$ and $k = \Delta t = 1/N$ be the parameters of the finite differences discretization of I. We denote by $R_j^n, W_j^n, Q_j^n, U_j^n$, and F_j^n, the discrete computed approximations of the mollified temperature function $u^\varepsilon(jh, nk)$, the mollified time derivative temperature $u_t^\varepsilon(jh, nk)$, the mollified space derivative temperature $u_x^\varepsilon(jh, nk)$, the mollified second space derivative temperature $u_{xx}^\varepsilon(jh, nk)$, and mollified source term function $f^\varepsilon(jh, nk)$, respectively. Here, the ε dependency on the discrete functions has been eliminated to simplify the notation.

Computation of F_j^n throughout the entire do-domain I requires the extension of the data to a slightly larger domain $I_\delta = [-p\delta_x, 1 + p\delta_x] \times [-p\delta_t, 1 + p\delta_t]$. For computational efficiency, the original two-dimensional problem is reduced to a sequence of one-dimensional problems by "marching" in the x (or t) direction and we only need to consider one-dimensional extensions. If needed, by storing the radius of mollification at each step, we can reconstruct $K \subseteq I$, the compact subset where the error estimate given by Theorem 1 is valid. For details, see Zhan and Murio [7] and the references therein.

For j = 1 to M - 1, the space marching scheme to compute W_j^n, Q_j^n, U_j^n, and F_j^n is defined by

$$W_j^n = \frac{R_j^{n+1} - R_j^{n-1}}{2k},\ n = 1, 2, ..., N-1,$$

$$Q_j^n = \frac{R_{j+1}^n - R_{j-1}^n}{2h},\ n = 1, 2, ..., N-1,$$

$$U_j^n = \frac{R_{j+1}^n - 2R_j^n + R_{j-1}^n}{h^2},\ n = 1, 2, ..., N-1,$$

$$F_j^n = W_j^n - a_x(jh, nk) Q_j^n - a(jh, nk) U_j^n,$$

$$n = 1, 2, ..., N-1.$$

The discretized measured approximations of the temperature data functions are modeled by adding random errors to the exact data functions. That is, the reconstructions are attempted on the whole domain $I = [0,1] \times [0,1]$, and

$$u^\varepsilon(jh, nk) = u_j^n + \varepsilon_j^n,$$

$j = 0, 1, ..., M$, $n = 0, 1, ..., N$,

where the ε_j^n's are Gaussian random variables with values in $[-\varepsilon, \varepsilon]$.

Examples

The numerical examples presented next cover an interesting variety of possible behaviors for the source term $f(x,t)$. The first example describes a forcing term that is oscillatory in space

while the second example illustrates a forcing term which is highly oscillatory in time. Example 3 corresponds to a smooth source term that decreases rapidly in time near the boundary $x = 0$ and very slowly near the boundary $x = 1$. Finally, example 4 involves a rather complicated non-smooth rapidly varying forcing term in the x - component.

Tables 1, 2, 3, and 4, illustrate the quantitative behavior of the numerical method. For each of the four examples, the average l_2 relative error norm -over 200 hundred random trials- corresponding to $f(x,t)$, $u(x,t)$, $u_x(x,t)$, $u_{xx}(x,t)$ and $u_t(x,t)$, are reported.

The qualitative behavior of the method is illustrated in Figures 1, 2, 3, and 4, where we show the exact and typical computed source terms associated with each one of the examples. All the graphs correspond to source terms reconstructed with parameters $h = \Delta x = \frac{1}{M} = k = \Delta t = \frac{1}{N}$, $M = 128$, $p = 3$ and $\varepsilon = 0.005$.

Example 1

Identify $f(x,t)$ in $u_t = (-xu_x)_x + f(x,t)$ if the exact data temperature is given by

$$u(x,t) = e^{x-t}\sin(10t).$$

In this example, the exact source term function is

$$f(x,t) = (x\sin 10t + 10\cos 10t)\, e^{x-t}.$$

Table 1 Relative l_2 error norms (M = N)

M	ϵ	f	u	u_x	u_{xx}	u_t
64	.005	.169	.009	.073	.243	.171
128	.005	.088	.008	.064	.203	.088
256	.005	.022	.009	.062	.199	.016
64	.010	.171	.010	.074	.243	.173
128	.010	.087	.011	.064	.203	.087
256	.010	.022	.010	.062	.199	.017

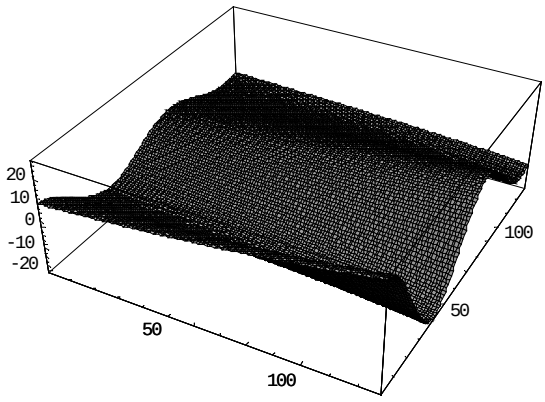


Figure 1a Exact source term

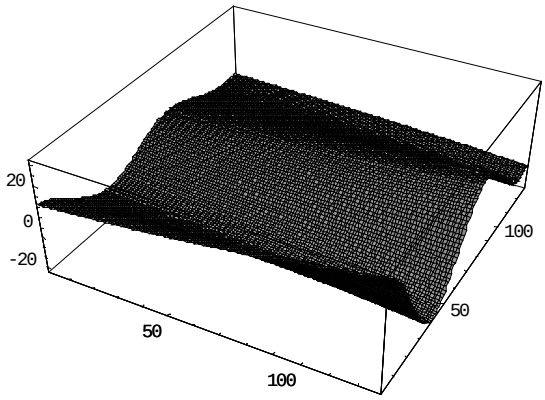


Figure 1b Computed source term

Example 2

Identify $f(x,t)$ in

$$u_t = ((\frac{15}{10} + \sin 20x)u_x)_x + f(x,t)$$

if the exact data temperature is given by

$$u(x,t) = e^{x-t}.$$

In this example, the exact source term function is

$$f(x,t) = -(\frac{25}{10} + 20\cos 20x + \sin 20x)e^{x-t}.$$

Table 2 Relative l_2 error norms (M = N)

M	ϵ	f	u	u_x	u_{xx}	u_t
64	.005	.171	.0085	.073	.243	.172
128	.005	.059	.0024	.058	.187	.028
256	.005	.022	.0087	.062	.199	.017
64	.010	.079	.0048	.075	.251	.075
128	.010	.059	.0049	.058	.187	.028
256	.010	.022	.0048	.062	.197	.017

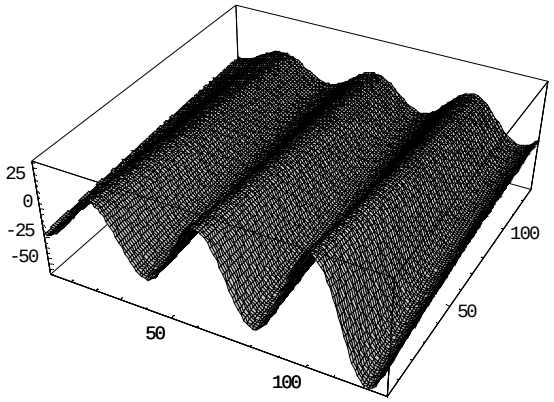


Figure 2a Exact source term

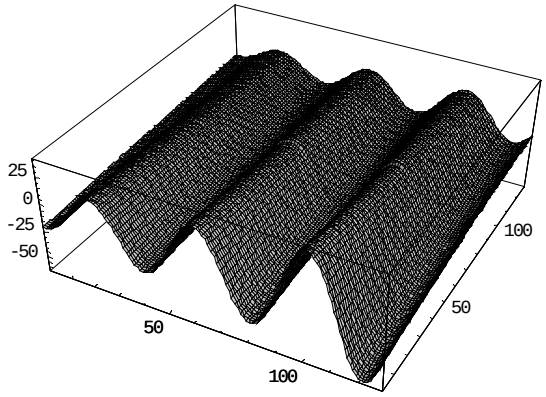


Figure 2b Computed source term

Example 3

Identify $f(x,t)$ in

$$u_t = ((1+4(x-\frac{1}{2})^2)u_x)_x + f(x,t)$$

if the exact data temperature is given by

$$u(x,t) = e^{x-t}.$$

In this example, the exact source term function is

$$f(x,t) = -(4(x-\frac{1}{2})^2 + 8(x-\frac{1}{2}) + 2)e^{x-t}.$$

Table 3 Relative l_2 error norms (M = N)

M	ϵ	f	u	u_x	u_{xx}	u_t
64	.005	.124	.0027	.075	.259	.059
128	.005	.088	.0025	.058	.187	.089
256	.005	.055	.0023	.055	.174	.081
64	.010	.119	.0051	.075	.259	.069
128	.010	.085	.0049	.058	.187	.069
256	.010	.063	.0047	.055	.174	.076

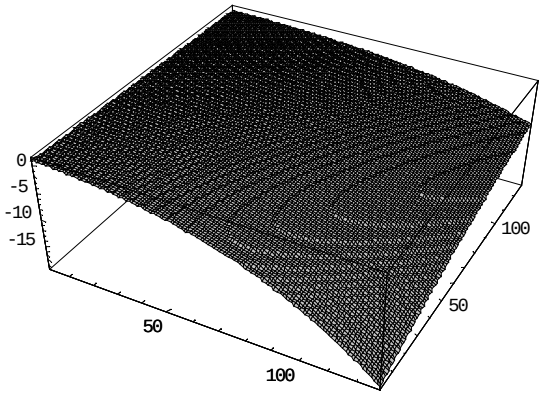


Figure 3a Exact source term

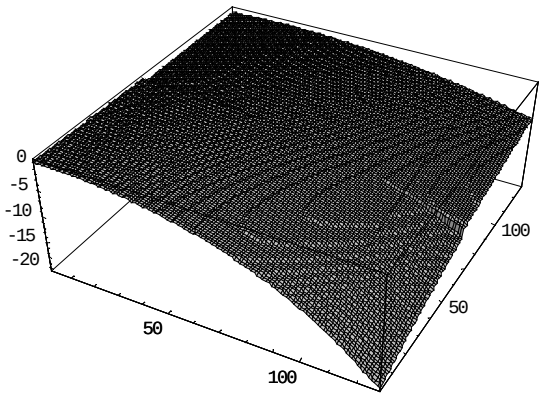


Figure 3b Computed source term

Example 4

Identify $f(x,t)$ in $u_t = (au_x)_x + f(x,t)$ if the exact diffusivity coefficient is given by

$$a(x) = \begin{cases} 1, & 0 \le x < \frac{1}{4}, \\ 4x, & \frac{1}{4} \le x < \frac{1}{2}, \\ 3-2x, & \frac{1}{2} \le x < \frac{3}{4}, \\ 1.5, & \frac{3}{4} \le x \le 1, \end{cases}$$

and the exact data function is

$$u(x,t) = e^{x-t}.$$

In this example, the exact source term function is

$$f(x,t) = \begin{cases} -2e^{x-t}, & 0 \le x < 1, \\ -(5+4x)e^{x-t}, & \frac{1}{4} \le x < \frac{1}{2}, \\ (-2+2x)e^{x-t}, & \frac{1}{2} \le x < \frac{3}{4}, \\ -\frac{25}{10}e^{x-t}, & \frac{3}{4} \le x \le 1. \end{cases}$$

Table 4 Relative l_2 error norms (M = N)

M	ϵ	f	u	u_x	u_{xx}	u_t
64	.005	.119	.0025	.075	.026	.070
128	.005	.101	.0026	.058	.187	.020
256	.005	.096	.0025	.055	.174	.059
64	.010	.118	.0050	.075	.026	.062
128	.010	.098	.0049	.058	.187	.045
256	.010	.092	.0036	.055	.174	.042

4. CONCLUSIONS

The simple approach and results offered in this presentation indicate that the methodology is a viable alternative to recover arbitrary source terms depending on space and time.

Extension of the procedure to higher dimensional cases is straightforward.

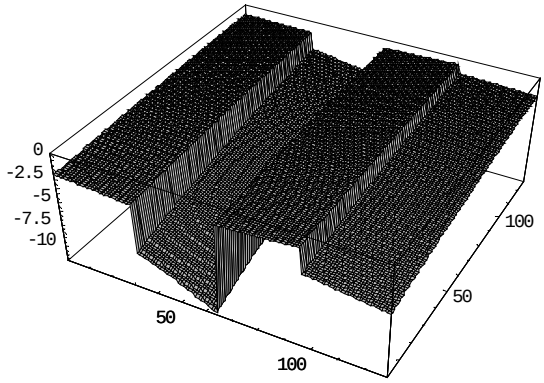


Figure 4a Exact source term

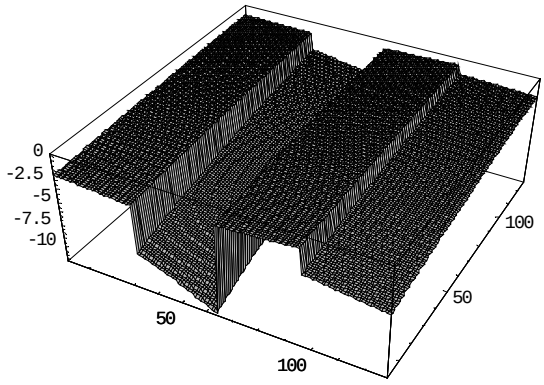


Figure 4b Computed source term

Figures 1-4 illustrate numerical results obtained with sensors placed at every single point of the spatial grid. If the number of sensor is smaller than $M+1$, or are not equally spaced, the data should be carefully interpolated before applying the algorithm.

ACKNOWLEDGMENT

The work of the second author, Diego A. Murio, was partially supported by a C. Taft fellowship.

5. REFERENCES

1. Cannon, J. R., and P. DuChateau, Inverse problems for an unknown source in heat equation, *Journal of Mathematical Analysis and Applications*, **75**, pp. 465-485, (1980)

2. Coles, C., and D. A. Murio, Simultaneous space diffusivity and source term reconstruction

in 2D IHCP, *Computers Math. Applic.* **42**, pp. 1549-1564, (2001)

3. Ewing, R., and T. Lin, Parameter identification problems in single-phase and two-phase flow, *International Series of Numerical Mathematics*, pp. 85-108, Birkhäuser Verlag, (1989)

4. Isakov, V., Inverse Source Problems, *American Mathematical Society*, Providence, (1990)

5. Nanda, A., and P. Das, Determination of the source term in the heat conduction equation , *Inverse Problems*, **12**, pp. 325-339, (1996)

6. Murio, D. A., C. E. Mejía and S. Zhan, Discrete mollification and automatic numerical differentiation, *Computers Math . Applic.* **35**, No. 5, pp. 1-16, (1998)

7. Zhan, S. and D. A. Murio, Surface fitting and numerical gradient computations by discrete mollification, *Computers Math. Applic*. **37**, No. 5, pp. 85-102, (1999)

4^{th} *International Conference on Inverse Problems in Engineering*
Rio de Janeiro, Brazil, 2002

LOCAL REGULARIZATION ALGORITHMS OF SOLVING COEFFICIENT INVERSE PROBLEMS FOR SOME DIFFERENTIAL EQUATIONS

Alexandre Grebennikov
Facultad de Ciencias Físico Matemáticas
Benémerita Universidad Autónoma de Puebla,
Puebla, Pue., Mexico
e-mail: agrebe@fcfm.buap.mx

ABSTRACT

The Inverse Problems under consideration consist in a reconstruction of the coefficients of differential equations. These coefficients are functions only of the space variables and characterize the properties of a media. One coefficient is included in the Laplace operator, written in the divergent form, another is the cofactor at time derivative of the solution. We suppose that the model has the source term, includes an initial condition and the Dirichlet or Neuman boundary conditions. This model, described by a parabolic equation, corresponds to the applied problems of heat-conduction and also to the identification of the characteristics of the porous media of confined aquifers. In the case of a stationary process we have an elliptic equation and the problem of the coefficient reconstruction corresponds to the Electrical Tomography. The measurements of the equation solution and the source term at discreet points, that usually do not form a regular net, are used as the input data. We propose Local Approach and the Full Spline Approximation Method (F.S.A.M.) for the numerical solution of these Inverse Problems. The theoretical justification of constructed algorithms and results of numerical experiments are given.

NOMENCLATURE

A	matrix
$f_x^{'}, \frac{\partial f}{\partial x}$	derivative of f on x
h	grid step
q, u	equation solution
Q	source term
n	number of points in a grid
r, φ	polar coordinates
S_k	recursive approximation spline
t	time
T, P	coefficients of the equation
x, y	Cartesian coordinates
Δ	Laplace operator
Ω	region
δ	error estimation
α	regularization parameter

Subscripts and Superscripts

l, i, j	subscripts
k	recursion index

INTRODUCTION

Let us consider an equation

$$\frac{\partial}{\partial x}\left(T\ q_x^{'}(x,y,t)\right) + \frac{\partial}{\partial y}\left(T\ q_y^{'}(x,y,t)\right) \quad (1)$$

$$+P\ q_t^{'}(x,y,t) + Q(x,y,t) = 0,$$

where $x, y \in \Omega$ - some region on a plane, $0 < t < t_0$, the functions $T = T(x,y)$ and $P = P(x,y)$ characterize the properties of a media, $Q = Q(x,y,t)$ is the source term. This model corresponds to the heat-conduction applied problems (in this case $T < 0$, $P < 0$) and to problems of identification of the porous media characteristics of confined aquifers. The model has also an initial condition and the Dirichlet or Neuman boundary conditions. As a rule, the statements of the corresponding inverse problems lead to nonlinear extremum problems [1], [2], [3]. We use another Local Approach, proposed in [4], [5], where for $T = -1$ the reconstruction of the coefficient P, depending only on time, was considered. In [5], [6], some results of the reconstruction of the coefficient T for given P, are presented with the detailed error analysis. We consider here Inverse Problems of the reconstruction of coefficients T and P in equation (1), using measurements of the solution of the equation and of the source term as the input data.

Concretely, the next problems are considered: 1) simultaneous reconstruction of coefficients T and P depending only on space variables (x, y); 2) reconstruction of the coefficient T in the stationary case. The input data are given in discrete and noised form, that makes the considered problems ill-posed. The regularization based on the Full Spline Approximation Method (F.S.A.M.) is proposed. It consists of five steps: 1) recalculation data to the regular net; 2) recursive pre-smoothing of the recalculated input data; 3) checking up the stop rule in the recursive pre-smoothing process; 4) pre-reconstruction, i.e. solving, possibly with the precondition, a system of the linear algebraic equations with respect to reconstructing coefficients; 5) post-smoothing of the pre-reconstructed coefficients. The F.S.A.M. differs from the previously proposed and justified by the author Spline Approximation Method (S.A.M.) [8], [9] by the presence of the precondition and the post-smoothing. Pre- and post-smoothing are realized by explicit spline approximation formulas [7]. The precondition is realized here by the Tikhonov regularization. We consider here the number of the recursions (number of the smoothings) and parameter in the Tikhonov regularization as two independent regularization parameters. Since the problems under consideration have the character of instability as a problem of second degree numerical differentiation of functions, we use here the F.S.A.M. with *cubic* splines [7]. Further we suppose that the exact input data, the initial and boundary conditions guarantee the existence and the uniqueness of the solutions for the considered Inverse Problems.

INVERSE PROBLEMS FOR THE TWO-DIMENSIONAL PARABOLIC DIFFERENTIAL EQUATIONS

Let us consider the problem of restoring the coefficients of an equation (1) of the parabolic type, when $(x, y) \in \Omega$, $t \in [0, t_0]$, the functions $T(x, y)$ and $P(x, y)$ characterize the properties of a media, $Q(x, y, t)$ is the source term. We consider an initial condition and Dirichlet or Neumann boundary conditions as zeros for simplicity. This equation corresponds to the applied problems of heat-conduction [1], [2], as well as of identifying the characteristics of the porous media of confined aquifers [13]. Local Approach here consists in the use of equation (1) in four moments of time, which gives principal possibility to determine functions $T(x, y), P(x, y)$.

Let us formulate the inverse problems for the exact data and for the noised discrete data in equation (1).

Problem 1 (Inverse Problem for the exact data). Let the exact functions $Q(x, y, t)$, $q(x, y, t)$, $q_x^{'}(x, y, t)$, $q_y^{'}(x, y, t)$, $\Delta q(x, y, t)$, $q_t^{'}(x, y, t)$ be given in four moments of time $t = t_l, l = l_1, l_2, l_3, l_4; 0 \le l_i \le nt, i = 1, ..., 4; nt \ge 6$. It is necessary to reconstruct functions $T(x, y), P(x, y)$.

We introduce the matrix $A = \{a_{k\,j}\}$, $a_{k1} = q_x^{'}(x, y, t_{l_k})$, $a_{k2} = q_y^{'}(x, y, t_{l_k})$, $a_{k3} = \Delta q(x, y, t_{l_k})$, $a_{k4} = q_t^{'}(x, y, t_{l_k})$, $k = 1, ..., 4$; and the vector functions $T^1(x, y) = (T_x^{'}(x, y), T_y^{'}(x, y), T(x, y), P(x, y))^T$, $Q_0 = (Q(x, y, t_{l_1}), Q(x, y, t_{l_2}), Q(x, y, t_{l_3}), Q(x, y, t_{l_4}))^T$. Solution of Problem 1 can be obtained from the linear algebraic system

$$AT^1 = -Q_0. \qquad (2)$$

In practice the input data as a rule are discrete and noisy. So we consider below another posing of the Inverse Problem. We suppose that the functions $q(x, y, t)$, $Q(x, y, t)$ are given by discrete and noisy values $q_{i,l} = q(p_i, t_l) + \delta_{i,l}^q, Q_{i,l} = Q(p_i, t_l) + \delta_{i,l}^Q$, in moments of time $t = t_l, l = 1, ..., nt$, in the points $p_i = \left(x^i, y^i\right), i = 1, ..., N$, of the irregular net. Errors $\delta_{i,l}^Q$, $\delta_{i,l}^q$, satisfy the estimations $\max_{i,l}\{|\delta_{i,l}^q|\} \le \delta^q$, $\max_{i,l}\{|\delta_{i,l}^Q|\} \le \delta^Q$, $\delta = \max\{\delta^q, \delta^Q\}$.

Problem 2 (Inverse Problem for the discrete noisy data). Let functions $Q(x, y, t)$, $q(x, y, t)$ be given by discrete and noisy data $\{q_{i,l}\}$, $\{Q_{i,l}\}$ described above. It is necessary to reconstruct approximately functions $T(x, y), P(x, y)$.

To formulate the method for solving Problem 2 we suppose that points p_i are the same for both functions and all moments of the time t_l , $l = 0, 1, ..., nt$. We will consider the region $\Omega \subseteq \Omega_0 = [0, 1] \times [0, 1]$, and introduce the grids: $\{x_i\} : x_i = (i - 1) \times h_x$, $h_x = 1/(n_x - 1)$, $i = -1, ..., n_x + 2$; $\{y_i\} : y_i = (i-1) \times h_y$, $h_y = 1/(n_y - 1)$, $i = -1, ..., n_y + 2$. We suppose that the grid $\{t_m\}$ is the uniform

grid: $t_m = (m-1) \times h_t$, $h_t = 1/(n_t - 1)$, $m = 1, ..., n_t$. We use for recalculation of given values of functions $q(x,y,t)$, $Q(x,y,t)$ to the regular net $\{x_i, y_j\}$ the irrational recuperation, that can be written for some function $f(x,y,t_l)$ in fixed moment t_l of time by formulas:

$$IR_l(x,y,f) = \frac{\sum_{i=1}^{N} d_i(x,y) f(p_i,t_l)}{\sum_{i=1}^{N} d_i(x,y)}, \qquad (3)$$

where

$$d_i(x,y) = \left(\left(x - x^i\right)^2 + \left(y - y^i\right)^2 \right)^{-2},$$

$i = 1, ..., N$. We designate

$$d(N) = \max_{1 \le i \le N} \min_{j \ne i} \|p_i - p_j\|_{R_2},$$

where R_2 is two-dimensional Euclidean space. We shall use also formulas of the Recursive Smoothing spline-method [8], [9] for the case of two variables functions for fixed moments t_l of time:

$$S_k(x,y,t_l,f) = \qquad (4)$$

$$\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} S_{k-1}(x_i, y_j, t_l, f) s_i(x) s_j(y),$$

where $k \ge 1$; $S_0(x_i, y_j, t_l, f) = IR_l(x_i, y_j, f)$, $s_i(u)$ are local basic cubic splines [10], constructed on the units $u_{i-2}, ..., u_{i+2}$; $i = 0, ..., nu + 1$; u is x, or u is y. We introduce the discrepancy (residual) function

$$\varrho_k(f) \equiv \left(\frac{1}{(n_x n_y)} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \Delta_{i,j}(f) \right)^{1/2},$$

where
$\Delta_{i,j}(f) = |S_k(x_i, y_j, t_l, f) - IR_l(x_i, y_j, f)|^2$.

We propose for the numerical solution of Problem 2 the Full Spline Approximation Method, which consists in calculating reconstruction splines T_σ, P_σ, $\sigma = \max\{h_t, h_x, h_y, \delta, d(N)\}$, with the next steps:

1) *recalculation* data to the regular net by formulas (3) for $f = q, Q$;

2) recursive by $k_f = 1, ...$ *pre-smoothing* of the recalculated input data by formula (4) for $f = q, Q$;

3) *stop rule comparing:* if the discrepancy function $\varrho_{k_f}(f)$ for $f = q, Q$ satisfy the estimation $\varrho_{k_f}(f) \le c\delta^f$, then $k_f := k_f + 1$, go to the item 2); if $\varrho_{k_f}(f) > c\delta^f$, then $K_f = k_f$; $c = const > 1$;

4) *pre-reconstruction,* i.e. calculating "smoothed" matrix A_σ and the right-hand side Q_σ for the system (2) by formulas: $A_\sigma = \{\widetilde{a}_{l\,j}\}$, $\widetilde{a}_{l\,1} = \frac{\partial S_{K_q}(x,y,t_l,q)}{\partial x}$, $\widetilde{a}_{l\,2} = \frac{\partial S_{K_q}(x,y,t_l,q)}{\partial y}$, $\widetilde{a}_{l\,3} = \Delta S_{K_q}(x,y,t_l,q)$, $\widetilde{a}_{l\,4} = \frac{\partial S_{K_q}(x,y,t_l,q)}{\partial t}$; $Q_\sigma = (S_{K_Q}(x,y,t_{l_1},Q), S_{K_Q}(x,y,t_{l_2},Q), S_{K_Q}(x,y,t_{l_3},Q), S_{K_Q}(x,y,t_{l_4},Q))^T$ and solving a preconditioned by the Tikhonov regularization system

$$(A_\sigma^* A_\sigma + \alpha E)\widetilde{T}_\sigma^1 = -A_\sigma^* Q_\sigma, \qquad (5)$$

where A_σ^* is the matrix conjugate to A_σ, $\alpha = \max\{\sqrt{\varepsilon}, \sqrt{\delta}\}$ is the regularization parameter, ε is the estimation of the round-up;

5) *post-smoothing,* i.e. calculation of reconstructions T_σ and P_σ by application of formulas (4) with the given number $k = \overline{K}$ times to the pre-reconstructed at the 4-th step functions $\widetilde{T}_\sigma$ and $\widetilde{P}_\sigma$ (the third and fourth components of the vector $\widetilde{T}_\sigma^1$); stop.

We suppose that in Problem 1 indexes l_i are chosen as maximal distant one from another, such as $2 \le l_1 < l_2 < l_3 < l_4 < nt$.

Theorem. We suppose:

1) the set $\{p_i\}$ is dense in Ω and the condition holds: $d(N) \to 0$, $N \to \infty$;

2) the function $q(x,y,t)$ has continuous derivatives of fourth order on x, y and of second order on t ;

3) the Problem 1 has unique solution $T^1(x,y)$.

Then for sufficient small σ the system (5) also has the unique solution $\widetilde{T}\,{}_\sigma^1(x,y)$ that converges uniformly to T^1, i.e.,

$$\max_{(x,y)\in\Omega} |\widetilde{T}_\sigma^1(x,y) - T^1(x,y)| \to 0, \sigma \to 0.$$

Proof. At first we note that the function $IR_l(x,y,f)$ interpolates the values $f(p_i,t_l)$ in points p_i and this function is exact on constant functions. Hence, $IR_l(x,y,f)$ has approximating properties. Then, if $\|f\|_{C[\Omega]} \le M = const$, then $\|IR_l(x,y,f)\|_{C[\Omega]} \le M$, that guarantees a lack of increasing of the uniform error in the input data at irrational approximation for sufficiently small $d(N)$. To say strongly, in the inequality of the discrepancy principle it is necessary to use points (x^i, y^i)

instead (x_i, y_j), and put $c = 1$. But mentioned above approximating properties of the function $IR_l(x, y, f)$ give us the possibility to compensate this change for small $d(N)$ by using $c > 1$. The regularizing properties of the Recursive Smoothing spline-method have been justified for the problem of numerical differentiation in [8], [9], [11] for the choice the number of smoothing as the regularization parameter from the discrepancy (residual) principle. Thus, using assumption 2) we obtain the uniform convergence of the corresponding first and second partial derivatives to the exact ones. This means that a matrix A_σ in the system (5), and the vector $-Q_\sigma$ in it's right-hand side converge to the exact matrix and the right-hand side of the system (2) accordingly, when $\sigma \to 0$. Then from the assumption 3) and observed convergence of the matrix A_σ it follows, that for sufficiently small σ the inverse matrix A_σ^{-1} exists. The system (2), as a rule, is ill-conditioned, that is why we resolve not the system $A_\sigma \widetilde{T}_\sigma^1 = -Q_\sigma$, but the preconditioned system (5). Well known properties of the Tikhonov regularization [12] together with the obtained above convergence A_σ and Q_σ give us the convergence $\widetilde{T}_\sigma^1$ to T^1. The post-smoothing with fixed number of iterations $\overline{K}$ (usually $\overline{K} = 2$ or 3) does not play the main part in this convergence, but, as a rule, make better the geometric characteristics of the pre-recuperated solution of inverse problems. These arguments complete the proof.

Constructed algorithm is realized as a complex of MATLAB programs. The input data in numerical example below have been formed as noised data of the exact solution $q(x, y, t) = \sin(t^3 y + x)$ of equation (1) on an irregular grid with the number of knots $N = 51$ in $\Omega_0 = [0, 1] \times [0, 1]$. The function $Q(x, y, t)$ was calculated by using the exact $T(x, y) = \exp(\sin(\pi x))$, $P(x, y) = \sin(\pi xy)$, $q(x, y, t)$ and then have been noised with additional random errors. We have used in Ω_0 the regular net with $n_x = n_y = 11$, $n_t = 6$, $t_0 = 0.5$ for the F.S.A.M. The estimation δ of a noise is over 1%. We present results of the $P(x, y)$ reconstruction. On the graphs of Fig. 1 we can see: 1) isolines of exact $q(x, y, t)$ and points of measurements marked by "stars"; 2) exact P; 3) P reconstructed without any regularization; 4) P reconstructed with the Tikhonov regularization only; 5) P reconstructed with pre-smoothing only; 6) P reconstructed by the F.S.A.M. We can see at graphs that the F.S.A.M. gives the best result.

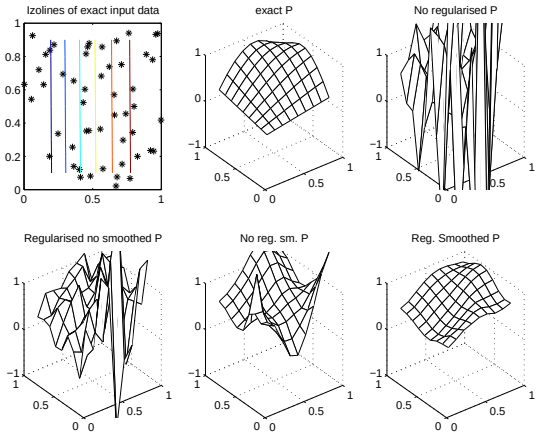


Figure 1: Regularization effect at the reconstruction of a coefficient P of the parabolic differential equation.

RECONSTRUCTION OF COEFFICIENT FOR SOME ELLIPTIC DIFFERENTIAL EQUATIONS

In the stationary case we have the simplified kind of the equation (1) with $q_t' \equiv 0$. The problem under consideration consists in reconstruction of the coefficient T only. We shall suppose that the domain Ω is a unit circle and present in the polar coordinate system (φ, r) the corresponding Laplace equation:

$$T_r' u_r' + \frac{1}{r^2} T_\varphi' u_\varphi' + T\left(u_{rr}'' + \frac{1}{r} u_r' + \frac{1}{r^2} u_{\varphi\varphi}''\right) = 0 \quad (6)$$

with the boundary condition $u(\varphi, 1) = f(\varphi)$. If the solution u is known and initial conditions for T are given, then in principle it is possible to solve the partial differential equation of the first order concerning T by the characteristics method [14]. The local approach for this type of the equations on the basis of the local spline approximation formulas was proposed in [15], [16]. If the input data are given as noised values of the function $u(\varphi, r)$ on some sufficiently thick net, it is possible to realize local approach with the F.S.A.M. This scheme can appear in the component quantification problem for the

fluids of complex mixture (for example, mixture of gas, oil and water). To obtain necessary measurements it is possible to introduce into the tube a set of the cylindrical electrodes.

To demonstrate the possibility of the local approach and the regularization properties of the F.S.A.M. we consider here the most simple case of the radial symmetric solution u such that $u = u(r)$ does not depend on the angle φ. Equation (1) transforms into the equation

$$\frac{1}{r}(Tru_r^{'}(r))_r^{'} = 0,$$

that leads to the relation

$$T(r) = \frac{T(1)u_r^{'}(1)}{ru_r^{'}(r)}.$$

We suppose that the input data present the noised values of the solution $\widetilde{u}_i = u(r_i) + \xi_i$, in the n points $r_i = ih$, $h = 1/n$, $|\xi_i| \leq \delta$, $i = 1, ..., n$; and also the exact values $u_r^{'}(1)$, $T(1)$ that we put equal to 1 for simplicity. We use the adopted for this case F.S.A.M., that includes three recursion steps: 1) pre-smoothing the input data by explicit approximation cubic splines S_k; 2) stop rule in the form of the residual principle; 3) post-smoothing $\overline{K}$ times the calculated pre-reconstruction. The theoretical justification of the regularization properties of this algorithm for sufficiently smooth u and T is similar to the one, presented above in the theorem, with the corresponding modifications. But, as a rule, T presents some piecewise constant function, corresponding to the electric properties of the mixture components. However, the proposed algorithm gives good results of the reconstruction of the coefficient T in this case too. Moreover, if the values of this constants $\{T_i\}$ are known a priori, we include this information into the algorithm as the last post-processing step, that consists in the projection of the post-smoothed result on the set $\{T_i\}$. This projection can be realized with respect to the absolute or the relative criteria. If the values $\{T_i\}$ have not very different scale, the absolute criterion gives good results, otherwise we need to use the relative criterion.

Let us present outcomes of some model numerical experiments. For the exact $T(r) = T_1 = 0.5$, $r \in [0, 0.3]$; $T(r) = T_2 = 2$, $r \in [0.3, 0.7]$; $T(r) = T_3 = 1$, $r \in [0.7, 1]$; we calculated the exact $u(r) = 2\ln r - 1.5\ln(0.3) + 0.5\ln(0.7) + 1$, $r \in [0, 0.3]$; $u(r) = 0.5\ln r + 0.5\ln(0.7) + 1$, $r \in [0.3, 0.7]$; $u(r) = \ln r + 1$, $r \in [0.7, 1]$. We used the values $u(r_i)$ with the additional random errors as the input data. In Fig. 2 the results of the coefficient T reconstruction for $n = 51$, $\delta = 0.05$ are presented. At all graphs the exact T is marked by the solid line. Graph (a): dotted line - reconstruction without regularization; graph (b): dotted line - reconstruction with the pre-smoothing only; graph (c): "+" line - F.S.A.M. reconstruction; graph (d): "stars" line - F.S.A.M. reconstruction with the post-processing (absolute criterion).

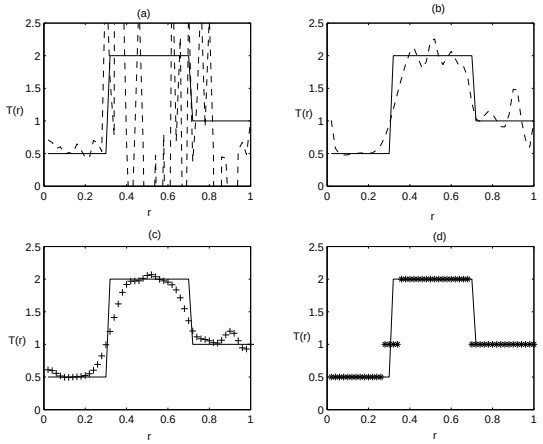


Figure 2: Regularization effect at the coefficient T reconstruction in the elliptic differential equation at the radial symmetry.

Now we consider the case of the angle symmetry, when the solution u and the coefficient T depend on the variable φ only. The equation (1) becomes the following one $(Tu_\varphi^{'}(\varphi))_\varphi^{'} = 0$, that leads to the relation

$$T(\varphi) = \frac{T(0)u_\varphi^{'}(0)}{u_\varphi^{'}(\varphi)}.$$

In the model numerical experiments we used the exact $T(\varphi) = T_1 = 20$, $\varphi \in [0, \pi/2]$; $T(\varphi) = T_2 = 1$, $\varphi \in [\pi/2, 2\pi]$. We calculated the exact $u(\varphi) = 0.05\varphi$, $\varphi \in [0, \pi/2]$; $u(\varphi) = 1, \varphi \in [\pi/2, 2\pi]$. The input data present the noised values of the solution $\widetilde{u}_i = u(\varphi_i) + \xi_i$, in the n points $\varphi_i = (i-1)h$, $h = 2\pi/(n-1)$, $|\xi_i| \leq \delta$, $i = 1, ..., n$; and also the exact value $u_\varphi^{'}(0) = 0.05$. We use the F.S.A.M. algorithm with the

relative criterion in the post-processing. In Fig. 3 the results of the coefficient T reconstruction for $n = 51$, $\delta = 0.05$ are presented. At all graphs the exact T is marked by the solid line. Graph (a): dotted line - reconstruction without regularization; graph (b): dotted line - reconstruction with the pre-smoothing only; graph (c): "+" line - F.S.A.M. reconstruction; graph (d): "stars" line - F.S.A.M. reconstruction with the post-processing (relative criterion).

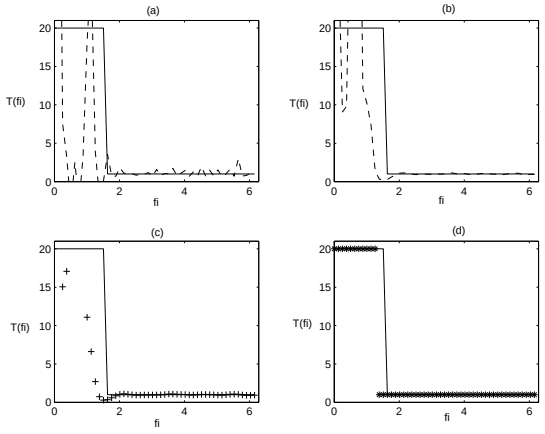


Figure 3: Regularization effect at the coefficient T reconstruction in the elliptic differential equation at the angle symmetry.

Let us consider the Electrical Tomography scheme, when the external electromagnetic field initiate some distribution of the potential inside the domain Ω. We introduce the $g-$function, that characterize the derivative of the produced potential $u_l'(x,y)$ along the straight line l, connecting two boundary points $P_1 = (x^1, y^1)$ and $P_0 = (x^0, y^0)$. The line l has the parametric presentation $p = x\cos\varphi + y\sin\varphi$, where $|p|$ is a length of the perpendicular, passed from the center of coordinates to the line l, φ is the angle between the axis x and this perpendicular. We suppose that we can make measurements of the difference $v(p,\varphi) = u(x^1, y^1) - u(x^0, y^0)$ of potentials in the boundary points. In this case, using the traditional scheme of the Radon transformation [17], we can obtain the corresponding integral equation of the first kind concerning the introduced $g-$function, that is related with the coefficient T. Let us consider Ω as the unit circle and the radial symmetric case, for which the function $v(p,\varphi)$ depends only on the variable p. Then the mentioned integral equation is Abel's equation:

$$\int_p^1 \frac{g(t)tdt}{\sqrt{t^2 - p^2}} = v(p),\ p \in [0,1]. \qquad (7)$$

$g-$function has here the explicit relation with $T(p)$ by the formula $g(p) = c/pT(p)$, $c = const$. Hence, the reconstruction of the $g-$function (we will call it $Eg-$ Tomography) give us the possibility to reconstruct $T(p)$. The Local Approach consists here in the application of the inverse Radon transformation:

$$g(p) = -\frac{1}{\pi p}\frac{d}{dp}\int_p^1 \frac{v(t)tdt}{\sqrt{t^2 - p^2}},\ p \in [0,1], \qquad (8)$$

and using as the input data the measured values of the difference of potentials $\widetilde{v}_i = v(p_i) + \xi_i$, in the boundary points of the n parallel lines, corresponding to $p_i = (i + 1/2)h$, $h = 1/n$, $|\xi_i| \le \delta$, $i = 0, 1, ..., n-1$. The approximate formula for calculation of the $g-$function is the following:

$$\begin{aligned} \widetilde{g}(p) &= \frac{\widetilde{v}_k}{\pi} d(p_{k+1}, p) + \\ &\quad \sum_{i=k+1}^{n-1} \frac{\widetilde{v}_i}{\pi}[d(p_{i+1}, p) - d(p_i, p)], \\ d(t,p) &= 1/\sqrt{t^2 - p^2}, \\ p &\in [p_k, p_{k+1}), k = 0, 1, ..., n-1. \end{aligned}$$

We applied the described above in this section F.S.A.M. algorithm to reconstruct numerically $g-$function and then function $T(p)$. We underline, that in this scheme we use as the input data the values of the simulated potential on the boundary only, not inside the circle. Let us present some results for the same model example, that we calculated above for the radial symmetric case.

The first kind of simulation of the function $v(p)$ have been realized for the known $g(p)$, calculating $v(p)$ by formula (7) and the approximation of the integral with the rectangular formula. The results for $n = 51$, $\delta = 0.05$ are demonstrated in Fig. 4 the quality of the proposed algorithm as the regularization procedure, realizing the inverse Radon transformation (8). At all graphs the exact T is marked

by the solid line. Graph (a): dotted line - reconstruction on the no noised simulated data $\{v(p_i)\}$ without regularization; graph (b): dotted line - reconstruction on noised simulated data without regularization; graph (c): "+" line - F.S.A.M. reconstruction on noised simulated data; graph (d): "stars" line - F.S.A.M. reconstruction on noised simulated data with the post-processing (absolute criterion).

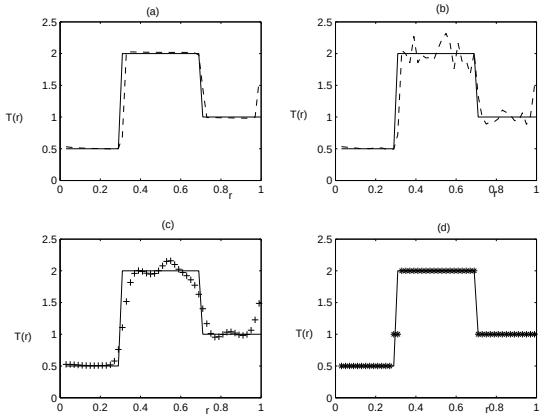


Figure 4: Regularization effect at the coefficient T reconstruction with the inverse Radon transformation at the radial symmetry.

The second kind of simulation consists in the construction of the model potential distribution in the domain Ω for the known $T(p)$ under the influence of the known external electric field. We considered the plane vector field $\overrightarrow{V}(x)$ parallel to axis x independent on y. The simulated relative exact values of the function $v(p)$ can be calculated by formulas: $v(p) = 2[(1/T_2 - 1/T_3)\overline{y} + (1/T_1 - 1/T_2)\overline{z} + \overline{x}/T_3]$, $p \in [0, 0.3]$; $v(p) = 2[(1/T_2 - 1/T_3)\overline{y} + \overline{x}/T_3]$, $p \in [0.3, 0.7]$; $v(p) = 2\overline{x}/T_3$, $p \in [0.7, 1]$, where $\overline{x} = (1 - p^2)^{1/2}$, $\overline{y} = (0.7^2 - p^2)^{1/2}$, $\overline{z} = (0.3^2 - p^2)^{1/2}$. In Fig. 5 the results of the coefficient T reconstruction for $n = 21$, $\delta = 0.05$ are presented as maps of isolines. Graph (a): the exact $T(x, y)$; graph (b): reconstruction on noised simulated data without regularization; graph (c): S.A.M. reconstruction on noised simulated data; graph (d): F.S.A.M. reconstruction on noised simulated data with the post-processing (absolute criterion).

We considered also scanning by the rotating field $\overrightarrow{V}(x)$ of structures without the radial

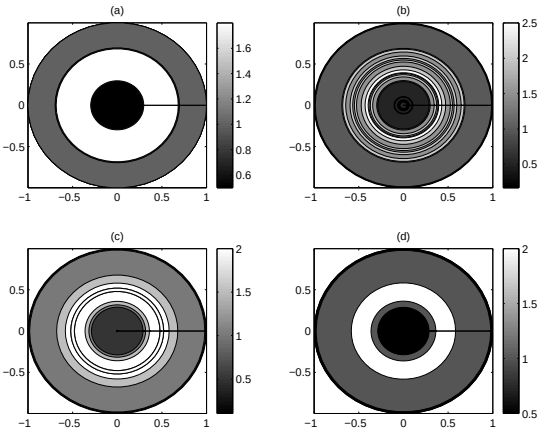


Figure 5: Regularization effect at the coefficient T reconstruction by Eg–Tomography in the radial symmetric case.

symmetry, for which the coefficient $T(x, y) = T_1, (x, y) \in \Omega_i \subset \Omega, i = 1, ..., m$; $T(x, y) = T_2, (x, y) \in \Omega/(\cup_{i=1}^{m} \Omega_i)$. Here T_1, T_2 are known constants. The input data for every fixed angle of scanning are values of potentials in N boundary points of the domain Ω. On the basis of the conception of the $Eg-$ Tomography scheme and the ray character of the considered field $\overrightarrow{V}(x)$ we developed algorithm for the reconstruction of such coefficients $T(x, y)$. Let us present the results of the model numerical experiments for exact simulated input data with $T_1 = 2, T_2 = 1$, $N = 61$ and different numbers $M = 6, 12, 36$ of scanning angles. In Fig. 6 we can see: graph (a) - the exact $T(x, y)$; graph (b), (c), (d) - reconstructed T for $M = 6, 12, 36$ correspondingly. We can see from the graphs of Fig. 2 -6 that proposed algorithms give the possibility to reconstruct desired coefficients of considered elliptic equations with good quality.

References

[1] J. V. Beck , B. Blackwell and S. R. St. Clair. *Inverse Heat Conduction Problems.* John Wiley and Sons, New York, 1985.

[2] O. M. Alifanov, Artyukhin E. A. and Rumyantsev S. V. *Extreme Methods for Solving Ill-posed Problems with Applications to Inverse Problems.* Begell House, Wallinford, UK,1995.

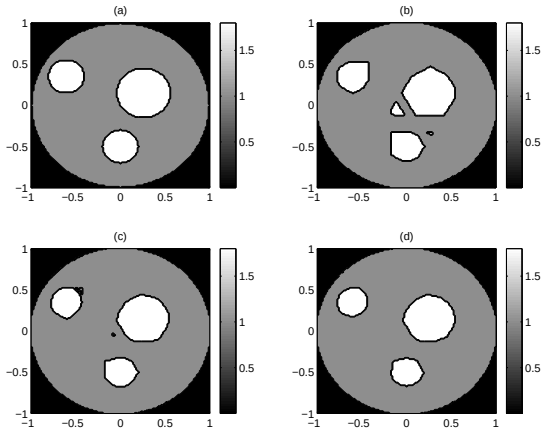


Figure 6: Reconstruction of the coefficient T by Eg–Tomography in the case without the radial symmetry.

[3] M. M. Lavrentiev, V. G. Romanov and S. P. Shishatsky. *Ill-posed problems of the mathematical physics and analysis.* "Nauka", Moscow, 1980.

[4] A. I. Grebennikov. Fast spline-algorithms for multidimensional data processing and solving coefficient inverse problems for heat conduction equation. *Numerical analysis: methods and algorithms.* Moscow State Univ. Publ. House, Moscow, 1998, p. 62-71.

[5] A. I. Grebennikov. Spline algorithms for data processing and solving some inverse problems. *Recent Advances in Numerical Methods and Applications II. Proceedings of the Fourth International Conference NMA'98.* World Scientific, Singapore, 1999, p. 375-383.

[6] A. I. Grebennikov. Local spline - approximation method of solving some coefficient inverse problems for differential equation of the parabolic type. *Inverse Problems in Engineering J.*, Vol. **9**, p. 455-469 (2001).

[7] A. I. Grebennikov. On explicit method of approximation of functions one and many variables by splines. *Comp. Math. Math. Phyz.*, **18**, N 4, p. 853-859 (1978).

[8] A. I. Grebennikov. Spline approximation method for solving some incorrectly posed problems. *Doclady Akad. Nauk SSSR,* **298**, N 3, p.533-537 (1988)

[9] A. I. Grebennikov. Spline approximation method for restoring functions. *Sov. J. Numer. Anal. Mathem. Modelling.* **4**, N 4, p. 1-15 (1989).

[10] I. J. Schoenberg. Contributions to problem of approximation of equidistant data by analytic functions. *Quart.-Appl. Math.*, **4**, p. 45-99, 112-141 (1946).

[11] V. A. Morozov and A. I. Grebennikov. *Methods of Solving Ill-Posed Problems: Algorithmic Aspect.* Moscow State Univ. Publ. House, Moscow, 1992, p. 320.

[12] A. N. Tikhonov and V. And. Arsenin. *Solutions of ill-posed problems.* Winston and Sons, Washington, 1977.

[13] S. Gomez, A. Perez and R. M. Alvares. The Multiscale Optimization for Parameter Identification in the Ariguanabo Aquifer. *Rep. of investigations*, IIMAS of UNAM, Mexico, 1998.

[14] Ph. Hartman. *Odinary Differential Equations.* John Willey and Sons, New York, 1964.

[15] A. I. Grebennikov. One approach to numerical solution of problems for some quasilinear partial differential equations of the first order. *Methods and algorithms in Numerical analysis.* Moscow State Univ. Publ. House, Moscow, 1982, p. 96-97.

[16] A. I. Grebennikov. Fast method of boundary problem solution for ordinary differential equations. *Methods and algorithms in Numerical analysis.* Moscow State Univ. Publ. House, Moscow, 1982, p. 84-95.

[17] J. Radon. Uber die Bestimmung von Funktionen durch ihre Integrawerte langs gewisser Mannigfaltigkeiten. Berichte Sachsische Academic der Wissenschaften, Leipzig. *Math.-Phys.* KI. N 69, p. 262-267 (1917).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

Function Optimization Using Extremal Dynamics

Fabiano Luis de Sousa
INPE-DMC
Av. dos Astronautas, 1758
S.J.Campos, 12227-010, Brazil
Email: fabiano@dem.inpe.br

Fernando Manuel Ramos
INPE-LAC
Av. dos Astronautas, 1758
S.J.Campos, 12227-010, Brazil
Email: fernando@lac.inpe.br

ABSTRACT

In this paper a new stochastic algorithm for function optimization is presented. Called Generalized Extremal Optimization, it was inspired by the theory of Self-Organized Criticality and is intended to be used in complex inverse design problems, where traditional gradient based optimization methods may become inefficient. Preliminary results from a set of test functions show that this algorithm can be competitive to other stochastic methods such as the genetic algorithms.

NOMENCLATURE

k Index of bit rank.
L Length of binary string that encodes the design variables.
l Length of binary string for one design variable.
N Number of design variables.
V Value of the objective function for a given binary string.
x Design variable.
ΔV Bit fitness.
τ Free adjustable parameter of the optimization algorithm.

INTRODUCTION

Stochastic algorithms inspired by nature have been successfully used for tackling optimization problems in engineering and science. Simulated Annealing (SA)[1] and Genetic Algorithms (GAs)[2] are probably the two methods most used. Their robustness and ability to be easily implemented to a broad class of problems, regardless of such difficulties as the presence of multiple local minima in the design space and the mixing of continuous and discrete variables, has made them good tools to tackle complex problems, for example, in the aerospace field[3-7]. The main disadvantage of these methods is that they usually need a great number of objective function evaluations to be effective. Hence, in problems where the calculation of the objective function is very time consuming, these methods may become impracticable. Nevertheless, the availability of fast computing resources or the use of hybrid techniques[8-10] has made the power of those algorithms available even to that kind of problems. There are today many derivatives of the SA and GAs methods, created to give more efficiency to the proposed original algorithms, but that keep essentially their same principles.

Recently, Boettcher and Percus[11] have proposed a new optimization method based on a simplified model of biological evolution developed to show the emergence of Self-Organized Criticality (SOC) in ecosystems.[12] Called Extremal Optimization (EO), it has been successfully applied to tackle hard problems in combinatorial optimization.

Although algorithms such as SA, GAs and the EO are inspired by natural processes, their practical implementation to optimization problems shares a common feature: the search for the optimal is done through a stochastic process that is "guided" by the setting of adjustable parameters. Since the proper setting of these parameters are very important to the performance of the algorithms, it is highly desirable that they have few of such parameters, so that the cost of finding the best set to a given optimization problem does not become a costly task in itself. The EO algorithm has only one adjustable parameter. This may be an "a priori" advantage over the SA and GA algorithms, since they use more than one.

In this paper the Generalized Extremal Optimization (GEO) algorithm is presented. The GEO algorithm is built over the EO method, but the way it is implemented allows it to be readily applied to a broad class of engineering problems. The algorithm is of easy implementation, does not make use of derivatives and can be applied to nonconvex or disjoint problems. It can also deal in principle with any kind of variable, either continuous, discrete or integer. All these features make it suitable to be used in complex inverse design problems, where traditional gradient methods could not be applied properly due to, for example, the presence of multiple local minima or use of mixed types of design variables. In this work the performance of the GEO algorithm is tested in a set of non-linear multimodal functions used commonly to test GAs. The performance of the GEO algorithm for these functions is compared with the ones for a standard GA and the Cooperative Co-

evolutionary GA (CCGA) proposed by Potter and De Jong.[13]

THE EXTREMAL OPTIMIZATION ALGORITHM

Self-organized criticality has been used to explain the behavior of complex systems in such different areas as geology, economy and biology.[14] The theory of SOC states that large interactive systems evolves naturally to a critical state where a single change in one of its elements generates "avalanches" that can reach any number of elements on the system. The probability distribution of the sizes "s" of these avalanches is described by a power law in the form $P(s) \sim s^{-\gamma}$, where γ is a positive parameter. That is, smaller avalanches are more likely to occur than big ones, but even avalanches as big as the whole system may occur with a non-negligible probability. To show that SOC could explain features of systems like the natural evolution, Bak and Sneepen[12] developed a simplified model of an ecosystem in which species are placed side by side on a line with periodic boundary conditions. To each species, a fitness number is assigned randomly, with uniform distribution, in the range [0,1]. The least adapted species, the one with the least fitness, is then forced to mutate, and a new random number assigned to it. The change in the fitness of the least adapted species alters the fitness landscape of their neighbors, and to cope with that new random numbers are also assigned to them, even if they are well adapted. After some iterations, the system evolves to a critical state where all species have fitness above a critical threshold. However, the dynamics of the system eventually causes a number of species to fall below the critical threshold in avalanches that can be as big as the whole system.

An optimization heuristic based on a dynamic search that embodies SOC would evolve solutions quickly, systematically mutating the worst individuals. At the same time this approach would preserve throughout the search process, the possibility of probing different regions of the design space (via avalanches), enabling the algorithm to escape local optima. Inspired by the SOC theory, the basic EO algorithm was proposed as follows:[11]

1.Initialize configuration C of design variables x_i at will; set $C_{best} = C$.

2. For the current configuration C,
   a) set a fitness F_i to each variable x_i,
   b) find j satisfying $F_j \leq F_i$ for all i,
   c) choose C' in a neighborhood N(C) of C so that x_j must change,
   d) accept C = C' unconditionally,
   e) if $F(C) < F(C_{best})$ then set $C_{best} = C$.

3. Repeat step (2) as long as desired.

4. Return C_{best} and $F(C_{best})$.

The above algorithm shows good performance on problems, such as graph partitioning, where it can choose new configurations randomly among neighborhoods of C, while satisfying step 2c. But when applied to other types of problems, it can lead to a deterministic search.[11] To overcome this, the algorithm was modified as follows: in step 2b the N variables x_i are ranked so that to the variable with the least fitness is assigned rank 1, and to the one with the best fitness rank N. Each time the algorithm passes through step 2c a variable is chosen to be mutated according to a probability distribution of the k ranks, given by:

$$P(k) = k^{-\tau}, \quad 1 \leq k \leq N, \tag{1}$$

where τ is a positive adjustable parameter. For $\tau \to 0$, the algorithm becomes a random walk, while for $\tau \to \infty$, we have a deterministic search. The introduction of the parameter τ, allows the algorithm to choose any variable to mutate, but privileging the ones with low fitness. This implementation of the EO method received the name τ-EO algorithm[11], and showed superior performance to the standard implementation even in cases where the basic EO algorithm would not lead to local minima.

As pointed out by Boettcher and Percus,[11] "a drawback of the EO method is that a general definition of fitness for the individual variables may prove ambiguous or even impossible". What means that for each new optimization problem assessed, a new way to assign the fitness to the design variables may have to be created. Moreover, to our knowledge it has been applied so far to combinatorial problems with no implementation to continuos functions. In order to make the EO method applicable to a broad class of design optimization problems, without concern to how the fitness of the design variables would be assigned and capable to tackle either continuos, discrete or integer variables, a generalization of the EO, called Generalized Extremal Optimization, was devised. In this new algorithm, the fitness assignment is not done directly to the design variables, but to a "population of species" that encodes the variables. Each species receives its fitness, and eventually mutates, following general rules. The GEO algorithm is described in the next Section.

THE GENERALIZED EXTREMAL OPTIMIZATION ALGORITHM

We devised the GEO algorithm using the same logic of the evolutionary model of Bak and Sneppen,[12] but applying the τ-EO approach to choose the species that will mutate. Following Bak and Sneppen,[12] L species are aligned and for each species is assigned a fitness value that will determine the species that are more prone to mutate. We can think of these species as bits that can assume the values of 0 or 1. Hence, the entire population would consist of a single binary string. The design variables of the optimization problem are encoded in this string that would be similar to a chromosome in a canonical GA, but with each bit considered as a species or individual, as shown in Figure 1.

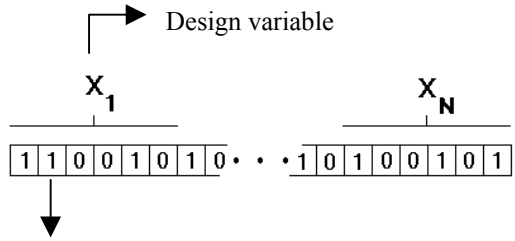


Figure 1 – Encoding of N design variables. In this example each design variable is represented by 6 bits.

To each species (bit) is assigned a fitness number that is proportional to the gain (or loss) the objective function value has in mutating (flipping) the bit. All bits are then ranked from rank 1, for the least adapted bit, to N for the best adapted. A bit is then chosen to mutate (flip) according to the probability distribution (1). This process is repeated until a given stopping criteria is reached and the best configuration of bits (the one that gives the best value for the objective function) found through the process is returned. In Figure 1

The practical implementation of the GEO algorithm to a function optimization problem is as follows:

1. Initialize randomly a binary string of length L that encodes N design variables of bit length l_j (j = 1, N). For the initial configuration C of bits, calculate the objective function value V and set C_{best} = C and V_{best} = V.

2. For each bit i of the string, at a given iteration:
   a) flip the bit (from 0 to 1 or 1 to 0) and calculate the objective function value V_i of the string configuration C_i,
   b) set the bit fitness as $\Delta V_i = (V_i - V_{best})$. It indicates the relative gain (or loss) that one has in mutating the bit, compared to the best objective function value found so far.
   c) return the bit to its original value.

3. Rank the bits according to their fitness values, from k = 1 for the least adapted bit to k = L for the best adapted. In a minimization problem, higher values of ΔV_i will have higher ranking, and otherwise for maximization problems. If two or more bits have the same fitness, rank them randomly.

4. Choose with equal probability a candidate bit i to mutate. Generate a random number RAN with uniform distribution in the range [0,1]. If the mutating probability $P_i(k)$ of the chosen bit is equal or greater than RAN the bit is confirmed to mutate. Otherwise, the process is repeated until a bit is confirmed to mutate.

5. For the bit i chosen to mutate set C = C_i and V = V_i.

6. If V < V_{best} (V > V_{best}, for a maximization problem) then set V_{best} = V and C_{best} = C.

7. Repeat steps 2 to 6 until a given stopping criteria is reached.

8. Return C_{best} and V_{best}.

Equality and inequality constraints can be easily incorporated to the algorithm simply setting a high (for a minimization problem) or low (for a maximization problem) fitness value to the bit that, when flipped, leads the configuration to an unfeasible region of the design space. Side constraints are directly applied through the encoding of the design variables. Note that the move to an infeasible region is not prohibited, since any bit has a chance to mutate according to the P(k) distribution. Moreover, no special condition is posed for the beginning of the search process, which can even start from an infeasible region.

A slightly different implementation of the GEO algorithm can be obtained, changing the way the bits are ranked and mutated. Instead of ranking all the bits according to steps 2-3, we can rank them separately for each variable. In this way the bits of each variable will have a rank ranging from 1 to l_j. In step 4 one bit of each variable is chosen to be flipped according to the probability distribution P(k). We will call this implementation hereinafter GEO_{var}. In the following Section the performance of the GEO algorithm is verified against a set of test functions.

RESULTS

The GEO algorithm and its variation GEO_{var} were applied to a set of test functions described in [13]. They are nonlinear, multimodal, multidimensional functions with variables bounded by side constraints. As in the GAs used in [13], each variable is encoded in 16 bits. All functions have one global optimum, where the value of the objective function is zero. As with any stochastic algorithm, the performance of GEO is influenced by its control parameter. In order to find the "best" value of τ applicable for each test function, we varied τ in the range [0.25,3.0] with steps of 0.25. For a given test function, the best value of τ was the one that lead to the best (minimal) value for the objective function, after a given number of function evaluations (NFE).

In Figures 2 to 6, the performance of the GEO algorithms for the set of test functions is shown together with the results for the GAs. All data points on the graphs below represent an average of 50 independent runs. The best objective function value found through the search is shown against the number of function evaluations.

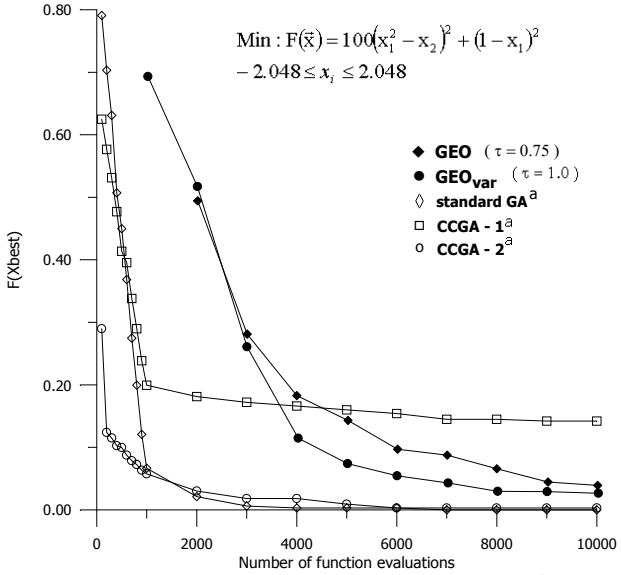


Figure 2 – Results for the Rosenbrock function. [a]From [13].

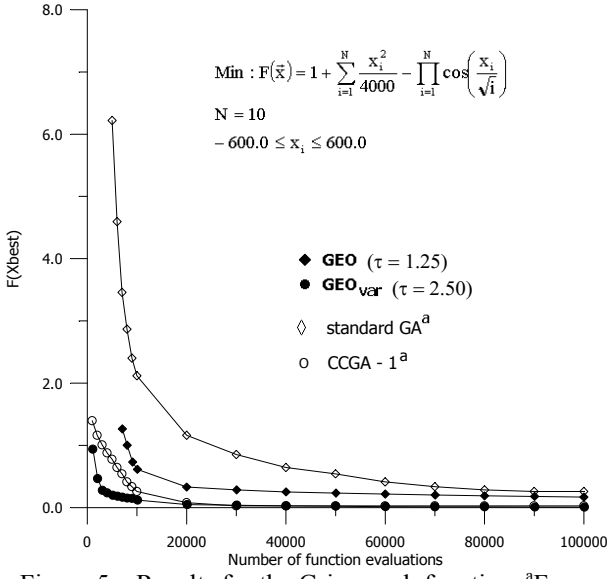


Figure 5 – Results for the Griewangk function. [a]From [13].

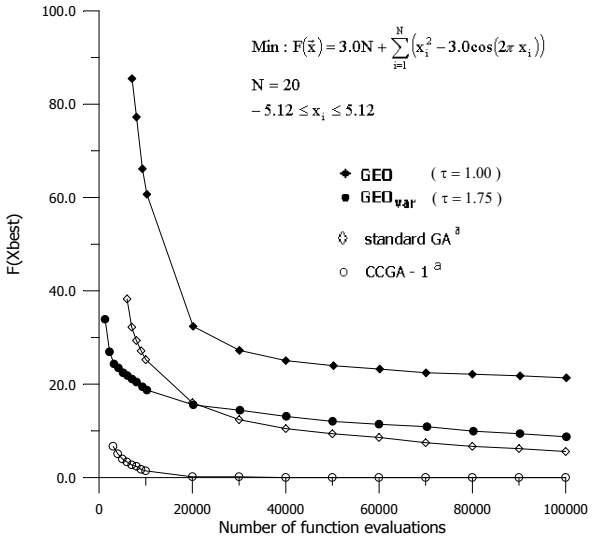


Figure 3 – Results for the Rastringin function. [a]From [13].

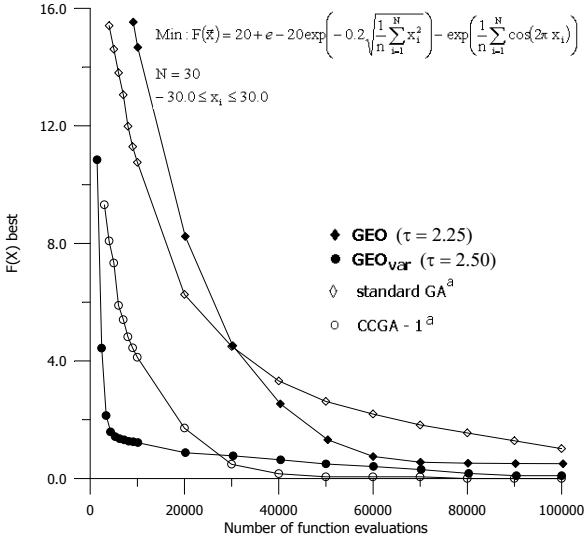


Figure 6 – Results for the Ackley function. [a]From [13].

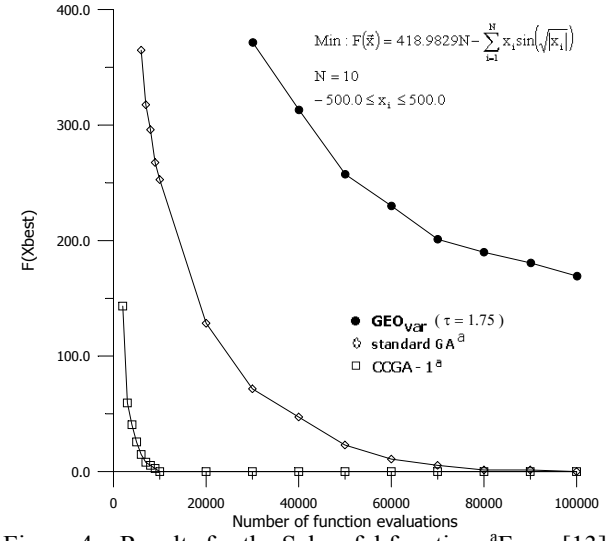


Figure 4 – Results for the Schwefel function. [a]From [13].

From the results shown throughout this Section, it can be seen that the GEO_{var} performed equally or better than the GEO for all functions. This indicates that, at each iteration, mutating one bit per variable may be advantageous compared to mutating only one bit for the whole string.

It can be also observed that, for a given test function, the value of τ that gave the best results was always lesser in the GEO algorithm than in the GEO_{var}. It must be also remarked, that the range where the "best" τ was found for both GEOs is not large, what means that the computational effort to "fine tune" τ is not really a burden for the method.

Finally, the results shown above indicate that the GEO can work successfully. Although it performed very poorly for the Schwefel function, when compared to the GAs, it was quite competitive for the other test functions, mainly when the variables were tackled simultaneously

(GEO_{var}). In fact, it must be remembered that does not exists a "best of all" optimization algorithm,[15] and it is not expected that the GEO algorithm would outperform all the other kinds of stochastic algorithms in all cases.

CONCLUSIONS

In this paper the Generalized Extremal Optimization algorithm was presented. Inspired by the theory of Self-Organized Criticality, it is an stochastic algorithm devised to tackle complex design optimization problems that presents such features as nonconvex design spaces or presence of different kinds of design variables. As an "a priori" advantage over other popular stochastic algorithms, it has only one adjustable parameter, and can be easily fine tuned to give its best performance on a given problem. Tested in a set of nonlinear, multimodal functions commonly used to assess the performance of stochastic algorithms, it showed to be a potential candidate to be incorporated into the designer's tool suitcase. Ongoing research is aimed at the study of the implementation of the GEO algorithm to constrained function optimization and its application to real inverse design problems.

ACKNOWLEDGEMENTS

F. M. Ramos acknowledges the support of CNPq-Brazil through the research grant 300171/97-8.

REFERENCES

1. Kirkpatrick, S., Gellat, C.D. and Vecchi, M.P., Optimization by Simulated Annealing, *Science*, Vol. 220, Number 4598, 1983, pp. 671-680.

2. Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, 1989.

3. Ahmed, Q., Krishnakumar, K. and Neidhoefer, J., Applications of evolutionary Algorithms to Aerospace Problems – A Survey, *Computational Methods in Applied Sciences'96*, John Wiley & Sons, 1996, pp. 237-242.

4. Schoonover, P.L., Crossley, W.A. and Heister, S.D., Application of a Genetic Algorithm to the Optimization of Hybrid Rockets, *Journal of Spacecraft and Rockets*, Vol. 37, No. 5, 2000, pp. 622-629.

5. Jones, B.R., Crossley, W.A. and Lyrintzis, A., Aerodynamic and Aeroacoustic Optimization of Rotocraft Airfoils via a Parallel Genetic Algorithm, *Journal of Aircraft*, Vol. 37, No. 6, 2000, pp. 1088-1096.

6. Wang, X. and Damodaran, M., Aerodynamic Shape Optimization Using Computational Fluid Dynamics and Parallel Simulated Annealing Algorithms, *AIAA Journal*, Vol. 39, No. 8, 2001, pp. 1500-1508.

7. Jilla, C.D. and Miller, D.W., Assessing the Performance of a Heuristic Simulated Annealing Algorithm for the Design of Distributed Satellite Systems, *Acta Astronautica*, Vol. 48, No. 5-12, 2001, pp. 529-543.

8. Vicini, A. and Quagliarella, D., Airfoil and Wing Design Through Hybrid Optimization Strategies, *AIAA Journal*, Vol. 37, No. 5, pp. 634-641, 1999.

9. Crain, T. Bishop, R.H. and Fowler, W., Interplanetary Flyby Mission Optimization Using a Hybrid Global-Local Search Method, *Journal of Spacecraft and Rockets*, Vol. 37, No. 4, pp. 468-474, 2000.

10. Desai, R. and Patil, R., SALO: Combining Simulated Annealing and Local Optimization for Efficient Global Optimization, *Los Alamos National Laboratory, TR LA-UR-95-2862*, Albuquerque, NM, 1995.

11. Boettcher, S. and Percus, A.G. Optimization with Extremal Dynamics. *Physical Review Letters*, Vol. 86, pp. 5211-5214, 2001.

12. Bak, P. and Sneppen, K. Punctuated Equilibrium and Criticality in a Simple Model of Evolution. *Physical Review Letters*, Vol. 71, Number 24, pp. 4083-4086, 1993.

13. Potter, A.P. and De Jong, K.A. A Cooperative Coevolutionary Approach to Function Optimization. *The Third Problem Solving From Nature*, Springer-Verlag, pp. 249-257, 1994.

14. Bak, P. *How Nature Works*, Copernicus, Springer-Verlag, 1999.

15. Wolpert, D.H. and Macready, W.G. No Free Lunch Theorems for Search, *Santa Fe Institute Technical Report*, SFI-TR-95-02-010, 1995.

COMPARISON OF VARIOUS METHODS FOR SOLVING THE CAUCHY PROBLEM IN LINEAR ELASTICITY

Liviu Marin, Lionel Elliott, Derek B. Ingham and Daniel Lesnic
Department of Applied Mathematics, University of Leeds,
Leeds, West Yorkshire, LS2 9JT,UK.
Emails: liviu@amsta.leeds.ac.uk, lionel@amsta.leeds.ac.uk,
amt6dbi@amsta.leeds.ac.uk, amt5ld@amsta.leeds.ac.uk

ABSTRACT

In the formulation of the Cauchy problem in linear elasticity the Lamé system of equations has to be solved subject to overspecified boundary conditions on both the displacement and traction vectors over a portion of the boundary of the solution domain, with the remaining portion of the boundary being underspecified. This classical Cauchy problem is ill-posed and direct inversion numerical techniques fail to produce a stable solution. Therefore, in this paper several boundary element regularization methods, such as iterative, conjugate gradient, Tikhonov regularization and singular value decomposition are developed and compared.

NOMENCLATURE

G	shear modulus
N, N_1, N_2	numbers of boundary elements
p	percentage of noise
$\boldsymbol{t}$	traction vector
$\widetilde{\boldsymbol{t}}$	specified boundary traction
$\boldsymbol{u}$	displacement vector
$\widetilde{\boldsymbol{u}}$	specified boundary displacement
$\|\|\cdot\|\|_2$	the Euclidean norm
$\|\|\cdot\|\|_{L^2}$	the norm of the space L^2

Greek Symbols

ε_{ij}	components of the strain tensor
Γ	the boundary of the domain Ω
Γ_1, Γ_2	parts of the boundary Γ
λ	regularisation parameter
ν	Poisson ratio
$\overline{\nu}$	equivalent Poisson ratio for plain strain/stress state
Ω	solution domain
σ	standard deviation
σ_{ij}	components of the stress tensor
θ	angular polar coordinate
ξ_i	singular values

Superscripts

(an)	analytical value
(k)	quantity at the k^{th} iteration
(num)	numerical value

1 INTRODUCTION

The Cauchy problem in elasticity has been studied by Yeih *et al.* [1] and Koya *et al.* [2], who have analysed its existence, uniqueness and continuous dependence on the data and have proposed a regularisation procedure, namely the fictitious boundary indirect method, based on the simple or double layer potential theory. Further, Marin *et al.* [3] have developed an alternating iterative algorithm which reduced the problem to solving a sequence of well-posed boundary value problems which were solved using the boundary element method (BEM), whilst [4-6] have used both the conjugate gradient method (CGM) and the Tikhonov regularisation method combined with the BEM.

The purpose of this paper is to describe and compare several boundary element regularisation methods, such as iterative, conjugate gradient, Tikhonov regularisation and singular value decomposition methods, for solving the Cauchy problem in isotropic linear elasticity. The regularisation is obtained by matching the number of iterations performed, the choice of the regular-

isation parameter, or the choice of the optimal truncation number to the level of the noise in the input data.

2 CAUCHY PROBLEM

Consider an isotropic linear elastic material which occupies a bounded domain $\Omega \subset \mathbb{R}^d$ with piecewise smooth boundary Γ. In the absence of body forces, the equilibrium equations for the displacement $\boldsymbol{u}$ are given by, [7],

$$\frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}(\boldsymbol{x})) = 0, \qquad \boldsymbol{x} \in \Omega, \quad i = \overline{1,d} \tag{1}$$

where σ_{ij} is the stress tensor, and the strain tensor ε_{ij} is given by

$$\varepsilon_{ij}(\boldsymbol{u}(\boldsymbol{x})) = \frac{1}{2}\left(\frac{\partial u_i(\boldsymbol{x})}{\partial x_j} + \frac{\partial u_j(\boldsymbol{x})}{\partial x_i}\right), \quad i,j = \overline{1,d}. \tag{2}$$

These tensors are related by the constitutive law

$$\sigma_{ij}(\boldsymbol{u}(\boldsymbol{x})) = 2G\varepsilon_{ij}(\boldsymbol{u}(\boldsymbol{x})) + \frac{2G\nu}{1-2\nu}\varepsilon_{kk}(\boldsymbol{u}(\boldsymbol{x}))\delta_{ij}, \quad i,j = \overline{1,d}, \tag{3}$$

with G and ν the shear modulus and Poisson ratio, respectively, and δ_{ij} the Kronecker delta tensor. We now let $\boldsymbol{n}(\boldsymbol{x})$ be the outward normal vector at Γ and $\boldsymbol{t}(\boldsymbol{x})$ be the traction vector at a point $\boldsymbol{x} \in \Gamma$ whose components are defined by

$$t_i(\boldsymbol{x}) = \sigma_{ij}(\boldsymbol{x})\, n_j(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma, \quad i = \overline{1,d}. \tag{4}$$

In the direct problem formulation, the knowledge of either the displacement or traction vectors on the whole boundary Γ gives the corresponding Dirichlet, Neumann, or mixed boundary conditions which enables us to determine the displacement vector in the domain Ω. Then, the strain tensor ε_{ij} can be calculated from (2) and the stress tensor is determined using (3). In contrast, in the inverse problem formulation both the displacement and traction vectors are specified on a part of the boundary Γ, say Γ_2, namely,

$$u_i(\boldsymbol{x}) = \widetilde{u}_i(\boldsymbol{x}), \; t_i(\boldsymbol{x}) = \widetilde{t}_i(\boldsymbol{x}), \; \boldsymbol{x} \in \Gamma_2, \; i = \overline{1,d}, \tag{5}$$

where $\widetilde{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{t}}$ are prescribed vector valued functions. In the above formulation of the boundary conditions (5), it can be seen that the boundary Γ_2 is overspecified by prescribing both the displacement $\boldsymbol{u}_{|\Gamma_2} = \widetilde{\boldsymbol{u}}$ and the traction $\boldsymbol{t}_{|\Gamma_2} = \widetilde{\boldsymbol{t}}$ vectors, whilst the boundary $\Gamma_1 = \Gamma - \Gamma_2$ is underspecified since both the displacement $\boldsymbol{u}_{|\Gamma_1}$ and the traction $\boldsymbol{t}_{|\Gamma_1}$ vectors are unknown and have to be determined. The problem given by equations (1) and (5), called the Cauchy problem, is much more difficult to solve both analytically and numerically than the direct problem, since the solution does not satisfy the general conditions of well-posedness. Although the problem may have a unique solution, it is well known that this solution is unstable with respect to small perturbations in the data on Γ_2 and therefore regularisation methods are required.

3 BEM

The Lamé system (1) in the two-dimensional case, i.e. $d = 2$, can be formulated in integral form, [8], namely,

$$C_{ij}(\boldsymbol{x})u_j(\boldsymbol{x}) + \int\!\!\!\!\!\!-_{\Gamma} T_{ij}(\boldsymbol{y},\boldsymbol{x})u_j(\boldsymbol{y})\, d\Gamma = \int_{\Gamma} U_{ij}(\boldsymbol{y},\boldsymbol{x})t_j(\boldsymbol{y})\, d\Gamma \tag{6}$$

for $i,j = 1,2$, $\boldsymbol{x} \in \overline{\Omega} = \Omega \cup \Gamma$, where the first integral is taken in the sense of the Cauchy principal value, $C_{ij}(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in \Omega$ and $C_{ij}(\boldsymbol{x}) = 1/2$ for $\boldsymbol{x} \in \Gamma$, and U_{ij} and T_{ij} are the fundamental displacements and tractions for the two-dimensional isotropic linear elasticity given by

$$\begin{aligned} U_{ij}(\boldsymbol{y},\boldsymbol{x}) &= C_1 C_2 \delta_{ij} \ln r(\boldsymbol{y},\boldsymbol{x}) \\ &\quad - C_1 \frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_i}\frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_j} \\ T_{ij}(\boldsymbol{y},\boldsymbol{x}) &= \frac{C_3}{r(\boldsymbol{y},\boldsymbol{x})}\left[\left(C_4\delta_{ij} \right.\right. \\ &\quad \left. + 2\frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_i}\frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_j}\right)\frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial n(\boldsymbol{y})} \\ &\quad \left. - C_4\left(\frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_i}n_j(\boldsymbol{y}) - \frac{\partial r(\boldsymbol{y},\boldsymbol{x})}{\partial y_j}n_i(\boldsymbol{y})\right)\right]. \end{aligned} \tag{7}$$

Here $r(\boldsymbol{y},\boldsymbol{x})$ represents the distance between the node/collocation point $\boldsymbol{x}$ and the field point $\boldsymbol{y}$ and the constants $C_1, \cdots, C_4$ are given by

$$\begin{aligned} C_1 &= -1/[8\pi G(1-\overline{\nu})], \quad C_2 = 3 - 4\overline{\nu}, \\ C_3 &= -1/[4\pi(1-\overline{\nu})], \quad C_4 = 1 - 2\overline{\nu}, \end{aligned} \tag{8}$$

where $\overline{\nu} = \nu$ for the plane strain state and $\overline{\nu} = \nu/(1+\nu)$ for the plane stress state.

A BEM with continuous linear boundary elements, [8], is employed in order to solve the Cauchy problem in linear elasticity by using the regularisation methods described in the next section. If the boundaries Γ_1 and Γ_2 are discretised into N_1 and N_2 continuous linear boundary elements, respectively, such that $N = N_1 + N_2$, then on applying the boundary integral equation (6) and the boundary conditions (5) at each node $x \in \Gamma$, we arrive at a system of $2N$ linear algebraic equations with $4N_1$ unknowns which can be generically written as

$$\mathbb{C}\boldsymbol{X} = \boldsymbol{F} \tag{9}$$

where the vector $\boldsymbol{F}$ is computed using the boundary conditions (5), the matrix $\mathbb{C}$ depends only on the material properties and the geometry of the boundary Γ and the vector $\boldsymbol{X}$ contains the unknown values of the displacements and the tractions on the boundary Γ_1.

4 REGULARISATION

4.1 Singular value decomposition

Consider the ill-conditioned system of linear algebraic equations (9), where $\mathbb{C} \in \mathbb{R}^{2N\times 4N_1}$, $\boldsymbol{X} \in \mathbb{R}^{4N_1}$, $\boldsymbol{F} \in \mathbb{R}^{2N}$ and assume for the moment that $N \geq 2N_1$. Then the singular value decomposition (SVD) of the matrix $\mathbb{C}$ is a decomposition of the form

$$\mathbb{C} = \mathbb{W}\mathbb{X}\mathbb{V}^T = \sum_{i=1}^{4N_1} \boldsymbol{w}_i \xi_i \boldsymbol{v}_i^T \tag{10}$$

where $\mathbb{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{4N_1}) \in \mathbb{R}^{2N\times 4N_1}$ and $\mathbb{V} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{4N_1}) \in \mathbb{R}^{4N_1\times 4N_1}$ are matrices with orthonormal columns, i.e. $\mathbb{W}^T\mathbb{W} = \mathbb{V}^T\mathbb{V} = \mathbb{I}_{4N_1}$, and $\mathbb{X} = \mathrm{diag}(\xi_1, \cdots, \xi_{4N_1})$ is a diagonal matrix with nonnegative diagonal elements appearing in the order

$$\xi_1 \geq \xi_2 \geq \cdots \geq \xi_{4N_1} \geq 0. \tag{11}$$

The numbers ξ_i are called the singular values of the matrix $\mathbb{C}$, whilst the vectors $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ are the left and the right singular vectors of the matrix $\mathbb{C}$, respectively. The SVD (10) can be defined for any N_1 and N_2 since, if $N_2 < N_1$, we can simply apply the SVD (10) to the matrix $\mathbb{C}^T$. In the ideal setting, without perturbations and rounding errrors, the treatment of the ill-conditioned system of equations (9) is straightforward, namely we simply ignore the SVD components associated with the zero singular values and compute the solution

$$\boldsymbol{X} = \sum_{i=1}^{rank(\mathbb{C})} \frac{\boldsymbol{w}_i^T \boldsymbol{F}}{\xi_i} \boldsymbol{v}_i. \tag{12}$$

In practice, $\mathbb{C}$ is never exactly mathematically rank deficient, but instead numerically rank deficient, i.e. it has one or more small nonzero singular values ξ_i for some i greater than some $r_\epsilon < \mathrm{rank}(\mathbb{C}) = 4N_1$. The small singular values inevitably give rise to difficulties and the solution $\boldsymbol{X}$ is dominated by the last $(4N_1 - r_\epsilon)$ components of equation (12).

The most common approach to regularise numerically rank deficient problems is to consider the given matrix $\mathbb{C}$ as a noisy representation of a mathematically rank deficient matrix and to replace $\mathbb{C}$ by a matrix that is close to $\mathbb{C}$ and mathematically rank deficient. The standard choice is the rank-n matrix $\mathbb{C}_n$ defined as

$$\mathbb{C}_n = \sum_{i=1}^{n} \boldsymbol{w}_i \xi_i \boldsymbol{v}_i^T \tag{13}$$

i.e. we replace the small nonzero singular values $\xi_{n+1}, \cdots, \xi_{4N_1}$ with exact zeros and it is referred to as the truncated SVD (TSVD) solution. The optimal truncation number n is chosen according to the discrepancy principle, [9], i.e. we choose the first n such that

$$\|\mathbb{C}_n \boldsymbol{X} - \boldsymbol{F}\|_2 \leq \varepsilon \tag{14}$$

where ε is a measure of the perturbations in the matrix $\mathbb{C}$ and in the Cauchy data $(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{t}})$ on the boundary Γ_2 and of the incompatibility of the exact solution of the system of equations (9).

4.2 Tikhonov regularisation

Consider again the ill-conditioned system of linear algebraic equations (9) whose Tikhonov regularised solution of zeroth-order is given by

$$\boldsymbol{X}_\lambda : T_\lambda(\boldsymbol{X}_\lambda) = \min_{\boldsymbol{X}\in\mathbb{R}^{4N_1}} T_\lambda(\boldsymbol{X}) \tag{15}$$

where T_λ represents the Tikhonov functional given by

$$T_\lambda(\boldsymbol{X}) = \|\mathbb{C}\boldsymbol{X} - \boldsymbol{F}\|_2^2 + \lambda\|\boldsymbol{X}\|_2^2 \tag{16}$$

where $\lambda > 0$ the regularisation parameter to be chosen. Formally, the Tikhonov regularised solution $\boldsymbol{X}_\lambda$ of the problem (15) is given as the solution of the regularised equation

$$\left(\mathbb{C}^T\mathbb{C} + \lambda\mathbb{I}\right)\boldsymbol{X}_\lambda = \mathbb{C}^T\boldsymbol{F}. \qquad (17)$$

Regularisation is necessary when solving inverse problems because the simple least squares solution, i.e. $\lambda = 0$, is dominated by contributions from data errors and computer rounding errors. By adding regularisation we are able to damp out these contributions and maintain the norm $\|\boldsymbol{X}\|_2$ to be of reasonable size. If too much regularisation, or damping, i.e. λ is large, is imposed on the solution then it will not fit the given data $\boldsymbol{F}$ properly and the residual norm $\|\mathbb{C}\boldsymbol{X} - \boldsymbol{F}\|_2$ will be too large. If too little regularisation is imposed on the solution, i.e. λ is small, then the fit will be good, but the solution will be dominated by the contributions from the data errors, and hence $\|\boldsymbol{X}\|_2$ will be too large. It is quite natural to plot the norm of the solution as a function of the norm of the residual, i.e. $(\|\mathbb{C}\boldsymbol{X}_\lambda - \boldsymbol{F}\|_2, \|\boldsymbol{X}_\lambda\|_2)$, parametrised by the regularisation parameter λ. This plot results in general in an L-curve graph which is really a trade-off curve between two quantities that both should be controlled. The optimal value of the regularisation parameter λ is obtained as the maximum point of the curvature of the L-curve, [10].

As with every practical method, the L-curve has its advantages and disadvantages. There are two main disadvantages or limitations of the L-curve criterion. The first disadvantage is concerned with the reconstruction of very smooth exact solutions, [11, pp.193-197,12]. For such solutions, [13] showed that the L-curve criterion will fail, and the smoother the solution, the worse the regularisation parameter λ computed by the L-curve criterion. However, it is not clear how often very smooth solutions arise in applications. The second limitation of the L-curve criterion is related to its asymptotic behaviour as the problem size $4N_1$ increases. As pointed out in [14], the regularisation parameter λ computed by the L-curve criterion may not behave consistently with the optimal parameter λ_{opt} as $4N_1$ increases. However, this ideal situation in which the same problem is discretised for increasing $4N_1$ may not arise so often in practice. Often the problem size $4N_1$ is fixed by the particular measurement setup, and if a larger $4N_1$ is required then a new experiment must be made. Apart from these two limitations, the advantages of the L-curve criterion are its robustness and ability to treat perturbations consisting of correlated noise, [10]. However, as commented by the referee, the so called "L-curve method" is not a regularizing algorithm and cannot be used for the solution of ill-posed continuous problems, but it can be used for numerical experiments dealing with discrete ill-conditioned systems of linear equations such as (17).

4.3 Conjugate gradient method

Since the boundary conditions on the boundary Γ_1 are to be determined, we consider the boundary displacement on the underspecified boundary Γ_1 as a control $\boldsymbol{v} \in L^2(\Gamma_1) \times L^2(\Gamma_1)$ in a direct problem formulation to fit the Cauchy data $\widetilde{\boldsymbol{u}} \in L^2(\Gamma_2) \times L^2(\Gamma_2)$ on the overspecified boundary Γ_2. Thus we consider the direct problem

$$\begin{cases} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}(\boldsymbol{x})) = 0, & \boldsymbol{x} \in \Omega \\ u_i(\boldsymbol{x}) = v_i(\boldsymbol{x}), & \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{u}(\boldsymbol{x}))n_j(\boldsymbol{x}) = \widetilde{t}_i(\boldsymbol{x}), & \boldsymbol{x} \in \Gamma_2 \end{cases} \qquad (18)$$

with $\widetilde{\boldsymbol{t}} \in L^2(\Gamma_2) \times L^2(\Gamma_2)$. Assuming that Γ is a smooth boundary consisting of two non-intersecting closed curves Γ_1 and Γ_2, we note that there is a unique solution $\boldsymbol{u}(\boldsymbol{v}, \widetilde{\boldsymbol{t}}) \in H^{1/2}(\Omega) \times H^{1/2}(\Omega)$ of the direct problem (18), [15]. Then we aim to find $\boldsymbol{v}$ such that

$$A\boldsymbol{v} := \boldsymbol{u}(\boldsymbol{v}, \widetilde{\boldsymbol{t}})_{|\Gamma_2} = \widetilde{\boldsymbol{u}}. \qquad (19)$$

To do so, we minimise the functional

$$J(\boldsymbol{v}) = \frac{1}{2}\|A\boldsymbol{v} - \widetilde{\boldsymbol{u}}\|_{L^2(\Gamma_2)\times L^2(\Gamma_2)} \qquad (20)$$

with respect to $\boldsymbol{v} \in L^2(\Gamma_1) \times L^2(\Gamma_1)$. The functional (20) is twice Fréchet differentiable and its first gradient has the form, [5],

$$J'(\boldsymbol{v}) = -\left(\sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{x}))n_j(\boldsymbol{x})\right)_{|\Gamma_1} \qquad (21)$$

where $\boldsymbol{\psi}$ is the solution of the adjoint problem

$$\begin{aligned} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{x})) = 0,\ \boldsymbol{x} \in \Omega \\ \psi_i(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{x}))n_j(\boldsymbol{x}) \\ = u_i(\boldsymbol{v}, \widetilde{\boldsymbol{t}})(\boldsymbol{x}) - \widetilde{u}_i(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_2. \end{aligned} \qquad (22)$$

Thus the CGM applied to our problem has the form of the following algorithm:

Step 1. Set $k = 0$. Choose $\boldsymbol{u}^{(0)} \in L^2(\Gamma_2) \times L^2(\Gamma_2)$.

Step 2. Solve the direct problem

$$\begin{cases} \dfrac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}(\boldsymbol{x})) = 0, & \boldsymbol{x} \in \Omega \\ u_i(\boldsymbol{x}) = u_i^{(k)}(\boldsymbol{x}), & \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{u}(\boldsymbol{x}))n_j(\boldsymbol{x}) = \widetilde{t}_i(\boldsymbol{x}), & \boldsymbol{x} \in \Gamma_2 \end{cases} \tag{23}$$

to determine the residual $\boldsymbol{r}^{(k)}$

$$\boldsymbol{r}^{(k)} = A\boldsymbol{u}^{(k)} - \widetilde{\boldsymbol{u}} = \boldsymbol{u}(\boldsymbol{u}^{(k)}, \widetilde{\boldsymbol{t}})_{|\Gamma_2} - \widetilde{\boldsymbol{u}}. \tag{24}$$

Step 3. Solve the adjoint problem

$$\begin{aligned} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{x})) = 0,\ \boldsymbol{x} \in \Omega \\ \psi_i(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{x}))n_j(\boldsymbol{x}) = r_i^{(k)}(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_2, \end{aligned} \tag{25}$$

to determine the gradient $\boldsymbol{g}^{(k)}$

$$g_i^{(k)}(\boldsymbol{x}) = \sigma_{ij}(\boldsymbol{\psi}(\boldsymbol{0}, \boldsymbol{r}^{(k)})(\boldsymbol{x}))n_j(\boldsymbol{x})_{|\Gamma_1}. \tag{26}$$

Calculate β_k and $\boldsymbol{d}^{(k)}$ as follows:

$$\begin{aligned} k = 0: \quad \beta_k = 0, \quad \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)} \\ k \geq 1: \quad \beta_k = \frac{\|\boldsymbol{g}^{(k)}\|^2_{L^2(\Gamma_1)\times L^2(\Gamma_1)}}{\|\boldsymbol{g}^{(k-1)}\|^2_{L^2(\Gamma_1)\times L^2(\Gamma_1)}}, \\ \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)} + \beta_k \boldsymbol{d}^{(k-1)}. \end{aligned} \tag{27}$$

Step 4. Solve the direct problem

$$\begin{cases} \dfrac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}(\boldsymbol{x})) = 0, & \boldsymbol{x} \in \Omega \\ u_i(\boldsymbol{x}) = d_i^{(k)}(\boldsymbol{x}), & \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{u}(\boldsymbol{x}))n_j(\boldsymbol{x}) = 0, & \boldsymbol{x} \in \Gamma_2 \end{cases} \tag{28}$$

to determine $A_0\boldsymbol{d}^{(k)}$ given by

$$A_0\boldsymbol{d}^{(k)} = \boldsymbol{u}(\boldsymbol{d}^{(k)}, \boldsymbol{0})_{|\Gamma_2}. \tag{29}$$

Compute α_k and $\boldsymbol{u}^{(k+1)}$ as

$$\begin{aligned} \alpha_k = \frac{\|\boldsymbol{g}^{(k)}\|^2_{L^2(\Gamma_1)\times L^2(\Gamma_1)}}{\|A_0\boldsymbol{d}^{(k)}\|^2_{L^2(\Gamma_2)\times L^2(\Gamma_2)}}, \\ \boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}. \end{aligned} \tag{30}$$

Step 5. Set $k = k + 1$. Repeat from **Step 2** until a stopping criterion is prescribed.

As a stopping criterion we choose the one suggested by Nemirovskii [16], namely we choose the first $k \in \mathbb{N}$ such that

$$\|\boldsymbol{r}^{(k)}\|_{L^2(\Gamma_2)\times L^2(\Gamma_2)} \leq \delta\varepsilon \tag{31}$$

where ε is a measure of the errors of the Cauchy data on Γ_2 and $\delta > 1$ is a constant which can be taken heuristically to be 1.1, as suggested by Hanke and Hansen [17].

4.4 Alternating iterative method

The alternating iterative algorithm, which was proposed by Kozlov *et al.* [18], consists of the following steps:

Step 1. Specify an initial approximation $\boldsymbol{t}^{(0)}(\boldsymbol{x}) = (t_1^{(0)}(\boldsymbol{x}), t_2^{(0)}(\boldsymbol{x}))$ for the tractions on Γ_1 and solve the well-posed mixed boundary value problem

$$\begin{aligned} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}^{(0)}(\boldsymbol{x})) = 0,\ \boldsymbol{x} \in \Omega \\ \sigma_{ij}(\boldsymbol{u}^{(0)}(\boldsymbol{x}))\ n_j(\boldsymbol{x}) = t_i^{(0)}(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_1 \\ u_i^{(0)}(\boldsymbol{x}) = \widetilde{u}_i(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_2 \end{aligned} \tag{32}$$

in order to determine $\boldsymbol{u}^{(0)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega$ and $\boldsymbol{u}^{(0)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \Gamma_1$.

Step 2. Having constructed the approximation $\boldsymbol{u}^{(2k)}$, $k \geq 0$, the well-posed mixed boundary value problem

$$\begin{aligned} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}^{(2k+1)}(\boldsymbol{x})) = 0,\ \boldsymbol{x} \in \Omega \\ u_i^{(2k+1)}(\boldsymbol{x}) = u_i^{(2k)}(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_1 \\ \sigma_{ij}(\boldsymbol{u}^{(2k+1)}(\boldsymbol{x}))\ n_j(\boldsymbol{x}) = \widetilde{t}_i(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_2 \end{aligned} \tag{33}$$

is solved to determine $\boldsymbol{u}^{(2k+1)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega$ and $\boldsymbol{t}^{(2k+1)}(\boldsymbol{x}) = \sigma_{ij}(\boldsymbol{u}^{(2k+1)}(\boldsymbol{x}))\ n_j(\boldsymbol{x})$ for $\boldsymbol{x} \in \Gamma_1$.

Step 3. Having constructed the vector-valued function $\boldsymbol{u}^{(2k+1)}$, $k \geq 0$, the well-posed mixed boundary value problem

$$\begin{aligned} \frac{\partial}{\partial x_j}\sigma_{ij}(\boldsymbol{u}^{(2k+2)}(\boldsymbol{x})) = 0,\ \boldsymbol{x} \in \Omega \\ \sigma_{ij}(\boldsymbol{u}^{(2k+2)}(\boldsymbol{x}))\ n_j(\boldsymbol{x}) = t_i^{(2k+1)}(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_1 \\ u_i^{(2k+2)}(\boldsymbol{x}) = \widetilde{u}_i(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma_2 \end{aligned} \tag{34}$$

is solved in order to determine $\boldsymbol{u}^{(2k+2)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \Omega$ and $\boldsymbol{u}^{(2k+2)}(\boldsymbol{x})$ for $\boldsymbol{x} \in \Gamma_1$.

Step 4. Repeat from **Step 2** until a prescribed

stopping criterion is satisfied.

Kozlov *et al.* [18] showed that if Γ is smooth, $\widetilde{\boldsymbol{u}} \in H^{1/2}(\Gamma_2) \times H^{1/2}(\Gamma_2)$ and $\widetilde{\boldsymbol{t}} \in \left(H^{1/2}(\Gamma_2) \times H^{1/2}(\Gamma_2) \right)^*$, then the alternating algorithm based on steps $1-4$ produces two sequences of approximate solutions $\{\boldsymbol{u}^{(2k)}(\boldsymbol{x})\}_{k\geq 0}$ and $\{\boldsymbol{u}^{(2k+1)}(\boldsymbol{x})\}_{k\geq 0}$ which both converge in $H^1(\Omega) \times H^1(\Omega)$ to the solution $\boldsymbol{u}(\boldsymbol{x})$ of the problem (1) and (5) for any initial guess $\boldsymbol{t}^{(0)} \in \left(H^{1/2}(\Gamma_1) \times H^{1/2}(\Gamma_1) \right)^*$.

As a stopping criterion we use again the discrepancy principle which ceases the iterative procedure as in (14).

5 COMPARISON OF THE METHODS

It is the purpose of this section to present and compare the numerical results for the Cauchy problem considered in this study which have been obtained using the four regularisation methods described in Section 4. In order to present the performances of the numerical methods proposed, we solve the Cauchy problem for a typical benchmark test example in a two-dimensional smooth geometry, namely the unit disc $\Omega = \{\boldsymbol{x} = (x_1, x_2) \mid x_1^2 + x_2^2 < 1\}$. We assume that the boundary Γ of the solution domain is divided into two parts, namely $\Gamma_1 = \{\boldsymbol{x} = (x_1, x_2) \mid \boldsymbol{x} \in \Gamma,\ \alpha_1 < \theta(\boldsymbol{x}) < \alpha_2\}$ and $\Gamma_2 = \{\boldsymbol{x} = (x_1, x_2) \mid \boldsymbol{x} \in \Gamma,\ 0 \leq \theta(\boldsymbol{x}) \leq \alpha_1\} \cup \{\boldsymbol{x} = (x_1, x_2) \mid \boldsymbol{x} \in \Gamma,\ \alpha_2 \leq \theta(\boldsymbol{x}) \leq 2\pi\}$, where $\theta(\boldsymbol{x})$ is the angular polar coordinate of $\boldsymbol{x}$ and $\alpha_1 = \pi/4$ and $\alpha_2 = 3\pi/4$. We consider an isotropic linear elastic medium characterized by the material constants $G = 3.35 \times 10^{10}$ N/m^2 and $\nu = 0.34$ corresponding to a copper alloy. The following analytical solution for the displacement and stress

$$u_i^{(an)}(x_1, x_2) = \frac{1-\nu}{2G(1+\nu)}\ \sigma_0\ x_i,$$
$$\sigma_{ij}^{(an)}(x_1, x_2) = \sigma_0\ \delta_{ij} \qquad (35)$$

is considered in the domain Ω, where $\sigma_0 = 1.5 \times 10^{10}$ N/m^2.

In order to investigate the stability and the regularisation properties of the numerical methods considered, the boundary data $\widetilde{\boldsymbol{u}}_{|\Gamma_2}$ has been perturbed as $\widetilde{u}_i^{\varepsilon}|_{\Gamma_2} = \widetilde{u}_i|_{\Gamma_2} + \delta\widetilde{u}_i$, where $\delta\widetilde{u}_i$ is a Gaussian random variable with mean zero and standard deviation $\sigma_i = \max_{\Gamma_2} |\widetilde{u}_i|\ p/100$, and p is the percentage of additive noise included in the input data $\widetilde{\boldsymbol{u}}_{|\Gamma_2}$ in order to simulate the inherent measurements errors. The numerical results presented in this section have been obtained using $N = 80$ and $N_2 = 3N_1 = 60$ continuous linear boundary elements. These values were found to be sufficiently large such that any further refinement of the mesh size did not significantly improve the accuracy of the results.

As a stopping criterion we have used the discrepancy principle (14) for the SVD and the L-curve method for the Tikhonov regularisation.

For the alternating iterative algorithm, a variable relaxation factor, with respect to the angular polar coordinate $\theta(\boldsymbol{x})$ given by, [3],

$$\widetilde{\rho}(\theta(\boldsymbol{x})) = a \sin \pi \left(\frac{\theta(\boldsymbol{x}) - \alpha_1}{\alpha_2 - \alpha_1} \right) \qquad (36)$$

where $a \in (0, 2]$, when passing from the step 2 to the step 3, is employed in order to improve its rate of convergence. As a stopping criterion we have used the discrepancy principle for the alternating iterative method and the stopping rule (31) for the CGM.

In order to compare the four regularisation methods considered, Figures 1(a) and (b) present on the same figures the numerical solution for the x_2 component of the displacement and the traction vectors, respectively, on the boundary Γ_1 obtained with each of these methods for $p = 2\%$ noise added into the displacement data $\widetilde{\boldsymbol{u}}_{|\Gamma_2}$. It can be seen from these figures that the most accurate solution is the one given by the alternating iterative algorithm. Both the SVD and the Tikhonov regularisation methods give reasonably good approximations for the displacement and the traction vectors on the underspecified boundary Γ_1, with the mention that the SVD solutions are more accurate. The numerical solution obtained by the CGM is poor in comparison with the numerical solutions obtained by the other methods described in this paper. However, for less severe examples and for which we have a better initial guess, it was found that the CGM also produces numerical solutions almost as accurate as the numerical solutions obtained by the Tikhonov regularisation method. The differences between the regularisation methods considered are even larger in the

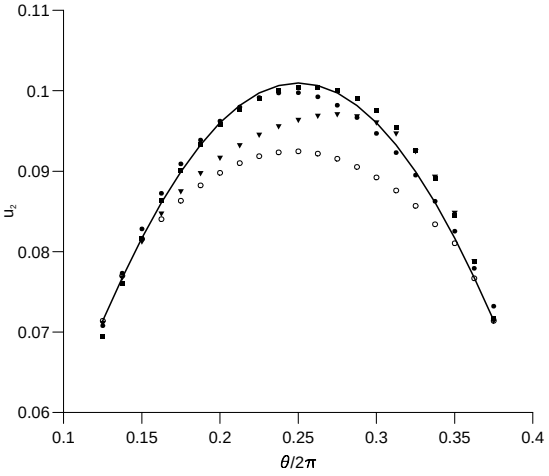
case of the numerical solution for the traction vector, as can be seen from Figure 1(b). In Table 1 we present the accuracy errors

$$e_u = ||\boldsymbol{u}^{(num)} - \boldsymbol{u}^{(an)}||_{L^2(\Gamma_1)\times L^2(\Gamma_1)},$$
$$e_t = ||\boldsymbol{t}^{(num)} - \boldsymbol{t}^{(an)}||_{L^2(\Gamma_1)\times L^2(\Gamma_1)} \qquad (37)$$

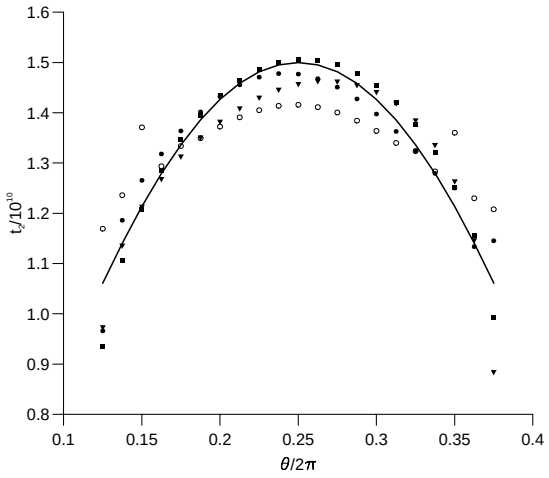
where $\boldsymbol{u}^{(an)}$ and $\boldsymbol{t}^{(an)}$ are the analytical displacement and traction vectors and $\boldsymbol{u}^{(num)}$ and $\boldsymbol{t}^{(num)}$ are the numerical displacement and traction vectors, respectively, obtained using the regularisation methods presented in Section 4 on the underspecified boundary Γ_1 for different levels of noise added into the input data. From this table it can be seen that the alternating iterative algorithm is the regularisation method which provides the most accurate numerical results, the SVD and the Tikhonov regularisation method produce reasonably good numerical approximation for both the displacements and the tractions, whilst the CGM produces less accurate numerical results.

Table 1: The accuracy errors e_u and e_t given by equation (37), obtained using the four regularisation methods described in Section 4 for various levels of noise added into the input data $\boldsymbol{u}^{(an)}|_{\Gamma_2}$, namely $p \in \{0, 1, 2\}$.

p	0%	1%	2%
e_u(CGM) $\times 10^3$	5.25	6.13	7.53
e_u(Tikhonov) $\times 10^3$	0.82	5.55	7.39
e_u(SVD) $\times 10^3$	0.73	3.81	3.96
e_u(Kozlov *et al.*) $\times 10^3$	0.57	1.95	3.02
e_t(CGM) $\times 10$	0.93	1.27	1.46
e_t(Tikhonov) $\times 10$	0.21	1.15	1.38
e_t(SVD) $\times 10$	0.20	0.98	0.99
e_t(Kozlov *et al.*) $\times 10$	0.19	0.45	0.64



(a)



(b)

Figure 1. (a) The analytical $u_2^{(an)}$ (—) and the numerical $u_2^{(num)}$ displacements, and (b) the analytical $t_2^{(an)}$ (—) and the numerical $t_2^{(num)}$ tractions, retrieved on the underspecified boundary Γ_1 by using various regularisation methods, namely, the alternating iterative algorithm (•), the SVD (■), the Tikhonov regularisation method (▼) and the CGM (○), for $p = 2\%$ noise.

6 CONCLUSIONS

In this paper four regularisation methods for the Cauchy problem in isotropic linear elasticity have been investigated. Three of the methods are general regularisation methods, whilst the fourth one is an alternating iterative algorithm developed for Cauchy problems for self-adjoint and positive-definite operators. It was found that the Cauchy problem in linear elasticity can be regularised by any of the methods considered since all of them produced a stable

numerical solution. However, the numerical solutions obtained by these methods differ in terms of accuracy. It has been found that the SVD method outperforms the Tikhonov regularisation method, whilst the latter method outperforms the CGM. However, all these three methods are second best compared to the alternating iterative algorithm. We note that the CGM is less accurate than the other methods considered. A possible reason for this is that in the CGM, the boundaries Γ_1 and Γ_2 should be disjoint non-intersecting closed curves and this is not the case for our test example. Overall, it can be concluded that the Cauchy problem in isotropic linear elasticity can be regularised by various methods, such as the regularisation methods presented in this paper, but more accurate results are obtained by particular methods which take into account the particular structure of the problem, such as the alternating iterative algorithm.

REFERENCES

1. W.C. Yeih, T. Koya and T. Mura, An inverse problem in elasticity with partially overspecified boundary conditions. I. Theoretical approach, *Trans. ASME J. Appl. Mech.*, **60**, 595 (1993).

2. T. Koya, W.C. Yeih and T. Mura, An inverse problem in elasticity with partially overspecified boundary conditions. II. Numerical details, *Trans. ASME J. Appl. Mech.*, **60**, 601 (1993).

3. L. Marin, L. Elliott, D.B. Ingham and D. Lesnic, Boundary element method for the Cauchy problem in linear elasticity, *Eng. Anal. Boundary Elements*, **25**, 783 (2001).

4. C.H. Huang and W.Y. Shih, A boundary element based solution of an inverse elasticity problem by conjugate gradient and regularization method, in *Proc. 7th Int. Offshore Polar Eng. Conf.*, Honolulu, USA, 1997, 338-395.

5. L. Marin, Dinh Nho Háo and D. Lesnic, Conjugate gradient-boundary element method for a Cauchy problem in the Lamé system, in *BETECH XIV*, (Brebbia, C.A. and Kassab, A.J., editors), WIT Press, Southampton, UK, 2001, 229-238.

6. L. Marin, L. Elliott, D.B. Ingham, and D. Lesnic, Boundary element based solution for the Cauchy problem in linear elasticity by the Tikhonov regularization method, in *The 3rd UK Conference on Boundary Integral Methods*, (Harris, P.J., editor), University of Brighton Press, Brighton, UK, 2001, 149-158.

7. L.D. Landau and E.M. Lifshits, *Theory of Elasticity*, Pergamon Press, Oxford, 1986.

8. C.A. Brebbia, J.F.C. Telles and L.C. Wrobel, *Boundary Element Techniques*, Springer Verlag, Berlin, 1984.

9. P.C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.

10. P.C. Hansen, The L-curve and its use in the numerical treatment of inverse problems, in *Computational Inverse Problems in Electrocardiology*, (Johnston, P., editor), WIT Press, Southampton, UK, 2001, 119-142.

11. A.N. Tikhonov, A.S. Leonov and A.G. Yagola, *Nonlinear Ill-Posed Problems*, Vol.1, Chapman & Hall, London, 1998.

12. A.G. Yagola, A. Leonov and V. Titarenko, Ill-posed problems and a priori information, in *Inverse Problems in Engineering Mechanics III*, (Tanaka, M. and Dulikravich, G., editors), Elsevier, 2002, 235-244.

13. M. Hanke, Limitations of the L-curve method in ill-posed problems, *BIT*, **36**, 287 (1996).

14. C.R. Vogel, Non-convergence of the L-curve regularization parameter selection method, *Inverse Problems*, **12**, 535 (1996).

15. J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Their Applications*, Springer-Verlag, Berlin, 1972.

16. A.S. Nemirovskii, The regularizing properties of the adjoint gradient method in ill-posed problems, *Comput. Maths. Math. Phys.*, **26**, 7 (1986).

17. M. Hanke and P.C. Hansen, Regularization methods for large-scale problems, *Surveys Math. Industry*, **3**, 253 (1993).

18. V.A. Kozlov, V.G. Maz'ya, and A.F. Fomin, An iterative method for solving the Cauchy problem for elliptic equations, *Comput. Maths. Math. Phys.*, **31**, 45 (1991).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

SECOND ORDER METHODS FOR INVERSE PROBLEMS: AN APPLICATION IN HYDROLOGY

François-Xavier Le Dimet, Pierre Ngnepieba
Laboratoire de Modélisation et Calcul, projet IDOPT
Université Joseph Fourier
38041 Grenoble, cedex, France
ledimet@imag.fr

ABSTRACT

The estimation of some unknown parameters is carried out by using least square methods requiring the computation of a gradient. We present a method for the derivation of second order quantities especially the products of the hessian with a vector. We will see how this information can be used for the estimation of the condition number of the Hessian and for sensitivity analysis .

SECOND ORDER ANALYSIS

Let us consider a model describing the evolution of fluid, we will assume that the model has been discretized with respect to time, it writes:

$$\begin{cases} \dfrac{dX}{dt} = F\left(X, K\right) \\ X(0) = U \end{cases}$$

X is the state variable describing the medium, K is some unknown parameter and U the initial condition which is also unknown. We assume that the medium has been observed between times 0 and T and we have an observation Xobs . C is an operator from the state space toward the space of observation. K and U are estimated by the minimization of the cost function J defined by :

$$J\left(U, K\right) = \frac{1}{2}\int_0^T \left\| C.X - X_{obs} \right\|^2 dt$$

This is the simplest form for the cost function. In practice it should contains some regularization term
The optimal values of U and K are solutions of the optimality system:

$$\begin{cases} \nabla_U J = 0 \\ \nabla_K J = 0 \end{cases}$$

The gradients are evaluated by introducing the adjoint model, P being the adjoint variable of the same dimensionality as X, defined by:

$$\begin{cases} \dfrac{dP}{dt} + \left[\dfrac{\partial F}{\partial X}\right]^t P = C^t\left(CX - X_{obs}\right) \\ P\left(T\right) = 0 \end{cases}$$

Then it comes :

$$\begin{cases} \nabla_U J = -\left[\dfrac{\partial F}{\partial K}\right]^t .P \\ \nabla_K J = -P\left(0\right) \end{cases}$$

The compution of the gradient, by a backward integration of the adjoint model, permits to carry out some descent type method to compute U and K
To compute H, the hessain matrix of J :

$$H\left(U, K\right) = \begin{pmatrix} H_{U,U} & H_{U,K} \\ H_{U,K} & H_{K,K} \end{pmatrix}$$

we introduce Q and R. If we consider the sytem

$$\begin{cases} \dfrac{dQ}{dt}+\left[\dfrac{\partial F}{\partial X}\right]^t Q-\left[\dfrac{\partial^2 F}{\partial X^2}R\right]^t P=C^tCR \\ \dfrac{dR}{dt}-\left[\dfrac{\partial F}{\partial X}\right]R=0 \\ Q(T)=0 \\ R(0)=V \end{cases}$$

If this is integrated, then it can be shown [1], that we have :

$$H_{U,U}.V=Q(0)$$
$$H_{U,V}.V=\left[\frac{\partial F}{\partial K}\right]^t.Q$$

In the same way, we will consider

$$\begin{cases} \dfrac{dQ}{dt}+\left[\dfrac{\partial F}{\partial X}\right]^t Q-\left[\dfrac{\partial^2 F}{\partial X^2}R\right]^t P=C^tCR \\ \dfrac{dR}{dt}-\left[\dfrac{\partial F}{\partial X}\right]R=V \\ Q(T)=0 \\ R(0)=0 \end{cases}$$

Then we will have :

$$H_{V,V}.V=\left[\frac{\partial F}{\partial K}\right]^t.Q$$

Thus it is possible to compute the product of the Hessian by a vector. The systems differ by the forcing terms and the initial or final conditions. The first equation can be deduced from the adjoint model by changing the right hand side, the second equation is obtained by a linearization of the model.
Of course the full Hessian can computed if we take for U and V the vectors of the canonical base. But the Hessian is by itself of little interest, what could be important is to access its spectral properties : largest and smallest eigenvalues, eigenvectors. These quantities can be computed without an explicit computation of the Hessian.

APPLICATION TO AN INFILTRATION MODEL

Identification

As an application we will consider a 1-D model of infiltration in an unsaturated ground. The state variable is (h) : water pressure., in a domain between the surface at z=0 and the bottom at z=Z.

$$\begin{cases} C(h)\dfrac{\partial h}{\partial t}=\dfrac{\partial}{\partial z}\left[K(h)\left(\dfrac{\partial h}{\partial z}-1\right)\right] \\ h(0,z)=h_{ini}(z) \\ h(t,0)=h_{surf}(t) \\ h(t,Z)=h_{bot}(z) \end{cases}$$

C(h) and K(h) are given by:

$$C(h)=\begin{cases} \dfrac{q_s(2-n)}{h_g}\left(\dfrac{h}{h_g}\right)^{n-1}\left[1+\left(\dfrac{h}{h_g}\right)^n\right]^{\frac{2}{n}-2}, h<0 \\ 0, h\geq 0 \end{cases}$$

$$K(h)=\begin{cases} K_s\left[1+\left(\dfrac{h}{h_g}\right)^n\right]^{h\left(\frac{2}{n}-1\right)}, h<0 \\ K_s, h\geq 0 \end{cases}$$

There are five parameters (K_s, h_g, q_s, h, n) and the three initial and boundary conditions to be identified.
If all the parameters of the model are known, then its is possible to compute a cumulated infiltration given by:

$$I_{cum}(t)=\int_0^Z (q(t,z)-q_{ini})dz$$

where q is the water content of the ground.

The observation bear on the cumulted infiltration $I_{obs}(t)$. Therefore the problem is to determine $U=(q_{ini},q_{surf},q_{bot})$ and $L=(K_s,h_g,q_s,h,n)$ minimizing the cost function J defined by :

$$J(U,L)=\frac{l_1}{2}\left\|U-U^e\right\|^2+\frac{l_2}{2}\left\|L-L_e\right\|^2$$
$$+\frac{\Delta t}{2}\sum_{j=0}^{M}\left(I_{cum}(t_j)-I_{obs}(t_j)\right)^2$$

Remarks:

1- the dependance with respect to the parameters is highly non linear.
2- The first two terms are used as regularization terms, U_e and L_e are a priori estimations of U and L
3- The model has been discretized. with a finite difference scheme in space with Z =1m, the grid size was 1cm. The temporal scheme was an implicit Euler scheme with T=2h and a time step of 1s.

P being the adjoint variable, the adjoint model is defined by:

$$\begin{cases}-\frac{\partial}{\partial t}(C.P)+\left[\frac{\partial C}{\partial h}\right].\left(P.\left[\frac{\partial h}{\partial t}\right]\right)\\ -\frac{\partial}{\partial z}\left[K.\frac{\partial P}{\partial z}\right]+\left[\frac{\partial K}{\partial h}\right].\left(\frac{\partial P}{\partial z}\left[\frac{\partial h}{\partial z}-1\right]\right)=\\ (I_{cum}-I_{obs})\frac{\partial I_{cum}}{\partial h}d(t-t_i)\\ P(t=T,z)=0\\ P(t,z=0)=0\\ P(t,z=Z)=0\end{cases}$$

From the backward integration of the adjoint system we deduce the gradient:

$$\nabla_L(U,L)=-\int_0^T\int_0^Z\left(\left[\frac{\partial C}{\partial L}\right].\left(P.\left(\frac{\partial h}{\partial t}\right)\right)\right)$$
$$-\left(\frac{\partial K}{\partial L}\right)\left(\frac{\partial P}{\partial z}.\left(\frac{\partial h}{\partial z}-1\right)\right)dtdz$$
$$+\sum_{i=0}^{M}\int_0^Z(I_{cum}-I_{obs})\left[\frac{\partial I_{cum}}{\partial L}\right]d(t-t_i)\,dz+l_2(L-L_e)$$

$$\nabla_{q_{ini}}(U,L)=\int_0^Z(C.P)_{t=0}-(I_{cum}-I_{obs})d(t-t_i)\,dz$$
$$+l_1(q_{ini}-q_{ini}^e)$$
$$\nabla_{q_{bot}}(U,L)=\int_0^T\left(K\frac{\partial P}{\partial z}\right)_{z=Z}dt+l_1(q_{bot}-q_{bot}^e)$$

It is well known that there is no commutativity between discretization and the derivation of the adjoint, therefore all the former calculations should also be carried out on the discrete model. A numerical experiment has been realized on a material known as Grenoble sand.
The method of optimization was a conjugate gradient method written in the code M1GC3 [2]

	Experimental values	Identified
K_s	4.528 (10-3)	1.7793 (10-3)
h_g	-16.40	-15.90
q_s	0.312	0.2879
h	6.73	4.97
n	2.79	2.34
q_{ini}	8.166 (10-2)	7.49(10-2)
q_{surf}	0.312	0.29
q_{bot}	8.16 (10-2)	0.3

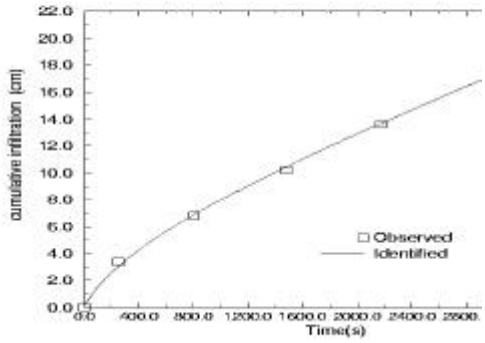


Fig 1: Adjustment of the optimal cumulated infiltration to the observations.

The numerical results show that if some parameters are correctly identified some others are more difficult to adjust. It is clear that the cost function is not convex with respect to the parameters to be identified.

Computation of the condition number of the Hessian

The convergence properties of the optimization algorithm are linked to the condition number of the Hessian, the condition number is the ratio (in module) of the largest eigenvalue to the smallest one, therefore the condition number is always greater or equal to 1 (the hessian being symetric its eigenvalues are real). A large condition number means that the problem is ill-conditionned.
The largest eigenvalue can be computed by iterated power method : if the largest eigenvalue is simple then the sequence defined by:

- V_0 is given
- V_{k+1} is defined by

$$\begin{cases} U_{k+1} = HV_k \\ V_{k+1} = \dfrac{U_{k+1}}{\|U_{k+1}\|} \\ l_{k+1} = \|U_{k+1}\| \end{cases}$$

then $l_k \to l_{\max}$, the spectral radius of H, when $k \to \infty$. Therefore to compute the spectral radius it is sufficient to be able to compute the product of the Hessian by a vector. To compute the smallest eigenvalue it is enough to point out that (in module) the smallest eigenvalue of a matrix is the largest of the inverse matrix. Therefore at each step of the former algorithm a linear system has to be solved. To solve a linear system does not require to know explicitly the matrix of the system with, for instance, a conjugate gradient method, it is enough to be able to compute the product hessian.vector.
For different values of the number of observations the condition number has been computed. Fig. 2 display the spectral radius of the Hessian, Fig. 3 the smallest eigenvalue and Fig. 4. the condition number. Even if the behaviour is not monotonic, the condition number increases when there is less observations.

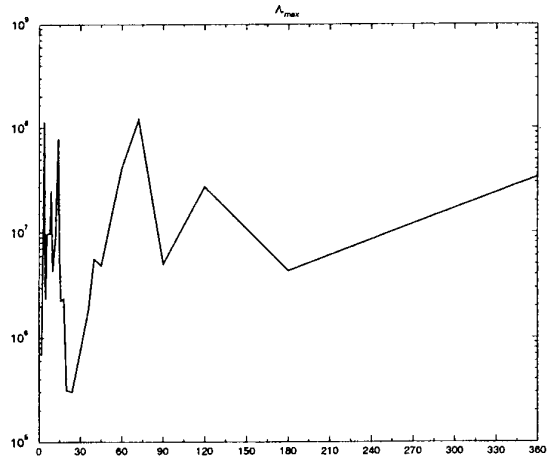


Fig 2: Largest eigenvalue as a function of the number of observations

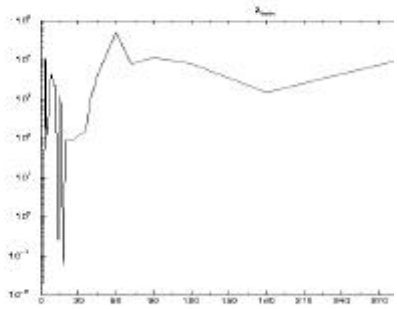


Fig 3: Smallest eigenvalue as a function of the number of observations

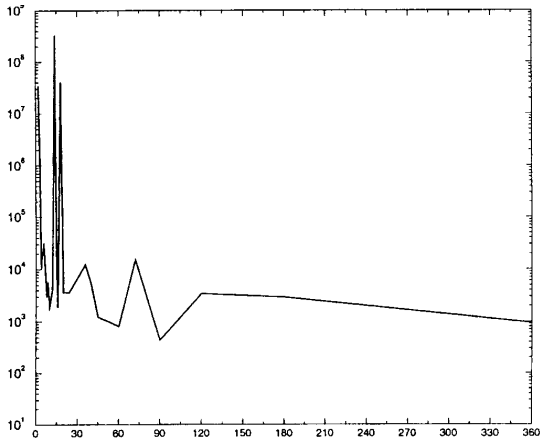


Fig4: Condition number as a function of the number of observations

SENSITIVITY STUDY

General sensitivity study

Let us consider a general model written :

$$M\left(X, Z\right) = 0 \ .$$

X is the state variable of the model and *Z* some parameters; We assume that Z being given then the model has a unique solution *X(Z).*

A sensitivity study is defined by a so-called response function *G(X,Z),* a real value function, X is the solution of the model associated to Z . Therefore this function is totally defined when has been fixed.. By definition the sensitivity is the gradient of G.

In many physical application the sensitivity is computed by finite difference : if $Z = \left(z_i\right) \quad i = 1,\ldots, N$, then the sensitivity is estimated by :

$$\nabla G = \left(\frac{\partial G}{\partial z_i}\right)$$

$$\left(\frac{G\left(X\left(Z + \alpha e_i\right), Z + \alpha e_i\right) - G\left(X\left(Z\right), Z\right)}{\alpha}\right)$$

e_i being the vectors of the canonical base. This method has several inconvenients:

- it requires N integrations of the model. In many geophysical applications M may be very large.
- the value of α is arbitrary. To get the correct value, the result should be independent of α , therefore several attempt may be necessary before getting the right value.

Introducing an adjoint model permits to compute the exact sensitivity in only one run of the adjoint model. To derive this sensitivity , let us introduce some perturbation h on Z, the Gateaux derivative are defined by:

$$\hat{X}(h) = \lim_{\alpha \to 0} \frac{X\left(Z + \alpha h\right) - X\left(Z\right)}{\alpha}$$

Deriving the model and the response function gives:

$$\frac{\partial M}{\partial X}\hat{X} + \frac{\partial M}{\partial Z}h = 0$$

$$\hat{G} = \frac{\partial G}{\partial X}\hat{X} + \frac{\partial G}{\partial Z}h$$

We will get the gradient of *G* by exhibiting the linear dependance of $\hat{G}$ with respect to h.

To do so we introduce *Q* the adjoint variable, with the same dimensionality as the state variable, then we take the inner product of the Gateaux derivative of the model with *Q*. It comes:

$$\left(Q, \frac{\partial M}{\partial X}\hat{X}\right) + \left(Q, \frac{\partial M}{\partial Z}h\right) = 0$$

It is clear that if the adjoint model is defined as the solution of:

$$\left[\frac{\partial M}{\partial X}\right]^t Q = \frac{\partial G}{\partial X}$$

Then we will obtain

$$\hat{G} = \left(-\left[\frac{\partial M}{\partial Z}\right]^t Q + \frac{\partial G}{\partial Z}, h \right)$$

and :

$$\nabla G = -\left[\frac{\partial M}{\partial Z}\right]^t Q + \frac{\partial G}{\partial Z}$$

Therefore the sensitivity is obtained in only one run of the adjoint model. The price to be paid is to write the adjoint code, with a complicated model it could be a tremendous task. Nevertheless some tools of automatic diifferentiation may be helpful.

Sensitivity in the presence of observations.

In many cases the input of a model are observations. If we are looking for the sensitivity with repect to these observations a difficulty comes from the fact that they does not appear explicitly in the model. The observations are included only in the optiamlity system. Therefore this last one should be considered as a generalzed model and the general sensitivity analysis should be carried out on the optimality system. Because O.S. will be derived we will introduce second derivatives in the sensitivity analysis.

Example

If the model is :

$$\begin{cases} \dfrac{dX}{dt} = F(X, K) \\ X(0) = U \end{cases}$$

Where the initial condition has been choosen by the minimization of a cost function:

$$J(U) = \frac{1}{2}\int_0^T \left\| CX - X_{obs} \right\|^2 dt$$

The adjoint model is:

$$\begin{cases} \dfrac{dP}{dt} + \left[\dfrac{\partial F}{\partial X}\right]^t P = C^t (CX - X_{obs}) \\ P(T) = 0 \end{cases}$$

The optimality condition is:

$$\nabla J(U) = -P(0) = 0 .$$

If the response function has the form:

$$W(K) = \int_0^T G(X)\, dt$$

Then the general sensitivity analysis is carried out on the optimality system.
Q and R two adjoints variables are introduced as the solution of the system:

$$\begin{cases} \dfrac{dQ}{dt} + \left[\dfrac{\partial F}{\partial X}\right]^t Q + \left[\dfrac{\partial^2 F}{\partial X^2} P\right]^t .R \\ -C^t CR = \left[\dfrac{\partial G}{\partial X}\right] \\ \dfrac{dR}{dt} - \left[\dfrac{\partial F}{\partial X}\right] R = 0 \\ Q(0) = 0 \\ Q(T) = 0 \end{cases}$$

Then the sensitivity is given by:

$$\nabla W(K) = -\int_0^T \left(\left[\frac{\partial F}{\partial K}\right]^t Q + \left[\frac{\partial^2 F}{\partial X \partial K} R\right]^t P \right) dt$$

We obtain a non standard system because one of the equation has two boundary condition and the second one, no boundary condition. It is possible to transform this problem into a problem of optimization for which a conjugate

gradient method can be use [2]. An iterative method has to be used to solve the system.

Therefore we see that in the presence of observartions requires to use second order information.
The equation which are used for computing the sensitivity are the close (from the coding point of view) ofthose used to compute the Hessian. Example

With the same physical model as above we have considered as response function the quadratic norm of the hydraulic conductivity.

$$EK\left(h_{obs}\right) = \int_0^T \int_0^Z K\left(h\right)^2 \, dz dt$$

The observation being a function of time, its gradient will be also a function of time. In Figure 5 , the norm of the sensitivity is represented as a function of time, it is assumed that there is an observation at each time step of the numerical scheme. The same quantity is displayed in Figure 6, but with an observation each minute. Both simulation last one hour.

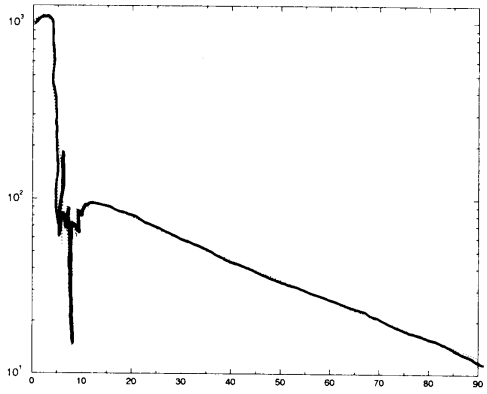


Fig5. Norm of the sensitivity. One observation at each time step. Unity of time =40s.

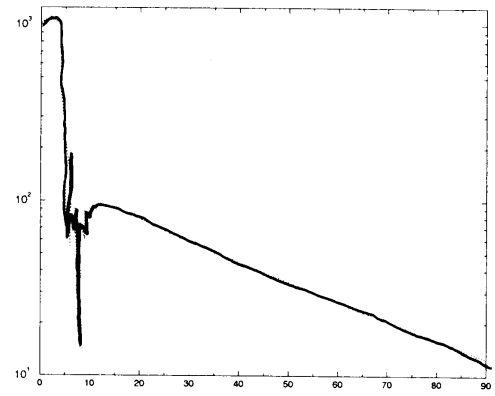


Fig5. Norm of the sensitivity. One observation each 60s. Unity of time =40s.

It is clear that, in this case, an evaluation of the sensitivity by finite differences would have been very costly from the computational viewpoint.

CONCLUSION

The access to second order information is important to improve the numerical algorithms and to estimate the propagation of uncertainties in the observations or on some other parameters of the model. This information is obtained through the second order adjoint, which can be considered as an important tool for inverse problems.

REFERENCE

[1]F .- X. Le Dimet, I. M. Navon and D. N. Daescu, Second Order Information in Data Assimilation , *Monthly Weather Review* , Vol. 130, No. 3 , 629-648 (2002)
[2]P. Ngnepieba, Assimilation de données et identification de paramètres : une application en hydrologie.Thèse de Doctorat de L'Université Joseph Fourier, Grenoble, France, (2001)
[3]J.-C. Gilbert, C. Lemaréchal,Some numerical experiment with varaible storage quasi-newton methods .*Mathematical Programming*, 4(, 407-435 (1989)

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

BACKWARD SPECIFICATION OF PRIOR IN BAYESIAN INFERENCE AS AN INVERSE PROBLEM

Andrei V. Gribok
Department of Nuclear Engineering
315 Pasqua Engineering Building
The University of Tennessee, Knoxville
Knoxville, TN 37996-2300
agribok@utk.edu

Aleksey M. Urmanov
Department of Nuclear Engineering
315 Pasqua Engineering Building
The University of Tennessee, Knoxville
Knoxville, TN 37996-2300
urmanov@utk.edu

J. Wesley Hines
Department of Nuclear Engineering
212 Pasqua Engineering Building
The University of Tennessee, Knoxville
Knoxville, TN 37996-2300
hines@utkux.utk.edu

Robert E. Uhrig
Department of Nuclear Engineering
315 Pasqua Engineering Building
The University of Tennessee, Knoxville
Knoxville, TN 37996-2300
ruhrig@utk.edu

ABSTRACT

Specification of prior distribution is one of the most important methodological as well practical problems in Bayesian inference. Although a number of approaches have been proposed, none of them is completely satisfactory from both theoretical and practical points of view. We propose a new method to infer prior distribution from a priori information which may be available from observations. The method consists of specifying a predictive distribution of the value of interest and then working backwards towards the prior distribution on the parameters. The method requires the solution of the Fredholm integral equation of the first kind, which can be effectively solved using Tikhonov regularization. Numerical examples for two cases of Bayesian inference are presented.

NOMENCLATURE

L-likelihood function
$\pi(\theta|\alpha)$-prior distribution of the parameter
$\pi(\theta|x,\alpha)$-posterior distribution of the parameter
$\pi(z|\alpha)$-prior predictive distribution
$\pi(z|x,\alpha)$-posterior predictive distribution
θ-parameter of binomial distribution
λ-regularization parameter
α,β-hyperpameters
N-number of Bernoulli trials
z-random variable
x-random variable
μ,σ^2-parameters of normal distribution
$B(\alpha,\beta)$-beta function

INTRODUCTION

Transferring prior beliefs into an exact mathematical form has been, and remains one, of the most controversial and challenging issues of Bayesian inference. The problem is twofold. The first one is how to specify our knowledge in the most succinct and tractable form and the second one is how to transfer prior knowledge of observable variables onto prior knowledge of parameters which are generally unobservable. A number of approaches have been developed, with the most notable ones being: conjugate priors, Jeffreys noninformative priors and empirical Bayesian methods [1,3]. Conjugate priors, although being widely used, can only be justified if enough information is available to believe that the true prior distribution belongs to the specified family; otherwise, the main justification for using conjugate prior is their mathematical tractability. Jeffreys noninformative prior uses the Fisher information matrix to place a maximally noninformative prior on the parameters, exploiting the fact that the Fisher information matrix is widely considered to be an indicator of the accuracy of a parameter estimate. However, this approach can only be effectively used in one-dimensional cases and does not satisfy the Likelihood principle [1]. Other problem with noninformative priors is that there might be a number of them for a given problem and there is no clear cut rule which noninformative prior has to be preferred. Empirical Bayesian methods use the marginal distribution of the value of interest to elicit prior distribution on the parameters. The empirical estimation of the prior is strictly speaking a violation of Bayes theorem because the same

data set is used for both: estimation of the likelihood and inferring the prior distribution. This approach effectively invalidates Bayes theorem due to the fact that:

$$P(q \mid D) \neq \frac{P(D \mid q) \cdot P(q \mid D)}{P(D)} \tag{1}$$

The formula (1) means that once the prior probability is conditioned on the current data set, the Bayes formula is no longer valid and we can not formally go ahead with Bayesian inference. Our approach is based on the observation that for many practical engineering problems the range of predicted values is known and hence through the predictive distribution this knowledge can be transferred to the prior distribution over parameters by solving the Fredholm integral equation of the first kind.

BAYESIAN INFERENCE AND BAYESIAN PREDICTIONS

The core of Bayesian inference is Bayes formula, which inverts information contained in a data set into an estimation of a parameter or model,

$$p(q \mid x, a) = \frac{L(x \mid q) p(q \mid a)}{\int_{\Theta} L(x \mid q) p(q \mid a) dq} \tag{2}$$

where $\pi(\theta|x,\alpha)$ is posterior distribution of the parameter θ conditioned on the current data set x and a hyperparameter α which defines the prior distribution $\pi(\theta|\alpha)$. $L(x|\theta)$ is the likelihood function which specifies the probability for the given data set x to occur conditioned on the parameter θ. Bayesian predictions can be based on both posterior and prior distributions of the parameter. Instrumental to performing Bayesian prediction is the likelihood of a future data set z, which is defined as $L(z|\theta)$. This likelihood assesses the plausibility for data z to occur in future experiments for a given value of the parameter θ. Combining this likelihood with the prior distribution on the parameters, we get what is called the prior predictive distribution:

$$p(z \mid a) = \int_{\Theta} L(z \mid q) p(q \mid a) dq \tag{3}$$

This reflects a distinct feature of Bayesian inference: it can produce predictions with no current data at hand, providing prior information is informative enough.

Combining the future likelihood and the posterior distribution we get the posterior predictive distribution:

$$p(z \mid x, a) = \int_{\Theta} L(z \mid q) p(q \mid x, a) dq \tag{4}$$

Equation (4) summarizes our inference about future values of z after have seen the data x. Integrals (3) and (4) have been used in Bayesian inference for a long time and are known under different names. As we already mentioned, if the likelihood of future data is used in (3) and (4), they are known as prior and posterior predictive distributions respectively [2]. If the current data set is used to estimate the likelihood, then integral (3) is known as the marginal distribution of x [3] or, in the neural networks community, as evidence [4]. We shall use the terms prior predictive distribution and marginal distribution interchangeably in this paper. There are a number of ways in which the marginal distribution is used to select a prior in Bayesian analysis. One of them is the maximum likelihood II approach [4] where the integral in (3) is maximized over the prior distribution $\pi(\theta|\alpha)$ for different values of the hyperparameter α. The moment approach [4] tries to relate moments of the prior distribution to moments of the marginal likelihood. The distance approach [4] is most closely related to the method that we propose. It prescribes to estimate the empirical marginal distribution from the historical data and then attempts to match the left-hand side of equitation (3) to this empirical prior using different priors in the right-hand side. However, this approach requires a complex optimization. It should be pointed out that all of the approaches that we mentioned attempt to restrict the class of priors which can be deduced from the integral relationship (3). However, they stop short of directly solving the integral equation (3) using regularization techniques. Our approach consists of solving the integral equation (3) using Tikhonov regularization [5] thus restricting the class of desired priors to smooth ones.

The focus of our analysis is the prior predictive distribution (3). Under the assumption that $\pi(z|\alpha)$ and $L(z|\theta)$ are known, Formula (3) represents a linear Fredholm integral equation of the first kind. In this case, the future likelihood represents the kernel, and the prior distribution over the parameter is the desired solution. It should be stressed that the predictive distribution is a function of an observable variable z, while the prior distribution is a function of an unobservable variable θ. The integral relationship (3) represents the forward problem of Bayesian inference, inference of predictive distribution when prior and likelihood are known. However to place restrictive informative prior on parameters one often has to solve equation (3) for prior distribution which is the inverse problem of Bayesian inference. In many practical engineering applications, the range of future observations is known from physical considerations. For example, the range of temperature, pressure and flow rate measurements in nuclear power plants is known if plant operates under normal conditions. Hence, we can place

rather informative restrictions on the predictive distribution of future observations. This information can come from physical and engineering judgments as well as from historical observations of the variable of interest. Once we deduce what the possible predictive distribution of future observations is, we can solve the integral equation (3) to get the prior distribution of the parameter θ. Doing this we effectively transform prior information about observable variables onto prior information about unobservable parameters.

However, the solution of the integral equation (3) will require the use of regularization because of the ill-posed nature of the problem. It should be pointed out that the predictive distribution of the future observation $\pi(z|\alpha)$ will always contain uncertainty or noise because of its empirical nature. Solving integral equation (3) by numerical methods will effectively transform ill-possedeness into ill-conditioning of the matrix $L(z|\theta)$. We apply Tikhonov regularization to solve this ill-conditioned system of equation.

Tikhonov regularization scheme in its general form can be written as:

$$\left\{ \min \left\| \int_{\Theta} L(z \mid q) p(q \mid a) dq - p(z \mid a) \right\|_2 + l^2 \int_{\Theta} \left(\frac{d^2 p(q \mid a)}{dq^2} \right)^2 dq \right\} \quad (5)$$

Tikhonov regularization imposes smoothness constrains on the sought solution which is, in our case, the probability density function. Imposing smoothness constrains on the probability density function (pdf) is a very natural restriction because all known and practical pdfs are smooth and differentiable.

Summarizing our approach we can outline three steps that should be performed in order to apply it:

1. Using prior information or engineering judgment, define marginal distribution of the variable of interest.
2. Define the likelihood of future measurements of the variable of interest.
3. Solve integral equation (3) for prior distribution of the parameter.

NUMERICAL EXAMPLES

Inferring the Value of the Parameter for a Binomial Distribution

We present two numerical examples of backward specification of prior by solving the integral equation. The first one deals with the inference of a parameter for a binomial distribution and the second one deals with the inference of the standard deviation for a normal distribution with known average.

The likelihood of a future data set z for a binomial distribution can be written as:

$$L(z \mid q) = C_z^N q^{\ z} (1-q)^{N-z} \quad (6)$$

If the number of trials N is fixed, then the likelihood (6) represents a function of two variables: z and θ. The prior predictive density of z would be:

$$p(z \mid a, b) = \int_0^1 C_z^N q^{\ z} (1-q)^{N-z} p(q \mid a, b) dq \quad (7)$$

or in terms of the Fredholm integral equation of the first kind:

$$g(z) = \int_0^1 K(q, z) f(q) dq \quad (8)$$

Assuming the beta distribution as a conjugate prior for binomial likelihood, we get:

$$p(z \mid a, b) = \int_0^1 C_z^N q^{\ z} (1-q)^{N-z} \frac{q^{a-1}(1-q)^{b-1}}{B(a, b)} dq \quad (9)$$

which after simplifications produces:

$$p(z \mid a, b) = \frac{C_z^N}{B(a, b)} B(z+a, N-z+b) \quad (10)$$

which is beta binomial distribution. Hence, the integral equation (7) has an exact solution in analytical form and we can estimate how close the regularized solution would be to the true one.

In order to progress from equation (9) to a system of linear equations, we use the midpoint rule for discretization. We discretize the likelihood for N=100, z=0…100 and p=0…1 with 100 samples. We consider z as the number of successes in 100 trials. The matrix representing the likelihood is 100X100. Thus the discretization leads to a square system of linear equations

$$b = Aq, \quad A \in R^{100x100} \quad (11)$$

The condition number of matrix A is $1.1*10^{18}$, pointing to severe ill-conditioning. We use Tikhonov regularization in standard form to solve this ill-conditioned system of linear equations:

$$q_l = \arg\min\left\{\|Aq - b\|^2 + l^2\|q\|^2\right\} \qquad (12)$$

The left-hand side $\pi(z|\alpha,\beta)$ and the exact solution $\pi(\theta|\alpha,\beta)$ of the integral equation (8) are shown in Figs.1 and 2

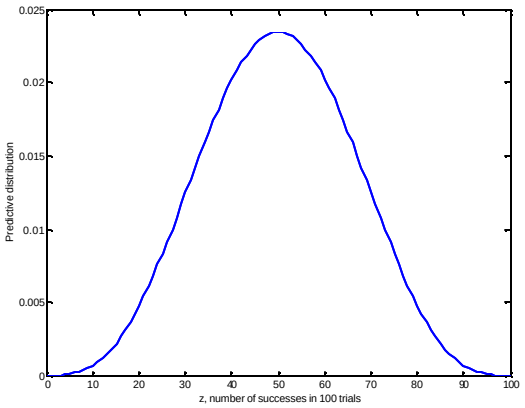


Fig. 1 Predictive distribution $\pi(z)$

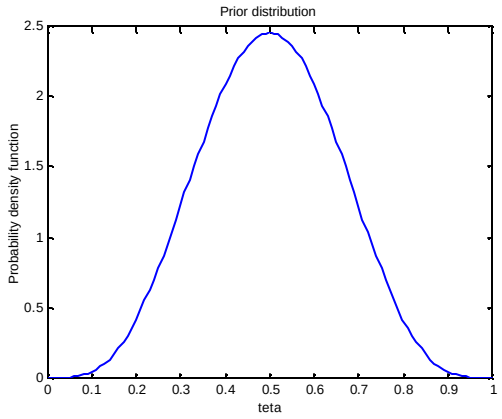


Fig.2 Prior distribution on parameter θ.

To obtain the predictive distribution in Fig. 1, we solved the forward problem (11) with prior distribution depicted in Fig.2 as θ.
The ordinary least squares (OLS) solution for the system (11) is presented in Fig. 3

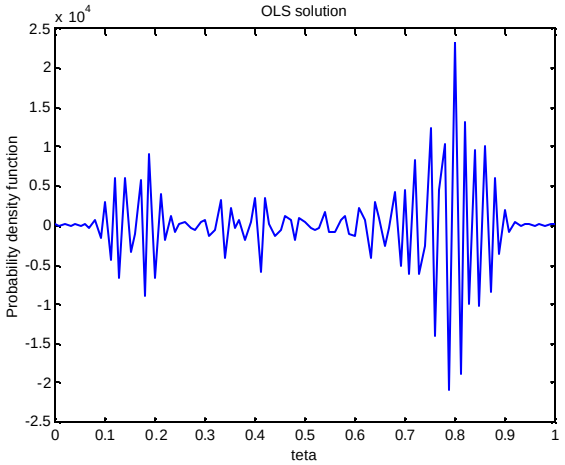


Fig.3 OLS solution.

As we can see, the OLS solution is very oscillatory and makes no sense. It bears no resemblence to the exact known solution shown in Fig.2. However, the regularized solution presented in Fig.4 is very close to the exact one in Fig.2 and can be used as the prior distribution.

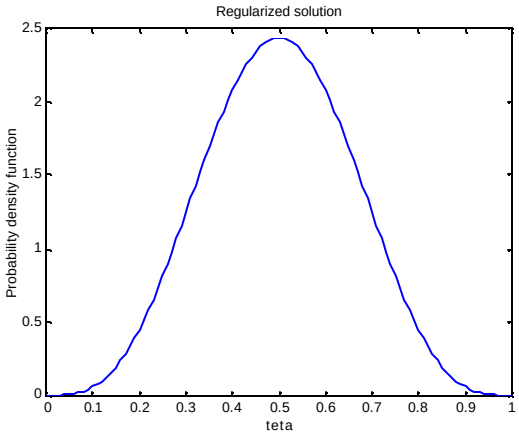


Fig. 4 Regularized solution

We used Morozov's discrepancy principle [6] to select regularization parameter $\lambda=8.5*10^{-5}$. However, the most interesting case represents a situation in which the predictive distribution is estimated from the data or from the priori knowledge, as in the case shown below.

Suppose we have some statistical data about the number of successes in 100 tosses in previous trials. We can use this historical data to estimate what can be called the empirical predictive distribution or marginal distribution, and using this distribution, we can solve for prior equation (9). The empirical predictive distribution being estimated from the data would contain a significant amount of noise, which would make the OLS solution of

equation (9) very unstable and irrelevant. An example of the empirical prior distribution estimated from the data is shown in Fig. 5.

The kernel density estimator, with a Gaussian kernel width of 10, was used to estimate this density from some historical data representing 5 trials of 100 tosses of a fair coin. The parameter of interest was the number of successes that was recorded as 61, 51, 60, 47, and 49 in simulations. As can be seen from Fig. 5, the marginal distribution of z is a bell shaped curve with mean value slightly higher than 50.

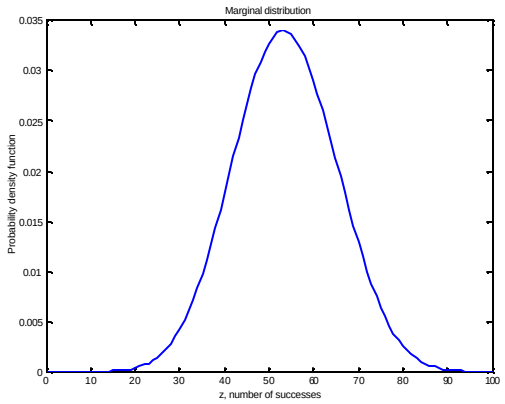


Fig. 5 Marginal distribution

Due to the large kernel width used to estimate the density from the empirical data, the curve has one mode. Using this empirical density as the left hand side of equation (9), we can again numerically solve it for the prior distribution. The unregularized solution is shown in Fig.6

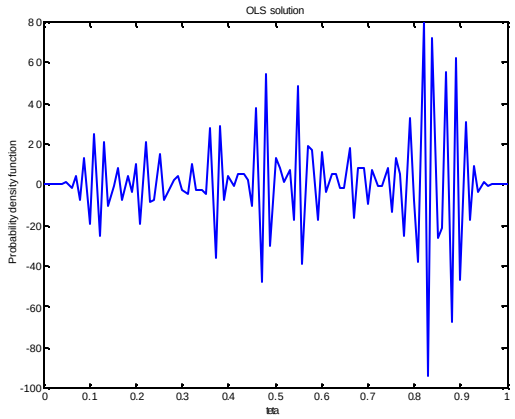


Fig.6 Ordinary least squares solution

As we can see, the solution is still very oscillatory and does not represent a real probability density function. However, the regularized solution depicted in Fig.7 looks like a proper probability density and can be used as a prior for future inference.

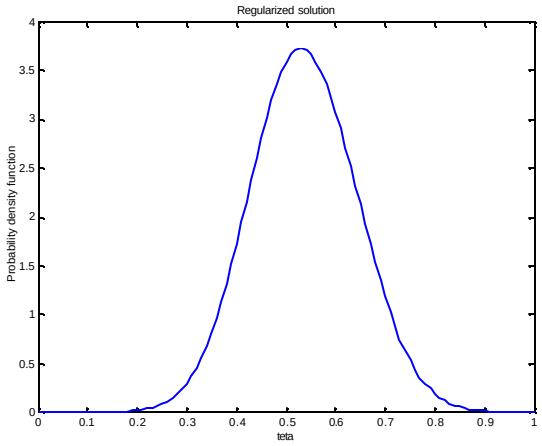


Fig.7 Regularized solution

In this case, the most remarkable feature of using regularization is that it makes the inference about the possible prior distribution virtually insensitive to the ambiguous nature of the kernel density estimator. The problem with empirical density estimators is that their results are very sensitive to the chosen parameters of the techniques. For example, the density estimated with kernel techniques depends very much on the kernel width. Fig. 8 shows the density of the same data set estimated with the kernel width chosen to be 3.

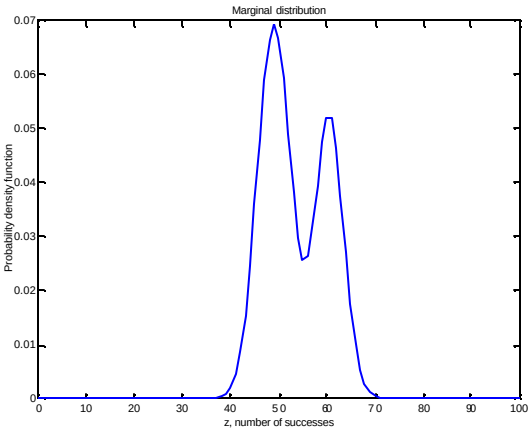


Fig.8 Marginal probability density function

The estimated density now has two modes which looks quite plausible in the light of the available data. The OLS and regularized solutions are shown in Figs 9 and 10.

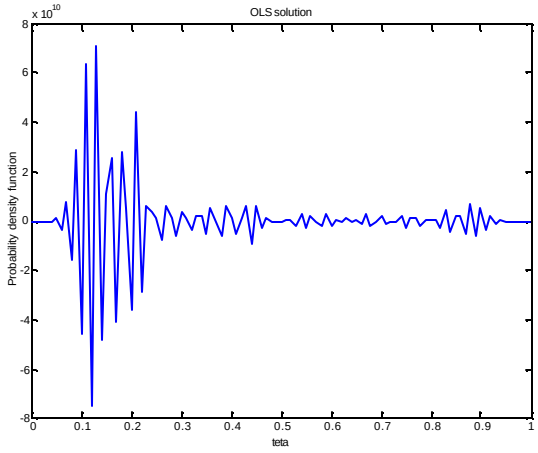


Fig. 9 OLS solution

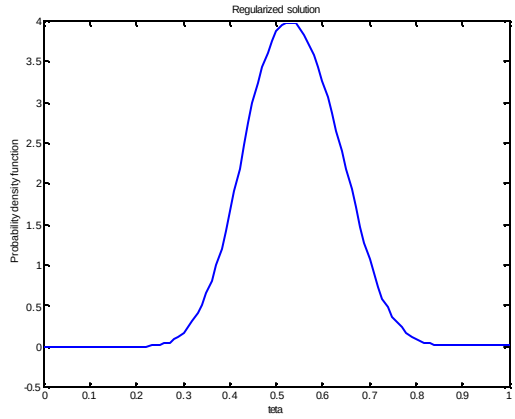


Fig. 10 Regularized solution

As can be seen from these figures, the OLS solution is again highly unreasonable and does not represent a real probability density function; however, the regularized solution is very close to the one obtained for the kernel width equal to 10 and shown in Fig. 7. The discrepancy principle was again used to choose the regularization parameters for these cases. It should be mentioned that in the last example, with the marginal distribution obtained from the data the first order, Tikhonov regularization was used with a smoothing operator representing an approximation of the first derivative.

Inference of Variance of Normal Distribution with Known Mean.

The second numerical example to be analyzed is the inference about the variance of a normal distribution when the mean value is known. In this case the likelihood of future data z can be written as:

$$L(z/s^2) = \frac{1}{\sqrt{2ps^2}} \exp(-\frac{(z-m)^2}{2s^2}) \qquad (13)$$

where μ is the known mean value. The corresponding conjugate prior density for variance is inverse-gamma and can be written as:

$$p(s^2|a,b) = \frac{\left(s^2\right)^{-(a+1)} \exp(-\frac{b}{s^2})}{b^{-a}\Gamma(a)} \qquad (14)$$

where α and β are two hyperparameters which define the shape and scale of prior distribution. Combining the likelihood and prior distribution we again obtain the prior predictive distribution:

$$p(z|a,b) = \int_{s^2} L(z|s^2)p(s^2|a,b)ds^2 \qquad (15)$$

Now assume that we have a data sample y generated from $N(\mu,\sigma^2)$. We can use this data sample to estimate the empirical distribution and use it as $\pi(z|\alpha,\beta)$. Having done this, we can again solve the integral equation (15) for the prior distribution $\pi(\sigma^2|\alpha,\beta)$ using Tikhonov regularization. Suppose we have a data sample of ten random values generated from y ~ N(0,1), y=(0.4855;-0.0050;-0.2762;1.2765;1.8634;-0.5226;0.1034;-0.8076;0.6804;-2.3646). The probability density function estimated from this sample is shown in Fig.11.

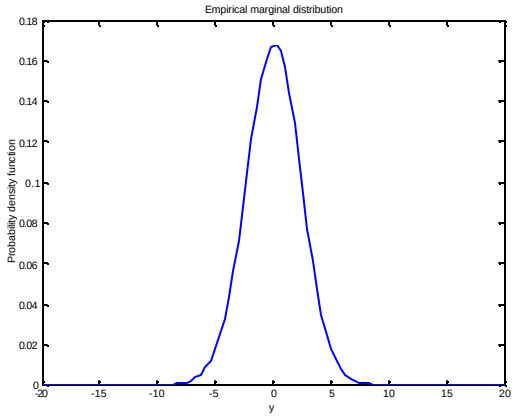


Fig. 11 Empirical marginal distribution

This probability density function is the only source of information about the random variable y that we have. The probability density function can be used as the empirical marginal distribution $\pi(z/\alpha,\beta)$ in the left-hand side of equation (15). Because the likelihood for the data is written

in (12), we can numerically solve the integral equation (15). The OLS solution is shown in Fig. 12

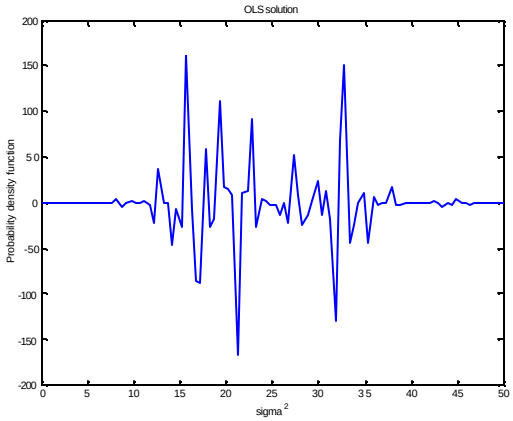


Fig.12 OLS solution

This solution cannot represent a real density function. However, the regularized solution is much more plausible and is very close to the inverse-gamma distribution. The regularized solution is shown in Fig.13.

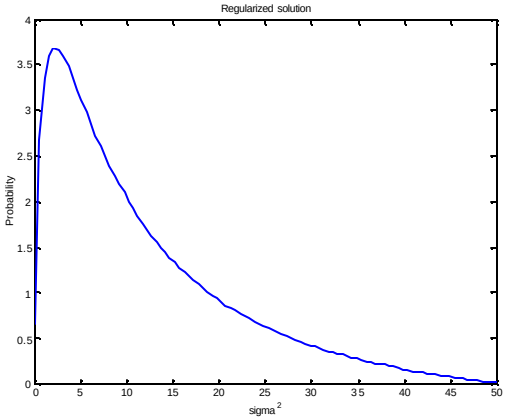


Fig. 13 Regularized solution

CONCLUSIONS

This paper presents a new inverse problem: inference of the prior distribution from the marginal or predictive distribution. The solution of this inverse problem requires the solution of the Fredholm integral equitation of the first kind, which can be effectively solved using Tikhonov regularization. The assumption about the smoothness of the sought solution is very legitimate in this case because the sought solution is a probability density function, which must be smooth by its nature. Two numerical examples for the inference of the prior distribution for the parameter were given: first of a binomial distribution and then for inference of the variance of normal distribution with known mean. The described approach may represent a valuable alternative to the selection of prior in practical applications and provides new insight into the nature of prior selection. One dimensional case is only analysed. In multidimensional case we would have to obtain prior for each individual parameter and then form the joint prior as a product of those individual priors using the argument about parameters independence.

REFERENCES

1. Christian P. Robert, *The Bayesian Choice. A Decision – Theoretic Motivation*. Springer-Verlag New York 1994.

2. J. Aitchison and I.R. Dunsmore, *Statistical Prediction Analysis*, Cambridge University Press, Cambridge, 1975

3. James O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer,1985, pp.94-95

4. D.J.C. MacKay, Bayesian Interpolation, *Neural Computation*, 4, (3) pp. 415-447

5. A.N. Tikhonov, 1963. Solution of incorrectly formulated problems and the regularization method, *Doklady Akad. Nauk USSR* 151, pp. 501-504.

6. V.A. Morozov, 1966. On the solution of functional equations by the method of regularization, *Soviet Math. Dokl.*, 7, pp.414-417

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

Optimal choice of descent steps in gradient type methods when applied to combined parameter and function or multi-function estimation

Tahar Loulou*
Ecole des Mines d'Albi-Carmaux
81000, Albi,
France
loulou@enstimac.fr

Eugene Artioukhine
IGE Parc Technologique
2 Avenue Jean Moulin, F-90000
Belfort, France
artyukh@ige.univ-fcomte.fr

Abstract

This paper discusses the implementation of the iteration algorithms for solving the general problem of recovering a complete set of thermal coefficients. It is well known that in the solution of inverse heat conduction problems it often becomes necessary to determine several independent functions or parameters at one time. An example of multi-function estimation is the inverse heat conduction problem, which uses transient temperature measurements to estimate the thermal dependent conductivity and specific heat of a given material. An example of combined parameter and function estimation is the determination of a constant heat transfer coefficient and time-wise varying heat source. Numerical algorithms based on gradient type-methods of minimization are often used in the estimation procedure. In such situations, these methods are less efficient and present low convergence rate. The use of a common descent parameter (step size) is at the origin of this problem. An optimal choice of vectorial descent parameter is introduced in this study and shows a considerable increase in the convergence rate. The developed algorithm was applied to different inverse heat conduction problems involving parameter and function or multi-function estimation. This approach appears to be effective for improving the computational efficiency of iterative algorithms for the two cases.

Nomenclature

$\mathcal{B}_i(t)$	trial function,
c_p	heat capacity of tissue,
$\boldsymbol{D}$	descent direction vector,
J	residual function,
$\nabla \boldsymbol{J}$	residual function gradient vector,
∇J_i	residual function gradient component,
k	thermal conductivity,
K	number of time step,
M	number of unknown parameters,
N	number of sensors,
q_i	component of unknown heat flux vector
$Q(t)$	unknown heat flux,
$Q_1(t)$	unknown heat flux,
$Q_2(t)$	unknown heat flux,
$T(x,t)$	computed temperature,
T_i	initial temperature,
$t,\ t_f$	time, final time,
$\boldsymbol{U}$	unknown vector,
$V(x,t)$	variation variable,
$x,\ L$	space, slab thickness
$Y_i(t)$	measured temperature,
–	–
β	parameter in descent direction,
δ	integrated measurement error,
Δ	small variation,
γ	descent parameter,
$\boldsymbol{\gamma}$	vector of descent parameter,
$\psi(x,t)$	adjoint variable,
ρ	density,
σ	standard deviation of measurement,
ω	random variable

Introduction

The subject of inverse problem has been an active area of research for the past several decades. This exciting field has found application in almost all disciplines of science and technology in general, and in heat transfer in particular. Different technics have been used to solve inverse problems including the conjugate gradient method [1, 2], the sequential estimation method [3, 4], the mollification method, and other several methods.

The present work deals with the implementation of the iteration algorithms for solving the general problem of recovering a complete set of thermal coefficients in a quasi-linear parabolic model. It is well known that in the solution of inverse heat conduction problems it often becomes necessary to determine several independent functions or parameters at one time. Such multi-parameter estimation problem arise in the solution of coefficient-type inverse problems. In the solution of inverse transfer

*To whom all correspondence should be addressed.

problems with one unknown (function or parameter) it has been found and proved very effective to use algorithms based on gradient type-methods of minimization. The use of these methods in a case when it is necessary to determine several independent variables becomes more difficult by the fact that the descent parameter (descent step) is chosen to be the same for all components of the direction of descent. Such a method of choosing a common step frequently leads to very slow or no convergence at all of the gradient-type methods. The convergence may be speeded up considerably by choosing different descent parameters for the different components of the gradient of the minimizing functional, i.e. to determine not only one common step but a vector of steps (descent parameter) from the condition that the target functional has a minimum with respect to this factor at each iteration. The developed algorithm was applied to different inverse heat conduction problems involving the estimation of combined parameter and function or two functions. The first problem deals with the estimation of constant thermal conductivity or specific heat and a time dependent heat flux. As second example two unknown surface heat fluxes are estimated simultaneously by utilizing temperature measurements collected inside a one dimensional slab. The third problem concerns the estimation of two plane heat sources within a finite wall. A comparison between the conjugate gradient method using a common descent step and the same method but with vectorial descent step in term of the convergence rate, the estimation error, and the CPU time is presented for each example. The developed approach is a modification of conventional optimization techniques of gradient type and appears to be effective enough for improving the computational efficiency of iterative algorithms for combined parameter and function or multi-function estimation problems.

This paper is divided in four major sections. The mathematical formulation of an inverse heat conduction problem and its resolution for estimating simultaneously one parameter and one function is shown in section two. The modification of the descent parameter from a common scalar for all parameters to be recovered to vector form is presented in section three. Numerical results of a systematic investigation of the method are given in section four with several examples. The last section presents some concluding remarks.

Inverse problem formulation

Generally, inverse heat conduction problems are solved by minimizing a residual functional $J(\boldsymbol{U})$ based on the ordinary least square norm and coupled with some stabilizing technic used in the iterative procedure of the estimation. The sum of the squared residuals between a given measured data and the responses of a model simulating the physical problem under investigation defines the least square norm. For continuous measured data, the residual functional is written as follows :

$$J(\boldsymbol{U}) = \sum_{i=1}^{N} \int_0^{t_f} \left[T(x_i, t; \boldsymbol{U}) - Y(x_i, t)\right]^2 dt \quad (1)$$

where $T(x_i, t; \boldsymbol{U})$ and $Y(x_i, t)$ are respectively the computed and the measured temperature, at the location x_i and over the time period $[0,\, t_f]$ corresponding to the duration of the experiment. Usually the measured temperatures are not continuous time function but are collected at known sensor locations and at discrete time steps, i.e. $t_k,\ k = 1, \dots K$. In the following sections, the computed and measured temperature are denoted $T_i^k = T(x_i, t_k)$ and $Y_i^k = Y(x_i, t_k)$.

The vector $\boldsymbol{U}$ can be a set of parameters and/or coefficients of basic functions used to approximate an unknown or more functions to be recovered by solving the inverse problem under consideration. As example let consider the following problem for estimating simultaneously one parameter and one function.

A slab of thickness L is initially at zero temperature. For time > 0, the boundary surface at $x = L$ is kept insulated, while that at $x = 0$ is subjected to prescribed heat flux $Q(t)$. The mathematical model for this one-dimensional transient heat conduction problem, with constant physical properties, is given as follows

$$\rho\, c_p \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2}, \qquad 0 < x < L \quad t > 0 \quad (2)$$

$$-k \frac{\partial T}{\partial x} = Q(t), \qquad x = 0 \qquad t > 0 \quad (3)$$

$$\frac{\partial T}{\partial x} = 0, \qquad x = L \qquad t > 0 \quad (4)$$

$$T(x, 0) = 0, \qquad 0 \le x \le L \quad (5)$$

Our objective is to estimate the unknown parameter k and the function $Q(t)$ from the transient temperature histories taken at two or more precise known sensor locations inside de slab i.e. $0 < x_i < L,\ i = 1, \dots N,\ N \ge 2$. For estimating a constant thermal conductivity k and a transient heat flux $Q(t)$, one way to construct $\boldsymbol{U}$ is :

$$\boldsymbol{U}^T = [k, q_1, q_2, \dots, q_m] \quad (6)$$

where the superscript T denotes the transpose and q_i are the coefficients of the following parametric representation of the unknown heat flux, i.e. :

$$Q(t) \,=\, \sum_{i=1}^{m} q_i \, \mathcal{B}_i(t) \tag{7}$$

The functions $\mathcal{B}_i(t)$ are any trial functions (polynomials, B-splines, ...), used to approximate the unknown function form of the heat flux $Q(t)$. In this special case, the total number of parameters to be recovered by the solution of the inverse problem is $M \,=\, 1 \,+\, m$.

As detailed in Özişik [1], Alifanov [2], and Jarny [5] the solution of an inverse problem with the conjugate gradient method involves the following basic steps : (a) the solution of the *direct problem*, (b) the solution of the *adjoint problem*, (c) the computation of the *gradient equation*, (d) the solution of the *variation problem*, (e) the choice of *stopping criterion*, and (f) the computational *algorithm*.

The minimization procedure of the functional (1) by utilizing the conjugate gradient method is built as follows [6] :

$$\boldsymbol{U}^{s+1} \,=\, \boldsymbol{U}^{s} \,+\, \gamma^s \, \boldsymbol{D}^s, \qquad s \,=\, 1,\, 2,\, \ldots \tag{8}$$

where the superscript s is the iteration number, γ^s is the *common descent parameter* given by :

$$\gamma^s = \frac{\displaystyle\sum_{i=1}^{N} \int_0^{t_f} \left[T(x_i,t) \,-\, Y(x_i,t)\right] \, V(x_i,t) dt}{\displaystyle\sum_{i=1}^{N} \int_0^{t_f} \left[V(x_i,t)\right]^2 dt} \tag{9}$$

The variable $V(x,t)$ is the solution of the *variation problem* in the case of estimating a function or the solution of the *sensitivity problem* when estimating a parameter.

For the problem under consideration, one can show [1, 2, 5] that the associated variation problem is given by :

$$\rho \, c_p \, \frac{\partial V}{\partial t} \,=\, k \, \frac{\partial^2 V}{\partial x^2} \,+\, \Delta k \, \frac{\partial^2 T}{\partial x^2} \tag{10}$$

$$0 < x < L \qquad t > 0$$

$$-\, k \, \frac{\partial V}{\partial x} \,-\, \Delta k \, \frac{\partial T}{\partial x} \,=\, \Delta Q(t) \tag{11}$$

$$x \,=\, 0 \qquad t > 0$$

$$\frac{\partial V}{\partial x} \,=\, 0, \qquad x \,=\, L \qquad t > 0 \tag{12}$$

$$V(x,0) \,=\, 0, \qquad 0 \le x \le L \tag{13}$$

In equation (8), the coefficient γ^s determines the step size in going from $\boldsymbol{U}^s$ to $\boldsymbol{U}^{s+1}$. It is computed by minimizing $J(\boldsymbol{U}^{s+1})$ given in equation (1) with respect to γ^s

$$\min_{\gamma^s} \int_0^{t_f} \left[T(\boldsymbol{U}^s \,+\, \gamma^s \, \boldsymbol{D}^s) \,-\, Y\right]^2 dt \tag{14}$$

Taylor series expansion are employed to develop an approximative formula to equation (14) and the obtained result is differentiated with respect to γ^s to get the expression (9).

In the step size expression, $\boldsymbol{D}^s$ represents the descent direction vector which is given by :

$$\boldsymbol{D}^s \,=\, -\,\nabla \boldsymbol{J}^s \,+\, \beta^s \, \boldsymbol{D}^{s-1} \tag{15}$$

where $\nabla \boldsymbol{J}$ is the gradient vector of $J(\boldsymbol{U})$ and the parameter β^s is given by :

$$\beta^s \,=\, \frac{<\, \nabla \boldsymbol{J}^s \,-\, \nabla \boldsymbol{J}^{s-1} \,,\, \nabla \boldsymbol{J}^s \,>}{<\, \nabla \boldsymbol{J}^s \,,\, \nabla \boldsymbol{J}^s \,>}, \quad \beta^0 = 0 \tag{16}$$

where $<\,,\,>$ is the scalar product defined in the space of real parameters. The above expression is known as Polak-Ribiere version of the conjugate gradient method [6]. Using the parametric form, the gradient of the residual functional (1) is given by the vector

$$\nabla \boldsymbol{J}^T \,=\, [\, \nabla J_1,\, \nabla J_2, \ldots,\, \nabla J_M \,] \tag{17}$$

For the special case mentioned above, the simultaneous estimation of k and $Q(t)$ and by considering the parametric representation of $Q(t)$, given in equation (7), it can be shown that the i^{th} component of the vector $\nabla \boldsymbol{J}$ has the following analytical expression :

$$\nabla J_i \,=\, \int_0^{t_f} \psi(0,t) \, \mathcal{B}_i(t) \, dt, \quad i = 1, \ldots m \tag{18}$$

It corresponds to the components of the gradient vector of $J(\boldsymbol{U})$ with respect to the parametric representation of the function $Q(t)$. The first component of the vector $\nabla \boldsymbol{J}$ corresponding to the gradient of $J(\boldsymbol{U})$ with respect to thermal conductivity k is given by :

$$\begin{aligned} \nabla J_1 \;&=\; \int_0^{t_f} \psi(0,t) \, \frac{\partial T(0,t)}{\partial x} \, dx \, dt \\ &+\; \int_0^{t_f} \int_0^{L} \psi(x,t) \, \frac{\partial^2 T(x,t)}{\partial x^2} \, dx \, dt \end{aligned} \tag{19}$$

where the variable $\psi(x,t)$ is solution of the so called *adjoint problem* and for this case, it is given by the following system :

$$-\, \rho \, c_p \, \frac{\partial \psi}{\partial t} \,=\, k \, \frac{\partial^2 \psi}{\partial x^2} \,+\, S(x,t) \tag{20}$$

$$0 < x < L \qquad 0 \le t < t_f$$

$$\frac{\partial \psi}{\partial x} = 0, \qquad x = 0 \quad 0 \le t < t_f \tag{21}$$

$$\frac{\partial \psi}{\partial x} = 0, \qquad x = L \quad 0 \le t < t_f \tag{22}$$

$$\psi(x, t_f) = 0, \qquad 0 \le x \le L \tag{23}$$

where the source term is given by $S(x,t) = 2\,[T(x_i,t) - Y(x_i,t)]\,\delta(x - x_i)$. δ represents the Dirac function. All the components of the inverse problem resolution are obtained. The iterative procedure can be applied to estimate k and $Q(t)$ following the numerical algorithm presented in [1, 2, 5].

Stopping criterion : In the absence of noise, the iterative process, equation (8), is repeated until each component of the vector $\boldsymbol{U}$ satisfies the following stopping criteria :

$$\left| \frac{u_i^{s+1} - u_i^s}{u_i^{s+1}} \right| \le \varepsilon, \qquad i = 1, \dots M \tag{24}$$

where ε is a small number ($10^{-4} \sim 10^{-6}$). In the event that the input temperatures are given with errors, the iterative process is stopped in accordance with the residual criterion [2], i.e. upon fulfillment of the following condition :

$$J(\boldsymbol{U}) \le \delta^2 \tag{25}$$

where δ^2 is given by

$$\delta^2 = \sum_{i=1}^{N} \int_0^{t_f} \sigma_i(t)\, dt \tag{26}$$

It represents the integrated error of the measured data at location x_i and having $\sigma_i(t)$ as standard deviation. Many iterative methods exhibit a *self-regularizing property* in the sense that early termination of the iterative process has a regularizing effect. In the iterative regularization method, the iteration index s plays the role of the regularizing parameter α used in Tikhonov's method [7], and the stopping rule ($J(\boldsymbol{U}) \le \delta^2$) plays the role of the parameter selection method.

Modification of the descent parameter

As reported in [8, 9, 10], the convergence of the conjugate gradient method may be altered by using the same descent parameter γ in the iterative process (8). Indeed, the preliminary numerical computations have shown that with the conventional choice of a descent parameter common to all unknown components of vector $\boldsymbol{U}$, the convergence of the presented method to the true values of the parameters depends strongly on the initial guess. Moreover, the convergence rate is strongly affected by the dependence between the separate unknowns. This problem is well discussed and explained in Beck's book [4] in term of sensitivity coefficient analysis. The parameter dependence which is known as *degree of correlation* is an inherent characteristic of any considered material and many parameter estimation technics can fail because of this characteristic. To overcome this difficulty, we develop in what follows a procedure presented in reference [2], for selecting the descent parameter in vector form with as many components as parameters and function or multi-function to be estimated. The descent vector will be denoted $\boldsymbol{\gamma}$.

For the test case considered above, the vector $\boldsymbol{\gamma}$ will contain two components γ_k and γ_Q (estimation of k and $Q(t)$). The basic idea is built on the linearity of the *variation* problem. Indeed, the total variation variable $V(x,t)$ defined in equations (10)-(13) can be regarded as the sum of two independent variation variables :

$$V(x,t) = V_1(x,t) + V_2(x,t) \tag{27}$$

where $V_1(x,t)$ is due to a small change in k, and $V_2(x,t)$ is due to the variation of $Q(t)$. Under this hypothesis, two "new" variation problems with respect to k and $Q(t)$ are introduced

$$\rho\, c_p \frac{\partial V_i}{\partial t} = k \frac{\partial^2 V_i}{\partial x^2} + Z_i(x,t) \tag{28}$$

$$0 < x < L, \qquad 0 < t \le t_f$$

$$-k \frac{\partial V_i}{\partial x} = X_i(t), \qquad x = 0 \quad t > 0 \tag{29}$$

$$\frac{\partial V_i}{\partial x} = 0, \qquad x = L \qquad t > 0 \tag{30}$$

$$V_i(x,0) = 0, \qquad 0 \le x \le L \tag{31}$$

where the two terms $Z_i(x,t)$ and $X_i(t)$ are given by :

$$Z_i(x,t) = \begin{cases} \dfrac{\partial^2 T(x,t)}{\partial x^2}\,\Delta k & \text{for } k\ (i=1) \\ 0. & \text{for } Q\ (i=2) \end{cases}$$

$$X_i(t) = \begin{cases} \dfrac{\partial T(0,t)}{\partial x}\,\Delta k & \text{for } k\ (i=1) \\ \Delta Q(t) & \text{for } Q\ (i=2) \end{cases}$$

As presented in references [1, 2] the descent parameter is obtained from the condition of minimizing the residual functional (1) with respect to the unknown to be recovered. The same approach

developed above (see expression (14)) to compute γ^s can be applied to calculate the descent vector components $\boldsymbol{\gamma}^T = [\gamma_k, \gamma_Q]$. The minimization of $J(\boldsymbol{U}^{s+1})$ with respect to $\boldsymbol{\gamma}^s$ is obtained from :

$$\min_{\boldsymbol{\gamma}^s} \int_0^{t_f} [T(\boldsymbol{U}^s + \boldsymbol{\gamma}^s \boldsymbol{D}^s) - Y]^2 \, dt \qquad (32)$$

and by using the linearity of the variation problem, i.e. the variation variable $V_i(x,t;\gamma_i D_i)$ is a linear function of γ_i which means :

$$V_i(x,t;\gamma_i D_i) = \gamma_i V_i(x,t;D_i), \quad i = 1,2 \qquad (33)$$

Equation (32) is differentiated with respect to each component of $\boldsymbol{\gamma}^s$ and the result is set equal to zero. Finally the differentiation results are rearranged to obtain the following set of linear algebraic equations :

$$\sum_{j=1}^{2} \alpha_{j,k} \gamma_j = \delta_k, \quad \text{where} \quad k = 1, \ldots 2 \qquad (34)$$

where $\alpha_{j,k}$ and δ_k are given by :

$$\alpha_{j,k} = \sum_{i=1}^{N} \int_0^{t_f} V_k(x_i,t) V_j(x_i,t) \, dt$$

$$\delta_k = -\sum_{i=1}^{N} \int_0^{t_f} [T(x_i,t) - Y(x_i,t)] \, V_k(x_i,t) \, dt$$

The descent parameter vector components $\boldsymbol{\gamma}^T = [\gamma_k, \gamma_Q]$ are obtained from the solution of equation system (34) by using any classical method of solving linear algebraic equations. We should mention here that the minimization algorithm remains the same except the step (d) (solution of the *variation problem*). In fact, instead of solving one variation problem, defined in equations (10)-(13), one should solve the "new" two variation problems defined in equations (28)-(31) to obtain respectively $V_1(x,t)$, and $V_2(x,t)$.

The determination of a vectorial descent parameter (step sizes) is the key-point to the inverse solution of simultaneously estimating combined parameters and functions or many functions because the rate of convergence can greatly improved in comparison with "*traditional ways*" of the conjugate gradient method which uses a common descent parameter to determine many parameters or functions or both.

Results and discussion

The accuracy and efficiency of the inverse analysis for simultaneously estimating a set of combined parameters and function or a set of functions is examined by conducting several test cases. All numerical simulations are performed for one-dimensional quasi-linear heat conduction problem in a slab of thickness $L = 1$ and over a time interval $t_f = 2$. Any one of several well-established analytical or numerical approaches can be used to solve the test cases under investigation. In this work we consider the finite difference method using an uniform space grid and a pure implicit time scheme. A mesh grid with 41 nodes in space and 101 in time is used for all the results presented below. A dimensionless space step of $\Delta x = 0.025$ and time step step $\Delta t = 0.02$ are used in the computations. In the iterative process, the maximum allowed number of iterations is $itmax = 500$. We prescribe the stopping criterion of 10^{-5} in expression (24) when the computation are run with errorless temperatures.

The simulated transient temperature data Y_i^k containing measurement errors are generated by adding random errors to the computed exact temperatures T_i^k as :

$$Y_i^k = T_i^k + \sigma \, \omega_{i,k} \quad \begin{cases} i = 1, \ldots N \\ k = 1, \ldots K \end{cases} \qquad (35)$$

where σ is the standard deviation of measurement errors which is assumed to be the same for all measurements, N is the number of sensors, and K is the number of measurements taken with each sensor i. For normally distributed random errors, there is a 99 % probability of the value of $\omega_{i,k}$ lying in the range :

$$-2.576 < \omega_{i,k} < +2.576 \qquad (36)$$

The values $\omega_{i,k}$ are generated randomly by the IMSL subroutine DRNNOR [11]. For each considered case two tests were performed, the first (1) with simulated measurements with standard deviation of $\sigma = 0$ (errorless measurements), the second (2) with $\sigma \neq 0$ (noisy data).

Estimation error : To quantify the relative error of the estimation procedure, the following definitions are introduced :

$$\varepsilon_p = \left| \frac{p - \bar{p}}{\bar{p}} \right| \times 100\% \qquad (37)$$

is the computation error for a given parameter p, and

$$\varepsilon_f = \frac{\int_0^{t_f} [\bar{f}(t) - f(t)]^2 dt}{\int_0^{t_f} [\bar{f}(t)]^2 dt} \times 100\% \qquad (38)$$

is the estimation error for a given function. The over-bar designates the exact parameter or function under hand. When the conjugate gradient method

is used with vectorial descent parameter, it is denoted VDP and CDP when a common descent parameter is employed.

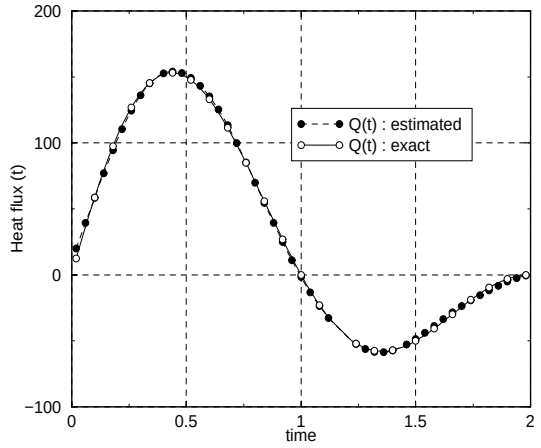


Figure 1: Computed and exact heat flux $Q(t)$ obtained with noisy data and estimated simultaneously with k

Test case 1 : Estimation of one parameter and one time dependent function. As first test example, we present the results obtained with the detailed problem given in the section *inverse problem formulation* and which consists in simultaneous estimation of k and $Q(t)$. The unknown heat flux $Q(t)$ is applied on the surface $x = 0$. The measurement data are collected by two (2) sensors placed at distinct locations $x_1 = 0.20$ and $x_2 = 0.80$ inside the slab. The specific heat is constant and set equal to 1. We present the inverse problem results for the following exact thermal conductivity and time-wise varying heat flux :

$$Q(t) = \frac{(t - t_f)\,\sin(-2\,\pi\,t/t_f)}{Q_{max}} \qquad k = 1 \quad (39)$$

where $Q_{max} = 10$. The unknown heat flux $Q(t)$ is parameterized according the equation (7) by utilizing cubic splines and taking $m = 12$ (number of trial functions).

The obtained results are summarized in table (1). The first remak is about the convergence with the used initial guess. The method using VDP converges in both cases of the utilized data, i.e. errorless and noisy. While the method with CDP doesn't converge at all. Even with noisy data, the obtained results with the modified descent parameter method are good and in acceptable agreement with the exact values. The recovered heat flux is plotted on figure (1). The error estimation is less than 1% for the shown cases. We should mention here that the conjugate gradient method with CDP converges when a better initial guess is used, for example $q_i^0 = 10.$ and $k^0 = 0.1$ but with high iteration number (> 500).

Test 1			
m	unknowns ⟶	$Q(t)$	k
VDP	initial guess	2.0	0.01
	meas. error σ	0.0	
	iteration number	**283**	
	CPU time	6.01	
	results	/	0.999982
	estimation error	0.00	0.00
CDP	initial guess	2.0	0.01
	meas. error σ	0.0	
	iteration number	no convergence	
m	unknowns ⟶	$Q(t)$	k
VDP	initial guess	2.0	0.01
	meas. error σ	1.0	
	iteration number	**96**	
	CPU time	2.05	
	results	fig. (1)	0.997005
	estimation error	0.04	0.29
CDP	initial guess	2.0	0.01
	meas. error σ	1.0	
	iteration number	no convergence	
m	unknowns ⟶	$Q(t)$	$\rho\,c_p$
VDP	initial guess	2.0	0.01
	meas. error σ	1.0	
	iteration number	**146**	
	CPU time	2.05	
	results	/	0.993915
	estimation error	0.11	0.60
CDP	initial guess	2.0	0.01
	meas. error σ	1.0	
	iteration number	no convergence	

Table 1: Results of estimating simultaneously $Q(t)$ and k and $Q(t)$ and $\rho\,c_p$ by utilizing errorless and noisy data.

With VDP method, the number of iteration drops to 147 with errorless data, and to 157 with noisy data when the following sensor locations $x_1 = 0.20$ and $x_2 = 1.00$ are considered. While we still having no convergence with the CDP method. This observation suggests that an experimental design investigation should be conducted to optimize the different factors involving in the estimation procedure : sensor location, experiment duration, optimal boundary condition, ...

On the same table, we show the results of estimating simultaneously the heat flux $Q(t)$ and $\rho\,c_p$ by using noisy data. The presented results are obtained after 146 iterations. With errorless temperatures, the exact results are reached after 80 it-

Test 2			
m	unknowns ⟶	$q_1(t)$	$q_2(t)$
VDP	initial guess	0.0	0.0
	meas. error σ	0.0	
	iteration number	**55**	
	CPU time	1.04	
	results	/	/
	estimation error	0.14	0.00
CDP	initial guess	0.0	0.0
	meas. error	0.0	
	iteration number	**250**	
	CPU time	3.13	
	results	/	/
	estimation error	0.18	0.00
m	unknowns ⟶	$q_1(t)$	$q_2(t)$
VDP	initial guess	0.0	0.0
	meas. error σ	0.01	
	iteration number	**51**	
	CPU time	2.07	
	results	fig.(2)	
	estimation error	0.30	0.00
CDP	initial guess	0.0	0.0
	meas. error σ	0.01	
	iteration number	**149**	
	CPU time	3.05	
	results	/	/
	estimation error	0.34	0.12

Table 2: Results of estimating simultaneously two heat fluxes $q_1(t)$ and $q_2(t)$ by utilizing errorless and noisy data.

erations (not shown to alleviate the table). The "best" sensor locations were found to be $x_1 = 0.30$ and $x_2 = 0.60$ for the considered heat flux shape and the utilized initial guess. The estimation error is of the same order of magnitude as the one observed in the previous case. In the second case, we have no convergence with or without errors in the simulated data when the CDP method is employed.

Test case 2 : Estimation of two time dependent heat fluxes. A slab of unit thickness is initially at zero temperature. For time > 0, the boundary surfaces at $x = 0$ and and $x = L$ are subject to two prescribed heat flux of strength $q_1(t)$ and $q_2(t)$.

The inverse heat conduction problem considered here is that of estimating simultaneously the two unknown time dependent surface heat flux $q_1(t)$ and $q_2(t)$ from the transient temperature recordings taken at two known sensor locations inside the considered domain $x_1 = 0.20$ and $x_2 = 0.80$. The maximum heat flux value is $Q_{max} = 10$. The two heat fluxes $q_1(t)$ and $q_2(t)$ have respectively triangular and rectangular shapes.

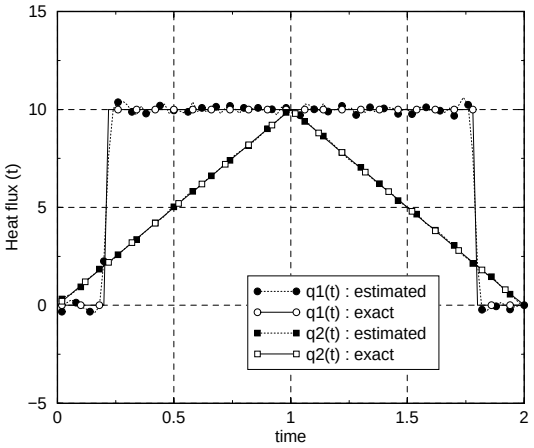


Figure 2: Computed and exact heat fluxes $q_1(t)$ and $q_2(t)$ obtained with noisy data

The results of the inverse estimation, with exact data and without any parametric representation of the two functions are presented in table (2). The results underline clearly the advantage of the vectorial descent parameter method when estimating simultaneously two time dependent heat fluxes. The ratio of iteration number between CDP method and VDP method for errorless data is about 5. This ratio drops to 3 in the case of noisy data. The CPU time is more important with the CDP method. The two estimated heat fluxes $q_1(t)$ and $q_2(t)$, with noisy data are plotted on figure (3). A comparison of the VDP and CDP methods reveals that accuracy is of the same order of magnitude, while the CDP method needs more iterations for convergence which results in an important CPU time.

Test case 3 : Estimation of two time dependent heat sources. A plate of thickness L initially at a uniform temperature $T_i = 0$ contains two plane heat sources of unknown strengths $S_1(t)$ and $S_2(t)$ placed at specified locations $x_1 = 0.20$ and $x_2 = 0.80$, respectively, inside the plate. For time $t > 0$, heat is generated by the sources at unknown rates, while the boundaries of the plate are kept insulated. Our goal is to estimate simultaneously the unknown strengths of the sources $S_1(t)$ and $S_2(t)$ from transient temperature measurements taken at both boundaries of the plate $x = 0$ and $x = L$. More details on the mathematical formulation of the above inverse problem, analytical derivation of the gradient, the computational algorithm and several numerical examples can be found in Silva Neto *et al.*[12] and are not repeated here for sake of brevity. In solving this problem we have used the same shapes and magnitude of the two sources presented in the above reference without any para-

Test 3			
m	unknowns ⟶	$S_1(t)$	$S_2(t)$
VDP	initial guess	0.0	0.0
	meas. error σ	0.10	
	iteration number	**34**	
	CPU time	1.88	
	results	/	/
	estimation error	0.89	4.95
CDP	initial guess	0.0	0.0
	meas. error σ	0.10	
	iteration number	**89**	
	CPU time	2.44	
	results	fig. (3)	
	estimation error	0.93	4.84

Table 3: Results of estimating simultaneously two heat sources $S_1(t)$ and $S_2(t)$ with noisy data.

metric form. The two heat sources $S_1(t)$ and $S_2(t)$ have rectangular shapes with a significantly different duration of energy releases.

Here again the VDP method was found to be 3 times faster than the CDP method in the case of noisy data. The agreement between the estimated and exact source strengths is good. The developed procedure remains valid in the case of multi-parameter estimation and one can show that it is reduced to the well known Newton-Gauss method [2, 4].

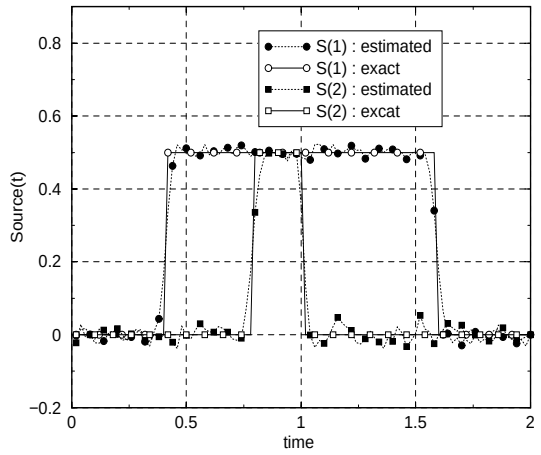


Figure 3: Computed and exact heat sources $S_1(t)$ and $S_2(t)$ obtained with noisy data

Conclusion

We have shown in this paper the application of the conjugate gradient method for estimating a set of parameters and functions or multi-functions with two kinds of descent parameter : **common** parameter or **vectorial** parameter. The obtained results illustrate the efficiency of the conjugate gradient method when applied with a vectorial descent parameter.

A comparison of the two variants of the conjugate gradient method (with VDP and with CDP) reveals that accuracy is of the same order of magnitude for both versions, while the CDP method needs more iterations for convergence resulting in more CPU time. An other important result is that the VDP method converges with wider deviation in the initial guess.

References

[1] Ozişik M.N. and Orlande H.R.B. *Inverse Heat Transfer : Fundamentals and Applications*. Taylor and Francis, Pennsylvania, 1999.

[2] Alifanov O.M., Artyukhin E.E. and Rumyantsev S.V. *Extreme Methods of Solving Ill-Posed Problems and theirs Applications to Inverse Heat Transfer Problems*. Begell House, New York, 1995.

[3] Beck J.V. Blackwell B. and St. Clair C.R. *Inverse Heat Conduction. Ill Posed Problems*. Wiley Interscience, New York, 1985.

[4] Beck J.V. and Arnold K.J. *Parameter Estimation in Engineering and Science*. Wiley Interscience, New York, 1977.

[5] Jarny Y., Özişik M.N. and Bardon J.P. A general optimization method using adjoint equation for solving multidimensional inverse heat conduction. *Int. J. Heat Mass Transfer*, 34(11):2911–2919, 1991.

[6] Polak E. *Computational Methods in Optimization*. Academic Press, New York, 1971.

[7] Tikhonov A.N. and Arsenin V.Y. *Solution of Ill-posed Problems*. Winston and Sons, Washington D.C., 1977.

[8] Artyukhin E.A., and Rumjantsev S.V. Optimal choice of descent steps in gradient methods of solution of inverse heat conduction problems. *Journal of Engineering Physics*, 39(2):865–869, 1980.

[9] Artyukhin E.A., and Okhapkin A.S. Determination of the parameters in the generalized heat-conduction equation from transient experimental data. *Journal of Engineering Physics*, 42(6):693–698, 1982.

[10] Artyukhin E.A., and Nenarokomov A.V. Coefficient inverse heat conduction problem. *Journal of Engineering Physics*, 53(3):1085–1090, 1987.

[11] IMSL. *Library Edition 10.0, User's Manual, Math/Library*. IMSL, 7500 Ballaire Blvd., Houston, Texas, 1987.

[12] Silva Neta A.J. and Özişik M.N. Inverse problem of simultaneously estimating the timewise-varying strenghts of two plane sources . *J. Applied Physics*, 73(5):2132–2137, March 1993.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

INTERIOR POINT ALGORITHMS FOR NONLINEAR CONSTRAINED LEAST SQUARES PROBLEMS

José Herskovits*,
Veranise Dubeux*
**Mechanical Engineering Program, COPPE*
Federal University of Rio de Janeiro, UFRJ
Rio de Janeiro, RJ, Brazil.
jose@ optimize.ufrj.br
veranise@ optimize.ufrj.br

Cristovão M. Mota Soares,**
Aurélio L. Araújo***
*** IDMEC/IST, Institute of Mechanical Engineering, Pole IST, Lisbon, Portugal.*
**** ESTIG, Polytechnic Institute of Bragança, Portugal.*
cmmsoares@alfa.ist.utl.pt
aaraujo@ipb.pt

ABSTRACT

We consider Nonlinear Least Squares problems with equality and inequality constraints and propose a numerical technique that integrates methods for unconstrained problems, based on Gauss-Newton algorithm, with FAIPA, the Feasible Arc Interior Point Algorithm for constrained optimization. We also present some numerical results on test problems available in the literature and compare them with the quasi-Newton version of FAIPA. We also describe an application to the identification of mechanical parameters of composite materials. The present algorithms are globally convergent, very robust and efficient.

INTRODUCTION

In this paper we consider Nonlinear Least Squares Problems with equality and inequality constraints, when nonlinear smooth functions are involved. Calling $x \equiv [x_1, x_2, \ldots, x_n]$ the design variables, $f(x)$ the objective function, $g(x) \equiv [g_1(x), g_2(x), \ldots, g_m(x)]$ the inequality constraints and $h(x) \equiv [h_1(x), h_2(x), \ldots, h_p(x)]$ the equality constraints, the problem can be denoted as:

$$\left.\begin{array}{ll} \underset{x}{\text{minimize}} & f(x),\ x \in R^n \\ \text{subject to} & g_i(x) \le 0;\ i = 1,\ldots,m \\ \text{and} & h_i(x) = 0;\ i = 1,\ldots,p \end{array}\right\} \qquad (1)$$

The function $f(x)$ is a sum of squares of the nonlinear functions $r_i(x)$; $i = 1,\ldots, s$.

$$f(x) = \frac{1}{2}\sum_{i=1}^{s}[r_i(x)]^2 = \frac{1}{2}\|r(x)\|_2^2 \qquad (2)$$

Problems of this type occur when fitting model functions to experimental data [1, 2]. In this case $r_i(x)$ is called a residual function. It represents the discrepancy between the true value and the approximate value, predicted by a nonlinear model. If the model is to have any validity, we can expect that $\|f(x^*)\|$ will be "small", and that s, the number of data points, will be much greater than n. We assume that $s > n$. Note that, if the set of equality constrains verifies regularity conditions [3], to have a solution it must be $p \le n$.

A large number of special purpose algorithms is available in the unconstrained case, but only very few methods were developed for the nonlinearly constrained case [4, 5, 6].

A numerical technique that integrates well-known methods for unconstrained problems in a general method for Nonlinear Constrained Optimization is presented in this paper. This method is the Feasible Arc Interior Point Algorithm, "FAIPA", that makes iterations in the primal and dual variables of the optimization problem to solve Karush-Kuhn-Tucker optimality conditions. Given an initial interior point, FAIPA defines a sequence of interior points with the objective reduced at each of the iterations. At each point, a feasible descent arc is obtained and an inexact line search is done along this arc. To compute the feasible arc, FAIPA solves three linear systems with the same matrix. These systems include the second derivative of the Lagrangian function. There is also a quasi-Newton version of FAIPA. In this one, the Hessian of the Lagrangian is replaced by a quasi-Newton approximation.

In the present algorithm, instead of the Hessian, we employ an approximation based on

Gauss-Newton method and some of their modifications. In the following sections we describe FAIPA, some existing methods for Least Square and we present the algorithm proposed here. Finally we describe the numerical results on some test problems and a practical application in solid mechanics.

FAIPA, THE FEASIBLE ARC INTERIOR POINT ALGORITHM

FAIPA, proposed by Herskovits [3, 7, 8], is an interior point method that solves general problems of nonlinear optimization. FAIPA makes interactions in the primal and dual variables of the optimization problem to solve Karush - Kuhn - Tucker (KKT) optimality conditions.

KKT conditions corresponding to Problem (1) can be written as follows:

$$\nabla f(x)+\nabla g(x)\lambda+\nabla h(x)\mu=0 \quad (3)$$
$$G(x)\lambda=0 \quad (4)$$
$$\lambda\geq 0 \quad (5)$$
$$g(x)\leq 0 \quad (6)$$
$$h(x)=0, \quad (7)$$
$$(8)$$

where $\lambda\in R^m$ and $\mu\in R^p$ are the Lagrange multipliers corresponding to the inequality and the equality constraints respectively, $G(x)\in R^{m\times m}$ denotes a diagonal matrix such that $G_{ii}(x)=g_i(x)$. In what follows we call $\Lambda\in \Re^{m\times m}$ a diagonal matrix with $\Lambda_{ii}=\lambda_i$.

FAIPA requires a feasible initial point and defines a sequence of feasible points, with a monotone reduction of the objective function.

The Feasible Arc Interior Point Algorithm to solve Problem (1) is described now:

FAIPA ALGORITHM

Parameter. $\alpha\in(0,1)$.

Data. $x\in\Omega_a^0$, $\lambda_i>0$, $\lambda\in R^m$, $\mu_i>0$, $\mu\in R^p$, $B\in R^{n\times n}$ symmetric and positive definite and $c_i=0$, $c\in R^p$.

Step 1. Computation of a feasible descent direction.

(i) Solve the linear system in (d_0, λ_0, μ_0):

$$\begin{bmatrix} B & \nabla g^t(x) & \nabla h^t(x)\\ \Lambda\nabla g(x) & G(x) & 0\\ \nabla h(x) & 0 & 0\end{bmatrix}\begin{bmatrix} d_0\\ \lambda_0\\ \mu_0\end{bmatrix}=-\begin{bmatrix}\nabla f(x)\\ 0\\ h(x)\end{bmatrix} \quad (9)$$

where $d_0\in R^n$, $\lambda_0\in R^m$, $\mu_0\in R^p$.

if $d_0=0$, stop.

(ii) Solve the linear system in (d_1, λ_1, μ_1):

$$\begin{bmatrix} B & \nabla g^t(x) & \nabla h^t(x)\\ \Lambda\nabla g(x) & G & 0\\ \nabla h(x) & 0 & 0\end{bmatrix}\begin{bmatrix} d_1\\ \lambda_1\\ \mu_1\end{bmatrix}=-\begin{bmatrix}0\\ \lambda\\ \mu\end{bmatrix} \quad (10)$$

where $d_1\in R^n$, $\lambda_1\in R^m$, $\mu_1\in R^p$.

(iii) If $c_i\leq\|\mu_{0i}\|$, make $c_i>1.2\|\mu_{0i}\|$, for $i=1,...,p$.

(iv) Let be

$$\phi(x,c)=f(x)+c^t|h(x)| \quad (11)$$

if $d_1^t\nabla\phi(x,c)>0$, set:

$$\rho=\inf\left[\|d_0\|^2;\frac{(1-\alpha)d_0^t\nabla\phi(x,c)}{d_1^t\nabla\phi(x,c)}\right] \quad (12)$$

else

$$\rho=\|d_0\|^2. \quad (13)$$

(v) Compute d

$$d=d_0+\rho d_1 \quad (14)$$

Step 2. Computation of a feasible descent arc.

(i) Let be

$$\tilde{w}_i^I=g_i(\mathrm{x}+\mathrm{d})-g_i(x)-\nabla g_i(x)^t d \quad (15)$$
where: $i=1,...,m$;

$$\tilde{w}_i^E=h_i(x+d)-h_i(x)-\nabla h_i(x)^t d \quad (16)$$
where: $i=1,...,p$.

(ii) Solve the linear system in $(\tilde{d}, \tilde{\lambda}, \tilde{\mu})$:

$$\begin{bmatrix} B & \nabla g^t(x) & \nabla h^t(x) \\ \Lambda \nabla g(x) & G(x) & 0 \\ \nabla h(x) & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{\lambda} \\ \tilde{\mu} \end{bmatrix} = - \begin{bmatrix} 0 \\ \lambda \tilde{w}^I \\ \mu \tilde{w}^E \end{bmatrix} \quad (17)$$

(iii) Find a step length t satisfying a given line search criterion on the auxiliary function $\phi(x,c)$ such that:

$$g_i(x + td + t^2 \tilde{d}) < 0 \text{ if } \tilde{\lambda}_i \geq 0, \quad (18)$$

or

$$g_i(x + td + t^2 \tilde{d}) < g_i(x) \quad (19)$$

otherwise.

Step 3. Updates.

(i) Set

$$x_{k+1} = x_k + td_k + t^2 \tilde{d}^2 \quad (20)$$

and define new values for: $w > 0$, $\lambda > 0$, $\mu > 0$ and B symmetric and positive definite.

(ii) Go to back to step1.

The size of linear systems (10) and (11) is equal to the sum of the number of variables plus the number of equality and inequality constrains. In [9] it is provide that (10) and (11) had a unique solution.

In Figure1 the Feasible Arc is represented in the case when there is an active inequality constraint, that is $g_i(x_k) = 0$. It is proved that it is possible to walk from x_k along the arc to get a new feasible point with a lower objective value. The algorithm has global convergence for any B symmetric and positive definite. However, taking $B \equiv H(x, \lambda, \mu)$, where

$$H(x,\lambda,\mu) = \nabla^2 f(x) + \sum_{i=1}^{m} \lambda_i \nabla^2 g_i(x) + \sum_{i=1}^{p} \mu_i \nabla^2 h_i(x) \quad (21)$$

is the second derivative of the Lagrangian, a Newton algorithm is obtained. A very efficient algorithm, without need of second derivatives computation, is obtained with B equal to a quasi-Newton approximation of $H(x, \lambda, \mu)$.

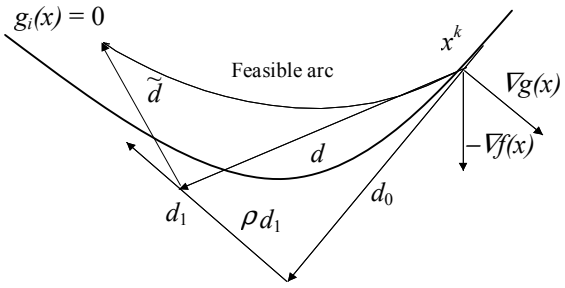


Figure 1: Feasible Arc.

ABOUT THE UNCONSTRAINED LEAST SQUARES PROBLEM

To understand the basic features of the algorithm present here, we consider the unconstrained nonlinear least square problem:

$$\min_{x \in \Re^n} f(x) = \frac{1}{2} \sum_{i=1}^{s} r_i(x)^2 \quad (22)$$

where $r(x)$ represent the residual vector.

The Jacobian Matrix of the residual is

$$J(x) = \begin{bmatrix} \frac{\partial r_1(x)}{\partial x_1} & \cdots & \frac{\partial r_1(x)}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial r_s(x)}{\partial x_1} & \cdots & \frac{\partial r_s(x)}{\partial x_n} \end{bmatrix}, \quad (23)$$

and the Hessian matrix of $f(x)$

$$\nabla^2 f(x) = J(x)^t J(x) + Q(x) \quad (24)$$

Where

$$Q(x) := \sum_{i=1}^{l} r_i(x) \nabla^2 r_i(x). \quad (25)$$

In the Gauss-Newton method, $Q(x)$ is ignored and the Hessian is simply approximate by

$$\nabla^2 f(x) \approx J(x)^t J(x). \tag{26}$$

The iterations for Gauss-Newton method are then

$$J(x_k)^t J(x_k)(x_{k+1} - x_k) = -J(x_k)^t r(x_k) \tag{27}$$

Gauss-Newton method is based on Newton's method and it can fail for the same reasons as Newton's method does. In particular, when $J(x_k)^t J(x_k)$ is not positive definite or when it is badly conditioned.

Gauss-Newton method assumes that, near of the solution, $J(x)^t J(x)$ is a good approximation to $\nabla^2 f(x)$, i. e. $Q(x)$ can be neglected. This assumption is not justified for problems with a large residual. A possible strategy in this case, is to include a quasi-Newton approximation M of the unknown second derivative term $Q(x)$ [4].

The search direction with a quasi-Newton approximation to $Q(x)$, called M, is given by:

$$[J(x_k)^t J(x_k) + M_k]d_k = -J(x_k)^t r(x_k). \tag{28}$$

Let be

$$s_k = (x_{k+1} - x_k) \tag{29}$$

$$y_k = J(x_{k+1})^t r(x_{k+1}) - J(x_k)^t r(x_k) \tag{30}$$

The following formula for M is based on the BFGS update [9]:

$$M_{K+1} = M_k - \frac{1}{s_k^t W_k s_k} W_k s_k s_k^t W_k + \frac{1}{y_k^t s_k} y_k y_k^t \tag{31}$$

where

$$W_k = J(x_{k+1})^t J(x_{k+1}) + M_k \tag{32}$$

It is proved that if it is ensured that $y_k{}^t s_k > 0$, then the updating formula has the property that if $J(x_{k+1})^t J(x_{k+1}) + M_k$ is a positive-definite matrix, then so it is $J(x_{k+1})^t J(x_{k+1}) + M_{k+1}$, see [4]. This property is used asymptotically when $J_k^t J_k$ is approximately equal to $J_{k+1}^t J_{k+1}$. However, $J(x_{k+1})^t J(x_{k+1}) + M_k$ can be singular or badly conditioned, resulting in a non-descent search direction, and the iteration fails. Levenberg – Marquardt method consists on adding a positive diagonal matrix εI where $\varepsilon > 0$ is taken big enough to have $J(x_{k+1})^t J(x_{k+1}) + M_k + \varepsilon I$ positive definite. The main difficulty to apply this technique is to get a way of choosing ε not very large in order to maintain as well as possible the speed of convergence of Gauss – Newton algorithm.

LEVENBERG – MARQUARDT METHOD WITH CHOLESKY DECOMPOSITION

Let be matrix $B_k = J(x_{k+1})^t J(x_{k+1}) + M_{k+1}$. If B is symmetric and positive-definite, it can be obtained a Cholesky factorization

$$B = LL^t \tag{33}$$

where L is lower-triangular matrix.

The modified Cholesky factorization is a numerically stable method to compute ε that produces a positive-definite matrix [4].

The elements of L can be expressed by a simple recurrence relation:

$$l_{ki} = \frac{b_{ki} - \sum_{j=1}^{i-1} l_{ij} l_{kj}}{l_{ii}} \tag{34}$$

for $i = 1, 2, \ldots, k\text{-}1$

and

$$l_{kk} = \sqrt{b_{kk} - \sum_{j=1}^{k-1} l_{kj}^2} \tag{35}$$

In the case when B is not positive definite, it is proved that one or more diagonal elements are such that

$$b_{kk} - \sum_{j=1}^{k+1} l_{kj}^2 \le 0. \tag{36}$$

In consequence, l_{kk} obtained in (35) is not a real number

Adding to b_{kk} a big enough positive number, a positive definite matrix B^+ is then obtained. This procedure is equivalent to Levenberg – Marquardt and allows to define very precisely the perturbation required to get a positive definite matrix.

ABOUT CONSTRAINED LEAST SQUARE PROBLEMS

The algorithm that we propose here is based on FAIPA. Instead of taking B equal to a quasi – Newton approximation of $H(x,\lambda,\mu)$, we construct a matrix that includes a Gauss-Newton approximation of the objective function.

$$H(x,\lambda,\mu) = \nabla^2 f_i(x) + \sum_{i=1}^{m} \lambda_i \nabla^2 g_i(x) + \sum_{i=1}^{p} \mu_i \nabla^2 h_i(x) \quad (37).$$

We employ the same update formula (31), but taking

$$y_k = \nabla l(x_{k+1},\lambda_{k+1},\mu_{k+1}) - \nabla l(x_k,\lambda_{k+1},\mu_{k+1}), \quad (38)$$

where

$$\nabla l(x,\lambda,\mu) = \nabla f(x) + \nabla g(x)\lambda + \nabla h(x)\mu. \quad (39)$$

In unconstrained optimization it is proved that $\nabla^2 f(x)$ is positive definite at a local minimum. When there are constraints, we have that in general $H(x,\lambda,\mu)$ is not positive-definite. In effect, it is only ensured that $H(x,\lambda,\mu)$ at a local solution is positive definite in the space tangent to the active constraints. However, FAIPA requires a positive definite matrix *B*.

We employ Levenberg-Marquardt method with Cholesky decomposition to obtain *B* positive definite.

NUMERICAL TESTS

We present some numerical results obtained with the algorithm for Constrained Least Squares problems, FAIPA_LS presented in this contribution. These results are compared with a quasi – Newton version of FAIPA.

We also describe an application to an inverse problem in solids mechanics.

Problem 25:

Source: Holzmann [10], Himmelblau [11].
Objective Function:

$$f(x) = \sum_{i=1}^{99} (r_i(x))^2$$

$$r_i(x) = -0.01i + \exp\left(-\frac{1}{x_1}(u_i - x_2)^{x_3}\right)$$

$$u_i = 25 + (-50\ln(0.01i))^{2/3}$$

$$i = 1,...,99.$$

Constraints:

$$0.1 \le x_1 \le 100$$
$$0 \le x_2 \le 25.6$$
$$0 \le x_3 \le 5$$

Start (feasible):

$$x_0 = (100,12.5,3)$$
$$f(x_0) = 32.835$$

Problem 57:

Source: Betts [12], Gould [13].
Objective Function:

$$f(x) = \sum_{i=1}^{44} (r_i(x))^2$$

$$r_i(x) = b_i - x_1 - (0.49 - x_1)\exp(x_2(a_i - 8))$$

$$i = 1,...,44.$$

a_i, b_i : appendix A of [14]

Constraints:

$$0.49 - x_1 x_2 - 0.09 \ge 0$$
$$0.4 \le x_1$$
$$-4 \le x_2$$

Start (feasible):

$$x_0 = (0.42,5)$$
$$f(x_0) = 0.030798602$$

Problem 70:

Source: Himmelblau [11, 14].
Objective Function:

$$f(x) = \sum_{i=1}^{19} (y_{i,cal} - y_{i,obs})^2$$

$$y_{i,cal} = \left(\frac{12x_1}{1+12x_2}\right)\left[x_3 b^{x_2}\left(\frac{x_2}{6.2832}\right)^{0.5}\left(\frac{c_i}{7.685}\right)^{x_2-1}\exp\left(\frac{x_2 - bc_i x_2}{7.658}\right)\right] + \left(\frac{12x_1}{1+12x_2}\right)\left[(1-x_3)\left(\frac{b}{x_4}\right)^{x_1}\left(\frac{x_1}{6.2832}\right)^{0.5}\left(\frac{c_i}{7.658}\right)^{x_1-1}\exp\left(\frac{x_1 - bcx_1}{7.658x_4}\right)\right]$$

$$b = x_3 + (1-x_3)x_4$$

c_i, $y_{i,obs}$: appendix A of [14]

Constraints:

$$0.49 - x_1 x_2 - 0.09 \geq 0$$
$$0.4 \leq x_1$$
$$-4 \leq x_2$$

Start (feasible):

$$x_0 = (0.42, 5)$$
$$f(x_0) = 0.030798602$$

Table 1. Numerical Results on Problems Hock/Schittkowski and Linear Equality Constrained Least Square Problem.

Problem 25 (n = 3, m = 6, p = 0)

Update of B	cfv	ofv	iter
FAIPA_qN	1.07239×10^{-5}	0.0	14
FAIPA_LS	3.89665×10^{-5}	0.0	13

Problem 57 (n = 2, m = 3, p = 0)

Update of B	cfv	ofv	iter
FAIPA_qN	0.0284597	0.0284596	21
FAIPA_LS	0.0284598	0.0284596	17

Problem 70 (n = 4, m = 9, p = 0)

Update of B	cfv	ofv	iter
FAIPA_qN	0.00749877	0.00749864	72
FAIPA_LS	0.00749847	0.00749864	33

Linear Problem (n = 4, m = 0, p = 3)

Update of B	cfv	ofv	iter
FAIPA_qN	3.08149×10^{-31}	0.0	5
FAIPA_LS	0.0	0.0	4

We report here our experience with 4 test problems. Three problems compiled by Hock et. al. [15] and the last one problem is a linear equality constrained least square (LSE) problem described in (http://www.netlib.org/lapack/lug/node85.html). The LSE problem is

$$\min_x \|Ax - b\| \text{ subject to } Bx = d$$

where

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 3 \\ 1 & 1 & 1 & -1 \end{bmatrix},\ b = \begin{bmatrix} 2 \\ 1 \\ 6 \\ 3 \\ 1 \end{bmatrix},$$

$$B = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix} \text{ and } d = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

The results are summarized in Table 1, where n is the number of variables, m the number of inequality constrains, p the number of equality constrains, iter is the number of iteration; cfv is the computed objective function value, ofv is the optimum function value, FAIPA_qN is the quasi-Newton version of FAIPA and FAIPA_LS, the present algorithm. All test problems were solved with the same value for the parameter α. This was taken: α = 0.7, as in the general version of FAIPA. The stop criterion adopted was a tolerance on the optimal objective function ε = 10^{-5}. The initial point, ofv and other characteristic of the test problems are described in Hock et. al. [15].

Identification of material parameters:

This example is intended to illustrate the application of the described optimization techniques to a class of inverse problems, namely the identification or estimation of material parameters in composite laminated plates made of two different materials. The problem consists of estimating the elastic properties of the two materials that make up the plate by fitting a set of experimentally measured undamped eigenvalues ($\tilde{\lambda}_i$) to those obtained through a higher order finite element model (λ_i).

The objective function is a weighted least squares estimator:

$$f(x)=\sum_{i=1}^{I} w_i\left(\frac{\tilde{\lambda}_i-\lambda_i(x)}{\tilde{\lambda}_i}\right)^2 \qquad (40)$$

where $w_i \in [0,1]$ expresses the confidence in the experimental data and, in this example it is taken as unity. The problem is then formulated as a non linear constrained minimization problem, where the design variables are non dimensional functions of the elastic properties of each material and the constraints are imposed in order to keep the constitutive matrix positive definite:

$$\begin{aligned} &min\ f(x)\geq 0 \\ &s.t.\quad g(x)\leq 0 \\ &x^l \leq x \leq x^u \end{aligned} \qquad (41)$$

Full details regarding this identification technique can be found in Araújo et al. [2,16].

The plate in this example is made of unidirectional layers of E glass and T300 carbon fibres in epoxy matrix. The pre-pregs used to build the plate were Structil 200g/m2 VEE220 R368, for the glass layers and Structil 350g/m2 CTE235 R367, for the carbon layers. The stacking sequence is $\left[90^0_{4C},\ 0^0_{3V}\right]_S$ and the rectangular plate dimensions and mass are a=191 mm, b=254 mm, h=3.89 mm and m=289.85g.

The initial estimates for the elastic properties of the glass and carbon layers correspond to the properties of typical unidirectional layers of these materials for 50% V_f:

E glass: ${}^0E_1 = 45GPa$; ${}^0E_2 = 4.5GPa$;
${}^0G_{12}={}^0G_{23}={}^0G_{13} = 3.7GPa$; ${}^0\nu_{12} = 0.28$.

T300 carbon: ${}^0E_1 = 117.2GPa$; ${}^0E_2 = 8.8GPa$;
${}^0G_{12}={}^0G_{23}={}^0G_{13} = 3.1GPa$; ${}^0\nu_{12} = 0.35$.

For the finite element discretisation a regular 12×16 mesh was used and the problem was solved in 17 iterations using the FAIPA (Wolfe criterion for line search) and the stopping criterion was the reduction of the penalty function (less than 1×10^{-6}). Results are presented in Tables 2 through 4. Residuals $r_{\omega i}$ on the natural frequencies were obtained from measured ($\tilde{\omega}_i = \sqrt{\tilde{\lambda}_i}/2\pi$) and identified ($\omega_i = \sqrt{\lambda_i(x)}/2\pi$) natural frequencies, using the following expression:

$$r_{\omega i} = \frac{\tilde{\omega}_i - \omega_i}{\tilde{\omega}_i}\times 100 \qquad (42)$$

A good agreement is sought between the identified global properties and their available strain gauge counterparts and one can conclude that the identified properties for each material are reasonably within what one could expect for these materials, except for the transverse shear modulus G_{13} and G_{23}, because the plate is not thick enough for these shear effects to be noticeable, hence any results for their identification are not truly reliable [2, 16]. Also, the different material densities were not taken into account in this example, which could in part explain the inability to fit the fifth natural frequency with sufficient accuracy.

Table 2. Identified global properties and strain gauge measurements

	Identified	Strain gauge
E_x [GPa]	17.0	—
E_y [GPa]	76.9	77.5
G_{xy} [GPa]	4.0	—
G_{xz} [GPa]	1.1	—
G_{yz} [GPa]	3.8	—
ν_{yx}	0.17	0.14-0.20

Table 3. Identified properties per material

	E glass	T300 carbon
E_1 [GPa]	44.6	100.1
E_2 [GPa]	4.7	7.7
G_{12} [GPa]	3.8	4.0
G_{13} [GPa]	3.4	3.7
G_{23} [GPa]	4.0	0.4
ν_{12}	0.27	0.45

Table 4. Experimental frequencies and residuals obtained after identification

i	$\tilde{\omega}_i$ [Hz]	$r_{\omega i}$ [%]
1	135.75	0.767
2	253.57	-0.036
3	373.81	0.147
4	489.90	0.801
5	567.68	-1.356
6	699.55	0.451
7	787.18	-0.542
8	809.46	-0.267
9	1195.0	-0.578
10	1310.0	0.508
11	1370.0	0.006

CONCLUSIONS

The present is a strong an efficient technique that extends to constrained problems the advantages of Gauss-Newton methods. The numerical results studied here show an improvement of the computer effort when compared with the classical quasi – Newton version of FAIPA.

We note that FAIPA is very robust and efficient and it was tested with more than 100 test problems in the literature and was applied in several practical applications [2, 16]. Unfortunately, we didn't find more Constrained Least Squares test problems.

ACKNOWLEDGMENTS

The authors wish to acknowledgments CNPq (Brazil), FAPERJ (Brazil) and FCT/ICCTI (Portugal) for the financial support provided to this research.

REFERENCES

1. A. L. Araújo, C. M. Mota Soares & M. J. Moreira de Freitas, *Characterization of Material Parameters of Composite Specimens Using Optimization and Experimental Data*. Composites: part B, Vol. 27B (2), p. 185-191, 1996.

2. A. L. Araújo, C. M. Mota Soares & M. J. Moreira de Freitas, P. Pedersen, J. Herskovits, *Combined numerical-experimental model for the identification of mechanical properties of laminated structures*. Composite Structures, Vol. 50, p. 363-372, 2000.

3. J. Herskovits, *A Feasible Arc Interior Point Technique for Nonlinear Optimization*, JOTA – Journal of Optimization Theory and Applications, Vol. 99, N1, p. 121-146, October, 1998.

4. P. E. Gill and W. Murray, *Algorithms for the Solution of the Nonlinear Least-Squares Problem*, SIAM Journal on Numerical Analysis, Vol. 15, No 5, p.977-992, October, 1978.

5. P. E. Gill, W. Murray and M. H. Wright. "*Practical Optimization*", A.P., 1981.

6. K. Schittkowski, *Solving Constrained Nonlinear least Square Problems by a General Purpose SQP-Method*, International Series of Numerical Mathematics, Birkhauseer Verlag Basel, Vol. 84 (c), p. 295-309, 1988.

7. J. Herskovits, *A View on Nonlinear Optimization*, Cap. Advances in Structural Optimization, J. Herskovits Ed., KLUWER Academic Publishers, Holland, p. 71-116, June, 1995.

8. J. Herskovits and G. Santos, *Feasible Arc Interior Point Algorithms for Nonlinear Optimization*, Fourth World Congress on Computational Mechanics, (in CD-ROM), Buenos Aires, Argentina, June, 1998.

9. D. G. Luenberger, *Linear and Nonlinear Programming*, 2ª ed., Sddilsson-Wesley, 1984.

10. G. Holzman, *Comparative analysis of nonlinear programming codes with the Weisman algorithm*, SRCC Report No. 113, University of Pittsburgh, Pittsburgh, 1969.

11. D. M. Himmelblau, *Applied nonlinear programming*, Mc-Graw Hill book-Company, New York, 1972.

12. J. T. Betts, *An accelerated multiplier method for nonlinear programming*, Journal of Optimization Theory and Applications, Vol. 21, No. 2, 147-174, 1977.

13. F. J. Gould, *Nonlinear tolerance programming, in: Numerical Methods for Nonlinear Optimization Theory and Applications*, Vol. 16, No. 1/2, 49, 66, 1975.

14. D. M. Himmelblau and Yates R. V. , *A new method of flow routing*, Water Resources Reseach, Vol. 4, p.1193, New York, 1968.

15. W. Hock, Schittkowski K. *Test Examples for Nonlinear Programming Codes*. Lecture Notes in Economics and mathematical Systems. No. 187. Springer-Verlag Berlin Heidelberg New York, 1981.

16. A. L. Araújo, C. M. Mota Soares, J. Herskovits, P. Pedersen, *Development of a finite element model for the identification of mechanical and piezoelectric properties through gradient optimization and experimental vibration data*. Composite Structures, to be published.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

Selection of Multiple Regularization Parameters in Local Ridge Regression Using Evolutionary Algorithms and Prediction Risk Optimization

J. Wesley Hines,
Andrei V. Gribok, **Aleksey M. Urmanov**
Department of Nuclear Engineering
The University of Tennessee
Knoxville, TN USA
jhines2@utk.edu, agribok@utk.edu, urmanov@utk.edu

Mark A. Buckner
Engineering Science and Technology Division
Oak Ridge National Laboratory
Oak Ridge, TN USA
buk@ornl.gov

ABSTRACT

This paper presents a new methodology for regularizing data-based predictive models. Traditional modeling using regression can produce unrepeatable, unstable, or noisy predictions when the inputs are highly correlated. Ridge regression is a regularization technique used to deal with those problems. A drawback of ridge regression is that it optimizes a single regularization parameter while the methodology presented in this paper optimizes several local regularization parameters that operate independently on each component. This method allows components with significant predictive power to be passed while components with low predictive power are damped. The optimal combination of regularization parameters are computed using an Evolutionary Strategy search technique with the objective function being a predictive error estimate. Examples are presented to demonstrate the advantages of this technique.

NOMENCLATURE

$X \in R^{n \times m}$	matrix of predictor variables
y	response variable
$b \in R^m$	vector of regression coefficients
σ^2	noise variance
$U \cdot diag(s_i) \cdot V^T$	SVD of X
λ^2	ridge parameter
λ_i	local ridge parameters

INTRODUCTION

In many predictive modeling engineering applications, the predictor data set is collinear. For some systems, such as predictive systems used to monitor process sensor calibrations, collinear predictors are necessary for building successful and robust inferential models [1]. Due to the presence of collinearity, traditional empirical modeling techniques such as ordinary least squares, neural network multi-layer perceptrons, and others that do not employ regularization produce very unstable and unrepeatable results [2]. Examples exist in most research fields.

To deal with instabilities due to collinear inputs, the method of regularization developed first by Tikhonov [3] was adopted in the form of ridge regression [4] or a more general class of penalized estimators [5]. When applying *ordinary least squares* (OLS) to a data set with collinear inputs, the coefficients are usually very large in magnitude. These large coefficients are caused by overfitting the training data and can amplify noise in the predictors and produce useless predictions.

This problem can be avoided by adding additional constraints to the usual sum of squared error objective function. The most common method, termed *ridge regression*, adds a term that also minimizes the magnitude of the regression coefficients. In his paper, Hoerl [4] proved that regardless of the conditioning, for finite data sets, there always exists a ridge estimate that decreases the mean squared error of the solution. This means that even if the data matrix is not badly ill-conditioned, one can still improve prediction accuracy by exploiting ridge regression rather than OLS. Adding the constraint will bias the estimate but reduce its variance making it more stable so that the probability that the ridge estimates falls in a certain vicinity of the true parameter value is higher than that of the OLS estimate.

There are two potential problems encountered when using ridge regression. The first problem is choosing an optimal ridge parameter and the second deals with assumptions inherent in the methodology. We will now briefly describe these potential problems.

The proper choice of the ridge parameter greatly affects the performance of ridge regression. Several methods of choosing a valid ridge parameter have found their way into engineering practice. The most common methods are the *Discrepancy Principle (DP)* [6]; *Mallows'* [7] *CL, Generalized Cross Validation (GCV*) [8], and the *L-curve method* [9]. Unfortunately, every parameter choice rule has its pitfalls. The high sensitivity of *CL* and *DP* to an underestimation of the noise level has limited their application to cases in which the noise level can be estimated with high fidelity [10]. On the other hand, noise-estimate-free *GCV* occasionally fails, presumably due to the presence of correlated noise [11]. The *L-curve* method is widely used; however, this method is nonconvergent [12]. All these methods directly or indirectly estimate the mean predictive error and select the ridge parameter so as to minimize the estimated mean predictive error.

The second problem deals with inherent assumptions of ridge regression. When implementing ridge regression, the components with associated singular values larger than the ridge parameter are considered to contain useful predictive information and are passed while components with singular values less than the regularization parameter are considered to contain noise or other useless information and are damped. The basic assumption that the components are arranged in order of predictive importance may not always hold. The components are arranged by their amount of variation and this may, or may not, lead to components arranged with respect to predictive ability. In fact, components can have a high variance with large singular values, but contain no predictive information. In this case ridge regression would needlessly pass this component, which results in degraded predictive performance. The other case is when a component with low variance and small singular value is unnecessarily damped. A more optimal technique would be to associate a ridge parameter with each component so that each component could be passed or damped with respect to its predictive capabilities rather than its amount of variation. This technique, called local ridge or generalized ridge regression [4], has found limited use because a method of optimizing the vector of local ridge parameters has not been found to be practical.

This paper presents an Evolutionary Algorithm method for optimizing the local ridge parameters to minimize Mallows' CL. CL was chosen because it has proven to be an unbiased estimate of prediction error [7]. The methodology section derives the local ridge solution and describes the evolutionary programming strategy. The developed methodology is then applied to the development of two predictive models. These two examples show the advantages of local ridge to pass components with small variance and high predictive capabilities and to damp components with high variation and little predictive value.

METHODOLOGY

This section will describe the methodologies used to implement the local ridge regression algorithm. It is broken down into two major sections: a section on evolutionary algorithms, and a section describing the objective function selected to be minimized.

Predictive Error Estimator as a Fitness Function

Consider the following linear regression problem

$$y = Xb + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma^2 I_n\right) \tag{1}$$

where $b \in R^m$ is the vector of regression coefficients to be determined using observed data (X,y); ε represents noise in the response y; $X_1,..., X_m$ are the explanatory variables or predictors. The *OLS* solution or maximum likelihood solution is given by

$$\hat{b}_{ols} = \left(X^T X\right)^{-1} X^T y, \tag{2}$$

where $\hat{b}_{ols}$ is the vector of regression coefficient estimates. The *OLS* solution is an unbiased estimate of the true solution, if such a solution exists. However, when the data matrix X is ill-conditioned, the *OLS* solution ($\hat{b}_{ols}$) becomes extremely unstable, i.e. it has a very large variance. This can be easily seen when the solution is written in the terms of a *singular value decomposition* (*SVD*) of the data matrix $X = U \cdot diag\left(s_i\right) \cdot V^T$

$$\hat{b}_{ols} = V \cdot diag\left(s_i^{-1}\right) \cdot U^T y = \sum_{i=1}^{m} s_i^{-1}\left(u_i^T y\right) \cdot v_i \qquad (3)$$

where u and v are called left and right eigenvectors of X and s_i are the singular values of the data matrix X.

The ill-conditioned matrix X has near zero last singular values. These last singular values usually correspond to the noise (or non-informative) components in X. When inverted, these near zero singular values drastically amplify the contribution of the noise components to the solution and destroy its predictive accuracy. Indeed, the variance-covariance matrix of the *OLS* solution is

$$Cov\left(\hat{b}_{ols}\right) = \sigma^2 \left(X^T X\right)^{-1} = \sigma^2 V \cdot diag\left(s_i^{-2}\right) \cdot V^T, \qquad (4)$$

where σ^2 is the noise variance in y. Near zero singular values result in a large variance of the solution, making it statistically insignificant.

When dealing with collinear data (ill-conditioned X), one can use ridge regression [4] to avoid the problem of instability. The ridge solution is obtained as

$$\hat{b} = \left(X^T X + \lambda^2 I_m\right)^{-1} X^T y \qquad (5)$$

where $\lambda \geq 0$ is the *ridge parameter* and I_m is the mxm identity matrix. In terms of the *SVD* of X, the ridge solution can be written as

$$\begin{aligned} \hat{b}_\lambda &= V \cdot diag\left(\frac{s_i^2}{s_i^2 + \lambda^2}\right) \cdot U^T y \\ &= \sum_{i=1}^{m} \frac{s_i^2}{s_i^2 + \lambda^2}\left(u_i^T y\right) \cdot v_i = \sum_{i=1}^{m} f_i \rho_i v_i \end{aligned} \qquad (6)$$

with

$$f_i = \frac{s_i^2}{s_i^2 + \lambda^2} \qquad (7)$$

referred to as the filter factors and $\rho_i = u_i^T y$ as the correlation coefficients. Notice that for $\lambda=0$, the ridge solution becomes the *OLS* solution; for $\lambda>0$, the solution is different. The filter factors determine if the information in the i^{th} component is incorporated into the solution or damped. If λ is large with respect to a singular value, the filter factor dampens the corresponding component while if λ is small with respect to a singular value, its corresponding component is passed. Therefore, a suitably large λ eliminates the destroying effect of the near zero singular values and makes the solution stable and statistically significant. The variance-covariance matrix in this case is

$$\begin{aligned} Cov\left(\hat{b}_\lambda\right) &= \sigma^2 \left(X^T X + \lambda^2 I_m\right)^{-1} X^T X \left(X^T X + \lambda^2 I_m\right)^{-1} \\ &= \sigma^2 V \cdot diag\left(\frac{s_i^2}{\left(s_i^2 + \lambda^2\right)^2}\right) \cdot V^T \end{aligned} \qquad (8)$$

When $\lambda \to \infty$, the variance of the corresponding solution goes to zero. Unfortunately, the decreasing variance is not the only consequence of using ridge regression. Shrinkage also introduces a bias into the solution which increases with increasing λ. It is shown in [4] that there always exists some λ_{opt} that optimally balances the bias and variance such that the *mean squared error* (*MSE*) of the solution, defined as

$$MSE(\lambda) = E\left\{\left(b_{true} - \hat{b}_\lambda\right)^T \left(b_{true} - \hat{b}_\lambda\right)\right\}, \qquad (9)$$

is less than that of the *OLS* solution. The only difficulty in computing such an optimal ridge parameter is that the *MSE* of the solution is not computable unless the true solution is known. However, one can use the *mean predictive error* (*MPE*) to select λ,

$$MPE(\lambda) = E\left\{\left(Xb_{true} - X\hat{b}_\lambda\right)^T \left(Xb_{true} - X\hat{b}_\lambda\right)\right\}, \qquad (10)$$

which can be successfully approximated using available observations. To approximate the *MPE* one can use Mallows' [9] *CL*

$$CL(\lambda) = \frac{\left(y - X\hat{b}_\lambda\right)^T \left(y - X\hat{b}_\lambda\right)}{n} + \frac{2\sigma^2}{n} trace(H(\lambda)), \qquad (11)$$

where the hat matrix ($H(\lambda)$) is defined as

$$H(\lambda) = X\left(X^T X + \lambda^2 I_m\right)^{-1} X^T. \qquad (12)$$

In standard ridge regression, (5) and (6), we use the same value of the ridge parameter for each component. It may be desirable to have an individual ridge parameter for each singular value (or component) and optimize the values of all these individual parameters in an attempt to

reduce the *MPE* further. This can be useful in situations when intermediate components are not related to the response, but due to a limited number of observations and possible random correlations, they still contribute to the solution, degrading the prediction accuracy. To eliminate a particular component from the solution, the following form of ridge regression with individual ridge parameters can be used

$$\hat{b}_{\lambda_i} = \left(X^T X + V\lambda_i^2 V^T\right)^{-1} X^T y = \sum_{i=1}^{m} \frac{s_i^2}{s_i^2 + \lambda_i^2}\left(u_i^T y\right)\cdot v_i \quad (13)$$

We refer to this variation of ridge regression as local ridge regression with λ_i being the local ridge parameters. A large λ_i with respect to its corresponding singular value prevents the corresponding *component* from contributing to the solution. As before we can chose λ_i's to minimize the *MPE* approximated by *CL* in the form

$$CL(\lambda_i) = \frac{\left(y - X\hat{b}_{\lambda_i}\right)^T\left(y - X\hat{b}_{\lambda_i}\right)}{n} + \frac{2\sigma^2}{n} trace\left(H(\lambda_i)\right), \quad (14)$$

where the hat matrix is defined as

$$H(\lambda_i) = X\left(X^T X + V\lambda_i^2 V^T\right)^{-1} X^T. \quad (15)$$

Unlike standard ridge regression in which one λ is optimized, this problem is a multidimensional optimization with a vector of λ_i's being optimized. The number of possible combinations of even a moderate number of real-valued ridge parameters becomes enormous even with a fairly coarse grid of the ridge parameters values. Orr [13] attempted to optimize each parameter by itself and repeated the optimizations until the solution converged. This method is time consuming and may be subject to local minima. *Evolutionary Algorithm (EA) optimization* is able to choose an optimal subset of regularization parameters that minimize *CL* (14) as the fitness function.

Evolutionary Algorithms to Optimize Local Ridge

Evolutionary Algorithms (EA) have been successfully applied to solve complex engineering optimization problems. Arguably the best know representatives are Genetic Algorithms (GA) and Evolutionary Strategies (ES) [14, 15]. Differential Evolution (DE) [16, 17] is a population-based, direct-search algorithm for global optimization. While originally designed to operate on continuous floating point variables DE has recently been extended to optimize a mixture of integer, discrete, and continuous variables as well as multiple linear and non-linear constraints [18].

DE has proven to be exceptionally simple (less than 30 lines of C-code) and robust for a variety of real-world optimization problems [17]. While the structure of DE is similar to other population based search algorithms, like ES and GAs, it differs in both its self-referential mutation scheme and its selection process. Here is the basic structure of DE.

Initialization. First, we start with an objective function *f(X)* to be optimized, where *X* is a vector of *D* parameters, X = (x_1,…,x_D). Our goal is to find the optimal values of the vector *X* that provide a minimum value of *f(X)*. DE operates on a population, P_G, of candidate vectors. The size of the population, *NP*, remains constant for all generations. Each member of the population is denoted as $X_{i,G}$, where *i* indexes the population and *G* is the particular generation, P_G = {$X_{1,G}$,…,$X_{i,G}$,…,$X_{NP,G}$}, *i* = 1,2,…*NP*, *G* = 1,…,G_{max}. The parameters of *X* are analogous to the chromosomes of an individual *i* in a generation *G*, $X_{i,G}$ = $x_{j,i,G}$, *i* = 1,2,…*NP*, *j*=1,2,…,*D*.

For most real-world engineering problems, the parameters of the objective function will be constrained by lower and upper boundary conditions $x_j^{(L)}$ and $x_j^{(U)}$ where *j=1,…,D*. Typically the initial population, P_0, is generated by randomly selecting parameter values between these lower and upper boundaries.

Mutation and Recombination. While a predefined probability distribution function drives mutation for most EAs, DE utilizes a self-referential mutation scheme based on the differences of randomly sampled objective vectors from the current population. The distribution of the differences is consequently determined by the distribution of the population itself. This means that any bias introduced in the way DE attempts to improve the population of objective vectors is implicitly driven by the objective function or problem being optimized.

DE uses both mutation and recombination to produce a second population of children or trial vectors. One *"child"* vector is created for each *"parent"* in a random manner. When *crossover*

occurs, a parameter of the *"child"* becomes a linear combination of three randomly chosen vectors, otherwise that parameter of the *"parent"* is passed along to the *"child"*. Another portion of the code ensures that each *"child"* vector differs from its *"parent"* in at least one parameter (chromosome). This is done for every *"parent"* vector in the current population. Several user specified control variables, such as crossover and mutation rates, affect the convergence properties and robustness of DE and often depend on the characteristics of the objective function. Guidelines for selecting the parameters are provided in [16, 17] and successful selection of the parameters can usually be obtained after a few trial iterations using differing values.

Selection. The selection scheme utilized by DE is also different from other ES and GAs. Each successive population, is selected from either the current *"parent"* population, or the *"child"* population. Each individual "child" in the trail population is compared with a single "parent" in the current population and the individual with the lower objective function "survives" and passes on into the next generation. This means that all the individuals in each successive generation are at least as good as their "parent" in the current generation. In contrast to other EAs, which compare a candidate individual to all other individuals in the population, DE only compares the candidate individual to a single member of the current population.

CASE STUDIES

This section presents two applications of the local ridge algorithm developed in the previous section. The first example is a predictive model using automobile data that shows unimportant, high variance components can be correctly damped with local ridge. The second example is a predictive model that estimates the value of a process parameter in a fossil power plant that demonstrates important, low variance components can be passed.

Automobile Example

The first example uses automobile data that can be found at the University of California Irvine, Repository of Machine Learning Database [18]. The dataset was first used in the 1983 American Statistical Association Exposition and later used by Quinlan [19] to predict automobile gas mileage. The data set has information from 392 automobiles with seven variables of interest provided in Table 1. In this example we will use the first six variables to predict the seventh variable: the car's acceleration.

Table 1. Automobile Data Set

Variable		Type
1	MPG	continuous
2	Cylinders	multi-valued discrete
3	Displacement	continuous
4	Horsepower	continuous
5	Weight	continuous
6	Year	multi-valued discrete
7	Acceleration	continuous

Performing a Principal Components Analysis (PCA) on the standardized data results in the following amounts of variation incorporated in each the six Principal Components (PC). The singular values are also listed.

Table 2. Principal Component Analysis

PC	Singular Value	% Variation
1	42.17	77.6
2	18.34	14.4
3	9.03	3.5
4	7.86	2.6
5	5.41	1.2
6	3.74	0.6

An analysis of the principal components show:

PC#1 is a weighted average of cylinders, displacement, power, and weight and negatively with MPG.
PC#2 is weighted towards the year.
PC#3 is weighted towards MPG.
PC#4 is a measure of the difference between the two variables cylinders and power.
PC#5 is weighted towards weight.
PC#6 is likely due to noise.

The condition number of X'X, which is the matrix that is inverted when calculating the OLS solution, is slightly ill-conditioned with a condition number of 130.

Table 3 presents the results of a correlation analysis of the principal components and the response variable: acceleration. In this table we see that the first component is highly correlated

with acceleration and that components 3, 4, and 5 have slight correlations with acceleration. This is to be expected since component two is year, and there were old and new cars with high and low accelerations. We may expect that ridge regression will pass the first five components and that by passing component two, the predictive performance will be slightly degraded.

Table 3. Correlation Analysis

Principal Component	Absolute Correlation
1	0.5510
2	0.0013
3	0.3625
4	0.3006
5	0.3014
6	0.0598

We will now evaluate various models using Mallow's CL as an estimator of the predictive error (eq. 12). Specifically, we will evaluate models using OLS, Principal Component Regression (PCR) with all and various combinations of PCs, and Ridge regression with the regularization component optimized to be 0.8286. This regularization parameter is optimal in the sense that it gives the minimum CL value.

The results in Table 4 show that the model with the minimum CL value (other than Local Ridge) is PCR with components 1, 3, 4, and 5. This agrees with the results expected from the correlation analysis, which expect components 2 and 6 to be removed.

Table 4. Prediction Results

Method	Estimate of Prediction Error
OLS	2.9746
All PCs [1 2 3 4 5 6]	2.9746
One PC [1]	5.3138
[1 3]	4.3329
[1 3 5]	3.6546
[1 3 4 5]	2.9729
Ridge (alpha 0.8286)	2.9739
Local Ridge	2.9578

Similar results occur when these predictive models are used with a validation set. In that case, the odd observations are used for training and the even observations are used for calculating the validation error. The predictive error corresponds to the estimates given by CL.

Note that the regularization coefficient of 0.8286 is significantly smaller than each of the singular values, and will therefore pass all of the components. This regularization parameter reduces the condition number from 130 to 120.

Table 5 lists the local ridge parameters obtained through the evolutionary algorithm optimization and their corresponding filter factors. We see that the 2nd component (Year) is properly damped out, and that the last component is partially damped.

Table 5. Principal Component Analysis

PC	Singular Values	Local Ridge Parameters	Local Ridge Filter Factors
1	42.67	2.416	0.9968
2	18.40	8899.3	0.0000
3	9.03	0.779	0.9926
4	7.86	0.819	0.9893
5	5.41	0.562	0.9893
6	3.75	2.286	0.7286

Figure 1 is a plot of ridge filter factors, local ridge filter factors, and correlation coefficients. Note the very low correlation of the 2nd component (year) with the response variable. Standard ridge regression allows that component to pass (filter factor near 1) while local ridge effectively damps it out of the solution (filter factor near 0).

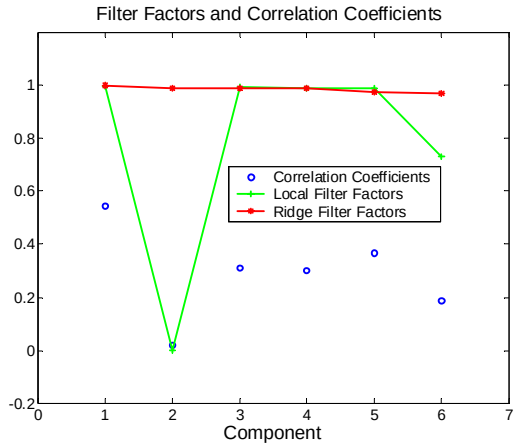


Figure 1. Filter Factors and Correlation Coefficients

Referring to Table 4, we see that the local ridge solution gives the best CL value, which is an estimate of predictive error. Therefore, the local ridge outperformed all other linear prediction models for the automobile example.

Process Sensor Estimation

The second example we want to discuss deals with the prediction of sensor values in power plants. The safe and economical operation of Fossil and Nuclear Power Plants (NPP) requires knowledge of the state of the plant, which is obtained by measuring critical plant parameters with sensors and their instrument chains. Traditional approaches used to validate that the sensors are operating correctly involve the use of redundant sensors coupled with periodic instrument calibration. Since few of the sensors are actually out of calibration, the end result is that many instruments are unnecessarily maintained. An alternative condition based technique is desirable.

When implementing condition based calibration methods, the instruments are calibrated only when they are determined to be out of calibration. On-line, real-time sensor calibration monitoring identifies faulty sensors which permits reduced maintenance efforts and increases component reliability.

Inferential sensing is the prediction of a sensor value through the use of correlated plant variables. Most calibration monitoring systems produce an inferred value and compare it to the sensor value to determine the sensor status. There are a number of techniques, which were proposed for on-line inferential sensing during recent years [20].

All of these methods use related sensors as inputs to estimate a model (sets of weights), which is subsequently used to infer the sensor's value based on the input values. A peculiar feature of any on-line sensor validation system is that this system should not only accurately infer the sensor's value but it should also be robust to moderate changes in input values. This means that the sensor validation system should resolve a subtle compromise between accuracy and robustness. Recently, the role of regularization in this process was realized [1, 2]. Although traditional regularization techniques perform well for these types of problems, sensor value prediction accuracy can be improved using multiple (local) regularization parameters. To demonstrate this, we used eighty-two variables, recorded at a TVA plant, arranged in a data matrix X (1000 x 82), as predictor variables to infer the value of a response variable Y, which is the sensor under surveillance. One thousand initial samples were used as training data and two thousand were left as a test set. The prediction accuracy was estimated for four techniques: ordinary least squares solution, regular ridge regression with regularization parameter selected to optimize CL, truncated singular value decomposition (TSVD) with an optimized truncation parameter and local ridge regression with local ridge parameters selected with DE optimization of CL as the cost function. The prediction MSEs for test data set are shown in Table 6.

Table 6 MSE for Different Techniques

	OLS	**Ridge: λ= 0.0273**	**TSVD k=20**	**Local Ridge**
MSE	0.586	0.579	0.415	0.292

As we can see, the prediction MSE is the lowest for local ridge regression with regularization parameters selected by DE. We should point out that DE is a stochastic optimization technique and is subject to random fluctuations. Several DE runs were performed on the same training data set and the result with smallest CL was selected to perform local ridge. The filter factors for regular and local ridge, along with correlation coefficients between response variable and components are shown in Fig. 2.

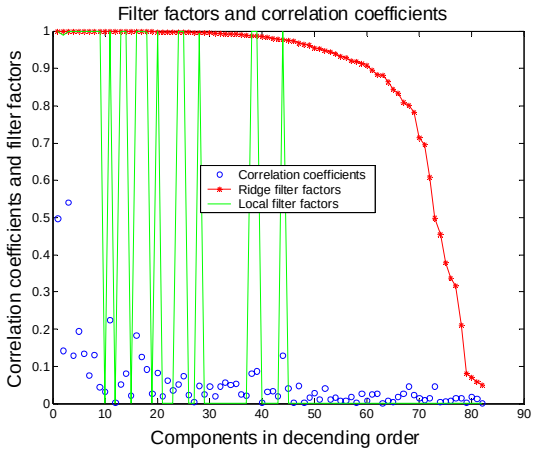


Figure 2. Filter Factors and Correlation Coefficients

It can be seen from Fig.2 that the DE optimized filter factors completely removed components starting from number 45. They also passed the first 10 principal components. However, in contrast to regular ridge regression or TSVD, it passed and damped middle range components selectively. It is important to notice that local ridge filters out components with small correlation coefficients and passes components with relatively significant correlation coefficient. In this case, it is also interesting to notice that local ridge filter factors either completely pass or completely dampen a component, thus performing "selective" TSVD.

CONCLUSION

This paper presented a methodology for implementing local ridge regression through optimizing Mallows' CL with Differential Evolution. Two example implementations of the algorithm on actual data show that this method provides better values of CL and better predictive performance than OLS or standard Ridge Regression. The use of Differential Evolution to optimize high dimensional local ridge optimization problems is both useful and practical.

REFERENCES

1. A.V. Gribok, I. Attieh, J.W. Hines and R.E. Uhrig, Regularization of feedwater flow rate evaluation for venturi meter fouling problems in nuclear power plants, *Nuclear Technology*, Vol.134, pp.3-14, (2001)

2. J.W. Hines, A.V. Gribok, I. Attieh, and R.E. Uhrig, Regularization methods for inferential sensing in nuclear power plants, *Fuzzy Systems and Soft Computing in Nuclear Engineering*, Ed. Da Ruan, Springer, (1999)

3. A.N. Tikhonov, Solution of incorrectly formulated problems and regularization method, *Soviet math. Dokl*. **4**, pp. 1053-1038, (1963)

4. A.E. Hoerl, and R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, Vol. 12, No. 1, pp. 55-82, (1970)

5. I.J. Good, Nonparametric roughness penalties for probability densities, *Biometrica*, 58, pp. 255-277, (1971)

6. V.A. Morozov, On the solution of functional equations by the method of regularization, *Soviet Math. Dokl.*, 7, pp.414-417, (1966)

7. C.L. Mallows, Some comments on CP", *Technometrics*, 15, No. 4, pp. 661-675, (1973)

8. G.H. Golub, M. Heath, and G. Wahba, Generalized cross-validation as a method for choosing a good ridge Parameter, *Technometrics*, Vol. 21, No. 2, pp.215-223, (1979)

9. P.C. Hansen, Analysis of discrete ill-posed problems by means of the l-curve, *SIAM Review*, Vol. 34, No. 4, pp. 561-580, (1992)

10. P.C. Hansen, Regularization, GSVD and Truncated GSVD, BIT 29, pp. 491-504, (1989)

11. G. Wahba, Spline models for observational data, *SIAM*, (1990)

12. A.S. Leonov and A.G. Yagola, The L-curve method always introduces a non-removable systematic error, *Moscow University Physics Bulletin*, Vol. 52, No. 6, pp.20-23, (1997)

13. M.J.L. Orr, Local smoothing of radial basis function networks, International Symposium on Artificial Neural Networks, Hsinchu, Taiwan, (1995)

14. H.P. Schwefel, *Numerical Optimization of Computer Models*, John Wiley & Sons Ltd, Chichester, U.K, (1977)

15. I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart, 1973.

16. R. Storn, and K. Price, Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization*, Kluwer Academic Publishers, 11(4), pp.341–359, (1997)

17. K. Price, An introduction to differential evolution, *New Ideas in Optimization*, McGraw-Hill, London, pp. 79–108, (1999).

18. C.L. Blake and C.J. Merz, UCI Repository of Machine Learning Databases, University of California, Irvine, Department of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html, (1998)

19. R. Quinlan, Combining instance-based and model-based learning, proceedings on the Tenth International Conference of Machine Learning, University of Massachusetts, Amherst, Morgan Kaufmann, pp. 236-243, (1993)

20. A.V. Gribok, J.W. Hines, A. Urmanov and R.E. Uhrig, Heuristic, systematic, and informational regularization for process monitoring, accepted for publication in the special issue of the *International Journal of Intelligent Systems* on Intelligent Systems for Process Monitoring, Wiley Publishers, (2002)

4^{th} *International Conference on Inverse Problems in Engineering*
Rio de Janeiro, Brazil, 2002

HOMOGENIZATION TECHNIQUE IN INVERSE PROBLEMS FOR BOUNDARY HEMIVARIATIONAL INEQUALITIES

Stanisław Migórski
Faculty of Mathematics, Physics and Computer Science
Institue of Computer Science, Jagiellonian University
Cracow, Poland
migorski@softlab.ii.uj.edu.pl

ABSTRACT

The purpose of the paper is to present a methodology which is useful in the derivation of approximate models with simpler geometry of some inverse problems. First we formulate a direct problem (being the boundary hemivariational inequality) which is given in a domain with a complicated geometry (e.g. perforated domains, layered structures). For such direct problem we consider the inverse one and we provide result on the existence of solutions. Next we establish the homogenization result for the direct problem. It turns out that in the homogenized inequality the complex boundary condition is replaced by a much simpler one. Finally, we study the asymptotic behavior of the set of solutions of the inverse problem. The main result shows that the solutions to the inverse problem for homogenized hemivariational inequality can be considered as reasonable approximations of the solutions of the original inverse problem.

NOMENCLATURE

a.e.	almost every (everywhere)
D	gradient operator
div	divergence operator
$\overline{E}$	closure of a set E
inf, sup	infimum, supremum operations
j^0	the Clarke directional derivative
K_{ad}	set of admissible parameters
$m(Y)$	the Lebesgue measure of a set Y
n	unit normal
$\mathbb{N}$	the set of natural numbers
$\mathbb{R}$	the set of real numbers
$S_\varepsilon(a)$	solution set corresponding to a
2^X	subsets of a space X
χ	characteristic function
Γ	boundary of a domain
∂j	generalized gradient of j
$\partial\Omega$	boundary of a domain Ω

INTRODUCTION

In this paper we present a continuation of our efforts on the development of models for mechanical structures in which it is necessary to deal with multivalued and nonmonotone laws. It is well known (see [1], [2] and [3]) that there is a large class of mechanical problems with nonconvex energy functions which are generally nonsmooth. They lead to nonmonotone, possibly multivalued constitutive laws or/and boundary conditions which can not be derived from convex superpotentials via the differentiation. Such mechanical problems can be successfully described by a type of variational expressions called hemivariational inequalities which were introduced by P.D. Panagiotopoulos, cf. [1] and [2]. For example, considering the contact between an elastic structure and a granular medium (or a composite material) we arrive to multivalued boundary conditions of the subdifferential type. In hemivariational inequalities the aformentioned laws are formulated via the notion of the generalized Clarke gradient [4].

On the other hand studying the problems of estimation of material parameters in mechanical systems we meet structures with complex geometry. The goal of this paper is to study the problem of identification of a discontinuous coefficient in a domain with a complicated geometry of boundary. On the corresponding parts of the boundary the mixed boundary conditions are assumed: the Dirichlet-Neumann conditions and a condition of the subdifferential type. Our purpose is not to present an efficient computational algorithm but rather to describe

a technique which can be useful for dealing with inverse problems for structures with a complicated boundary. This technique can be applied before passing to the numerical issues. We show that the inverse problem obtained by the boundary homogenization procedure for a hemivariational inequality can be considered as a reasonable approximation of the initial complicated inverse problem.

We consider the elliptic boundary hemivariational inequality of the form: find $u \in V$ such that

$$\begin{cases} -\operatorname{div}(a(x)Du) + u = f \ \text{ in } \ \Omega \\ u = 0 \ \text{ on } \ \Gamma_1 \\ \dfrac{\partial u}{\partial n_a} = g \ \text{ on } \ \Gamma_2, \\ -\dfrac{\partial u(x)}{\partial n_a} \in \partial j(x, u(x)) \ \text{ a.e. on } \ \Gamma_3, \end{cases} \tag{1}$$

where Ω is a bounded domain in $\mathbb{R}^N$, V is a closed subspace of $H^1(\Omega)$ such that $H_0^1(\Omega) \subset V$, $\partial u/\partial n_a = \sum_{i=1}^N a(x) D_i u \, n_i$ denotes the conormal derivative of u associated to a, n is the outward normal to $\partial\Omega$, ∂j is the Clarke subdifferential of a locally Lipschitz function $j\colon \mathbb{R} \to \mathbb{R}$ and $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ with $\Gamma_1 \neq \emptyset$.

The inverse problem for the above hemivariational inequality system consists in finding a coefficient a in a set K_{ad} which solves

$$\min_{a \in K_{ad}} \min_{u(a)} F(u(a)),$$

where $u = u(a)$ is a weak solution of (1) corresponding to the coefficient a and F is a prescribed cost criterion.

We give a rigorous mathematical result on the asymptotic behavior of the direct problem when the small parameter describing the geometrical structure of the boundary tends to zero. The previous results in this direction were obtained by Damlamian and Li Ta-tsien [5] for elliptic differential equations and by Migorski and Ochal [6] for elliptic hemivariational inequalities. We mention also that the method of interior homogenization (i.e. when the coefficients are of the form $a_\varepsilon(x) = a(x/\varepsilon)$ with a being a periodic function and $\varepsilon \to 0$) can not be used here since the highly oscillating coefficients are not uniformly bounded variation (see the choice of the set of admissible parameters K_{ad} below and [6]).

The reader is referred to [7], [8], [9], [10] for the corresponding optimal control problems for hemivariational inequalities and to [11], [12], [13], [14] for the stability of inverse and parameter identification problems.

HEMIVARIATIONAL INEQUALITY MODEL

We recall some definitions which are useful in the next sections. We will denote by Ω a bounded open subset of $\mathbb{R}^N$ with Lipschitz continuous boundary $\partial\Omega$. Given $f \in L^1(\Omega)$ the variation of f is defined by (cf. Giusti [15])

$$\int_\Omega |Df| = \sup\{\int_\Omega f \ \operatorname{div} g \, dx \ : \ g \in C_0^1(\Omega; \mathbb{R}^N), \ |g(x)| \leq 1 \text{ for } x \in \Omega\}.$$

If $\int_\Omega |Df| < +\infty$ that is the variation of f is finite, we say that f has bounded variation. The space of functions $f \in L^1(\Omega)$ with bounded variation is denoted by $BV(\Omega)$. Equipped with the norm $||f|| = ||f||_{L^1} + \int_\Omega |Df|$, $BV(\Omega)$ becomes a Banach space.

In this note we consider the set of admissible parameters of the form:

$$K_{ad} = \left\{ a \in A_{ad} \ : \ \int_\Omega |Da| \leq C \right\}, \tag{2}$$

where $A_{ad} = \{a \in L^\infty(\Omega) \ : \ 0 < c_1 \leq a(x) \leq c_2, \text{ a.e. in } \Omega\}$ and $C > 0$.

Concerning these two sets we recall their properties which are needed in the sequel.

Remark 1 *(a) The topologies of $L^1(\Omega)$ and $L^2(\Omega)$ coincide on the set A_{ad}, i.e. for all $a \in A_{ad}$, we have*

$$||a||_{L^1} \leq const ||a||_{L^2} \ \text{ and } \ ||a||_{L^2} \leq c_2 ||a||_{L^1}.$$

(b) The set K_{ad} is compact in $L^1(\Omega)$ for every constant $C > 0$. This is a consequence of the fact that the set of functions uniformly bounded in the $BV(\Omega)$ norm is relatively compact in $L^1(\Omega)$.

For the properties (a) and (b) we refer to Gutman [16] and to Giusti [15] (Theorem 1.19), respectively.

We recall the definitions of the generalized directional derivative and the generalized gradient of Clarke for locally Lipschitz function $\varphi: X \to \mathbb{R}$, where X is a Banach space (see Clarke [4], Chapter 2). The generalized directional derivative of φ at $x \in X$ in the direction $v \in X$, denoted by $\varphi^0(x;v)$, is defined by

$$\varphi^0(x;v) = \limsup_{y \to x \ \lambda \downarrow 0} \frac{\varphi(y+\lambda v) - \varphi(y)}{\lambda}.$$

The generalized gradient of φ at x, denoted by $\partial\varphi(x)$, is a subset of a dual space X^* given by

$$\partial\varphi(x) = \{\zeta \in X^* : \langle \zeta, v\rangle_{X^* \times X} \leq \varphi^0(x;v) \quad \text{for all } v \in X\}.$$

Recall also that given a Banach space X equipped with a topology τ and sets $\{M_n\}_{n\in\mathbb{N}} \subseteq 2^X$, the sequential Kuratowski upper limit is defined by

$$K_{seq}(\tau\text{–}X)\limsup M_n =$$
$$= \{x \in X : \exists \{n_\nu\}, x_{n_\nu} \in M_{n_\nu}, x_{n_\nu} \to x \text{ in } \tau\text{–}X, \text{ as } \nu \to +\infty\}.$$

The space X with the weak topology is denoted by w-X.

Now we assume that the boundary of Ω consists of three disjoint open subsets such that

$$\partial\Omega = \overline{\partial_1\Omega} \cup \overline{\partial_2\Omega} \cup \overline{\partial_3\Omega}, \quad \partial_1\Omega \neq \emptyset.$$

We consider the following problem: find a function $u: \Omega \to \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a(x)Du) + u = f \quad \text{in } \Omega \\ u = 0 \quad \text{on } \partial_1\Omega \\ \dfrac{\partial u}{\partial n_a} = g \quad \text{on } \partial_2\Omega \\ -\dfrac{\partial u(x)}{\partial n_a} \in \partial j(x, u(x)) \quad \text{on } \partial_3\Omega \end{cases} \tag{3}$$

where $\dfrac{\partial u}{\partial n_a} = a(x)Du \cdot n$ denotes the conormal derivative of u associated to a.

In order to give the variational formulation of the above problem let $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \partial_1\Omega\}$. This space is a closed subspace of $H^1(\Omega)$ provided V is equipped with the topology induced by $H^1(\Omega)$. It is well known that $||v|| = ||Dv||_{L^2(\Omega;\mathbb{R}^N)}$ is an equivalent norm on V.

Let $v \in V$. Multiplying the equation in Ω by v, integrating over Ω and applying the Green theorem, we have

$$\int_\Omega (a(x)Du\,Dv + uv)\,dx -$$
$$-\int_{\partial\Omega} \frac{\partial u}{\partial n_a} v\,d\sigma(x) = \int_\Omega fv\,dx.$$

Next taking into account the boundary conditions on $\partial_2\Omega$ and $\partial_3\Omega$, we obtain the following variational form of (3): find $u \in V$ such that

$$\begin{cases} \alpha(u,v) + \displaystyle\int_{\partial_3\Omega} \zeta v\,d\sigma(x) = \langle l, v\rangle \\ \qquad\qquad\qquad\qquad \text{for all } v \in V \\ \zeta(x) \in \partial j(x, u(x)) \ \text{ a.e. on } \partial_3\Omega \end{cases} \tag{4}$$

where

$$\alpha(u,v) = \int_\Omega (a(x)Du\,Dv + uv)\,dx \quad \text{and}$$
$$\langle l, v\rangle = \int_\Omega fv\,dx + \int_{\partial_2\Omega} gv\,d\sigma(x).$$

Using the definition of the Clarke subdifferential, the latter formulation reduces to the form of hemivariational inequality: find $u \in V$ such that

$$\alpha(u, v-u) + \int_{\partial_3\Omega} j^0(x, u(x); v(x) - u(x))\,d\sigma(x)$$
$$\geq \langle l, v-u\rangle \quad \text{for all} \quad v \in V.$$

We show that the problem admits a solution under mild hypotheses on the data. Namely we make the following assumptions:

$\underline{H(a)}$: $\quad a \in A_{ad}$.

$\underline{(H_0)}$: $\quad f \in L^2(\Omega)$, $g \in H^{-1/2}(\partial_2\Omega)$.

$\underline{H(j)}$: $\quad j: \partial_3\Omega \times \mathbb{R} \to \mathbb{R}$ is such that $j(\cdot, \xi)$ is measurable, $j(x, \cdot)$ is locally Lipschitz, $j(x,0) \in L^1(\partial_3\Omega)$,

$|\partial j(x,\xi)| \leq c_3(1 + |\xi|)$ a.e. $x \in \partial_3\Omega$ and for all $\xi \in \mathbb{R}$ with $c_3 > 0$,

$j^0(x, \xi; -\xi) \leq c_4(1 + |\xi|)$ a.e. $x \in \partial_3\Omega$ and for all $\xi \in \mathbb{R}$ with $c_4 \geq 0$.

Lemma 2 *Under hypotheses $H(a)$, (H_0) and $H(j)$, the problem (4) admits a solution. Furthermore, there is a positive constant c such that*

$$||u||_V \leq c\left(1 + ||f||_{L^2(\Omega)} + ||g||_{H^{-1/2}(\partial_2\Omega)}\right).$$

Proof. Let us define the functional $J: L^2(\partial_3\Omega) \to \mathbb{R}$ by

$$J(v) = \int_{\partial_3\Omega} j(x, v(x))\, d\sigma(x).$$

The conditions $H(j)$ imply that J is locally Lipschitz on $L^2(\partial_3\Omega)$. Recall also that the trace operator $v \to v|_{\partial\Omega}$ is a linear continuous operator from V to $H^{1/2}(\partial\Omega)$ and that the embedding of $H^{1/2}(\partial\Omega)$ to $L^2(\partial\Omega)$ is compact (see e.g. Zeidler [17]). Therefore we may apply Theorem 4.26 (with $\int_\Omega j^0(x,u;v-u)dx$ replaced by $\int_{\partial_3\Omega} j^0(x,u;v-u)\,d\sigma(x)$) of Naniewicz and Panagiotopoulos [3] and deduce that the hemivariational inequality (4) has at least one solution. From the hypothesis $H(a)$ we have $\alpha(u,u) \geq c_1\|u\|^2$. So using $H(j)$, from (4) we get

$$c_1||u||^2 \leq \int_{\partial_3\Omega} j^0(x,u;-u)\,d\sigma(x) +$$

$$+ \int_\Omega fu\,dx + \int_{\partial_2\Omega} gu\,d\sigma(x) \leq$$

$$\leq \widetilde{c}_1||u||_{L^2} + ||f||_{L^2}||u||_{L^2} + \widetilde{c}_2||g||_{H^{-1/2}}||u|| \leq$$

$$\leq \widetilde{c}_3\left(1 + ||f||_{L^2} + ||g||_{H^{-1/2}}\right)||u||$$

with suitable positive constants $\widetilde{c}_1$, $\widetilde{c}_2$ and $\widetilde{c}_3$. This completes the proof. □

Remark 3 *Lemma 2 still holds if the sign condition $j^0(x,\xi;-\xi) \leq c_4(1+|\xi|)$ in $H(j)$ is replaced by a weaker one*

$$j^0(x,\xi;-\xi) \leq \beta(x)(1+|\xi|^s)$$

where $0 \leq s < 2$, $\beta \in L^{2/(2-s)}(\partial_3\Omega)$, $\beta \geq 0$. This follows from the fact that then the bilinear form α is V-coercive; see also Section 4.3 of [3].

Remark 4 *Given $\beta \in L^\infty_{loc}(\mathbb{R})$, we denote by $\widehat{\beta}: \mathbb{R} \to 2^{\mathbb{R}}$ a multifunction obtained from β by "filling in the gaps" at its discontinuity points, i.e. $\widehat{\beta}(\xi) = [\underline{\beta}(\xi), \overline{\beta}(\xi)]$, where*

$$\underline{\beta}(\xi) = \lim_{\delta\to 0^+} \operatorname{ess\,inf}_{|t-\xi|\leq\delta} \beta(t),$$

$$\overline{\beta}(\xi) = \lim_{\delta\to 0^+} \operatorname{ess\,sup}_{|t-\xi|\leq\delta} \beta(t)$$

and $[\cdot,\cdot]$ denotes the interval. Here ess inf *and* ess sup *stand for the essential infimum and supremum, respectively, over the interval $[\xi-\delta, \xi+\delta]$. It is well known (see Chang [18]) that a locally Lipschitz function $j: \mathbb{R} \to \mathbb{R}$ can be determinated up to an additive constant by the relation $j(\xi) = \int_0^\xi \beta(s)\,ds$ and that $\partial j(\xi) \subset \widehat{\beta}(\xi)$. Moreover, if $\beta(\xi \pm 0)$ exist for every $\xi \in \mathbb{R}$, then $\partial j(\xi) = \widehat{\beta}(\xi)$. We refer to [19] for additional hypothesis on the function β under which the solution of the hemivariational inequality (4) is unique. In this case the inverse problems $(IP)_\varepsilon$ and (IP) considered below reduce to the usual minimization ones.*

STATEMENT OF INVERSE PROBLEM

The aim of this section is to give an existence result for the inverse problem under consideration.

Let the boundary Γ of Ω consist of three parts, i.e. $\partial\Omega = \Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$, Γ_3 has a positive measure and for every $\varepsilon > 0$ let Γ_1 be divided into two subsets Γ_1^ε and $\widetilde{\Gamma_1^\varepsilon}$. We consider the following inverse problem: given data f, g and a cost functional F defined on $H^1(\Omega)$, find a coefficient $a \in K_{ad}$, K_{ad} being the set of admissible parameters defined by (2) such that

$$\min_{a\in K_{ad}} \min_{u(a)\in S_\varepsilon(a)} F(u(a)) = m, \qquad (IP)_\varepsilon$$

where $S_\varepsilon(a)$ is the set of solutions to the problem:

$$\begin{cases} -\operatorname{div}(a(x)Du) + u = f \quad \text{in } \Omega \\ \dfrac{\partial u}{\partial n_a} = g \quad \text{on } \Gamma_1^\varepsilon \\ u = 0 \quad \text{on } \widetilde{\Gamma_1^\varepsilon} \\ -\dfrac{\partial u(x)}{\partial n_a} \in \partial j(x,u(x)) \quad \text{a.e. on } \Gamma_2 \\ u = 0 \quad \text{on } \Gamma_3. \end{cases} \qquad (DP)_\varepsilon$$

The hypotheses are the following.

$\underline{(H_0)_1}$: $\quad f \in L^2(\Omega),\ g \in H^{-1/2}(\Gamma_1^\varepsilon).$

$\underline{H(F)}$: $F: H^1(\Omega) \to \mathbb{R}$ is weakly lower semicontinuous.

$\underline{H(j)_1}$: $j: \Gamma_2 \times \mathbb{R} \to \mathbb{R}$ is such that $j(\cdot, \xi)$ is measurable, $j(x, \cdot)$ is locally Lipschitz, $j(x, 0) \in L^1(\Gamma_2)$,

$|\partial j(x, r)| \leq c_3(1 + |r|)$ a.e. $x \in \Gamma_2$ and for all $r \in \mathbb{R}$ with $c_3 > 0$,

$j^0(x, r; -r) \leq c_4(1 + |r|)$ a.e. $x \in \partial_3\Omega$ and for all $r \in \mathbb{R}$ with $c_4 \geq 0$.

Theorem 5 *If hypotheses $H(j)_1$, $H(F)$ and $(H_0)_1$ hold, then the inverse problem $(IP)_\varepsilon$ possesses a solution for every fixed $\varepsilon > 0$ and every admissible set K_{ad} of the form (2).*

Proof. We apply the direct method of the calculus of variations. Let $\{(a_k, u_k)\}_k$ be a minimizing sequence for the inverse problem $(IP)_\varepsilon$ with $a_k \in K_{ad}$ and $u_k = u(a_k) \in S_\varepsilon(a_k)$ (recall that by Lemma 2 for every $a \in A_{ad}$ the set $S_\varepsilon(a)$ is nonempty). In what follows since ε is assumed to be fixed we omit the dependence of $u(a)$ on ε. From Remark 1(b) we know that K_{ad} is compact. Denoting subsequences with the same index as original sequences, we can find a subsequence of $\{a_k\}$ such that

$$a_k \to \widehat{a}_0 \quad \text{in } L^1(\Omega), \text{ as } k \to \infty$$

and by Remark 1(a), we have

$$a_k \to \widehat{a} \quad \text{in } L^2(\Omega), \text{ as } k \to \infty \tag{5}$$

with $a_0 \in K_{ad}$. Since $u_k \in S_\varepsilon(a_k)$ we have (cf. (4))

$$\int_\Omega (a_k(x) Du_k Dv + u_k v)\, dx + \int_{\Gamma_2} \xi_k v\, d\sigma(x) = \\ = \int_\Omega fv\, dx + \int_{\Gamma_1} gv\, d\sigma(x) \tag{6}$$

for all $v \in V_\varepsilon$, $V_\varepsilon = \{v \in H^1(\Omega) : v = 0 \text{ on } \widetilde{\Gamma_1^\varepsilon} \cup \Gamma_3\}$ and

$$\xi_k(x) \in \partial j(x, u_k(x)) \text{ a.e. } x \in \Gamma_2. \tag{7}$$

It follows from Lemma 2 that the sequence $\{u_k\}$ is bounded in V_ε independently of k. Also note that from $H(j)_1$ we have $|\xi_k(x)| \leq c_3(1 + |u_k(x)|)$ a.e. on Γ_2. Hence

$$\|\xi_k\|_{L^2(\Gamma_2)} \leq \widehat{c}(1 + \|u_k\|_{L^2(\Gamma_2)}) \leq \widehat{c}_1(1 + \|u_k\|_{V_\varepsilon})$$

with $\widehat{c}$, $\widehat{c}_1 > 0$ independent of k. Therefore the sequence $\{\xi_k\}$ remains in a bounded set in $L^2(\Gamma_2)$. Using again the compactness of the trace operator and passing to a subsequence, if necessary, we may assume that

$$\begin{cases} u_k \to u_0 \ \text{ weakly in } H^1(\Omega), \\ \qquad\quad \text{in } L^2(\Gamma_2) \text{ and a.e. in } \Gamma_2 \\ \xi_k \to \xi_0 \ \text{ weakly in } L^2(\Gamma_2) \end{cases} \tag{8}$$

with $u_0 \in V_\varepsilon$, $\xi_0 \in L^2(\Gamma_2)$. Then, going back to (6) and using convergences (5) and (8), in the limit, as $k \to \infty$, we get

$$\int_\Omega (a_0(x) Du_0 Dv + u_0 v)\, dx + \int_{\Gamma_2} \xi_0 v\, d\sigma(x) = \\ = \int_\Omega fv\, dx + \int_{\Gamma_1} gv\, d\sigma(x)$$

for all $v \in V_\varepsilon$. In order to infer that $u_0 \in S_\varepsilon(a_0)$ we have to prove that

$$\xi_0(x) \in \partial j(x, u_0(x)) \text{ a.e. } x \in \Gamma_2. \tag{9}$$

Indeed, since the values of ∂j are nonempty, compact and convex subsets of $\mathbb{R}$ and $\partial j(x, \cdot): \mathbb{R} \to 2^{\mathbb{R}}$ has a sequentially closed graph (cf. Clarke [4]), we deduce (cf. e.g. Denkowski et al. [19]) that $\partial j(x, \cdot)$ is also upper semicontinuous. This property together with (8) allows to apply Convergence Theorem (see Chapter 1.4 of Aubin and Cellina [20]) and from (7) we obtain (9). Hence it follows that $u_0 = u(a_0) \in S(a_0)$ and so the pair (a_0, u_0) is admissible for $(IP)_\varepsilon$. Finally note that from $H(F)$ and (8) we have

$$F(u_0) \leq \liminf_{k \to \infty} F(u_k) = m$$

and hence $a_0 \in K_{ad}$ is the desired optimal parameter. □

BOUNDARY HOMOGENIZATION

The goal is to study the asymptotic behavior of the sets of solutions to the direct problem $(DP)_\varepsilon$ as $\varepsilon \to 0$. We will find the form of the limit problem (DP) which is obtained from $(DP)_\varepsilon$. It turns out that the limit problem does not depend on the function g appearing in the Neumann boundary condition on Γ_1^ε.

Let us denote by χ_ε the characteristic function of $\widetilde{\Gamma_1^\varepsilon}$ on Γ_1, i.e. $\chi_\varepsilon = 1$ on $\widetilde{\Gamma_1^\varepsilon}$ and $\chi_\varepsilon = 0$

on Γ_1^ε. The crucial hypothesis on the geometrical structure of the partition of Γ_1 is as follows.

(H_b) : for any weak-$*$ convergent subsequence of $\{\chi_\varepsilon\}$ in $L^\infty(\Gamma_1)$ its limit function is different from zero almost everywhere on Γ_1.

Theorem 6 *If the hypotheses $H(j)_1$, $(H_0)_1$ and (H_b) hold, $u_\varepsilon \in S_\varepsilon(a_\varepsilon)$, $u_\varepsilon \to u$ weakly in $H^1(\Omega)$, a_ε, $a \in A_{ad}$ and $a_\varepsilon \to a$ in $L^2(\Omega)$, then $u \in S(a)$, where $S(a)$ denotes the solution set to the following limit problem:*

$$\left\{\begin{array}{l} -\operatorname{div}(a(x)Du) + u = f \quad \text{in } \Omega \\ u = 0 \ \text{ on } \Gamma_1 \cup \Gamma_3 \\ -\dfrac{\partial u(x)}{\partial n_a} \in \partial j(x, u(x)) \ \text{ a.e. on } \Gamma_2. \end{array}\right. \tag{DP}$$

Proof. Let $u_\varepsilon \in V_\varepsilon \subset H^1(\Omega)$ be a solution to $(DP)_\varepsilon$, where $V_\varepsilon = \{v \in H^1(\Omega) : v = 0 \text{ on } \widetilde{\Gamma_1^\varepsilon} \cup \Gamma_3\}$. So we have

$$\int_\Omega (a_\varepsilon(x) Du_\varepsilon Dv + u_\varepsilon v)\, dx + \int_{\Gamma_2} \xi_\varepsilon v\, d\sigma(x) =$$

$$= \int_\Omega fv\, dx + \int_{\Gamma_1} gv\, d\sigma(x) \tag{10}$$

for every $v \in V_\varepsilon$ and

$$\xi_\varepsilon(x) \in \partial j(x, u_\varepsilon(x)) \ \text{ a.e. } x \in \Gamma_2. \tag{11}$$

Assume also that $u_\varepsilon \to u$ weakly in $H^1(\Omega)$. From the hypothesis $H(j)_1$ we know that $\{\xi_\varepsilon\}$ is bounded in $L^2(\Gamma_2)$. So we may assume that

$$\xi_\varepsilon \to \xi \ \text{ weakly in } L^2(\Gamma_2) \tag{12}$$

with $\xi \in L^2(\Gamma_2)$. On the other hand, by the compactness of the trace mapping $H^1(\Omega) \to L^2(\Gamma)$, we obtain

$$u_\varepsilon|_{\Gamma_1} \to u|_{\Gamma_1} \ \text{ in } \ L^2(\Gamma_1), \tag{13}$$

$$u_\varepsilon|_{\Gamma_2} \to u|_{\Gamma_2} \ \text{ in } \ L^2(\Gamma_2), \tag{14}$$

as $\varepsilon \to 0$. From hypothesis (H_b) we have that there exists $\chi \in L^\infty(\Gamma_1)$, $\chi \neq 0$ a.e. on Γ_1 such that

$$\chi_\varepsilon \to \chi \ \text{ weakly} * \text{ in } \ L^\infty(\Gamma_1). \tag{15}$$

Next we observe that $u_\varepsilon|_{\Gamma_1}\, \chi_\varepsilon = 0$ a.e. on Γ_1. Thus from (13) and (15) we have $u|_{\Gamma_1}\, \chi = 0$ which implies $u|_{\Gamma_1} = 0$. It is easy to see that the condition $u_\varepsilon|_{\Gamma_3} = 0$ gives $u|_{\Gamma_3} = 0$. Hence $u \in V$, with $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_1 \cup \Gamma_3\}$. Of course we have $V \subset V_\varepsilon$ and (10) implies

$$\int_\Omega (a_\varepsilon(x) Du_\varepsilon Dv + u_\varepsilon v)\, dx +$$

$$+ \int_{\Gamma_2} \xi_\varepsilon v\, d\sigma(x) = \int_\Omega fv\, dx$$

for all $v \in V$. Using the convergences $a_\varepsilon \to a$ in $L^2(\Omega)$, $u_\varepsilon \to u$ weakly in $H^1(\Omega)$ and (12) we pass to the limit in the above inequality and obtain

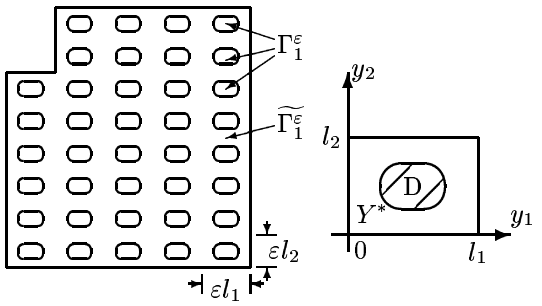
$$\int_\Omega (a(x) DuDv + uv)\, dx +$$

$$+ \int_{\Gamma_2} \xi v\, d\sigma(x) = \int_\Omega fv\, dx, \quad \forall\, v \in V.$$

Finally, a straightforward application of Convergence Theorem (see Aubin and Cellina [20]) to the inclusion (11) implies

$$\xi(x) \in \partial j(x, u(x)) \ \text{ a.e. } x \in \Gamma_2$$

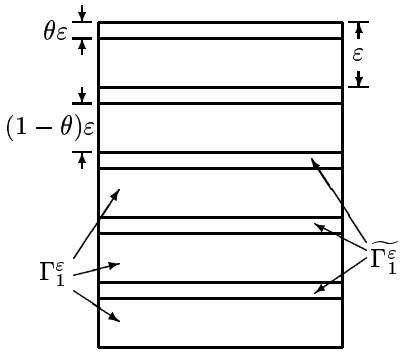
(the passage to the limit is possible due to (12) and (14)). Then clearly $u \in V$ solves (DP) which completes the proof. □



1. The ε-homothetic structure

Remark 7 *The assumption (H_b) imposes a restriction on the geometrical structure of $\widetilde{\Gamma_1^\varepsilon}$, as $\varepsilon \to 0$ and it implies that there exists a positive constant m such that for each $\varepsilon > 0$ the measure of $\widetilde{\Gamma_1^\varepsilon}$ in Γ_1 is not less than m. In particular, (H_b) is easily verified in the following two cases as depicted in Figure 1 and Figure*

2. In Figure 1, the boundary $\Gamma_1 \subset \mathbb{R}^2$ is ε-homothetic, periodic set obtained from the representative cell $Y = [0, l_1) \times [0, l_2)$. The part Γ_1^ε consists of the "holes" on the surface Γ_1 while $\widetilde{\Gamma_1^\varepsilon}$ is the complement of Γ_1^ε, $Y^* = Y \setminus \overline{D}$ and $\theta = m(Y^*)/m(Y)$. Figure 2 represents the boundary Γ_1 which has the layered structure. The hypothesis (H_b) is satisfied in both of these cases: the whole sequence χ_ε converges weakly-$*$ in $L^\infty(\Gamma_1)$ to θ. The result is the same for both structures, only the proportions of each partition count.



2. The layered structure

We are now in a position to formulate the inverse problem for the homogenized hemivariational inequality: find $a \in K_{ad}$ which solves the following minimization problem

$$\min_{a \in K_{ad}} \min_{u(a) \in S(a)} F(u(a)), \qquad (IP)$$

where $S(a)$ denotes the solution set of (DP).

First we state a result analogous to Theorem 5.

Theorem 8 *If the hypotheses $H(j)_1$, $H(F)$ hold and $f \in L^2(\Omega)$, then the inverse problem (IP) admits a solution on every admissible set K_{ad} of the form (2).*

For each positive ε, let us define the sets $\mathcal{N}_\varepsilon$ and $\mathcal{N}$ of solutions to $(IP)_\varepsilon$ and (IP), respectively, i.e.

$$\mathcal{N}_\varepsilon = \{(a^*, u(a^*)) \in K_{ad} \times S_\varepsilon(a^*) : F(u(a^*)) \leq F(u(a)) \text{ for all } a \in K_{ad}\}$$

and similarily for $\mathcal{N}$.

The main result is the following.

Theorem 9 *Under hypotheses $H(j)_1$, $H(F)$, $(H_0)_1$ and (H_b), we have*

$$K_{seq}(L^2(\Omega) \times (w\text{-}H^1(\Omega))) \limsup_{\varepsilon \to 0} \mathcal{N}_\varepsilon \subset \mathcal{N}.$$

Proof. Let $(a^*, u(a^*)) \in \limsup_{\varepsilon \to 0} \mathcal{N}_\varepsilon$. So $(a^*, u(a^*)) \in K_{ad} \times H^1(\Omega)$ and by definition of the upper limit there is a sequence $\{(a_\varepsilon^*, u_\varepsilon^*)\}_{\varepsilon > 0}$ such that $(a_\varepsilon^*, u_\varepsilon^*) \in \mathcal{N}_\varepsilon$ for every $\varepsilon > 0$ and $a_\varepsilon^* \to a^*$ in $L^2(\Omega)$,

$$u_\varepsilon^* \to u^* \text{ weakly in } H^1(\Omega),$$

where $u^* \in H^1(\Omega)$ and $u_\varepsilon^* = u(a_\varepsilon^*)$. Hence $u_\varepsilon^* \in S_\varepsilon(a_\varepsilon^*)$ and $F(u_\varepsilon^*) \leq F(u(a))$ for all $a \in K_{ad}$. From Theorem 6 we obtain $u^* \in S(a^*)$ and $u^* = u(a^*)$. By hypothesis $H(F)$ we have $F(u^*) \leq \liminf_\varepsilon F(u_\varepsilon^*) \leq F(u(a))$, for all $a \in K_{ad}$, which means that $(a^*, u(a^*)) \in \mathcal{N}$. □

We conclude this paper by pointed out that result analogous to Theorem 9 can be proved when the direct problem has the following form

$$\begin{cases} -\operatorname{div}(a(x)Du) + u = f \quad \text{in } \Omega \\ -\dfrac{\partial u(x)}{\partial n_a} \in \partial j(x, u(x)) \quad \text{a.e. on } \Gamma_1^\varepsilon \\ u = 0 \quad \text{on } \widetilde{\Gamma_1^\varepsilon} \\ \dfrac{\partial u}{\partial n_a} = \varphi \quad \text{on } \Gamma_2 \\ u = 0 \quad \text{on } \Gamma_3. \end{cases} \qquad (16)$$

In this case the boundary homogenization result for (16) reads as follows. By $M_\varepsilon(a)$ we denote the solution set of (16).

Theorem 10 *If the hypotheses $H(j)_1$, (H_b) hold, $f \in L^2(\Omega)$, $\varphi \in H^{1/2}(\Gamma_2)$, $u_\varepsilon \in M_\varepsilon(a_\varepsilon)$, $u_\varepsilon \to u$ weakly in $H^1(\Omega)$, a_ε, $a \in A_{ad}$ and $a_\varepsilon \to a$ in $L^2(\Omega)$, then $u \in M(a)$, where $M(a)$ is the set of solution to the homogenized problem:*

$$\begin{cases} -\operatorname{div}(a(x)Du) + u = f \quad \text{in } \Omega \\ u = 0 \quad \text{on } \Gamma_1 \cup \Gamma_3 \\ \dfrac{\partial u}{\partial n_a} = \varphi \quad \text{on } \Gamma_2. \end{cases} \qquad (17)$$

We underline that the limit problem for (16) is not a hemivariational inequality but the

boundary value problem for elliptic equation. The problem (17) does not depend on the function j appearing in the boundary condition on Γ_1^ε.

ACKNOWLEDGMENTS

Research supported in part by the State Committee for Scientific Research of the Republic of Poland (KBN) under Grants No. 2 P03A 004 19 and 7 T07A 047 18.

REFERENCES

1. P.D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications. Convex and Nonconvex Energy Functions*, Birkhäuser, Basel, 1985.

2. P. D. Panagiotopoulos, *Hemivariational Inequalities, Applications in Mechanics and Engineering*, Springer-Verlag, Berlin, 1993.

3. Z. Naniewicz and P. D. Panagiopopoulos, *Mathematical Theory of Hemivariational Inequalities and Applications,* Marcel Dekker, Inc., New York - Basel - Hong Kong, 1995.

4. F. H. Clarke, *Optimization and Nonsmooth Analysis,* Wiley - Interscience, New York, 1983.

5. A. Damlamian and Li Ta-tsien, Boundary homogenization for elliptic problems, *J. Math Pures et Appl.,* **66** (1987), 351-361.

6. S. Migórski and A. Ochal, *Inverse coefficient problem for elliptic hemivariational inequality,* in: *Nonsmooth/Nonconvex Mechanics: Modeling, Analysis and Numerical Methods*, D. Y. Gao, R. W. Ogden, G. E. Stavroulakis, Eds., pp. 247–262, Nonconvex Optimization and its Applications, Vol. 50, Kluwer, Dordrecht, Boston, London, 2001.

7. J. Haslinger and P. D. Panagiotopoulos, Optimal control of systems governed by hemivariational inequalities. Existence and approximation results, *Nonlinear Analysis, Theory, Methods, and Applications,* **24** (1995), 105–119.

8. M. Miettinen and J. Haslinger, Approximation of optimal control problems of hemivariational inequalities, *Numer. Funct. Anal. and Optimiz.,* **13** (1992), 43–68.

9. Z. Denkowski and S. Migórski, Optimal shape design problems for a class of systems described by hemivariational inequalities, *Journal of Global Optimization,* **12** (1998), 37–59.

10. S. Migórski and A. Ochal, Optimal control of parabolic hemivariational inequalities, *J. Global Optim.*, **17** (2000), 285-300.

11. S. Migórski, Stability of parameter identification problems with applications to nonlinear evolution systems, *Dynamics Systems Appl.,* **2** (1993), 387-404.

12. S. Migórski, Sensitivity analysis of inverse problems with applications to nonlinear systems, *Dynamic Systems and Applications*, **8** (1999), 73-89.

13. S. Migórski, Identification of nonlinear heat transfer laws in problems modeled by hemivariational inequalities, in: *Inverse Problems in Engineering Mechanics II,* M. Tanaka and G. S. Dulikravich, Eds., Elsevier Science B.V., 1998, 27–37.

14. S. Migórski, Parameter identification for evolution hemivariational inequalities and applications, in: *Inverse Problems in Engineering Mechanics III,* M. Tanaka and G. S. Dulikravich, Eds., Elsevier Science Ltd., 2002, 211–218.

15. E. Giusti, *Minimal Surfaces and Functions of Bounded Variation,* Birkhäuser, Boston, Basel, Stuttgart, 1984.

16. S. Gutman, Identification of discontinuous parameters in flow equations, *SIAM J. Control Optim.,* **28** (1990), 1049-1060.

17. E. Zeidler, *Nonlinear Functional Analysis and Applications II A/B,* Springer, New York, 1990.

18. K. C. Chang, Variational methods for nondifferentiable functionals and applications to partial differential equations, *J. Math. Anal. Appl.,* **80** (1981), 102–129.

19. Z. Denkowski, S. Migórski and N. S. Papageorgiou, *An Introduction to Nonlinear Analysis and its Applications*, Kluwer/Plenum, Boston, Dordrecht, London, vol. 1 and vol. 2, 2002, in press.

20. J.-P. Aubin and A. Cellina, *Differential Inclusions,* Springer-Verlag, Berlin - Heidelberg - New York - Tokyo, 1984.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

A COMBINATION OF GENETIC ALGORITHM AND NEURAL NETWORK FOR DIAGNOSING ARTERIOSCLEROTIC LESIONS

Pedro P.B. de Oliveira
Universidade Presbit. Mackenzie
R. da Consolação 896, Consolação
01302-907 São Paulo, SP – Brazil
pedrob@mackenzie.br

Osmar Vogler
Instituto Tecnológico de Aeronáutica
Pça. M.E. Gomes 50, V. das Acácias
12228-901 S.J. Campos, SP – Brazil
vogler@ele.ita.br

Cláudia E. da Matta
Centro Universitário Salesiano
Rua Dom Bosco 284, Centro
13600-900 Lorena, SP – Brazil
claudia@lo.unisal.br

ABSTRACT

Cardiovascular diseases are the largest cause of death in industrialised countries; hence, early diagnostic of bad conditions in patients can dramatically increase their chance of survival. One way of performing such an early detection is based on a laser system that is introduced into the patient's coronary so as to excite its inner walls, and an optical catheter that carries the resulting radiation to a Raman spectrometer at its other end. With the spectra obtained it is then possible to automatically diagnose the condition of the coronary. Here we report on an algorithm to perform such an automatic diagnosing. The approach relies on a quantisation of the intensity levels of the Raman spectra, and on a genetic algorithm, coupled to an artificial neural network, that is meant to learn to discriminate between three conditions of human coronaries: normal, atheromatous and calcified. While the neural network is the actual diagnosing system, the evolutionary algorithm is used to select the frequencies of the spectra that the neural network should account for. The best networks obtained have achieved 100% success rate, a remarkable result that rivals all its forerunners found in the literature, while preserving a simple solution scheme.

INTRODUCTION

The largest cause of mortality in industrialised countries are cardiovascular diseases, *arteriosclerosis* being the worst of them, as it affects important arteries that conducts blood to the heart (the coronaries, in this case) or to the brain. Healthy arteries are flexible and their cross-section area are sufficient to allow circulation of the required amounts of blood to those organs. However, smoking, stress, colestherol, ageing, etc, favour lipidic deposition onto the arteries' inner walls. At a certain point in this process, the so-called *atheroma* (or *atheromatous* tissue) is said to have been formed, characterising a preliminary stage of an unhealthy artery. This situation may become even worse, as *calcification* of the arteries become more likely, making them stiffer, with consequent loss of their elasticity. As a consequence of both conditions, *arteriosclerosis* has come about, what dramatically hinders the amount of blood that can circulate through the arteries.

The traditional procedure for diagnosing the health stage of an artery – normal, atheromatous or calcified – would involve histological analyses. More recently, technological advances gave rise to a new, faster and less invasive method that allows the diagnostic to be made *in vivo*, by introducing an optical catheter into the patient´s artery ([1]), linked to some spectroscopy technique, so that the diagnostic is obtained out of the collected spectra. In the case of when *Raman* spectroscopy ([2]) is used, the tissue is irradiated by an 830nm (infrared) laser beam and, as a result of the laser-tissue interaction, elastic and inelastic radiation is scattered. Optic sensors then capture the radiation (through other fiber optics), which is then processed, filtering out the elastic radiation (fluorescence) and noise, leaving only the signal component due to the inelastic scattering, which is the so-called *Raman radiation*. This signal provides information about the substances that constitute the target tissues.

This approach relies on a database of spectra (of the intensities of the Raman radiation at various frequencies or wavelengths), created out of an ensemble of coronaries, distributed in the three conditions mentioned above. With a subset of the samples classified by a human expert, one can use them as a reference for the automatic classification of the others. In the present work,

the latter consists of learning to identify the coronary conditions of the classified spectra by means of artificial neural networks ([3]), together with using an evolutionary computation approach ([4]) to search for the set of frequencies in the spectra that, for their relevance, should be taken into account during the learning and hence, the diagnostic processes. Our method extends the one in [5], improving it in various ways, thus achieving a significant more accurate automatic classification of human coronaries.

In the next section the method employed is described in detail, first by presenting the neural network architecture and general characteristics, and then, by presenting the genetic algorithm parameterisation. Subsequently, the results obtained are reported and discussed, and the last section provides concluding remarks.

THE METHOD

Preprocessing

Each Raman spectrum is defined by a distribution of intensity levels of radiation for various frequencies (or wavelengths); see Figure 2 for some examples. Before presenting these data to the neural network, each spectrum is first normalised in respect to the largest intensity value present in it. Then, the intensity values of the spectra are quantised; in the present work a quantisation level of 0.25 was used, meaning that the quantised spectra would have only five possible intensity values, from 0 to 1. Such a quantisation aims at facilitating the neural network learning, as it dramatically decreases the amount of data variation the network has to cope with during learning, even though the amount of data itself does not change. Naturally, the quantisation level has to be defined so as to change the general shape of the spectra, while still preserving their identity.

In addition to the latter preprocessing of the intensity values of the spectra, some preprocessing in the frequency range also takes place. Firstly, although the raw frequencies produced by the Raman spectrometer vary between 600 to 1800 cm^{-1}, only 754 frequencies are used for the sake of classification; this number derives from selecting the frequency window of interest, together with a frequency calibration procedure ([6]). Naturally, such a size would not be convenient neither viable to be used for training the neural network; hence, a frequency selection is required for rendering training feasible, which reduces the number of frequencies used to a value N<754. This is precisely the role of the genetic algorithm that is associated with our approach, as will be clear below.

A Neural Network

Neural networks have been used in most diverse areas of technology, with emphasis on pattern classification. Feedforward networks ([3]) display an appealing general aspect, with a simple architecture and well-known training algorithms, such as the backpropagation algorithm, where a strong (supervised) learning scheme is employed, in that input patterns are presented to the network, together with their correct, corresponding output patterns. Tolerance to data errors, example-based learning, and ability for data classification are the main characteristics that make neural networks a good candidate for pattern classification problems.

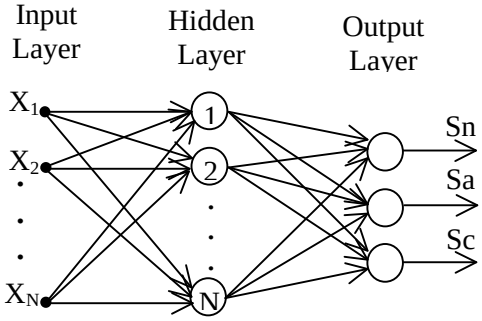


Figure 1: The neural network architecture.

The neural network architecture used in this work is shown in Figure 1. It is a feedforward, completely connected network, with one hidden layer.

The input layer size depends on the amount of variables selected by the method utilised to reduce to complexity of the problem; in the current case the network has to process N input data points, that is, the intensity values associated with each of the N selected frequencies that form a spectrum. The notation $X = (X_1, X_2 \ldots X_{N-1}, X_N)$ is, therefore, the data vector representing a subset of a spectrum, which is applied to all N nodes of the input layer.

The output layer possesses three nodes that classify the spectrum presented to the neural network in: Normal, Atheromatous and Calcified. *Sn*, *Sa* and *Sc* are the three nodes of the output

layer, representing those three coronary conditions, respectively.

A convention was established for the output values of the output nodes, as given by Table 1. Notice that each condition is defined by +1 in the corresponding output node, and -1 in the others.

Table 1: Expected values for the outputs.

Sa	Sn	Sc	Tissue
-1	+1	-1	Normal
+1	-1	-1	Atheromatous
-1	-1	+1	Calcified

Configuration. The identification of the network configuration followed [5], and was divided in three stages:
- Selection of the appropriate neural paradigm for the aplication: feedforward network, as a standard choice, that would allow trying standard variations of the backpropagation algorithm.
- Determination of the network topology to be used: a network with one hidden layer, with the same number of nodes as the input layer has yielded good performance and was preserved. The actual number (N) of input nodes is made fix during a run; however, various N-values have been tested.
- Determination of the parameters for the training algorithm and activation functions. This stage yields a great impact on the performance of the resulting system. The chosen activation function for all the network nodes, differently from [5], was *tansig* – the hyperbolic tangent sigmoid transfer function – which is a faster implementation of the hyperbolic tangent. This function came up as a natural choice for the output nodes, considering the concepts they should represent are in the range [-1, +1], the same one for the limit values of the function; also, the network was verified to provide superior performance when this function was also used in the hidden nodes.

Notice that, instead of using three output nodes for representing the three artery conditions of interest, only two nodes might have been sufficient. However, within the convention of Table 1, a mistaken measurement or a very noisy one, which should not be classified in any of the three artery conditions, leaves five alternatives for error, while there would be only one in the two-node alternative. Hence, the current choice can be seen as a safety measure in the diagnosis, as compared to the other alternative.

Training. A sample of 35 Raman spectra of human coronaries from real subjects was used, with the following characteristic distribution: 22 healthy, 5 atheromatous and 8 calcified; for the sake of simplifying the explanation of the training process, let us assume that all these spectra have already undergone the required preprocessing, including frequency selection (which, by the way, has not yet been described).

Initialisation of the weights in the network is randomly made. At this point the network can be trained; three training algorithms were tried out: standard and resilient backpropagation, and Levenberg-Marquardt. The steepest-descent of the standard backpropagation was too slow and could not converge; on its part, the Levenberg-Marquardt algorithm, usually considered quicker, demanded an excessive large amount of memory, thus hindering its use. The algorithm that presented more advantages was *resilient backpropagation*, which featured a reasonable trade-off between memory requirements and speed of convergence.

Every neural network was trained with the target of achieving 10^{-8} of mean square error between the outputs and the expected values, or until reaching 600 epochs of training (where each epoch is defined by the entire training set). Defining an acceptable error level at training is crucial for the network to be able to perform correctly when under test; it was observed that, for the present problem, a network with an error of 10^{-5} does have, in general, a very poor test performance.

Testing. The next step was testing the network, so as to determine the network performance with a data set not previously subjected to the network. For such, an ensemble of new 42 Raman spectra of human coronaries from real subjects was used, with the following characteristic distribution: 26 healthy, 6 atheromatous and 10 calcified. For best performance during the test, the intensity levels of these spectra were also quantised.

In order to measure the test error of a network, a quadratic error measure was used, as follows. Let us assume each test spectrum to be represented by T_i. When testing the *i*-th spectrum, its output vector in the network is $[o_i(a)\ o_i(n)\ o_i(c)]$, where the indexes *a*, *n* and *c*, refer to the

output nodes *Sa*, *Sn* and *Sc*, respectively, and the three o_i's are the rounded values of the actual network outputs. On the other hand, the correct output would be $[O_i(a)\ O_i(n)\ O_i(c)]$. The error associated with spectrum T_i can then be defined as $\varepsilon_{Ti} = [\varepsilon_i(a)\ \varepsilon_i(n)\ \varepsilon_i(c)]$, which can be rewritten as $\varepsilon_{Ti} = [(o_i(a)-O_i(a))^2\ (o_i(n)-O_i(n))^2\ (o_i(c)-O_i(c))^2]$. Finally, the total error of the entire ensemble of test spectra becomes $\varepsilon = \Sigma_i\ (\varepsilon_i(a) + \varepsilon_i(n) + \varepsilon_i(c))$, i=1, 2...,42.

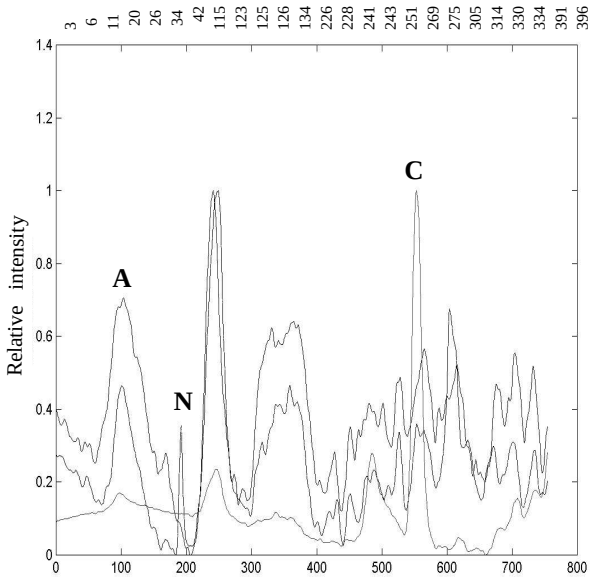


Figure 2: Three coronary spectra – **C**: calcified; **N**: normal; **A**: atheromatous – together with the respective 50 frequencies selected by the genetic algorithm.

The Genetic Algorithm

Jointly with neural networks, evolutionary computation techniques – and genetic algorithms, in particular – have also been utilised in virtually any area ([4]). Their robustness and conceptual simplicity are appealing features for being used a as a powerful search process. Even in situations in which a mathematical model is not available, or in which the associated search surfaces are very complex for traditional optimisers, evolutionary algorithms still display good chances of finding the global maximum.

As mentioned earlier, the role of the genetic algorithm here is to select the set of frequencies in the spectra that should be considered relevant for the neural network to rely on, so that they learn the patterns that characterise the three coronary conditions of interest.

Coding. When using a genetic algorithm, there is the necessity of coding a candidate solution (a chromosome) for each element of the population that will undergo the evolutionary process. In the present case, every chromosome represents one possible set of frequencies. Considering the coding has to represent N selected frequencies (out of the possible 754) every chromosome is represented as a 754-bit long string, with N positions set to bit 1, corresponding to each selected frequency; all the other bits in the chromosome are set to 0. The chromosome can then be regarded as a *mask* that is used onto a spectrum, so as to select the N frequencies (in fact, their corresponding intensities) that will be used to train or test a neural network. In Figure 2, only the listed frequencies would have the 1-bit in the corresponding positions of the associated mask, entailing that only the corresponding spectrum intensities would be used by a neural network.

The initial population is randomly generated, only imposing that each chromosome has to contain N 1-bits. Subsequent populations are produced through the usual genetic operators of mutation, crossover and elitism.

Elitism. At each generation, *elitism* is employed, that is, some of the best chromosomes are directly transferred to the next generation, without alteration. The use of elitism ensures a constant growth of the best fitness in the population, along the generations. However, as the network is initialised with random weights at each new generation, a chromosome with a good performance that is copied to the next generation may be evaluated in a different way, since the neural network would have a new, randomly generated configuration. Consequently, the original high fitness of a chromosome may not be preserved in the subsequent generation. The way this problem is circumvented is to preserve the neural networks associated with the best chromosomes, that is, the chromosomes are kept, in the next generation, together with their respective networks, with their weights after training.

Using such a kind of elitism, with 10% of the chromosomes, the best results were obtained.

Crossover. In order to perform the crossover, the mating pairs are selected by a standard fitness-proportional scheme (roulette-wheel selection). The mating pairs are then subjected to crossover at 60% rate. If crossover is performed, the offspring are transferred to the new population; otherwise, the mating pair is simply copied to the new population.

Notice that, with the current representation scheme of the chromosomes, if a standard crossover is performed (by simply swapping parts between two parents), very likely the offspring would have a number of 1-bits different of N, which would impair the predefined mask size associated with the search. In order to prevent that, a special crossover operator was devised which ensures that all offspring have the same mask size as their parents.

Mutation. The standard mutation rate was 10% of the population, and implemented through one pairwise random swap of bit positions in the mask.

It should also be remarked that the elite does not participate in crossover nor is it subjected to mutation.

Fitness Function. Each chromosome is evaluated according to the outcome of a neural network, when it is tested in the set of non-classified sample spectra; naturally, different chomosomes yield distinct fitness evaluations, as they represent different sets of frequencies that the neural network should consider during training and testing. The fitness of each chromosome is given simply by $1/\varepsilon$, where ε is the network's quadratic error measure, as defined in the previous section (naturally, preventing $\varepsilon=0$). As a consequence, in the begining of the search the population is roughly uniform in performance and selective pressure is very low, allowing the exploration of the search space; as evolution goes on, the creation of better chromosomes increases selective pressure more and more, allowing the best individuals to better exploit their regions of the search space.

Notice that the evaluation scheme used entails a discretisation of the fitness values. This derives from the fitness always being the result of sums of inversions of multiples of 4 (remember, for instance, that if the network output differs in only one the three components of the expected output vector, for one single test spectrum, then $\varepsilon=0+0+2^2=4$, thus leading fitness to 0.25).

RESULTS

The algorithm was executed for various neural network configurations, and variants of genetic operations. In general, the algorithm is computationally intensive, but yields significant performance.

Table 2 synthesises the results obtained. The first column refers to the use or not of quantisation in the input data; N is the mask size, that is, the number of frequencies the genetic algorithm has to select; the third column refers to the number of generations the genetic algorithm was allowed to run; the next column refers to the use or not of elitism; and the last column presents the overall succes rate over the 42 test spectra.

Table 2: Classification results obtained, with 30 chromosomes in the population. When used, quantisation level is 0.25 and elitism rate is 10%.

Quantisation	N	#Gens.	Elitism	Success
No	50	—	No	73%
No	50	300	No	80%
Yes	50	300	Yes	100%
Yes	100	100	Yes	100%

The first row corresponds to the situation of randomly choosing the frequencies and presenting them directly to the neural network, without applying the concept of evolutionary computation.

In the second row, the genetic algorithm is introduced, using crossover and mutation only, and with no quantisation of the spectra intensities; as a consequence, the results improve. The algorithm configuration used here is a slightly improved version over the approach in [5].

The third row extends the second case, now applying elitism and quantising the intensity levels for each frequency; as a consequence, classification of the samples improves, as does network convergence, thus yielding the excelent results represented by 100% success rate. The last row just shows a variation over the latter, in that the number of frequencies selected by the genetic algorithm was increased, but allowing the algorithm to run for fewer generations; this combination preserved the excelent performance already achieved.

A comparison between the results shown in the second and third rows of Table 2 is made

more evident in Figure 4, where it becomes clear the best evolution achieved by the current method, over its closest predecessor, represented by [5].

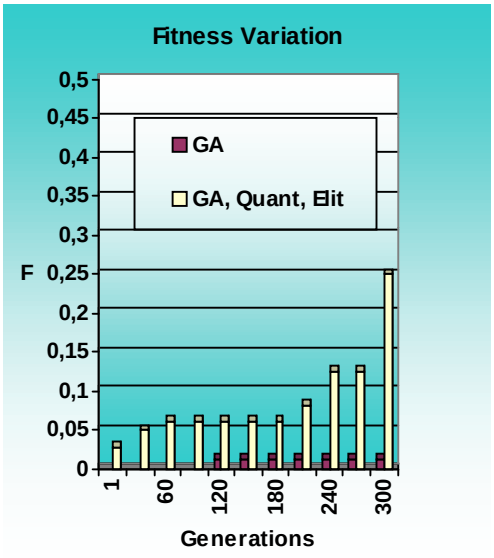


Figure 4: Fitness variation for 300 generations, as a comparison of the algorithm configurations shown in the second and third rows of Table 2.

Notice that the 73% success rate in the first results shown in Table 2 is surprisingly high. However, this does not mean that the problem at issue is too simple. First, remember that a neural network – a classifier *per se* – is underlying even the process that yielded those first results. Second, that supposedly high success rate is not sufficient in a medical situation like the one the current problem is related with, where the only acceptable possibility is close to 100% success rate. Third, one of the great challenges in the diagnosis of coronary injuries – even for a human expert – is the ability to discriminate between normal and atheromatous tissue spectra, since both possess overall similar aspects (see Figure 2); so, the tricky part of the diagnosis are really the subtleties in a spectrum that should provide the clues to discriminate between those two conditions.

Notice also that there is clear evidence that the spectrum quantisation is important for neural network training. So, by using the spectra without quantisation, the results get considerably worse, as the success rate fall from 100% down to about 80%.

Quantisation is also important for training in that it speeds up the network reaching a minimum error. For instance, at 0.25 quantisation level the network reaches the required training error within approximately 10 times faster than without quantisation; in other words, while the former situation requires about 70 epochs, for reaching a mean squared error of 10^{-8}, in the latter, not even 600 epochs are sufficient. Additionally, at 0.5 quantisation level the network requires 35 epochs.

Increasing the number of selected frequencies does not necessarily entail fitness build up, as the 100-bit mask size results showed (bottom row of Table 2). Also, because no significant fitness variation was perceived by changing mask size in the range from 30 to 100 bits, this shows that, in this range, a certain degree of robustness is exhibited by the classification procedure. However, as mask size decreases, a problem becomes more and more noticeable: the mean squared error of the neural network ceases to decrease; in extreme cases, such as with a 10-bit mask, no network manages to achieve the imposed 10^{-8} training error.

CONCLUDING REMARKS

An algorithm like the one discussed herein is meant to be the software core of a new, real-time system for diagnosing arteriosclerosis. This type of system would be extremely useful in the analysis of biological signals, as they require outstanding reliability of the detection algorithm.

This article did not try to cover all the possibilities offered by the problem, but only to consider a new, general and reliable solution. Alternative solutions can be found in the literature, similar or not to ours, based upon neural networks, evolutionary algorithms, wavelets, discriminant networks, and principal component analysis ([5], [6], [7], [8], [9] and [10]). While the present approach shares various features with those works – including the usage of the same data sets – and, to some extent, is a follow up of them, our results seem more promising than those obtained by our predecessors.

A considerable advance of the present approach was demonstrated by the success in the identification of all spectra of unknown classification. But recently, a similar performance has been reported ([9] and [10]). A key distinction between the two is their sophisticated wavelet

preprocessing of the input data, a contrast with our much simpler preprocessing due to the spectrum quantisation; also, our approach uses a more effective evolutionary computation algorithm, epitomised by the role of elitism, as well as a more natural and uniform neural network architecture.

In fact, the major thrust of the approach we presented is accuracy of classification, together with conceptual simplicity. The work extends the approach in [5], mainly by introducing the quantisation scheme of the spectra, by using elitism in the genetic algorithm, and by uniformising the neural network architecture through the same activation function in all the nodes.

Although the success of our method was demonstrated with spectra from real coronary samples, the spectra utilised were obtained with an exposure time of the samples to the Raman spectrometer, of around 0.5 sec. This success is certainly remarkable in its own sake, and practical from the medical standpoint; nonetheless, some medical situations may require about a tenth of this exposure time, what would produce noisier spectra. The performance of the current method in these noisier conditions is yet to be properly evaluated, as it is yet to be with the competing approaches.

We believe that quantisation can provide a simple and effective way to minimise the effect of noise; after all, a quantised noisy spectrum could even be identical to its noise-free version. However, in conditions with higher amounts of noise, quantisation has to be well supervised, since a high quantisation level (i.e., various intensity values) causes the neural network to learn too slowly; on the other hand, a small quantisation level causes loss of information from the spectrum.

No doubt, the winning approach will be the one with more robustness to handle noisier spectra, obtained with smaller exposure times. Lately, our method and the one in [10] have been probed under those stricter conditions, by subjecting the algorithms to noise-corrupted spectra, obtained out of the artificial injection of noise, at various levels, into the original spectra we used. Comparing their resulting discrimination ability, it is already clear to us the superiority of the approach described herein. However, these results are still informal and go beyond present purposes; details of such a new step in comparing the approaches will be published elsewhere.

The use of artificial neural networks for learning to correlate Raman spectra with some human trait abounds in the literature. Typically, the networks involved undergo a supervised training procedure not directly with the original Raman spectra obtained, but with a reduced input space, obtained out of principal components analysis; the resulting set of smaller feature vectors are then used in the supervised training process of the networks. For instance, [11] describes how multilayer perceptrons learn to correlate glucose concentration of the blood, with Raman spectra of the aqueous humor of the eye, so as to learn the Bayesian probabilities that glucose concentration lies in one of 3 ranges of physiological interest (hypoglycemic, normal or hyperglycemic). Similarly, in [12], multilayer perceptrons are trained for performing classification of skin lesions in 5 different classes, out of Raman spectra obtained directly from the lesions. Furthermore, in [13] multilayer perceptrons are trained with the reduced dimensionality Raman spectra of a group of clinical bacterial isolates (the spectra obtained from the actual whole-organisms!) associated with urinary tract infection, and manage to classify unseen samples in one of 5 possible classes, with slightly more than 80% success rate. In contrast with [11] and [12] – but, in tune with the work we report herein – in [13] multilayer perceptrons and radial-basis function networks are also used with the full Raman spectra; however, the results are worse than those with the reduced spectra. This contrast precisely clarifies the role of the genetic algorithm in the present approach: it reduces the dimensionality of the original spectra, not by transforming it, as multivariate methods do, but simply by filtering out those frequency components that should better not be accounted for in the classification process.

Finally, although one can think of non-invasive systems related to the consequences of arteriosclerosis – like the one in [14], used for hypertension detection – invasive systems like the one described here are, as far as we are aware of, still (and unfortunatelly) required, as a way of detecting the actual presence of arteriosclerotic lesions.

ACKNOWLEDGEMENTS

The presentation of this paper was made possible thanks to a grant provided by FAPESP (Proc. 02/00686-5), to which we are very grateful.

C.E.M. thanks Roberto K.H. Galvão and Atair R. Neto for discussions and fruitful advise. O.V. is grateful to IP&D-UNIVAP, for the provision of computational and physical infrastructure. We all thank Alderico R. Paula Jr. for various conversations on key issues of the paper.

REFERENCES

1. R. Kortum, A. Mehta, G. Hayes and R. Cothren. Spectral analysis of atherosclerosis using an optical fiber laser catheter**,** *Americam Heart Journal*, Vol. 118, p.381-391, 1989.

2. S. Sathaiah, L. Silveira Jr., C.A.G. Pasqualucci, R.A. Zângaro, C. Chavantes and M.T.T. Pacheco. Diagnosis of human coronary artery with near infrared Raman spectroscopy, *Proc. of the XV Int. Conf. on Raman Spectroscopy*, USA, p. 1120, 1996.

3. J.M. Zurada. *Introduction to artificial neural systems*, Boston: PWS Publishing Company, 1995.

4. D.E. Goldberg. *Genetic algorithms in search optimization and machine learning*, Addison-Wesley, 1989.

5. C.E. Matta. *A neural network based system for diagnosing arteriosclerotic lesions.* Tech. Report, School of Computer Science, Universidade do Vale do Paraíba, São José dos Campos, SP, Brazil, 2000. In Portuguese.

6. H. Sidaoui, R.A. Zângaro and M.T.T. Pacheco. Automatic system for handling spectral signals in real-time detection of atherosclerotic lesions. *Proc. of the XII Brazilian Congress on Automatics*, Vol. II, Uberlândia-MG, Brazil, p. 411-415, 1998. In Portuguese.

7. A.R. Paula Jr., C.M.F. Peris, H. Sidaoui and S. Sathaiah. Digital processing of Raman spectra for diagnosis of atherosclerosis. *Third Int. Caracas Conf. on Device, Circuits and Systems*, Cancún, Mexico, p. S75-1 to S75-6, 2000.

8. A.R. Paula Jr., C.E. Matta, E.R.M. Teixeira, L. Silveira Jr., P.R. Galhanone and R.K.H. Galvão. Raman spectra preprocessing for detection of atheromas through neural networks. *Proc. of the V Brazilian Conf. on Neural Networks*, Rio de Janeiro, Brazil, p. 259-264, 2001. In Portuguese.

9. A.R. Paula Jr., C.M.F. Peris and H. Sidaoui. Raman radiation preprocessing through wavelets and neural networks, for atheroma detection. *Proc. of the V Brazilian Symp. on Automatics*, Canela-RS, Brazil, p.1 (CD-ROM), 2001. In Portuguese.

10. A.R. Paula Jr. and S. Sathaiah. Raman spectral classification of atherosclerosis using neural networks and discriminant analysis. Accepted for *IEEE Caracas Int. Conf. on Device, Circuits and Systems*, to be held in Cancún, Mexico, 2002.

11. M. Storrie-Lombardi, J. Lambert and M. Borchert. Determining glucose levels from NIR Raman spectra of eyes. NASA Tech Briefs Online, NPO-20414, April 2000. Available from: http://www.nasatech.com/Briefs/Apr00/NPO20414.html. [Accessed Dec. 2001].

12. S. Sigurdsson, J. Larsen, L.K. Hansen, P.A. Philipsen and H.C. Wulf. Outlier estimation and detection: application to skin lesion classification. *Proc. of the IEEE 2002 Int. Conf. on Acoustics,Speech and Signal Processing*, IEEE Press, 2002.

13. R. Goodacre, E.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell and P.J. Rooney. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology*, 144, 1998, 1157–1170.

14. K.V. Chandrinos, M. Pilu, R.B. Fisher, and P.E. Trahanias. Image processing techniques for the quantification of atherosclerotic changes, VIII Mediterranean Conference on Medical and Biological Engineering and Computing, (MEDICON98), Lemesos, Cyprus, p. 14-19, 1998.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

NUMERICAL SOLUTION OF NONLINEAR SYNTHESIS PROBLEMS OF RADIATING SYSTEMS ACCORDING TO THE PRESCRIBED POWER DIRECTIVITY PATTERN

P. O. Savenko
Pidstryhach Institute for Applied Problems of Mechanics and Mathematics,
National Academy of Sciences of Ukraine,
Lviv, Ukraine
savenko@iapmm.lviv.ua

ABSRACT

The method of numerical solution of nonlinear inverse problems of the synthesis theory of radiating systems according to the given directivity pattern (DP) by power is stated. A variational statement of the problem, in which the mean square deviation of the prescribed and synthesized DP and restrictions on the norm of excitation sources is considered. The existence theorem of quasi-solutions is proved, the Euler equation for their finding is obtained. The conditions are determined and convergence of the used iterative processes is proved at numerical solution of the problem. On the basis of methods of branching theory of the nonlinear equations solutions it is shown, that for the nonlinear synthesis problem the bifurcation of solutions is characteristic. The equations for finding the bifurcation points are obtained. The quantity and characteristic properties in the space of real continuous functions, are determined. The numerical example of synthesis is given.

INTRODUCTION

One of the practically important classes of the problems originating on a design stage of audio and electrodynamics emanating systems, are the inverse problems (problem of synthesis), permitting to discover constructive optimal solutions [1-5]. The joining beginning of the inverse problems of acoustics and electrodynamics is the adequacy of mathematical models circumscribing various wave processes. Abstracting from a concrete type of radiating system on the operator level the inverse problem of finding the optimal distribution of external sources generating the field satisfying the given requirements to the characteristic of radiation power directivity pattern, is considered. Such problems are nonlinear and essentially ill-posed one. They are characterized by the nonuniqueness of solutions. The least investigated in the given class of the problems are the problems of an amount of existing solutions and their qualitative characteristics. The variational problems on search of quasi-solutions with usage of the smoothing functionals providing the best mean square approximation of DP synthesized to the given one are stated. The existence theorems of quasi-solutions are proved. Further, the problem of search of solutions is reduced to numerical solution and research of the Euler equation being a nonlinear equation with the operator of Hammerstein type. The appropriate iterative processes are constructed. The conditions are defined and their convergence is proved. By example of a linear antenna and linear antenna array it is shown, that for the given class of the problems bifurcation of solutions is characteristic. Their main properties are defined depending on the value of the parameter of regularization and properties of the prescribed power directivity pattern. It allows to localize existing solutions and in appropriate way to select initial approximation for obtaining solution of that or other type. The offered algorithms can be used in the process of solving the synthesis problem of various types of antennas and antenna arrays, including the mutual influence of sources.

STATEMENT OF INVERSE PROBLEM, THE EXISTENCE OF QUASI-SOLUTIONS

It is known [6], that the problem of electromagnetic field excitation in the unbounded homogeneous isotropic space (with dielectric

permeability ε and magnetic permeability μ) by external sources of electromagnetic oscillations, which are localized in some area $\overline{V} \in \mathbb{R}^3$ and vary in time according to the law $e^{i\omega t}$ (ω is the oscillation frequency), is reduced to the system of Maxwell equations with respect to $\boldsymbol{E}$, $\boldsymbol{H}$ which are the vectors of complex amplitudes of voltages of electrical and magnetic fields. Asymptotic of solutions of this system for $r \to \infty$ in a spherical coordinate system has the following form:

$$\begin{cases} \boldsymbol{E}(r,\vartheta,\varphi) = -i\omega\mu \dfrac{e^{-ikr}}{4\pi r}\{0,\ f_\vartheta(\vartheta,\varphi),\ f_\varphi(\vartheta,\varphi)\}, \\ \boldsymbol{H}(r,\vartheta,\varphi) = ik \dfrac{e^{-ikr}}{4\pi r}\{0,\ f_\varphi(\vartheta,\varphi), -f_\vartheta(\vartheta,\varphi)\}, \end{cases} \quad (1)$$

where $f_\vartheta(\vartheta,\varphi)$, $f_\varphi(\vartheta,\varphi)$ are the components of vector diagram of directness $\boldsymbol{f} = f_\vartheta \boldsymbol{i}_\vartheta + f_\varphi \boldsymbol{i}_\varphi$ of radiating system by field. The functions f_ϑ, f_φ, as a rule, are the integrated characteristics of the currents (fields) passing in the aperture of radiating system; their form and properties depend on the type and geometry of radiating system. The value

$$N(\vartheta,\varphi) = |\boldsymbol{f}(\vartheta,\varphi)|^2 = |f_\vartheta(\vartheta,\varphi)|^2 + |f_\varphi(\vartheta,\varphi)|^2 \quad (2)$$

characterizes the angular distribution of density of power flow and it is called the directivity pattern of radiating system by power.

Abstracting from the concrete type of radiating system, we present the function $\boldsymbol{f}(\vartheta,\varphi)$ with the help of the linear operator $A = \{A_\vartheta, A_\varphi\}$:

$$\boldsymbol{f} = A\boldsymbol{I} \qquad (f_\nu = A_\nu \boldsymbol{I},\quad \nu = \vartheta,\varphi), \quad (3)$$

which operates from some functional complex space H_I, to which functions of external currents (or fields), belong, into a functional complex space $C_f^{[2]}$ to which a set of realized DP belongs.

We consider the synthesis problem of the prescribed DP $N_0(\vartheta,\varphi)$ by power. In the elementary aspect it may be formulated as the problem of solutions determination of the first kind nonlinear operational equation

$$|A\boldsymbol{I}|^2 \equiv |A_\vartheta \boldsymbol{I}|^2 + |A_\varphi \boldsymbol{I}|^2 = N_0, \quad (4)$$

where $N_0(\vartheta,\varphi)$ is a real nonnegative function continuous on the compact $\overline{\Omega} \in \mathbb{R}^2$ (or $\overline{\Omega} \in \mathbb{R}^1$) (thus $\max\limits_{(\vartheta,\varphi)\in\Omega} N_0(\vartheta,\varphi) = 1$) which cannot belong to the set of values of the nonlinear operator $|A\boldsymbol{I}|^2$. It is known [7, 8], that the problem (4) is essentially ill-posed. Thus problem of finding the quasi-solutions of the equation (4) in variational statement, is considered.

Let's introduce into consideration the Gilbert space $H_I = L^2[\overline{V}] \oplus L^2[\overline{V}] \oplus L^2[\overline{V}]$ which is a complex space of square integrable vector-valued functions defined on the compact $\overline{V}$, and $C_f^{[2]} = C[\overline{\Omega}] \oplus C[\overline{\Omega}]$ which is a complex space of vector-valued continuous functions on $\overline{\Omega}$ for real arguments, equipped with a scalar product.

In the space $C_f^{[2]}$ alongside with the Chebyshev norm $\|\boldsymbol{f}\|_C = \max\limits_{(\vartheta,\varphi)\in\overline{\Omega}} |\boldsymbol{f}(\vartheta,\varphi)|$, where $|\boldsymbol{f}(\vartheta,\varphi)| = \left(|\boldsymbol{f}_\vartheta(\vartheta,\varphi)|^2 + |\boldsymbol{f}_\varphi(\vartheta,\varphi)|^2\right)^{1/2}$, we shall introduce the mean square metric, generated by a scalar product and norm:

$$(\boldsymbol{f}_1, \boldsymbol{f}_2) = \left(f_\vartheta^{(1)}, f_\vartheta^{(2)}\right) + \left(f_\varphi^{(1)}, f_\varphi^{(2)}\right) \equiv$$

$$\equiv \int_{\overline{\Omega}} \left[f_\vartheta^{(1)}(\vartheta,\varphi)\overline{f_\vartheta^{(2)}(\vartheta,\varphi)} + f_\varphi^{(1)}(\vartheta,\varphi)\overline{f_\varphi^{(2)}(\vartheta,\varphi)} \right] \sin\vartheta d\vartheta d\varphi,$$

$$\|\boldsymbol{f}\|_{C_f^{[2]}} = (\boldsymbol{f}, \boldsymbol{f})^{1/2} = \left(\|f_\vartheta\|^2 + \|f_\varphi\|^2\right)^{1/2},$$

$$\rho_{C_f^{[2]}}(\boldsymbol{f}_1, \boldsymbol{f}_2) = \|\boldsymbol{f}_1 - \boldsymbol{f}_2\|_{C_f^{[2]}}.$$

It is supposed, that the set of zeros of operator A consists only of zero element, i.e. $N(A) = \theta$ The problem about the best mean square approximation of nonnegative real function $N_0(\vartheta,\varphi)$, continuous on area $\overline{\Omega}$, by function

$$|\boldsymbol{f}(\vartheta,\varphi)|^2 \quad (\boldsymbol{f}(\vartheta,\varphi) = A\boldsymbol{I} \in R(A),$$

$\boldsymbol{I} = \{I_x, I_y, I_z\} \in H_I$) is stated. We formulate it as the minimization problem of a smoothing functional

$$\sigma_\beta(\boldsymbol{I}) = \left\| N_0 - |A\boldsymbol{I}|^2 \right\|_{C_f^{(2)}}^2 + \beta\|\boldsymbol{I}\|_{H_I}^2 \equiv$$

$$\equiv \left\| N_0 - |\boldsymbol{f}|^2 \right\|_{C_f^{(2)}}^2 + \beta\|\boldsymbol{I}\|_{H_I}^2 \quad (5)$$

on the space H_I, where $\beta > 0$ is a real weight parameter.

Theorem 1. *Let the linear operator* $A: H_I \to C_f^{[2]}$ *be quite continuous,* $N_0(\vartheta,\varphi)$ *is*

given nonnegative function continuous on $\overline{\Omega}$, *and* $\|N_0(\vartheta,\varphi)\|_C = 1$.

Then there exists at least one point of absolute minimum of functional $\sigma_\beta(\boldsymbol{I})$ *in* H_I *and every minimizing sequence contains a subsequence that weakly converges to one of the points of absolute minimum.*

Since H_I is a reflexive Banach space, to prove the theorem it is enough to show [9] the fulfilling of the following conditions:

(**i**) $\sigma_\beta(\boldsymbol{I})$ is a weakly lower semicontinuous functional,

(**ii**) $\lim\limits_{\|\boldsymbol{I}\|_{H_I}\to\infty} \sigma_\beta(\boldsymbol{I}) = +\infty$.

NUMERICAL SOLUTION OF THE PROBLEM

To find numerically the points of minimum and research of their qualitative characteristics we use the Euler equation

$$\boldsymbol{I} = B(\boldsymbol{I}) \equiv \frac{2}{\beta} A^*\left(N_0 \cdot A\boldsymbol{I}\right) - \frac{2}{\beta} A^*\left(|A\boldsymbol{I}|^2 \cdot A\boldsymbol{I}\right) \quad (6)$$

in the space H_I. The equation (6) is the nonlinear equation containing in the right part (except for linear) the nonlinear Hammerstein type operator.

Applying operator A to both sides of the equation (6) and taking into account, that $N(A) = \theta$, we obtain an equation for the synthesized DP in space $C_f^{[2]}$ that is equivalent to (6):

$$\boldsymbol{f} = D(\boldsymbol{f}) \equiv \frac{2}{\beta} AA^*\left(N_0 \cdot \boldsymbol{f}\right) - \frac{2}{\beta} AA^*\left(|\boldsymbol{f}|^2 \cdot \boldsymbol{f}\right). \quad (7)$$

Corollary 1. *Since the functional* σ_β *is Gateaux differentiable on* H_I, *has at least one valley and posses the* $m-$*property (valley is an interior point of some convex set, belonging to* H_I*), the equation* (6) *in the space* H_I *and equation* (7) *in the space* $C_f^{[2]}$ *have, at least, one solution each.*

Lemma 1. *Under the conditions of theorem 1, for limited values of parameter* β $(0 < \beta < +\infty)$

$$D(\boldsymbol{f}) = \frac{2}{\beta} AA^*\left(N_0 \cdot \boldsymbol{f}\right) - \frac{2}{\beta} AA^*\left(|\boldsymbol{f}|^2 \cdot \boldsymbol{f}\right) \quad (8)$$

is a completely continuous operator in the space $C_f^{[2]}$.

From lemma 1 it follows, that for limited values of parameter β the operator $D(\boldsymbol{f})$ maps each limited set into relatively compact set in the space $C_f^{[2]}$. Since for elements of the relatively compact subset of the normalized space, the strong convergence and weak one coincide [10], from the theorem 1 and lemma 1 follows

Corollary 2. *If* $\{\boldsymbol{I}_n\}$ *is a minimizing sequence of functional* $\sigma_\beta(I)$ *which weakly converges to a point of minimum* $\boldsymbol{I}_*$, *then sequence* $\{\boldsymbol{f}_n = A\boldsymbol{I}_n\} \in C_f^{[2]}$ *converges uniformly to* $\boldsymbol{f}_* = A\boldsymbol{I}_*$ *in* $C_f^{[2]}$.

At the beginning we consider an iterative process that calculates the solutions of equation (6) for the function of excitation sources distribution $\boldsymbol{I} \in H_I = L^2[\overline{V}] \oplus L^2[\overline{V}] \oplus L^2[\overline{V}]$. The equation (6) we rewrite as

$$\left(E - \frac{2}{\beta} A^* N_0 A\right)\boldsymbol{I} = -\frac{2}{\beta} A^*\left(|A\boldsymbol{I}|^2 \cdot A\boldsymbol{I}\right), \quad (9)$$

where $E : H_I \to H_I$ is an identity operator. If $\beta > 2\|A^* N_0 A\|$, there exists an inverse operator [11] $\left(E - \frac{2}{\beta} A^* N_0 A\right)^{-1}$. Using this operator the equation (9) takes the form

$$\boldsymbol{I} = \tilde{B}(\boldsymbol{I}) \equiv -\frac{2}{\beta}\left(E - \frac{2}{\beta} A^* N_0 A\right)^{-1} A^*\left(|A\boldsymbol{I}|^2 \cdot A\boldsymbol{I}\right). \quad (10)$$

Let us show, that the solution of equation (10) may be obtained as a limit of successive approximations of the following iterative process [12]

$$\boldsymbol{I}_{n+1} = \vartheta \boldsymbol{I}_n + (1-\vartheta)\tilde{B}(\boldsymbol{I}_n) \quad (n = 0,1,2,...), \quad (11)$$

where $\vartheta \in (0,1)$.

Since the Gilbert space H_I is the Banach strictly convex space, then to prove the convergence of iterative process (11) it is enough to show [12], that $\tilde{B}(\boldsymbol{I})$ is quite continuous and non-expanding operator satisfying the condition $\tilde{B}(S_r) \subset S_r$, where

$$S_r = \left\{\boldsymbol{I} : \|\boldsymbol{I}\|_{L_2} \le r\right\}, \; r = \left(\frac{1 - \mu\|A^* N_0 A\|}{3\mu\|A\|^4}\right)^{1/2},$$

$$(\mu = 2/\beta). \qquad (12)$$

The proof of the named properties of the operator $\tilde{B}(\boldsymbol{I})$ follows from the lemmas given below.

Lemma 2. *Let* $A: H_I \to C(\overline{\Omega})$ *is a linear completely continuous operator,* $\beta > 2\left\|A^* N_0 A\right\|$. *Then* $\tilde{B}(\boldsymbol{I})$ *is a non-expanding operator on* $S_r \subset H_I$, *i.e. for any* $\boldsymbol{I}_1, \boldsymbol{I}_2 \in S_r$ *the inequality* $\left\|\tilde{B}(\boldsymbol{I}_1) - \tilde{B}(\boldsymbol{I}_2)\right\|_{L_2} \le \left\|\boldsymbol{I}_1 - \boldsymbol{I}_2\right\|_{L_2}$ *holds.*

Lemma 3. *Let* $A: H_I \to C(\overline{\Omega})$ *is a linear completely continuous operator,* $\beta > 2\left\|A^* N_0 A\right\|$. *Then* $\tilde{B}(\boldsymbol{I}): H_I \to H_I$ *is a completely continuous operator satisfying the condition* $\tilde{B}(S_r) \subset S_r$.

We notice, that the successive approximations (11), depending on the choice of initial approximation, may converge to various solutions of the equation (10).

RESEARCH OF SOLUTIONS STRUCTURE

Nonlinear operational equations (6), (7) have nonunique solution. Let's consider the structure of the solution of the equation (7) by an example of the linear radiator synthesis problem. It is known [6], that DP of linear radiator of length $2a$, directed along an axis OZ and placed in unbounded isotropic and homogeneous space with exactness to constant multiplier is described by the formula

$$f(s) = AI \equiv \sqrt{\frac{c}{2\pi}} \int_{-1}^{1} I(z) e^{iczs} dz, \qquad (13)$$

where $s = \sin\vartheta'/\sin\alpha$ is generalized angular coordinate, $c = ka\sin\alpha$ is the real dimensionless parameter describing the electrical length of radiator, $k = 2\pi/\lambda$ is the wave number in vacuum, λ is length of a wave, 2α is the corner, in which it is necessary to direct the maximal portion of irradiating power. It is assumed, that $\lambda >> a$, and the corners ϑ', α are counted from a plane XOY. The DP by power is determined by the expression $N(s) = |AI|^2 \equiv |f(s)|^2$.

The formula (13) is considered also as mapping from the complex space of square integrable functions $H_I = L_2[-1,1]$ into the complex space $C[-1,1]$ of continuous functions for real argument, which is carried out by the linear completely continuous integrated operator A.

The conjugate operator A^* we find from equality $(f, AI) = (A^* f, I)$:

$$A^* f \equiv \sqrt{\frac{c}{2\pi}} \int_{-1}^{1} f(s) e^{-iczs} ds$$

Taking into account the form of operators A, A^*, we get the developed form of equations (6), (7) in corresponding spaces:

$$I(z) \equiv \frac{2}{\beta}\sqrt{\frac{c}{2\pi}} \int_{-1}^{1} \left\{ \left[N_0(s) - \frac{c}{2\pi} \left| \int_{-1}^{1} I(z') e^{icz's} dz' \right|^2 \right] \times \right.$$
$$\left. \times \int_{-1}^{1} I(z') e^{icz's} dz' \right\} e^{-iczs} ds, \qquad (14)$$

$$f(s) \equiv \frac{2}{\beta} \int_{-1}^{1} K(s,t,c) N_0(t) f(t) dt -$$
$$- \frac{2}{\beta} \int_{-1}^{1} K(s,t,c) |f(t)|^2 f(t) dt, \qquad (15)$$

where

$$K(s,t,c) = \frac{c}{2\pi} \int_{-1}^{1} e^{icz(s-t)} dz \equiv \frac{\sin c(s-t)}{\pi(s-t)}. \qquad (16)$$

The equation (15) is easier, than the equation (14), since while its determination it is possible to lead integration (16) in an obvious form. Therefore its solutions are investigated.

Let's consider the solutions structure of the equation (15), depending on value of parameter β. For this purpose, we replace the equation (15) by equivalent system, using the equality $f(s) = u(s) + iv(s)$:

$$\begin{cases} u(s) = B_1(u,v) \equiv \dfrac{2}{\beta} \displaystyle\int_{-1}^{1} K(s,t,c) N_0(t) u(t) dt - \\ \quad - \dfrac{2}{\beta} \displaystyle\int_{-1}^{1} K(s,t,c) \left[u^2(t) + v^2(t)\right] u(t) dt, \\ v(s) = B_2(u,v) \equiv \dfrac{2}{\beta} \displaystyle\int_{-1}^{1} K(s,t,c) N_0(t) v(t) dt - \\ \quad - \dfrac{2}{\beta} \displaystyle\int_{-1}^{1} K(s,t,c) \left[u^2(t) + v^2(t)\right] v(t) dt. \end{cases} \qquad (17)$$

We show the two important properties of system solutions (17), which are directly checked:

1°. If $u_*(s), v_*(s)$ is the solution of system, then $\begin{pmatrix} u(s) \\ v(s) \end{pmatrix} = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix} \times \begin{pmatrix} u_*(s) \\ v_*(s) \end{pmatrix}$ is also its solution, where γ is any real constant, i.e. the system (12) has one-parametrical families of solutions.

2°. When the prescribed function $N_0(s)$ is even one, the integrated operators B_1, B_2 of the system (17), map the even functions $u(s), v(s)$ to even ones. This property allows to determine invariant sets in space $C[-1,1]$ and thus to locate the existing solutions.

It is obvious, that at any $\beta > 0$ one of the solutions of system (17) is the trivial solution: $u(s) \equiv 0$, $v(s) \equiv 0$.

At the beginning, the system (17) is considered for $N_0(s) \equiv 1$ in real space. In this case it transforms to one equation of a form

$$u(s) = \mu \int_{-1}^{1} K(s,t,c)u(t)dt - \mu \int_{-1}^{1} K(s,t,c)u^3(t)dt, \quad (18)$$

where $\mu = 2/\beta$. The problem on determination such values of parameter $\mu_n = 2/\beta_n$ ($n = 0,1,2,...$) and all continuous, different from trivial solutions $\omega_n(s)$, satisfying the condition

$$\max_{s\in[-1,1]} |\omega_n(s)| \to 0 \quad \text{when} \quad \eta = \mu - \mu_n \to 0,$$

is considered. According to [13] the values of spectral parameter $\lambda_n = \frac{1}{\mu_n} = \frac{\beta_n}{2}$ $(n = 0,1,2,...)$, which are eigenvalues of linear homogeneous equation

$$\lambda\varphi(s) = \int_{-1}^{1} \frac{\sin c(s-t)}{\pi(s-t)} \varphi(t)dt, \quad (19)$$

can be the branch points of equation (18).

The eigenfunctions of equation (19) are the extended spheroidal wave functions $\varphi_n(s) = S_{0n}(c,s)$ $(n = 0,1,2,...)$ of zero order [14]. They form the complete orthogonal system in an interval $[-1,1]$. Since the kernel $K(s,t,c)$ is the symmetric and positive one, the eigenvalues λ_n of equation (19) are real and positive. They monotonously decrease with the growth of n: $\lambda_0 > \lambda_1 > \lambda_2 > \ldots$. A sequence of values of parameter $\beta_n = 2\lambda_n$, as possible bifurcation points of equation (18), also forms the sequence, monotonously decreasing and aspiring to zero: $\beta_0 > \beta_1 > \beta_2 > \ldots$.

The ***regular case***, when $\tilde{\mu}$ does not coincide with one of characteristic values of equation (19), is considered. It is shown, that in this case the nonlinear equation (18) has only trivial solution.

When μ_n is the characteristic values of equation (19) (multipleness of characteristic value is equal to unit), ***the case of one-dimensional branching*** of solutions take place. Assuming $\mu = \mu_n + \eta$, and using the methods of branching theory of nonlinear equations solutions we obtain, that in points $\beta_n = 2\lambda_n = 2/\mu_n$ $(n = 0,1,2,...)$ under condition $\eta > 0$ the two real solutions of equation (18), branch off from trivial one, which in the first approximation have the form:

$$u_{1,2}^{(n)}(s) = \pm\sqrt{2/\beta_n} S_{on}(c,s)\eta^{1/2} + o(\eta^{1/2}). \quad (20)$$

Since the functions $S_{0n}(c,s)$ are even when n is even and they are odd when n is odd, the branched off solutions possess (at the first approximation) the same property.

The function $S_{00}(c,s)$ has the maximal concentration of power in an interval of visibility and it is the least jet function in a class W_a [6]. In particular, $S_{00}(c,s)$ is the least jet total DP, and odd function $S_{01}(c,s)$ is the difference DP. From this follows, that while synthesizing DP close to total one the parameter β in functional (5) should be chosen from the condition $\beta \geq \beta_0 = 2\lambda_0$, and while synthesizing the difference DP this condition has a form $\beta \geq \beta_1 = 2\lambda_1$.

Since the system (17) is symmetric with respect to unknown functions *u, v*, the similar results are obtained for function $v(s)$ in class of only imaginary functions.

Let's consider the case, when the given DP $N_0(s) \neq 1$. The problem of determining the bifurcation points of solutions is reduced to solving of linear equation

$$\varphi(s) = \mu \int_{-1}^{1} N_0(t)K(s,t,c)\varphi(t)dt. \quad (21)$$

It is shown that eigenvalues of equation are real and positive, and they form a monotonously decreasing sequence, and eigenfunctions are orthogonal. They are defined by a numerical way, using both the mechanical quadrature and the Danylevsky method [15].

It is shown, that the solutions of nonlinear equation

$$u(s)=\mu\int_{-1}^{1}N_0(t)K(s,t,c)u(t)dt-\mu\int_{-1}^{1}K(s,t,c)u^3(t)dt \quad (22)$$

in the bifurcation points have a form (at the first approximation) analogous to (20).

For even functions $N_0(s)$ eigenfunctions $\varphi_n(s)$ of the equation (21) are even when n is even and they are odd when n is odd. Hence, the characteristic properties of the branching off nontrivial solutions in the points $\beta_n=2\lambda_n$ are analogous to above mentioned ones for $N_0(s)=1$.

The general structure of the real solutions for any even DP can be schematically represented by solutions "tree" (fig. 1). Its "trunk" corresponds to the trivial solution, and branches correspond to branching off solutions. From figure it is seen

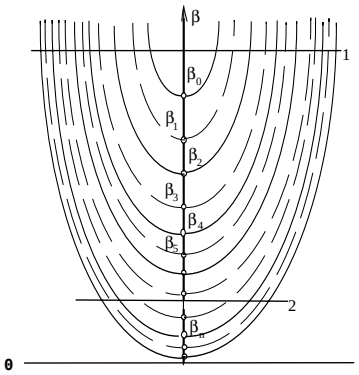


Fig. 1. A tree of the solutions of the equation (22)

(horizontal straight line 1), that for a choice of parameter β in functional $\sigma_\beta(\boldsymbol{I})$ it is expedient to find eigenvalues of the equation (21) and to put $\beta\approx 2\lambda_0$. In this case the solution of the nonlinear equation (22) includes the first eigenfunctions of the corresponding to it linear equation (21), which have the least jet factor. For small values β (horizontal straight line 2) the solution of the nonlinear equation (22) can be presented only through eigenfunctions $\varphi_n(s)$ with a high index, which are quickly oscillation ones.

The analogous results are received for the synthesis problem of the linear equidistant antenna array, DP of which is the Fourier discrete transformation.

Consider a numerical example of two lobe DP synthesis $N_0(s)=|\sin(\pi s)|$. The prescribed amplitude pattern and synthesized DP adequate to the solutions with various types of functions parity of $u(s)$, $v(s)$ are shown in fig. 2.

The current amplitude distributions, creating these DP are given in fig. 3. Let's pay attention to the solution with number 1. From a fig. 3 it is seen, that the amplitude distribution of a current is nonsymmetrical with respect to the antenna center. However, corresponding to it DP by power is symmetrical.

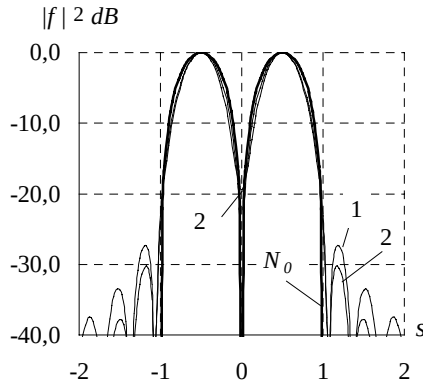


Fig. 2. Given and synthesized DP for $N_0(s)=|\sin\pi s|$, corresponding to various types of the solutions

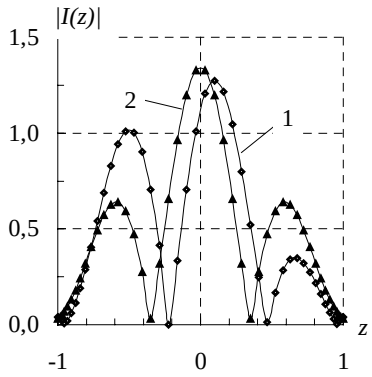


Fig. 3. Optimum currents creating DP, shown in fig. 2.

CONCLUSION

The supposed method of numerical solution of the synthesis problems is also applicable to the solving of the synthesis problems of various types of radiating systems. For this purpose it is necessary to determine a form of the operator A for concrete type of radiating system and, using the equality $(AI, f) = (I, A^* f)$, to find the form of conjugated operator A^*.

It is simply to be convinced that the operators AA^* for many types of radiating systems, having a symmetrical radiating aperture posses the property preservation of parity for the even prescribed DP N_0. This property allows to locate the solutions and to choose by corresponding way the initial approximation to obtain the solution of this or that type.

The stated technique can be also used for synthesis of antenna arrays using various by exactness mathematical models for the solving the direct problem. In this case currents $\boldsymbol{I}$ on the radiators are connected with stimulating them external voltages $\boldsymbol{U}$ by system of the linear equations $\boldsymbol{ZI} = \boldsymbol{U}$, where $\boldsymbol{Z}$ is a matrix or matrix-integral operator. If there exists a stable solution of this system, then having put $\boldsymbol{I} = \boldsymbol{Z}^{-1}\boldsymbol{U}$ and having presented DP of arrays as $\boldsymbol{f} = A\boldsymbol{Z}^{-1}\boldsymbol{U}$, analogously to stated above it is possible to solve the problems of constructive synthesis.

REFERENCES

1. M. I. Andriychuk, N. N. Voitovich, P. A. Savenko, V. P. Tkachuk, *The Antenna Synthesis According to the Amplitude Radiation Pattern. Numerical Methods and Algorithms*, Naukova Dumka, Kiev, 1993, p. 256.
2. P. A. Savenko. Numerical Solution of a Class of Nonlinear Problems in Synthesis of Radiating Systems. *Computational Mathematics and Mathematical Physics*, **40**, No. 6, pp. 889-899 (2000).
3. P. A. Savenko, L. M. Pasnak. Synthesis of a Linear Nonuniform Antenna Array of Coupled Dipoles by a Given Amplitude Pattern. *Journal of Communications Technology and Electronics*, **42**, No. 5, pp. 521-526 (1997).
4. P. A. Savenko, M. D. Tkach. Numerical Solution to the Nonlinear Synthesis Problem for a Microstrip Antenna Array with Allowance for the Mutual Coupling of Array Radiators. *Journal of Communications Technology and Electronics*, **46**, No. 1, pp. 50-57 (2001).
5. Savenko P. O., Anokhin V. J. Synthesis of Amplitude-Phase Distribution and Shape of a Plane Antenna Aperture for a Given Power Pattern // *IEEE Trans. on Antennas and Propag*. **45**, No. 4. – pp. 744-747 (1997).
6. G. T. Markov, B. M. Petrov G. P. Grudinskaya, *Electrodynamics and Spreading of Radiowaves*, Sov. Radio, Moskow, 1979, p. 374.
7. A. N. Tikhonov, V. Y. Arsenin, *The Methods of Solution of Incorrect Problems*, Nauka, Moskow, 1979, p. 295.
8. A. N. Tikhonov, A. V. Goncharskyi, V. V. Stepanov, and A. G. Yagola, *Regularizing Algoritms and A Priori Information*, Nauka, Moskow, 1983, p. 200.
9. M. M. Vainberg, *A Variational Method and a Method of Monotonous Operators*, Nauka, Moskow, 1972, p. 415.
10. M. M. Vainberg, *The Functional Analysis*, Nauka, Moskow, 1979, p. 128.
11. V. A. Trenogin, *The Functional Analysis*, Nauka, Moskow, 1980, p. 525.
12. M. A. Krasnosel'skij, G. M. Vajnikko, P. P. Zabreyko, i. e. *The Approached Solution of the Operational Equations*, Nauka, Moskow, 1969, p. 456.
13. M. M. Vainberg, V. A. Trenogin, *The Theory of Branching of Solutions of the Nonlinear Equations*, Nauka, Moskow, 1969, p. 525.
14. B. M. Minkovich, V. P. Yakovlev, *The Theory of Synthesis of Antennas*, Sov. Radio, Moskow, 1969, p. 296.
15. B. P.Demidovich, I. A. Maron, *The Bases of Calculating Mathematics*, Nauka, Moskow, 1970, p. 664.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

A PARALLEL GLOBAL OPTIMIZATION ALGORITHM FOR SOLVING INVERSE PROBLEMS

Henryk Telega
Institute of Computer Science
Jagiellonian University
Kraków, Poland
telega@ii.uj.edu.pl

ABSTRACT

In this paper two improved versions of *Genetic Clustering* (*GC*) algorithm [1] are described. *GC* is a parallel global optimization algorithm that was designed in order to solve such parameter inverse problems in which an approximation of certain level sets (central parts of basins of attractions of local minimizers) is required. The approximation of these sets can be useful when some additional criteria of optimization are considered after main results of parameter identification are obtained. In spite of some good properties of *GC*, tests have shown that *GC* is not effective for problems with more than 4 dimensions.

Two modifications of *GC* are proposed in order to overcome the dimensionality limitation. In the first modification clusters are remembered as ellipsoids. The second modification is based on the idea of cluster recognition with the use of Kohonen Self Organizing Maps (SOM) neural networks [2].

INTRODUCTION

One of sources of difficulties that are encountered in parameter inverse problems – apart of bad conditioning – is the existence of many solutions. The both mentioned properties make such problems to be ill posed. The paper focuses on parameter inverse problems that are formulated as global optimization tasks. Moreover, the algorithms that are considered are especially suitable for problems, in which an approximation of certain level sets (central parts of basins of attractions of local minimizers) is required. The approximation of these sets can be useful when some additional criteria of optimization are considered after main results of parameter identification are obtained. Such criteria can express in some way for instance the availability and/or the cost of materials. When one knows approximations of central parts of the basins he can give an approximate answer to the question: how much one can change the value of a parameter with "not too high" change of the objective.

A hybrid genetic parallel algorithm called *GC* (*Genetic Clustering*) was proposed in [1] in order to solve such problems.

The *GC* strategy is inspired with clustering methods in global optimization [3], [4], [5] and genetic algorithms [6]. *GC* finds local minimizers and also gives additional information – central parts of basins of attraction of local minimizers can be approximated.

PARALLEL GENETIC CLUSTERING (*GC*)

The aim of original version of *GC* algorithm is to find all local minima that have adequately large basins of attraction with a sufficiently large objective variability. The algorithm also gives rough information about the basins. The basins (or their central parts) are approximated by sets of hypercubes – these sets will be called clusters. Additionally, the algorithm deals with large differences between values of local minima and also with large plateaus.

An outline of a version of the *GC* algorithm is recalled below. Some asymptotic properties of the algorithm have been derived from the Markov theory of the Simple Genetic Algorithm [7]. The stop criterion has been justified in [8].

The genetic approach to clustering consists in implementing Simple Genetic Algorithm (SGA) in the global phase [3]. The idea of sampling by running SGA follows from the observation that genetic algorithms transform measures in a regular way so they deliver information about sets rather than about isolated points. In its nature

Genetic Algorithm constitutes a dynamical system that transforms measures. This fact allows us to expect good properties of genetic clustering, because density of measure contains information that is useful in clusters recognition.

GC consists in four operations that are executed consecutively: genetic sampling, subclusters' recognition, subclusters' aggregation and fitness modification. These operations are performed in a loop until the global stop criterion is satisfied.

The outline of the *GC* is as follows:

1. Divide the feasible set *D* into *p* subdomains (each subdomain is divided into small hypercubes that constitute the grid).
2. Set all subdomains to be "active".
3. `REPEAT`

Parallel in "active" subdomains:

- 3.1 Generate initial population from uniform distribution.
- 3.2 Evaluate fitness function *f* outside recognized clusters.
- 3.3 Modify fitness function ($f \leftarrow MAX$ on clusters).
- 3.4 Steps of simple genetic algorithm (*SGA*) - evaluate new generations until the complex stop criterion is satisfied:
  - a) subclusters can be recognized, or
  - b) *GA* recognizes plateau outside known clusters (then the subdomain is set to be "passive").

  Subclusters are parts of clusters that can be recognized after point 3.4.
- 3.5 Subclusters recognition. (output: new information about clusters and new "passive" subdomains). The seed of a subcluster is this cell (hypercube) of the grid that contains "the best" individual. All neighbor cells that contain more individuals than a certain threshold are added to the subcluster. One local method is started in each subcluster.
- 3.6 Join "proper" subclusters into clusters.

`UNTIL` all subdomains are "passive" `OR` satisfactory set of clusters is found.

The Simple Genetic Algorith (*SGA*) was chosen in the genetic phase of *GC*, because it allowed us to obtain some theoretical results concerning stop criterion and asymptotic behavior of *GC*.

The fitness modification results in repelling individuals from clusters (subclusters) that are already known. A single basin of attraction can be recognized in one or several steps of the loop. The domain *D* is divided into hypercubes of a volume θ. After the *SGA* is stopped, new subclusters can be detected by the analysis of density of individuals in the hypercubes. The hypercube that contains the best individual is selected as the seed of a new subcluster. Neighbor hypercubes, with the density of individuals $\rho > \rho_t$ (ρ_t is an arbitrary constant), are attached to the cluster. A rough local optimization method is started in each new subcluster and the result of this optimization is retained. If the local method that starts from a new subcluster ends in the already recognized cluster, then the subcluster is attached to the cluster.

The stop criterion distinguishes two basic kinds of *SGA* behavior. The first one is that *SGA* "finds" clusters after few generations. The second one is that *SGA* converges to the uniform distribution of individuals. This corresponds to the recognition of a plateau (or areas where fitness has small variability) outside of the already known clusters. Other cases are treated as the situation when *SGA* does not fit to the particular problem and a refinement of *SGA* parameters is suggested.

The stopping strategy is as follows:

Check stagnation of a sequence of some estimator of probabilistic distribution density. If this criterion is satisfied, then check if an arbitrary number of hypercubes has the density of individuals below an arbitrary threshold ρ_t. If so, then begin clustering procedure, otherwise, check if individuals are uniformly distributed in *D*.

In original version of *GC* each subdomain is divided into hypercubes that constitute a static grid. Each cell of the grid that belongs to some cluster "remembers" the number of the cluster.

PROPERTIES OF *GC*

Each population with a finite number of binary coded individuals can be identified with a vector which *i*-th coordinate stands for the occurrence frequency of the *i*-th individual in the population. Lets by *r* denote the length of an individual. The frequency vector belongs to the unit simplex Λ in $\Re^{r-1}$. All possible populations of

the size n correspond to the finite subset S_n in Λ [6].

The finite population SGA constitutes a stationary Markov chain with states from S_n. By non-zero mutation it is ergodic, and there exists a weak limit

$$\pi_n^k \underset{k\to\infty}{\overset{w}{\to}} \pi_n \qquad (1)$$

of probability distributions π_n^k on S_n in k-th generation [6].

In the case of an infinite population $n=\infty$ *SGA* is a deterministic dynamic system with states in Λ governed by the genetic search operator $\Gamma:\Lambda\to\Lambda$. The sequence of the limit probability distributions π_n has a weak limit distribution π^* when the size of population goes to infinity $n\to\infty$. Moreover if Γ is focused, and K is its set of fixed points then $\pi^*(K)=1$ [6].

Let F_ε the ε-envelope of the set K

$$F_\varepsilon = \{x\in\Lambda; \quad \exists y\in K; \quad d(x,y)<\varepsilon\} \qquad (2)$$

where d stands for the distance function in $\Re^{r\text{-}1}$.

Lemma [8]*:* $\forall\varepsilon>0\ \forall\varsigma>0\ \exists N>0$ such that $\forall n>N\ \exists G(n)$ and $\forall k>G(n)\ \pi_n^k(F_\varepsilon)>1-\varsigma$.
It means, that if the population is sufficiently large and a sufficiently large number of generations were performed, then the population is arbitrary close to the fixed one with the arbitrary large probability.

It is assumed here that SGA parameters are so chosen that Γ is focused. In particular small but non-zero mutation is assumed. The desired form of the fixed points set is the finite collection of isolated points in Λ. Moreover each local minimizer of the fitness function is represented in K. It corresponds to the population highly concentrated on its neighborhood.
Conjecture: only minimizers that have sufficiently large attractors (larger than the cell size) with the sufficiently high fitness variation can be found.

Algorithm detects situation in which the population is sufficiently concentrated in attractors so that the density cluster recognition is possible. The state in which arbitrary rate of grid cells contain the assumed number (less than the average) of individuals can be handled as the local stop criterion. By Lemma the above situation is asymptotically highly probable if there exists at least one attractor out of the cluster union.

The chart of modified fitness function becomes sufficiently flat at the end of computations. It corresponds to the unique fixed point of Γ at the center of Λ.

If the sufficiently large population that starts from the center of Λ (uniform distribution of individuals) does not leave its neighborhood sufficiently long this implies that the center of Λ is the fixed point of Γ (with the arbitrary large probability). It corresponds to the situation, that the probability of finding new local minimizers is arbitrary small.

One can say, that there is an analogy between the way in which mutation and crossover rates in *SGA* imply *GC* algorithm and the way in which the reduction phase implies *DC* and *SL* clustering algorithms described in [4], [5]. Both factors cause that some minima can be undetected. However, unlike the *DC* and *SL* with the reduction phase, the GA constitutes a filter that eliminates local minima with small fitness variability and shallow basins of attraction. GA strategy is also less sensitive on fitness values in local minimizers. Such filtering property can be useful in some cases. Another interesting feature of *GC* is such that it should be especially convenient for functions with large areas of small variability (areas "similar to" plateaus) which can be difficult for other methods.

Genetic clustering offers some interesting properties like:

- Approximation of basins for all "remarkable" local minimizers,
- Filtering local minima with sufficiently small and "shallow" attractors,
- Good time complexity in cases of objectives with large plateaus.
- The global stop criterion of this strategy can be mathematically justified; this is rarely met in the case of strategies based on evolutionary algorithms.

Tests described in [1], [8] have shown that *GC* can be effective in solving some inverse problems including for instance the problem of optimal pretraction design of a network structure made of

elastic unconnected fibers fastened at their ends to a square rigid frame.

However, *GC* is not effective for problems with more than 4 dimensions. This follows from the representation of clusters – they are remembered as unions of small hypercubes that constitute a regular mesh in the domain of searches.

In order to overcome the above limitation three modifications to the *GC* algorithm are proposed. They will be described in following sections.

CLUSTERS REPRESENTATION WITH THE USE OF ELLIPSOIDS

One of methods that can be proposed to overcome the problem with high dimensionality is to represent clusters by ellipsoids. The similar approach to clusters in global optimization is known in so called *Density Clustering* (*DC*) rule described in [3]. Some good properties of this version of *DC* are proven in [5]. The version of *DC* proposed by Rinnooy Kan and Timmer assumes that the reduction phase is applied, that means the initial sample is drawn from the uniform distribution over *D* and all sample points where the value of the objective function is below certain threshold are rejected [5]. A key assumption is that the objective function is well approximated by a quadratic function in neighborhoods of local minimizers. This implies that level sets (and clusters) are approximated by ellipsoids. Clusters are recognized iteratively in the following way: the seed point $\bar{x}$ of a cluster is the result of local optimization started from the unclustered best point of the reduced sample (the unclustered point with the smallest value of the objective function). Lets by T_0 denote the set $\{\bar{x}\}$ with the seed of the cluster. In consecutive steps next points of the sample are joined to the cluster. These points belong to subsets T_i of *D*, *i=1,2,…*, $T_i \subset T_{i+1}, i = 1,2,\ldots$. These subsets correspond to certain level sets. When $f \in C^2$ we can approximate level sets by

$$T_i = \left\{ x \in D \middle| (x-\bar{x})^T H(x_s)(x-\bar{x}) \le r_i^2 \right\},$$

where *H* denotes hessian.

All points that are within T_i which is described by a critical distance $r_i(\bar{x})$ of the seed are joined to the cluster. The distance $d(x_1, x_2)$ is defined as follows: for points x_1, x_2 from a neighborhood of $\bar{x}$

$$d(x_1, x_2) = \left[(x_1 - x_2)^T H(\bar{x})(x_1 - x_2) \right]^{\frac{1}{2}},$$

(an approximation of hessian can be obtained as a byproduct of quasi-Newton local methods). The parameter $r_i(\bar{x})$ is increased stepwisely (with increasing *i*) until there is no unclustered point from the reduced sample within $r_i(\bar{x})$. Rinnooy Kan and Timmer gave the formula for the critical distance:

$$r_i(\bar{x}) = \pi^{-\frac{1}{2}} \left(i\Gamma\left(1+\tfrac{d}{2}\right) \det(H(\bar{x}))^{\frac{1}{2}} m(D) \frac{\sigma \log kN}{kN} \right)^{\frac{1}{d}}, \quad (3)$$

where Γ is the Gamma function, *m* denotes the Lebesgue measure, *N* is the sample size and σ is a constant. The whole process of sampling, reduction and clusters recognition is repeated (*k* denotes the number of the iteration) until a global stop criterion is satisfied. The formula (3) assures that the probability of erroneous termination of the cluster recognition procedure (the procedure is terminated too early, see [5] for details) in step *i* decreases polynomially fast with increasing *k*.

This version of *DC* has also other advantages:

- It has the property of asymptotic correctness in the sense that it finds global minimum with the probability 1 as *k* increases to infinity.
- It is possible (and relatively easy) to apply bayesian stopping rules [3], [5].

The main drawbacks are obvious:

- The success of the method depends on how well the assumed approximation is.
- In fact each recognized cluster can contain more than one minimizer.

Applying similar approach to *GC* can diminish disadvantages that are caused by high dimensionality. Clusters are parametrized by the central point and radiuses. Each point generated by *SGA* can be classified as belonging to some cluster or not, so the idea of fitness modification can be almost unchanged.

The Bayesian stopping rules derived for *DC* cannot be applied directly to *GC*, because these rules assume uniform distribution of sample points. Also such good properties of *DC* as mentioned above cannot be directly attributed to *GC*. Analogous estimations for *GC* are still open problems, because it is difficult to predict and calculate the exact distribution of points after some genetic epochs.

Introducing such cluster representation to *GC* causes also that stopping strategy from *GC* should be modified. Under the assumption that clusters cannot intersect, the criterion "all subdomains are passive" in real cases should be removed.

The critical distance $r_i(\bar{x})$ has not the same meaning as in *DC*, but it can be probed as if the concentration of points would be caused by the reduction phase with sample points distributed uniformly.

CLUSTERS RECOGNITION AND REPRESENTATION WITH THE USE OF NEURAL NETWORKS

Clustering methods are known also as methods that help in constructing categories or taxonomies. Special kind of neural networks – so called self-organizing maps *SOM* ([2] and bibliography cited there, [9]) were proposed as tools for exploratory data analysis, in particular for visualization of high dimensional data items.

We propose to join *GC* with the mechanism analogous to *SOM* in order to recognize and remember clusters. Additionally the method can visualize clusters in some way.

SOM is a special kind of neural network with competitive learning. Competitive learning is an adaptive process in which the neurons gradually became sensitive to different input categories [2], for instance clusters of points. After the process of learning is finished, neurons become specialized – they "represent" different categories (clusters). The mechanism which allows neurons to specialize bases on a competition among them. After an input data x arrives, this neuron wins which better "represents" the data. Moreover neurons can "learn data" (it will be described below).

In *SOM* neurons are located on a discrete lattice that constitute the "self-organizing map". During the learning process the winning neuron and its neighbors on the lattice are allowed to learn.

The input data is represented in neurons by a vector w_i (reference vector), whose components correspond to synaptic weights. Neurons can be indexed with k. The winner neuron is determined from the formula:

$$k = k(x) = \arg\min_i \left\{ \|x - w_i\|^2 \right\} \tag{4}$$

That means the winner is this neuron, whose reference vector is closest to the input data x. This neuron and its neighbors modify their reference vectors according do the formula (5).

$$w_i(t+1) = w_i(t) + h_{ki}(t)[x(t) - w_i(t)] \tag{5}$$

Neighbors are determined by so called neighborhood kernel function h_{ki}.

In the simplest case the neighborhood function can be defined as follows:

$$h_{ij} = \begin{cases} 1 & \|r_i - r_j\| \le \lambda \\ 0 & \|r_i - r_j\| > \lambda \end{cases} \tag{6}$$

or

$$h_{ij} = \exp\left(-\frac{\|r_i - r_j\|^2}{2\lambda^2} \right) \tag{7}$$

where r_i and r_j are vectors that represent location of neurons in the lattice, t denotes time and λ is a constant.

In general the neighborhood function can also be variable in time – "wide" at the beginning of the learning process and decreasing slowly during learning.

In such approach clusters can be represented as reference vectors of neurons. The number of clusters does not need to be estimated in advance, the maximum number is equal to number of neurons. Learned neurons can categorize any further sample points to clusters.

Additionally a method of visualization of clusters was proposed by Kohonen [9]. The distances between the reference vectors of neighbor neurons can visualize the clusters structure on a two dimensional map. Details can be found in [2].

TESTS

The modified version of the *GC* in which clusters were represented as ellipsoids was tested on the same problems as the original version of *GC* (see [1], [8]).

The representative results of these tests will be presented for the 8-dimensional global optimization problem given by formula (9). Also the 16-dimensional version was tested. The original version of *GC* was tested on the 2-dimensional case of this problem (see [1]). A parameter inverse problem with a similar cost function and analogous complexity was presented in [1]. It was optimal pretraction design of a network structure made of elastic unconnected fibers fastened at their ends to a square rigid frame.

The linearized and homogenized governing equation (see [1]):

$$\begin{cases} -\mathrm{div}(\sigma\, Dw) = q & \text{in } \Omega, \\ w = 0 & \text{on } \partial\,\Omega \end{cases} \tag{8}$$

delivers the relationship between the compliance $w(x)$, pretraction tensor $\sigma(x) = diag\left(\sigma_1(x_2), \sigma_2(x_1)\right)$, and the transversal loading density $q(x)$ on the frame area Ω.

Given q try to find $\sigma^* \in \left(L^\infty(\Omega)\right)^2$ such that $\mathrm{F}(\sigma^*) \le \mathrm{F}(\sigma)$ for all $\sigma \in \left(L^\infty(\Omega)\right)^2$, and $w_{\sigma^*} \in H_0^1(\Omega)$ satisfies the state equation (8).

The cost functional $\mathrm{F}(\sigma) = \mathrm{E}(\sigma) + \mathrm{P}(\sigma)$, where $\mathrm{E}(\sigma) = \int_\Omega \sigma \left| Dw_\sigma \right|^2 dx$ is the stored energy of the network, and $\mathrm{P}(\sigma)$ denotes a penalty that forces pretractions suitable for an available assortment of fibers. $\mathrm{P}(\sigma)$ is a multimodal, nonnegative function which reaches zero for many admissible pretractions.

Moreover, the following constraints are defined: $0 < \lambda \le \sigma_1, \sigma_2 \le \Lambda$ in Ω, $\int_\Omega (\sigma_1 + \sigma_2) dx = S$ with $S \in [2\lambda, 2\Lambda]$.

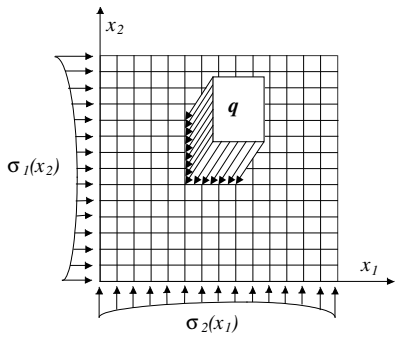


Figure 1. Schema of the fiber construction

The case of a balanced loading $\int_\Omega q dx = 0$ was considered. It is proved that under above assumptions there is more than one minimizer σ^*.

The adequate test global optimization problem can be as follows (see [1]):

$$\begin{aligned} f(x) = & \sum_{i=1}^{8} 0.01 \sin(0.05 x_i) + \\ & -c * \begin{pmatrix} x_j > 60 \ AND \\ x_j < 70 \ \ j = 1,...8 \end{pmatrix} * \sum_{i=1}^{8} x_i^2 + \\ & -c * \begin{pmatrix} x_{2j+1} > -50 \ AND \\ x_{2j+1} < -20 \ AND \ \ j = 0,1,2,3 \\ x_{2j+2} > -70 \ AND \\ x_{2j+2} < -40 \end{pmatrix} * \\ & * \sum_{i=1}^{8} x_i^2 \end{aligned} \tag{9}$$

where $f : R^8 \to R$, x_i, i=1,…8 stand for the components of x. The formula $\left(x_j > a \ AND \ x_j < b \ , j = ...\right)$ stands for one, if the condition in brackets is true for all given j, otherwise it stands for zero.

The constant c for 8-dimensional problem was equal to 0.00024. The domain of searches was a hypercube given by

$$-100 < x_i < 100 \ , \ i=1,\ldots,8 \tag{10}$$

The function f has two distinct "deep" local minima and many "shallow" local minima. One of the deep minima is the global minimum. Two-dimensional version of f is presented on Figure 1.

It is assumed that clusters are well approximated by spheres. For the 8-dimensional case the population of SGA was 500. Each time twenty generations were processed before clustering was applied. Only two distinct "deep" minima were discovered and two clusters were located.

Table 1. Results for 8-dim. problem

	minimum 1 ***f(x)*=5.11**	**minimum 2** ***f(x)*=2.61**
component of *x*	**value**	**value**
1	-50,0	70,0
2	-70,0	70,0
3	-50,0	70,0
4	-69,7	69,9
5	-49,9	69,7
6	-66,5	69,8
7	-49,9	69,5
8	-66,6	69,5

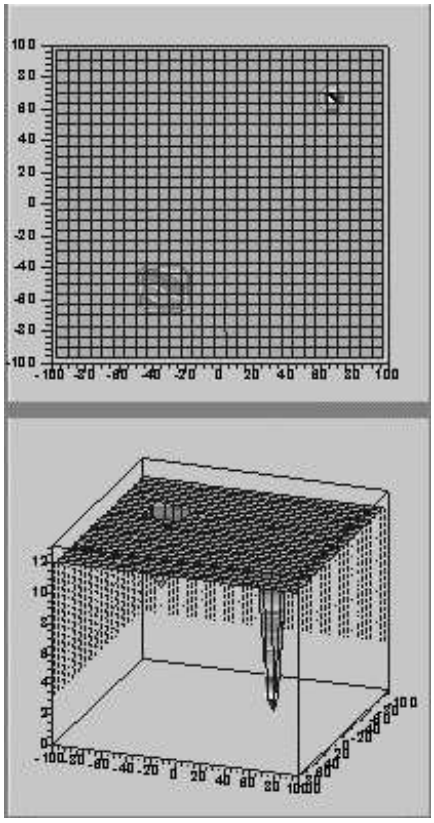


Figure 1. Two dimensional version of the 8-dimensional test function.

The scalar version of the algorithm was applied (without the division of the domain).

The results are presented in Table 1. The radius of the cluster 1 (minimum 1) was 32.2, the radius for the cluster 2 (minimum 2) was 8.37. The number of function calls in local searches (MIGRAD method from the CERN ROOT-Minuit package) was 1752 (968 for cluster 1 and 784 for cluster 2).

These results are much better than for original version of *GC* in which a great number of cells must be considered (this number depends of the chosen resolution).

However, clusters are not recognized so precisely as in *GC*. They do not contain whole basins of attraction of local minima. Moreover, they also include parts of the domain that should not be included. A modification is required when one would like to use points from clusters. For each such point the value of the objective function should be calculated (in tests the maximum value of the objective function was also remembered for each cluster).

When one wants to recognize clusters more precisely he or she can apply the original version of *GC* with the smaller domain that includes recognized cluster/clusters.

Tests showed, that the proposed algorithm maintains the filter property.

The maximum 16-dimensional case was tested successfully, however this is not the largest possible dimension. It seams, that problems can occur with the local strategies when the dimension is bigger than several tens.

The version with neural network representation of clusters is being tested now.

First tests made for two-dimensional case of the presented problem have shown that this method of clustering data in *GC* is costly in time and its superiority could be seen for problems with not too low dimensionality. One of ways to accelerate computations is to parallelize algorithm that simulates neural networks, another is to construct a hardware neural network.

ACKNOWLEDGMENTS

Research supported in part by the State Committee for Scientific Research of the Republic of Poland (KBN) under Grant No. 7 T07A 047 18.

REFERENCES

1. H. Telega, *Adaptive Parallel Genetic Clustering in Parameter Inverse Problems*, Inverse Problems in Engineering Mechanics III, M. Tanaka, G. S. Dulikravich Eds., Elsevier 2002, p. 315-325.

2. S. Kaski, *Data Exploration Using Self-Organizing Maps*, Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, 1997.

3. R. Horst, Pardalos M. Panos, *Handbook of Global Optimization*, Kluwer Ac. Publisher 1995

4. A.H.G. Rinnooy Kan, G. T. Timmer *Stochastic global optimization methods. Part I: Clustering methods,* Mathematical Programming **39**, 1987, p. 27-56.

5. A. H. G. Rinnoy Kan, G. T. Timmer, *Stochastic global optimization. Part 2: Multi level methods*. Mathematical Programming **39**, 1987, p. 57-78.

6. M. D. Vose, The Simple Genetic Algorithm - Foundation and Theory. The MIT Press, 1999.

7. Cabib E., Schaefer R., Telega H. [1998] A parallel genetic clustering for inverse problems.

Lecture Notes in Computer Science **1541**, Springer 1998, p. 551-556.

8. H. Telega, PhD Thesis, AGH Krakow, Poland (in polish), 1999.

9. T. Kohonen, *Self Organizing Maps*, Springer, Berlin, 1995.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

Curvature steps and geodesic moves for nonlinear least squares descent algorithms

Guy Chavent
Ceremade, Université Paris-Dauphine and Inria-Rocquencourt†*

Abstract

We adress in this paper the choice of both the step and the curve of the parameter space to be used in the line search search part of descent algorithms for the minimization of least squares objective functions.

Our analysis is based on the curvature of the path of the data space followed during the line search.

We define first a new and easy to compute "maximum curvature step", which gives a guaranteed value to the residual at the next iterate, and satisfies a linear decrease condition with $\omega = \frac{1}{2}$.

Then we optimize the "worst possible situation", by moving from one iterate to the next along a geodesic of the output set.

Preliminary numerical comparisons of the proposed algorithm with the Gauss-Newton algorithm are presented.

Nomenclature

x optimization variable
y descent direction
k upperscript for the k-th iteration
α optimization step
$F(x) \in \mathbb{R}^q$ residual
$f(x) \stackrel{\text{def}}{=} \frac{1}{2}\|F(x)\|^2 \in \mathbb{R}$ objective function
p path on the output set
ν arc length along p
$r(\nu)$ norm of residual at arc length ν on p
$\rho(\nu)$ radius of curvature of p at arc lenth ν
R lower bound to $\rho(\nu)$ on (part of) p
M subscript for "**M**aximum curvature path"
$\bar{}$ denotes the first stationary point of r

*75775 Paris Cedex16, France
†BP105, 78153 Le Chesnay Cedex, France

1 Introduction

Descent algorithms for the resolution of

$$\hat{x} \text{ minimizes } f(x) \stackrel{\text{def}}{=} \frac{1}{2}\|F(x)\|^2 \text{ over } \mathbb{R}^n \quad (1)$$

perform a line search in the parameter space by moving away from the current estimate x^k in a direction y^k, until ideally the first stationary point of f is attained. If we denote by p be the path of the data space followed during this step, this amounts to search for the first stationary value $\bar{r}$ of the residual $r = \|F\|$ along p.

We propose in this paper a new approach to line search, based on the curvature of the path p. Of course, one does not know the shape of p! But one can compute easily its radius of curvature ρ^k at the origin $F(x^k)$; then $R = \kappa\,\rho^k$ is, for $\kappa \leq 1$, a lower bound to the radius of cuvature of p in some neighborhood of its origin $F(x^k)$. This lower bound R will be the key to our study.

So we analyse first in section 2 the properties of paths p which leave a given point p_0 in a given direction v_0 with a curvature bounded by $1/R$. We show that, among all these paths, there is one (and generally only one) "worst path" p_M for which the residual norm $\bar{r}_M$ at the first stationary point is maximum, and simultaneously the arclength $\bar{\nu}_M$ at the first stationary point is minimum.

Then we use in section 3 these results to define a "maximum curvature step" α_M^k for the computation of x^{k+1} from x^k and y^k, which corresponds to moving forward on p up to the arclength $\bar{\nu}_M$ of the first stationary point on the worst path p_M.

This step is conservative : under the hypothesis that the radius of curvature of p has stayed above R over the $[0, \bar{\nu}_M]$ interval, one can be

sure that one has not passed the first stationary point, and that the residual has decreased at least below $\bar{r}_M$. It is also optimal in the sense that it is the largest step which ensures these two properties.

Section 4 is devoted to the choice of the curve g of the parameter space along which to move from x^k to x^{k+1}. Based on the observation that the guaranteed residual $\bar{r}_M$ after a curvature step is a decreasing function of R, we are led to replace the line search by a search along a curve $\alpha \rightarrow g(\alpha)$ such that the corresponding path p has the smallest curvature. As this path is constrained to stay on the "output set"

$$D = \{F(x) \in \mathbb{R}^q \mid \text{for all } x \in \mathbb{R}^n\} .$$

this amounts to chose $g = g_G$ such that p is a geodesic of D. We show that moving along g_G optimizes the worst possible case; once the maximum curvature step α_M^k has been computed, this can be implemented at no additional computational cost by following the second order approximation g_G^{app} to g.

Finally, section 5 presents a few preliminary numerical tests on examples of increasing nonlinearity, which include comparisons with the classical Gauss-Newton algorithm with a Backtracking or Quadratic line search.

2 The first stationary point along a path with bounded curvature

We denote in this section by p paths of $\mathbb{R}^q$ parameterized by their arc length ν, with $W^{2,\infty}$ regularity, and by $v(\nu) = p'(\nu)$ and $a(\nu) = p''(\nu)$ the corresponding velocity and acceleration. We consider as given the origin and the initial direction of the path:

$$p_0 \in \mathbb{R}^q \ , \ v_0 \in \mathbb{R}^q \ , \tag{2}$$

where v_0 is supposed to be a descent direction for the residual $r(\nu) = \|p(\nu)\|$:

$$< p_0, v_0 > \leq 0 \ . \tag{3}$$

Then given $R \in]0, \infty]$, we denote by:

$$\begin{aligned} \mathcal{P} \quad = \quad & \{ \ p \in W^{2,\infty}(\mathbb{R}^+) \ \text{ such that:} \\ & p(0) = p_0 \ , \ v(0) = v_0 \text{ and } \|a(\nu)\| \leq 1/R\} \end{aligned} \tag{4}$$

the set of all paths p of $\mathcal{P}$ which leave the p_0 in the descent direction v_0, and whose curvature is smaller than $1/R$.

To a path $p \in \mathcal{P}$ we associate the first stationary point $\bar{\nu}$ of its residual $r(\nu) = \|p(\nu)\|$, defined by:

$$\bar{\nu} = Inf\{\nu \geq 0 \text{ such that } \frac{d}{d\nu}(r^2) = 0\} \ , \tag{5}$$

and the corresponding value of the residual:

$$\bar{r} = r(\bar{\nu}) \tag{6}$$

In order to study how $\bar{\nu}$ depends on p, we single out among all paths of $\mathcal{P}$ the path p_M defined by:

$$\begin{aligned} & p_M \text{ turns steadily } \textit{away from zero} \\ & \text{with the } \textit{maximum curvature } 1/R. \end{aligned} \tag{7}$$

and we want to show that p_M is the *worst path* in the sense that it

- hits its first stationary point $\bar{\nu}_M$ sooner,
- with a residual $\bar{r}_M = r(\bar{\nu}_M)$ larger

than any other path p of $\mathcal{P}$.

The following properties hold (see [1])

Proposition 2.1 *The arclength $\bar{\nu}$ at which a path p attains its first stationary point satisfies:*

$$\bar{\nu}_M \ \leq \ \bar{\nu} \quad \forall p \in \mathcal{P} \ , \tag{8}$$

$$\bar{r} \leq r(\nu) \leq r_M(\nu) \quad \forall \nu \in [0, \bar{\nu}_M] \quad \forall p \in \mathcal{P} \ . \tag{9}$$

Proposition 2.2 *The arc length $\bar{\nu}_M(R)$ and residual $\bar{r}_M(R)$ at the first stationary point along the "worst path" p_M with radius R are given by:*

$$\bar{\nu}_M(R) = R \arctan \frac{\bar{\nu}_L}{R + \bar{r}_L} \ , \tag{10}$$

$$\bar{r}_M(R) = ((R + r_L)^2 + {\nu_L}^2)^{\frac{1}{2}} - R \ , \tag{11}$$

where $\bar{\nu}_L$ and $\bar{r}_L$ are the arc length and residual at the first stationary point along the linearized path $p_L(\nu) = p_0 + v_0 \nu$, given by:

$$\bar{\nu}_L = r_0 \cos\gamma_0 \quad , \quad \bar{r}_L = r_0 \sin\gamma_0 \ . \tag{12}$$

The residual $\bar{r}_M$ given by (11) decreases from r_0 (for $R = 0$, infinite curvature) to $\bar{r}_L = r_0 \sin\gamma_0$ (for $R = \infty$, zero curvature).

The above results have been established under the hypothesis that one knew a global lower bound R of the radius of curvature along p. This will not be the case in the applications to optimization we have in mind, where such an estimate will be available only on a neighborhood of $\nu = 0$. So we replace $\mathcal{P}$ by:

$$\widetilde{\mathcal{P}} = \{\, p \in W^{2,\infty}(\mathbb{R}+) \,|\, p(0) = p_0 \,,\, v(0) = v_0 \} \,, \tag{13}$$

and the following local properties hold:

Proposition 2.3 *Let $p \in \widetilde{\mathcal{P}}$ and $R > 0$ satisfy:*

$$\rho(\nu) \geq R \quad \forall \nu \in [0, \bar{\nu}_M(R)] \,, \tag{14}$$

where ρ is the radius of curvature along p. Then:

$$\bar{\nu} \geq \bar{\nu}_M(R) \quad , \tag{15}$$

$$\bar{r} \leq r(\bar{\nu}_M(R)) \leq \bar{r}_M(R) \quad . \tag{16}$$

The best estimates, i.e. the largest value of $\bar{\nu}_M(R)$ and the smallest value of $\bar{r}_M(R)$, are obtained for the largest R which satisfies (14), that is for $\widetilde{R}$ solution of;

$$\tilde{\nu} = \bar{\nu}_M(\widetilde{R}) \quad , \quad \widetilde{R} = R_m(\tilde{\nu}) \,. \tag{17}$$

where:

$$R_m(\nu) = Inf\{\ \rho(\tau)\ ,\ 0 \leq \tau \leq \nu\} \quad . \tag{18}$$

denotes the smallest radius of curvature on p up to arclength ν.

3 A maximum curvature step for descent algorithms

We consider in this section the resolution of the least squares problem (1) by a descent algorithm of the form

$$x^{k+1} = g(\alpha^k) \,, \tag{19}$$

where $\alpha \rightsquigarrow g(\alpha)$ describes the curve of the parameter space along which one moves from x^k to x^{k+1}. It is chosen such that:

$$g(0) = x^k \quad , \quad g'(0) = y^k \,, \tag{20}$$

where:

1. x^k is the current iterate,
2. y^k is a descent direction for f at x^k, computed from $\nabla J(x^k)$ and the previous descent direction(s) (Conjugate Gradient, Quasi-Newton algorithms...), or from F^k and $J^k = F'(x^k)$ (Quasi-Newton algorithms...),
3. α^k is the step on the curve g, which is required to satisfy the so called *linear decrease condition*:

$$f(g(\alpha^k)) \leq f(x^k) + \alpha^k \omega f'(x^k).y^k \,, \tag{21}$$

for some

$$\omega \in]0, 1/2] \tag{22}$$

to be specified by the user.

It will be convenient to denote by

$$z^k = g''(0) \tag{23}$$

the initial acceleration on the curve g. The usual situation where one moves from x^k to x^{k+1} along a straight line of the parameter space corresponds to the choice

$$g_S(\alpha) \stackrel{\text{def}}{=} x^k + \alpha y^k \,. \tag{24}$$

We consider in this section the curve g as given, for example -but not not necessarily- by (24), and discuss the choice of the step α^k. We recall first in sections 3.1 and 3.2 the classical *Back Tracking* and *Quadratic* steps (cf for example [2]), which we have implemented in our numerical tests for comparison purpose. Then we introduce in section 3.3 a new *Maximum Curvature* step by application of the results of section 2 on the first stationary point of a path.

3.1 The Back Tracking step

One choses first an initial guess α of the step. If the linear decrease condition (21) is not satisfied, one replaces α by $\mu\alpha$ for some $\mu \in]0, 1]$ and checks again. The Back Tracking step α^k_{BT} is then the first α for which condition (21) is satisfied.

In Gauss-Newton type algorithms, where y^k is determined by requiring that $x^k + y^k$ solves the linearized problem, $\alpha = 1$ is a reasonable first guess (condition (21) would then be satisfied with $\omega = 1/2$ if F were affine and $g = g_S$ had been chosen).

3.2 The Quadratic step

This algorithm is also a back tracking algorithm, but with a different step reduction strategy when the linear decrease condition (21) is not satisfied: instead of replacing α by $\mu\alpha$, one first uses $f(g(\alpha))$, $f(x^k)$ and $f'(x^k).y^k$, which have just been evaluated to check (21), to compute a quadratical approximation of f along the curve g, and the value $\tilde{\alpha}$ where it achieves its minimum.

When F happens to be affine, and one moves along the straight line $g = g_S$, the function $\alpha \rightsquigarrow f(g(\alpha))$ is quadratical: in this case, $\tilde{\alpha}$ produces exactly the minimum of f on the half line, and (21) is satisfied with $\omega = 1/2$.

But F is nonlinear, so $\tilde{\alpha}$ might be close to zero, or larger than α. In order to avoid such situations, the new α is taken as the projection of $\tilde{\alpha}$ on the interval $[\tau\alpha, (1-\tau)\alpha]$ for some $\tau \in [0, 1]$.

Then α_Q is the first α for which (21) holds.

3.3 The maximum curvature step

We associate to g a path $\tilde{p}$ of the data space defined by:

$$\tilde{p}(\alpha) = F(g(\alpha)) \quad \forall \alpha \geq 0 \ , \tag{25}$$

and denote by

$$\nu(\alpha) = \int_0^\alpha \|F'(g(\tau)).g'(\tau)\| \ d\tau \tag{26}$$

the arclength function along $\tilde{p}$, and by p the reparameterization of $\tilde{p}$ by the arclength ν. The curve g and the mapping F are supposed regular enough for p to have a finite curvature which varies continuously with ν, i.e. to satisfy:

$$p \in \mathcal{C}^{2,\infty}(\mathbb{R}^+) \ . \tag{27}$$

We propose here to use the specific least-squares structure (1) of f to define a new "maximum curvature step" α_M^k which will produce a guaranteed decrease of f and satisfy (21) with $\omega \simeq 1/2$: rather than evaluating one single real number $f''(x^k).(y^k, y^k)$ as it is done in the quadratic step, one could as well evaluate at a similar cost the vector $\tilde{p}''(0)$, which gives information on the shape of the path p in the data space, and allows to use results of section 2.

With the notations:

$$\begin{aligned} F^k &= \tilde{p}(0) = \ F(x^k) \ , \\ V^k &= \tilde{p}'(0) = F'(x^k).y^k \ , \\ A^k &= \tilde{p}''(0) = F''(x^k).(y^k, y^k) + F'(x^k).z^k \end{aligned} \tag{28}$$

the quantities associated to p in section 2 are given by:

$$\begin{aligned} &p_0 = F^k \ , \ v_0 = V^k/\|V^k\| \ , \ r_0 = \|F^k\| \ , \\ &f'(x^k).y^k = -\|V^k\| < p_0, v_0 > \ , \\ &\bar{\nu}_L = - < F^k, v_0 > \ , \\ &\bar{r}_L = (\|F^k\|^2 - < F^k, v_0 >^2)^{1/2} \ , \end{aligned} \tag{29}$$

and the radius of cuvature ρ^k of p at $\nu = 0$ is given by:

$$(1/\rho^k)^2 = \|A^k\|^2 - < A^k, v_0 >^2 \ . \tag{30}$$

In order to apply the results of proposition 2.3, one needs a lower bound R of the radius of curvature of p on a neighborhood of $\nu = 0$. But we know the radius of curvature ρ^k at $\nu = 0$. Hence it is natural to take R of the form:

$$R = \kappa^{k+\frac{1}{2}} \rho^k \quad \text{with} \quad 0 \leq \kappa^{k+\frac{1}{2}} \leq 1 \ , \tag{31}$$

where $\kappa^{k+\frac{1}{2}}$ is a security factor which accounts for the possible increase of the curvature along the path. We can now define the *maximum curvature step* α_M^k *along* g by:

$$\nu(\alpha_M^k) = \bar{\nu}_M(R) = R \arctan \frac{\bar{\nu}_L}{R + \bar{r}_L} \tag{32}$$

where ν is the arc length function defined in (26), and $\bar{\nu}_M(R)$ is the arclength of the first stationary point on the "worst path" p_M with curvature $1/R$, defined in (10).

Theorem 3.1 *If $\chi^{k+\frac{1}{2}}$ is small enough for p and R to satisfy (14), the maximum curvature step α_M^k defined by (32) satifies:*

$$\nu(\alpha_M^k) = \ \bar{\nu}_M(R) \ \leq \bar{\nu} \ , \tag{33}$$

$$\begin{aligned} f(g(\alpha_M^k)) \leq \ \tfrac{1}{2}\bar{r}_M(R)^2 \ &\leq f(x^k) \\ &+ \alpha_M^k \frac{1}{2} \frac{\nu(\alpha_M^k)}{\nu_L(\alpha_M^k)} f'(x^k).y^k \ , \end{aligned} \tag{34}$$

where $\bar{\nu}_M(R)$ and $\bar{r}_M(R)$ are given by (10) (11), and $\alpha \rightsquigarrow \nu_L(\alpha)$ is the arclength along the tangent to p at the origin, defined by:

$$\nu_L(\alpha) = \alpha \ \|V^k\| \ . \tag{35}$$

The linear decrease condition (21) is hence satisfied with

$$\omega = \frac{1}{2}\,\frac{\nu(\alpha_M^k)}{\nu_L(\alpha_M^k)} \;\simeq\; \frac{1}{2}\;, \tag{36}$$

We discuss now the implementation of the minimum curvature step. In order to determine α_M^k, we replace in equation (32) the arclength function $\alpha \rightsquigarrow \nu(\alpha)$ by either its linear approximation:

$$\nu_L(\alpha) = \alpha\ \|V^k\| \ \text{ (as in (35))}, \tag{37}$$

in which case α_M^k is given by:

$$\alpha_M^k \|V^k\| = \bar{\nu}_M(R)\;, \tag{38}$$

or by its quadratic approximation:

$$\nu_Q(\alpha) = \alpha\ \|V^k\| + \frac{\alpha^2}{2} < \frac{V^k}{\|V^k\|}, A^k > \;, \tag{39}$$

in which case α_M^k is the root of smallest absolute value of the second degree equation (32).

Then in order to ensure that hypothesis (14) of theorem 3.1 holds, one would have to check that the radius of curvature $\rho(\nu)$ of p remains larger than $R = \kappa^{k+\frac{1}{2}}\rho^k$ over the $[0, \bar{\nu}_M(R)]$ interval. However, this would be computationally expensive, and we shall content ourselves with checking for a consequence of (14) by theorem 3.1, namely that the linear decrease condition (21) is satisfied at each iteration for some ω smaller than 1/2. If this test fails, then $\kappa^{k+\frac{1}{2}}$ is multiplied by some $\mu \leq 1$, and the kth iteration is performed again. In the implementation of this algorithm used in section 5, the security factor κ was initialized to 1 at the beginning of each iteration, but other strategies can be considered.

When only $f(x^k)$ and $\nabla f(x^k)$ are computationally available, the cost incurred by the computation of α_M^k is that required to evaluate (exactly, or approximately by finite differences) the two directional derivatives V^k and A^k of the forward model F in the same direction y^k (two evaluations of F).

When $f(x^k)$ and the Jacobian $J^k \stackrel{\text{def}}{=} F'(x^k)$ are available, then V^k can be computed by the matrix product $J^k y^k$, so the only cost incurred by α_M^k is that of the evaluation of A^k (one evaluation of F).

4 Moving along Geodesics

We consider in this section the case where both $F(x^k)$ and its Jacobian $J^k = F'(x^k)$ are available for the computation of x^{k+1}, and discuss the choice of the curve $\alpha \rightsquigarrow g(\alpha)$ used to move from x^k to x^{k+1} with a maximum curvature step α_M^k

There is of course no hope of finding the curve g which gives the best decrease to f. But the maximum curvature step, based on the worst case analysis of section 2, has been shown in theorem 3.1 to ensure that $f(x^{k+1})$ is smaller than the "guaranteed residual" $\frac{1}{2}\bar{r}_M(R)^2$. So we will choose the curve g which gives the best guaranteed residual.

As we know from proposition 2.2, $\bar{r}_M$is a decreasing function of R. This leads to choose for g the curve g_G whose image p by F has the smallest possible curvature, i.e., as p is contrained to stay on $D = F(\mathbb{R}^n)$, such that p is a geodesic of D. Such a function $g(\alpha)$ is the solution of the differential equation (see [3] [4] for example):

$$J^T(g)J(g)\,g'' + J^T(g)\,F''(g).(g', g') = 0\;, \tag{40}$$

$$g(0) = x^k \quad , \quad g'(0) = y^k\;. \tag{41}$$

and the arclength function $\nu(\alpha)$ simplifies to:

$$\nu_G(\alpha) = \alpha\|V^k\| = \nu_L(\alpha)\;, \tag{42}$$

where $\nu_L(\alpha)$ is the arclength along the tangent to p at the origin, defined in (35).

Theorem 4.1 *Let p_G (resp. p_S) be the path on $F(\mathbb{R}^n)$ associated to the curve g_G (resp. g_S) defined by (40) (41) (resp. (24)), and $\tilde{\nu}_G$, $\widetilde{R}_G$ (resp $\tilde{\nu}_S$, $\widetilde{R}_S$) the corresponding solutions of (17).*

1. *The radius of curvature at the origin on p_G and p_S satisfy:*

$$\rho_G^k \geq \rho_S^k\;, \tag{43}$$

with a strict inequality as soon as $F''(x_k)(y_k, y_k)$ is not orthogonal to the tangent plane to $F(C)$ at $F(x_k)$, which is the generic situation.

2. *Because the radius of curvature along p_G and p_S can vary in an unpredictable way, the optimal step $\tilde{\nu}_G$ along p_G is not necessarily larger than $\tilde{\nu}_S$.*

3. *However, when the strict inequality holds in equation (43) (which is true in general), and when the linearized step $\bar{\alpha}_L$ is small enough (i.e. when one is close enough to convergence), then*

$$\widetilde{\nu}_G > \widetilde{\nu}_S \quad , \quad \widetilde{R}_G > \widetilde{R}_S \ , \tag{44}$$

and the guaranteed residual along the "geodesic" g_G is necessarily smaller than the one along the "straight line" g_S for any $R_S \leq \widetilde{R}_S$, provided R_G is chosen close enough to $\widetilde{R}_G$

From a computational point of view, following the geodesic is an expensive operation, as it requires the resolution of the differential equation (41) by a numerical scheme. If an Euler scheme were used, for example, each step would have the same cost as the computation of the Gauss-Newton direction y^k (compare (40) and (44)), plus the cost of the second directional derivative $F''(g).(g', g')$.

Moreover, we see from theorem 4.1 that the geodesic is, from the point of view of achievable guaranteed residuals, better than the half line only on the neighborhood $[0, \nu_{max}]$ of $\nu = 0$.

Hence we shall replace g_G by its second order approximation g_G^{app} defined by:

$$g_G^{app}(\alpha) = x^k + \alpha y^k + \frac{\alpha^2}{2} z^k \ , \tag{45}$$

where $z^k = g''(0)$ is the solution of (compare with (44)):

$$J^{k^T} J^k z^k + J^{k^T} F''(x_k).(y_k, y_k) = 0 \ . \tag{46}$$

As we have seen in section 3, the implementation of the maximum curvature step requires the computation of $F''(x_k).(y_k, y_k)$ anyway for the evaluation of ρ^k (see equations (28) (30)). Hence the only cost incurred by moving along the geodesic rather than along a straight line is that of the resolution of the linear system (46), which is the same as the Gauss-Newton system (44), but with $F''(x_k).(y_k, y_k)$ in the right-hand side instead of F^k.

5 Numerical results

We have performed some preliminary tests, which we present now. We shall invert the function:

$$F = \Phi - d \ , \tag{47}$$

where $\Phi : \quad \Omega =]-1, +\infty[\times \mathbb{R} \subset \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ is a regularized version of the Powell example, defined by:

$$\Phi(x_1, x_2) = \begin{pmatrix} x_1 \\ \frac{10\, x_1}{x_1 + 1} + 2 x_2^2 \\ \epsilon\, x_2 \end{pmatrix} \ , \tag{48}$$

and where d is the data to be inverted:

$$d = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \tag{49}$$

The vector d does not belong to $\Phi(\Omega)$, so the minimum residual is strictly positive. The last component is set to zero, which correspond to the situation where no information on x_2 is available.

When the regularization parameter ϵ goes to zero, the function Φ tends to the function of the Powell example, which has a singularity along the line $x_2 = 0$. We have used the values $\epsilon = 0.1$, which corresponds to a relatively smooth problem, and $\epsilon = 0.01$, which corresponds to a quite stiff problem, where the curvature of $\Phi(\Omega)$ varies very quickly when x_2 changes sign.

The solution of the minimization problem is, $\forall \epsilon > 0$:

$$\hat{x} = \begin{pmatrix} 0.125 \\ 0 \end{pmatrix} \ , \tag{50}$$

which, when $\epsilon \to 0$, tends to be on the singularity of $\Phi(\Omega)$.

The minimization of $f(x) = \frac{1}{2}\|F(x)\|^2$ has been performed by three variations of the Gauss-Newton algorithm; at each iteration, the descent direction y^k is the Gauss-Newton direction computed by (44), only the way x^{k+1} is computed from x^k, y^k and J^k changes:

the GN/BT algorithm: one moves from x^k to y^k straight in the direction y^k, with a step determined by the back tracking algorithm of section 3.1. This is a classical algorithm, which we use as a first reference. The parameters are the initial guess of the step at each iteration, set to $\alpha = 1$, the reduction factor, set to $\mu = 0.5$, and the coefficient ω in the linear decrease condition (21), set to $\omega = 10^{-4}$.

the GN/Q algorithm: same as above, but with a step determined by the quadratic algorithm of section 3.2. This algorithm is expected to be more efficient for smooth problems, so it can be a more demanding reference. The parameters are the initial guess of the step, still set to $\alpha = 1$, the security coefficient for the projection, set to $\tau = 10^{-2}$, and the coefficient ω, still set to $\omega = 10^{-4}$.

the GN/MC/G algorithm: one moves from x^k to y^k along the approximate geodesic g_G^{app} defined in (45) of section 4, with the maximum curvature step $\alpha_{M,G}^k$ defined by (32) of section 3.3 applied to g_G^{app}. The parameters are the initial value of the security factor for the radius of curvature at each iteration, set to $\kappa = 1$, and the reduction factor, set to $\mu = 0.5$.

Each algorithm was started at one of the two initial points:

$$x_1^0 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \text{or} \quad x_2^0 = \begin{pmatrix} 6 \\ 5 \end{pmatrix} , \tag{51}$$

and was stopped when the norm of the gradient was small:

$$\| {J^k}^T F^k \| \leq 10^{-4} . \tag{52}$$

We present in tables 1 to 4 the comparative results in four situations. Each table displays:

- the number of Gauss-Newton iterations,
- the total number of reductions of the step α (for the GN/BT and GN/Q algorithms) or the security factor κ for the radius of curvature (for the GN/MC/G algorithm),
- the total number of function evaluations (computation of the Jacobian J^k and the directional derivative $F''(x^k)(y^k, y^k)$ have been counted for one),
- the average value of the step α^k per Gauss Newton iteration,
- the solution $[x_1, x_2]$ found by the algorithm,
- the exit type of the algorithm: **N** when it stops because condition (52) is satisfied, **O** when it stops because the maximum number of iterations is attained.

The first comparison was made on a rather smooth problem ($\epsilon = 0.1$) and with the initial guess $x_1^0 = [2, 1]$. The results are as follows:

Algorithm	GN Iter	Reductions	Funct eval
GN/BT	325	1174	2099
GN/Q	74	142	216
GN/MC/G	160	552	871
Algorithm	**Mean step**	**Solution**	**Exit type**
GN/BT	.060	$[.1249; 10^{-6}]$	N
GN/Q	.128	$[.1249; 10^{-5}]$	N
GN/MC/G	.018	$[.1249; 10^{-9}]$	N

Table 1

As expected, the GN/Q algorithm is the most effective on this smooth problem. The GN/MC/G algorithm is better in this case than the GN/BT algorithm.

The next table show how the algorithms performs on the same problem ($\epsilon = 0.1$), but with the worse initial guess $x_2^0 = [6, 5]$:

Algorithm	GN Iter	Reductions	Funct eval
GN/BT	394	2550	2156
GN/Q	87	166	253
GN/MC/G	158	597	912
Algorithm	**Mean step**	**Solution**	**Exit type**
GN/BT	.065	$[.1249; 10^{-6}]$	N
GN/Q	.090	$[.1249; 10^{-5}]$	N
GN/MC/G	.129	$[.1249; 10^{-7}]$	N

Table 2

This time we see that the GN/MC/G algorithm takes the advantage over GN/BT: it does less function evaluations and also less iterations. But the GN/Q algorithm is still the best performer.

We test now the case of more strongly non linear problems: we divide ϵ by 20, so one has now $\epsilon = 0.005$. We begin with the closest ini-

tial guess $x_1^0 = [2, 1]$. The results are:

Algorithm	GN Iter	Reductions	Funct eval
GN/BT	10000	149590	159590
GN/Q	10000	29991	39991
GN/MC/G	1951	4467	8368
Algorithm	**Mean step**	**Solution**	**Exit type**
GN/BT	.002	$[.7012; 310^{-3}]$	O
GN/Q	.4417	$[.5583; 510^{-3}]$	O
GN/MC/G	.0007	$[.1249; 510^{-5}]$	N

Table 3

The problem being less regularized, we see that the GN/BT and GN/Q algorithms do not reach the solution in 10000 iterations. The GN/MC/G algorithm finds the sulution in less than 2000 iterations.

The last comparison is made with the same regularization parameter $\epsilon = 0.005$, but with the worse initial guess $x_2^0 = [6, 5]$:

Algorithm	GN Iter	Reductions	Funct eval
GN/BT	10000	149582	159582
GN/Q	10000	31166	41166
GN/MC/G	1725	3893	7342
Algorithm	**Mean step**	**Solution**	**Exit type**
GN/BT	.002	$[.7013; 310^{-3}]$	O
GN/Q	.578	$[.6027; -310^{-5}]$	O
GN/MC/G	.002	$[.1249; 210^{-5}]$	N

Table 4

Once again, the GN/MC/G algorithm is the only one which finds the solution in less than 10000 iterations.

Conclusion

The proposed algorithm, which is based on the optimization of the worst situation, seems to behave in a very robust way over a wide range of situation and nonlinearity. Unsurprisingly, it tends to outperform the two reference algorithms in situation of strong nonlinearity. Further numerical experimentation is required to assess its practical interest.

Acknowledgment

The author thanks Jean Charles Gilbert and Paul Armand for fruitful discussions, and Yves Laroque for the numerical tests.

References

[1] G. Chavent: *Curvature steps and geodesic moves for nonlinear least squares descent algorithms*, Inria Report, 2002.

[2] J- C. Gilbert: *Optimisation différentiable: Théorie et Algorithmes*, in preparation.

[3] M. P. do Carmo: *Riemannian Geometry*, Boston, Birkhauser, 1992.

[4] T. Rapcsac: *Smooth nonlinear optimization in* $\mathbb{R}^n$, Kluwer Academic Publishers, 1998.

[5] C. Udriste: *Convex functions and optimization methods on Riemannian manifolds*, Kluwer Academic Publishers, 1994.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

A ROCKET TRACKING ANALYSIS USING CONSTANT-GAIN FILTERS

Mauricio A. Pinheiro Rosa, Francisco A. Braz Filho, Lamartine N. F. Guimarães, Alexandre D. Caldeira, Eduardo M. Borges and Jonas Rubini Jr.
Instituto de Estudos Avançados – IEAv
Centro Técnico Aeroespacial – CTA
Caixa Postal 6044
Sâo José dos Campos – SP – Brasil
pinheiro@ieav.cta.br

ABSTRACT

A methodology has been developed to evaluate suitable gains for the α–β–γ filter algorithm presently implemented in the rocket tracking system of the brazillian center of rocket launching at Alcântara (CLA) for rockets with no complete previous flight data such as the VLS rocket – the brazillian satellite launcher vehicle. The methodology is based on a minimization criterion of the standard deviation of the error between nominal and estimated values of selected parameters. The analysis has also the objective to show that the selection of the most suitable division of the entire flight in only two phases (2-phase filter), as is the case for the present filter algorithm, may depend on several factors which makes this task very difficult. The results for a multiple phase α–β–γ filter (N-phase filter) show that the use of this filter basically eliminates the just mentioned problem with the 2-phase filter.

INTRODUCTION

One of the main concerns in Flight Safety during rocket tracking is the ability to follow the trajectory of the rocket hit location at the earth's surface (impact point) assuming a rocket balistic flight at every time instant. The mission should be aborted always when this trajectory indicates that the impact point is about to cross the boundaries of a previously defined safety region.

Considering that the radars can only measure the rocket position, the impact point is an estimated variable since it is calculated from the estimated rocket position and velocity. This estimation can be accomplished by employing filtering techniques to the radar noisy measurement signals.

This work's main motivation was to establish a computational methodology of the filter gains for the ADOUR and ATLAS radar signals of the rocket tracking system of the brazilian center of rocket launching at Alcântara (CLA). The methodology should be able to calculate appropriate filter gains not only for already available previous rocket flight data but also, and more importantly, for rockets which do not have yet a previous complete flight data set as is the case for the brazillian satellite launcher vehicle (VLS). Previous flight data of other rockets such as VS30 and VS30-ORION have been used for establishing the methodology.

The ADOUR and ATLAS radars of the CLA tracking system are located about 6 and 30 kilometers for from the launching ramp, respectively. Although the ATLAS radar is more precise than the ADOUR one, the latter is important at the beginning of flight due to its proximity to the ramp and also because the ATLAS radar can only see the target a few seconds after the rocket take off because of natural barriers between this radar and the rocket during the first instants of flight. A complete description of the CLA rocket tracking system is presented in References 1 and 2.

The filter presently implemented in the rocket tracking system at CLA is of the α–β–γ type with constant gains for each flight phase and the additional characteristics of a total of only two phases for the entire flight (2-phase filter) and the same filter gains for all the three spatial coordinates, where flight phases are previously selected time intervals of the entire flight. The computational methodology and most of the results presented herein consider these filter limitations although a brief analysis of the possible improvement in parameter estimation is also investigated if a N-phase filter is used.

In this work are briefly described the most important characteristics of the α–β–γ filter and also the computer tools used to analyze the data and the results. A fair description of the available data and of the filter gain computational methodology analyses is presented. Also, a typical VLS rocket flight has been selected to illustrate the application of the present computational methodology.

THE α–β–γ FILTER

The recursion formula for the estimate of the current state vector at instant t_{k+1} is given by

$$\hat{\boldsymbol{x}}(k+1) = \hat{\boldsymbol{x}}_p(k+1) + \boldsymbol{W}(k+1)\left[z(k+1) - \hat{z}_p(k+1)\right], \qquad (1)$$

where:

$\hat{\boldsymbol{x}}(k+1)$ = state vector estimate update,

$\hat{\boldsymbol{x}}_p(k+1)$ = predicted state vector estimate,

$\boldsymbol{W}(k+1)$ = filter gain vector estimate,

$z(k+1)$ = measured target coordinate position, and

$\hat{z}_p(k+1)$ = predicted measured target coordinate position.

For target motion in several coordinates, it is costumary to use kinematics models assumed independent across coordinates leading, therefore, to decoupled filtering. Thus, the following filter description will be presented for a single coordinate and assumed valid for all target motion coordinates.

The discrete stochastic target kinematics model adopted for the present filter algorithms implemented in the rocket tracking system at CLA is a Wiener process acceleration model [3] which can be written as

$$\boldsymbol{x}(k+1) = \boldsymbol{F}\,\boldsymbol{x}(k) + \boldsymbol{\Gamma}\,\mu(k), \qquad (2)$$

where

$$\boldsymbol{x}(k) = \begin{bmatrix} s(t_k) \\ v(t_k) \\ a(t_k) \end{bmatrix}, \qquad (3)$$

$$\boldsymbol{F} = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \qquad (4)$$

and

$$\boldsymbol{\Gamma} = \begin{bmatrix} T^2/2 \\ T \\ 1 \end{bmatrix}. \qquad (5)$$

where s, v and a are the rocket coordinate position, velocity and acceleration, respectively, and T is the radar data sampling time. In this model the white process noise $\mu(k)$ is the acceleration increment during the k-th sampling period and is assumed to be a *zero-mean white sequence* – the acceleration is a discrete time Wiener process.

The target position measurement model at a specific coordinate is given by

$$z(k) = \boldsymbol{H}(k)\boldsymbol{x}(k) + w(k), \qquad (6)$$

where

$$\boldsymbol{H}(k) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \qquad (7)$$

and $w(k)$ represents the measurement noises.

The predicted state vector, $\hat{\boldsymbol{x}}_p(k+1)$, is calculated as a function of the last updated estimate of the state vector, $\hat{\boldsymbol{x}}(k)$, using the equation

$$\hat{\boldsymbol{x}}_p(k+1) = \boldsymbol{F}\,\hat{\boldsymbol{x}}(k), \qquad (8)$$

whereas the predicted measurement value, $\hat{z}_p(k+1)$, is calculated as function of the predicted state vector, $\hat{\boldsymbol{x}}_p(k+1)$, using the equation

$$\hat{\boldsymbol{z}}_p(k+1) = \boldsymbol{H}\,\hat{\boldsymbol{x}}_p(k+1) = \boldsymbol{H}\boldsymbol{F}\,\hat{\boldsymbol{x}}(k). \qquad (9)$$

The vector gain of α–β–γ filter presents the following notation

$$\boldsymbol{W} \equiv \left[\alpha; \quad \frac{\beta}{T}; \quad \frac{\gamma}{2T^2} \right]^t , \tag{10}$$

where α, β/T e $\gamma/2T^2$ are the filter gains for the target position, velocity and acceleration coordinate variables, respectively, and α, β and γ are the constants to be calculated. Hereafter, although in not a very precise manner, the parameters α, β and γ will be occasionally named filter gains for the sake of clarity.

Defining

$$\hat{\boldsymbol{x}}(k) = \begin{bmatrix} S_k \\ V_k \\ A_k \end{bmatrix}, \tag{11}$$

$$\hat{\boldsymbol{x}}_p(k) = \begin{bmatrix} SP_k \\ VP_k \\ AP_k \end{bmatrix}, \tag{12}$$

$$\hat{z}_p(k) = ZP_k \tag{13}$$

and using the $\boldsymbol{F}$ and $\boldsymbol{H}$ matrices given by Eqs. (4) and (7), respectively, the following relationships can be obtained from Eqs. (1) through (9)

$$ZP_{k+1} = SP_{k+1}, \tag{14a}$$

$$SP_{k+1} = S_k + V_k T + A_k \frac{T^2}{2}, \tag{14b}$$

$$VP_{k+1} = V_k + A_k T \tag{14c}$$

and

$$AP_{k+1} = A_k . \tag{14d}$$

Also, the following updated state vector estimate equations, for a specific coordinate, can be obtained using Eqs. (1) and (10)

$$S_{k+1} = SP_{k+1} + \alpha(Z_{k+1} - ZP_{k+1}), \tag{15a}$$

$$V_{k+1} = VP_{k+1} + \frac{\beta}{T}(Z_{k+1} - ZP_{k+1}) \tag{15b}$$

and

$$A_{k+1} = AP_{k+1} + \frac{\gamma}{2T^2}(Z_{k+1} - ZP_{k+1}) . \tag{15c}$$

This set of equations, i.e., Eqs. (14a) through (15c), is the one to be solved for estimating the state vector variables for a specific coordinate.

The present complete filter algorithm implemented at CLA considers identical filter gains for all coordinates (*x, y* and *z*) and only two sets of filter gains for the entire rocket flight, one for each flight phase, no matters the number of rocket propulsion stages.

FILTER GAIN COMPUTATIONAL METHODOLOGY

In this section is presented the methodology adopted for the calculation of the filter gains for the algorithm presently implemented at CLA. The criterion of minimization of standard deviation of the errors between estimated and nominal values of a specific parameter is used to select the best set of filter gains. For this purpose, a computer code has been developed [4] for the calculation of standard deviations of the errors between estimated (filtered) and nominal parameters such as position and velocity coordinates, distance and impact point of the rocket, for each flight phase and specific set of filter gains. The overall algorithm utilizes the golden search method [5] for the calculation of the set of filter gains that minimizes the chosen parameter for the selected flight phase. In general, due to its great importance in flight safety, the rocket impact point is chosen as the parameter to be minimized. With the purpose of reducing the number of independent variables, namely, α, β and γ, it can be shown [3,6] that the state estimation covariance matrix converges to a steady-state value and explicit expression relating the steady-state filter gains can be obtained such as

$$\beta = 2(2 - \alpha) - 4\sqrt{1 - \alpha} \tag{16}$$

and

$$\gamma = \frac{\beta^2}{\alpha} . \tag{17}$$

For instance, if α is known, the other two parametes can be readly calculated from these equations.

Since the best filter gain magnitudes can be quite differents if the rocket is in a propulsed or balistic flight phase, the division of the flight in specific time intervals (phases) should be such that each phase of the flight is at least predominantly propulsed or balistic [1]. Therefore, for rockets with previous flight data record already available, the filter gains to be used in future missions should be the ones, which satisfy the just described criterion, using these data. In the oother hand, for rocket flights with no previous flight data record available, SAGADA data for the specific flight are used insteady of the real ones. In a very simplistic way, it can be said that SAGADA data are the ones obtained from the radars when they are forced to follow a fictious target with the same nominal trajectory as the real rocket. In this case, firstly a comparison between the results obtained with real flight and SAGADA data of other rocket flights with acceleration profiles as close as possible to the one to be calculated is performed. Then the results of these comparisons for predominantly propulsed and balistic phases are taken into account for the filter gain calculations using the SAGADA data of the rocket flight to be analyzed. For instance, in order to estimate the filter gains for the VLS rocket, for which there is no complete previous flight data record, two previous real flight data will be used to provide a comparison between real flight and SAGADA data. One set of data came from a VS30 rocket flight and the other from a VS30-ORION rocket flight. Besides, since there is only a single set of previous flight data available for the VS30-ORION rocket up to now, the analysis that follows will also be used for tunning up the filters for future missions of this type of rocket.

Figure 1 shows the nominal longitudinal rocket accelerations for a VS30 and a VS30-ORION flight. The VS30 rocket is a single-stage rocket whereas the VS30-ORION is a two-stage one with the two propulsion stages separated by a short balistic period. The VS30-ORION rocket is basically a VS30 with an additional stage. As can be seen in Figure 2, the VS30 and VS30-ORION rocket flights take about 6 and 9 minutes duration, respectively, therefore, both flights have very long balistic phases after the propulsion phase.

As already mentioned, the present algorithm implemented at the CLA rocket tracking system permits the division of the entire flight in two phases only. Therefore, for the VS30 rocket flight one phase was taken for the duration of the propulsion of the rocket (from 0 to 31 seconds) and the other phase for the remaining balistic part of the flight.

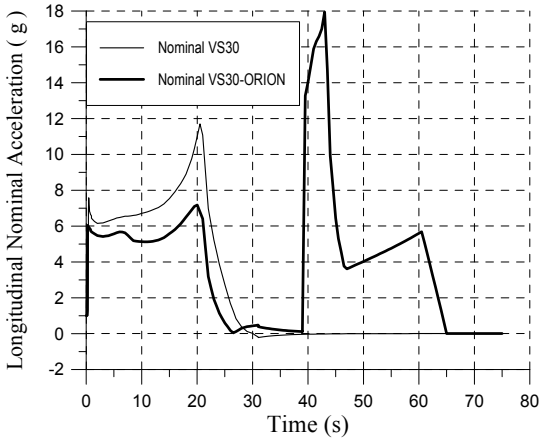


Figure 1 Rocket relative longitudinal acceleration.

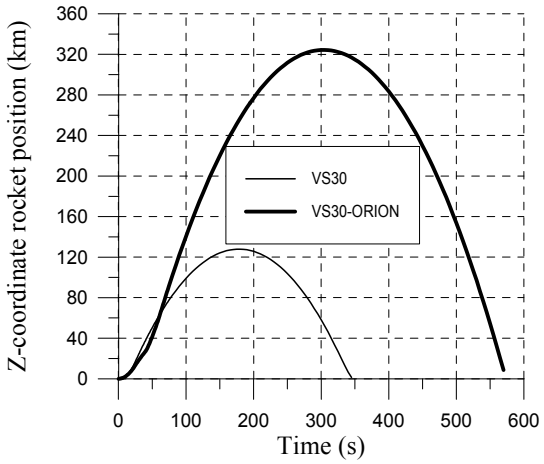


Figure 2. The z-coordinate rocket position for the VS30 and VS30-ORION rockets.

Although for the VS30-ORION flight is not so straightforward the choice of the phase intervals as for the VS30 rocket, taking into account that the balistic time interval between both propulsion stages is short and the fact that the accuracy of estimating the rocket fall location at the end of the flight is also important, the best flight division was to take one phase from the beginning of flight up to the end of the second stage (from 0 to 65 seconds) and the other phase from this time on which is only a balistic part of the flight. In order to illustrate the usual behavior of the standard deviation of the errors between nominal and estimated impact point as a function of filter

gains, Figure 3 shows the results for the first phase of the VS30 flight for the ATLAS radar. Tables 1 and 2 show the calculated α filter gain, which minimizes the standard deviations of the error in the impact point, for both phases of each flight for the ATLAS and ADOUR radars, respectively. Also shown in this table is the relative error between the α gains calculated with the real flight and SAGADA data.

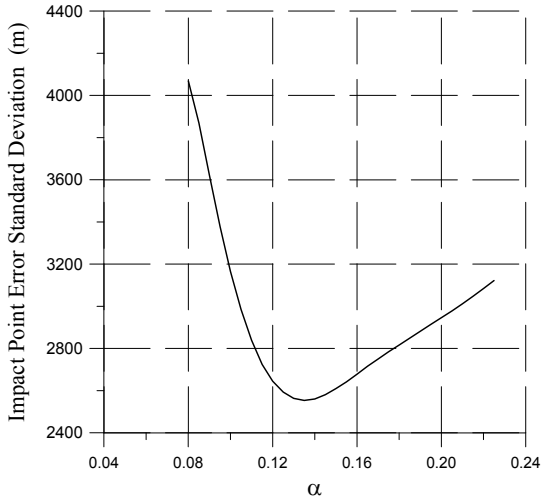


Figure 3. A typical variation of the impact point error standard deviation with respect to the filter gains.

Table 1: Values of the best α filter gain for Real and SAGADA data and the associated relative error for the ATLAS radar.

		ATLAS radar	
		VS30-ORION	VS30
Phase-1	Real	0.251	0.135
	SAGADA	0.251	0.135
	Error (%)	**0**	**0**
Phase-2	REAL	0.059	0.085
	SAGADA	0.059	0.076
	Error (%)	**0**	**+12**

The associated errors presented in Table 1 and 2 indicate the consistency of the results for both flights. For the ATLAS radar, the results also show that there is basically no need to make any corrections to the α filter gain calculated with SAGADA data. For the ADOUR radar by its turn ,the results show that for both flight phases the gain calculated with SAGADA data should be reduced for a better estimation. Since radars undergo maintainance frequently and the VS30-ORION data are the most recent ones, the corrections indicated by this flight data will be used to estimate best filter gains, for the ADOUR radar, for future flights with no previous flight data record available. It is important to mention that these correction factors should be revaluated whenever new flight data become available.

Table 2: Values of the best α filter gain for Real and SAGADA data and the associated relative error for the ATLAS radar.

		ADOUR radar	
		VS30-ORION	VS30
Phase-1	Real	0.115	0.105
	SAGADA	0.167	0.124
	Error (%)	**-30**	**-17**
Phase-2	REAL	0.040	0.052
	SAGADA	0.044	0.077
	Error (%)	**-10**	**-35**

Figure 4.a and 4.b show the evolution of the absolute error between filtered and nominal impact point during the first phase of the VS30-ORION flight for the best selected set of gains compared to arbitrary ones with lower and higher values of α gain, respectively. The value 0.251 is the best-calculated α gain for the specific flight phase. The results shown in this figure emphasize the importance of obtaining accurate estimates for the best set of filter gains. The wrong choice of the filter gains may yield such a poor impact point estimate that even a normal mission could be aborted for safety reasons. Another way to compare the results for different sets of gains is through the visualization of the evolution in time of the rocket impact point trajectory in a monitor screen in a very close way the Flight Safety staff does during real flights. For this purpose, a computer code has been developed such that in the monitor screen can be seen the contour of the world map around the launching ramp, the boundaries of a safety region, and the evolution in time of the nominal and estimated rocket and impact point trajectories. The kind of plot shown in Figure 4 is very helpful to quantify the error in the impact point estimate whereas the simulation in the monitor screen is very important to check the distance of the estimated impact point from the boundaries of the safety region as well as to verify if the peaks seen in the plot of Figure 4, for instance, are because the estimated impact point trajectory is getting far from the nominal one or is

simply delayed or advanced compared to the nominal values. Therefore, these are two very important tools in the analysis of the best filter gains. For instance, even within a flight phase, a certain part of the flight can be more importante than others. Since the best set of gains is calculated considering the entire flight phase duration, this may not be the most appropriate for that specific part of the flight. In such a case, the described tools are very helpful to visualize the difference between the nominal and estimated trajectories and also to quantify it.

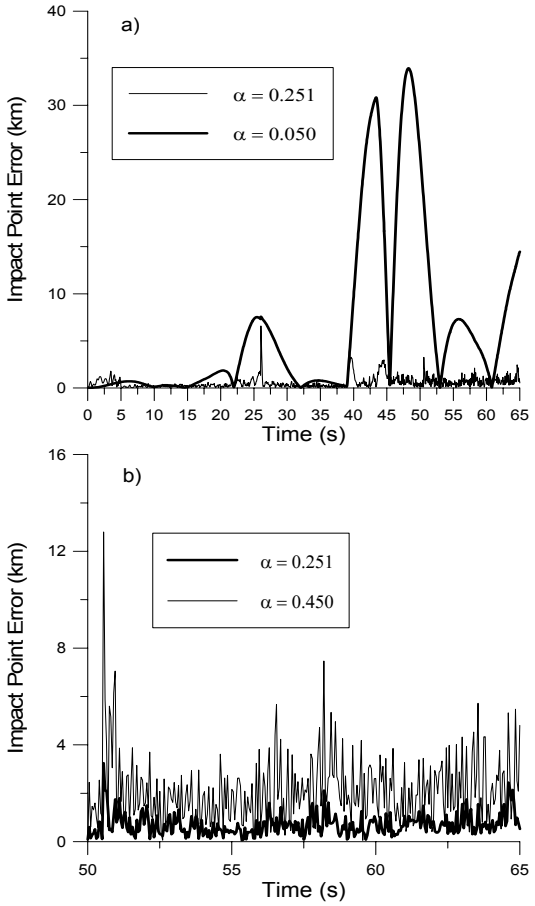


Figure 4. Comparison of the mpact point absolute error evolution in time for the VS30-ORION rocket calculated with the ATLAS radar data for different sets of filter gains.

As can be seen in Figure 5, the estimated rocket impact point calculated with the ATLAS radar data is, in general, much more accurate than the one calculated with the ADOUR ones. This makes the ATLAS radar the most important for rocket tracking at CLA. As mentioned before, althoug the ADOUR radar is less precise than the ATLAS radar, the former is closer to the launching ramp, thus the ADOUR radar usually yields better results during the first instants of flight as can also be seen in this figure. Therefore, this radar is very important for rocket tracking at the beginning of flight and also serves as a backup radar for the remainig of the flight.

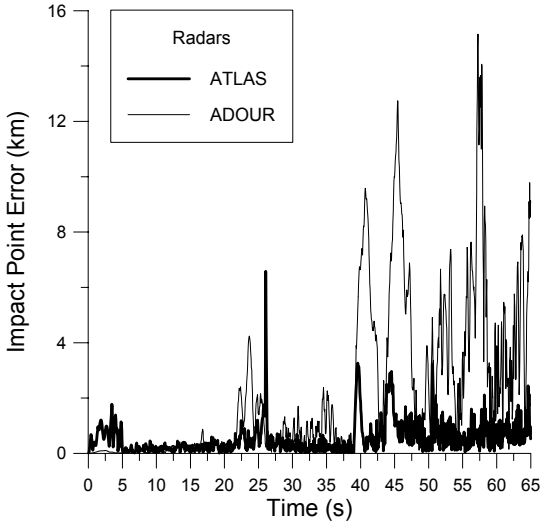


Figure 5. A comparison of the impact point error evolution in time calculated with both radar data for the first phase of the VS30-ORION flight.

RESULTS OF THE METHODOLOGY APPLICATION

The VLS rocket flight will be used to illustrate the application of the described filter gain computational methodology for rockets which does not have yet complete previous flight data record. As already mentioned, the VLS rocket is a satellite launcher vehicle, therefore, differently from the other rockets, the VLS does not fall on earth but stays in a earth orbit by the end of the flight. As can be seen in Figure 6, this rocket has three consecutive propulsion stages up to 193 seconds of flight followed by a long balistic period up to 500 seconds of flight when the fourth and last propulsion stage starts. Basically, by the end of the fourth propulsion stage the rocket should be already in its expected orbit. By inspection of the rocket nominal acceleration in Figure 6 and considering all that have been discussed in the methodology description, a reasonable way to divide the flight in only two phases should be to choose the first phase as the time interval consisted of the three first propulsion stages, that is up to 193 seconds, and

the remaining of the flight as the second phase. As can be seen in Figure 6, the second phase contains a long balistic flight followed by the fourth propulsion stage.

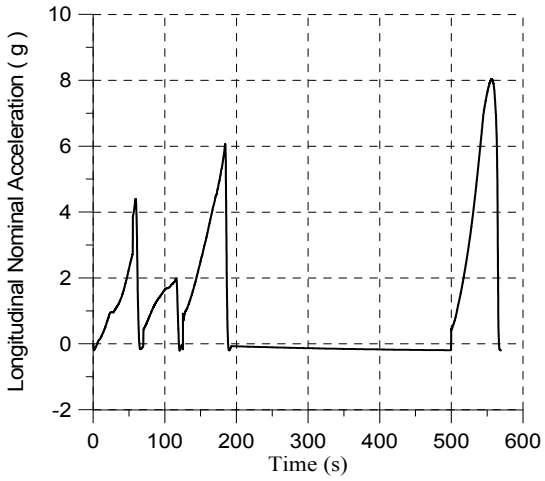


Figure 6. The VLS rocket longitudinal nominal acceleration.

Tables 3 and 4 present the calculated α-gain for the SAGADA data from the ATLAS and ADOUR radars, respectively, for 2, 5 and 15-phase filters. Also shown in the last column of these tables are the impact point error standard deviations for each phase of the 15-phase filter. For the 5-phase filter, each propulsion stage is a single flight phase and the balistic period is another one. For the 15-phase filter, each phase in the 5-phase filter is divided into three equally spaced intervals to yield the 15 phases. Tables 3 and 4 show that for the duration of the the first three propulsion stages not much improvement in the impact point estimates is obtained by considering each propulsion stage as a flight phase because all the three stages have very close values of calculated gains. When the propulsion stages are further divided as for the 15-phase filter, it can be seen that the calculated gains are quite different even within the same propulsion stage. For instance, considering the beginning of the flight up to 45 seconds, which is a critical part of the flight because the impact point trajectory is still very close to the boundaries of the safety region, the 2-phase filter has α=0.106 whereas the calculation for the 15-phase filter suggest α~0.050 as the best gain for that period of flight. Another important point noticed from the data in these tables is that, in the second phase of the 2-phase filter, the estimate in the last 24 seconds (last phase of the 15-phase filter) has a huge weight in the calculation of filter gain, as can be seen by its standard deviation value, although at that time the impact point estimate is not so important since the rocket should be basically in orbit already. Therofore, for the 2-phase filter like the one implemented at CLA, the best choice of the flight phases and the decision about which data is really important for the analysis is not a straighforward task. The Flight Safety staff certainly has a very important role in taking these decisions since they can provide helpful additional information. Clearly, these problems could be attenuated if the use of a N-phase filter was possible. Neverthless, considering the above discussion, Table 5 presents some possible sets of a 2-phase filter gain for the VLS rocket calculated with the suggested methodology. In this table, the gains are already corrected by the factors indicated in Tables 1 and 2. The corresponding β and γ parameters should be calculated from Eqs. (16) and (17).

Table 3. Filter gains calculated for three different number of phase filters using SAGADA data of the ATLAS radar for a VLS flight.

ATLAS Radar				
Final Instant (s)	**α**			**σ (m)**
	2 Phase Filter	**5 Phase Filter**	**15 Phase Filter**	
22	0.106	0.118	0.044	113
45			0.052	279
68			0.125	1456
86		0.092	0.073	490
105			0.050	471
124			0.113	1764
147		0.111	0.090	2854
170			0.066	1640
193			0.116	10861
295	0.091	0.036	0.049	7119
397			0.016	1089
499			0.011	892
522		0.098	0.037	4301
545			0.041	6480
569			0.102	81962

Table 4. Filter gains calculated for three different number of phase filters using SAGADA data of the ADOUR radar for a VLS flight.

ADOUR Radar				
Final Instant (s)	**α**			**σ (m)**
	2 Phase Filter	**5 Phase Filter**	**15 Phase Filter**	
22	0.116	0.112	0.052	106
45			0.044	248
68			0.115	1450
86		0.101	0.064	653
105			0.046	292
124			0.114	1723
147		0.125	0.090	2794
170			0.047	1570
193			0.127	11579
295	0.085	0.027	0.039	4423
397			0.010	909
499			0.010	1175
522		0.094	0.034	5631
545			0.038	10778
569			0.100	101505

Case-1 refers to the conventional flight division whereas Case-2 through Case-4 take into account information from the 15-phase filter results to divide flight into only two phases.

Table 5. Possible values of the α gain for a VLS rocket flight for both radars.

Case	Phase final instant (s)	ATLAS radar	ADOUR radar
1	193	0.106	0.081
	569	0.091	0.078
2	193	0.106	0.081
	500	0.033	0.023
3	50	0.058	0.047
	569	0.090	0.073
4	50	0.058	0.047
	500	0.061	0.051

COMMENTS

A methodology for the calculation of suitable sets of filter gains for the tracking of rockets with and without previous flight data record has been presented. The most difficult task in the use of such a methodology, for filter algorithms which permit only a reduced number of phases, is the choice of the most appropriate flight division. A N-phase filter algorithm eliminates completely this difficulty. Besides, this type of filter requires very few changes to the present algorithm, therefore it is been considered for implementation at CLA. For the α–β–γ filter type, in order to assure good efficiency, the entire mission should be normal, it no matters how many flight phases are considered, since the filter gains are set before the flight. So, a possible improvement to the filtering system should be to implement a filter algorithm, such as the Kalman filter [3,6], which revaluate the filter gains every time instant according to the flight evolution

ACKNOWLEDGEMENTS

The authors wish to express their acknowledgements to CLA and IAE (brazillian Space and Aeronautics Institute) for their help and support for the development of this work.

REFERENCES

1. A. D. Caldeira, E. M. Borges, F. A. Braz Filho, J. Rubini Jr, L. Guimarães and M. A. P. Rosa, Analysis of the filter gains of the rocket tracking system at the Alcântara Launching Center for the VLS1-V02 rocket flight, Relatório de Pesquisa IEAv, 2000 (in portuguese)

2. E. M. Borges; F. A. Braz Filho, A. D. Caldeira, L. Guimarães and M. A. P. Rosa, Preliminary report of a filter gain analysis for the ATLAS and ADOUR rocket tracking radars, Nota Técnica IEAv, August 2000 (in portuguese).

3. Yaakov Bar-Shalon and Xiao-Rong Li, *Estimation and Tracking: Principles, Techniques and Software*, Artech House Inc., 1993.

4. F. A. Braz Filho, A. D. Caldeira, Mauricio A. P. Rosa, E. M. Borges, L. Guimarães and J. Rubini Jr., A computer code for the calculation of N-phase filter gains, Nota Técnica IEAv, 2001 (in portuguese).

5. W. H. Press et. al., *Numerical Recipes*, Cambridge University Press, 1992.

6. A. Gelb, J. F. Kasper, Jr., R: A. Nash, Jr., C. F. Price, A. A. Sutherland, Jr., *Applied Optimal Estimation,* The M.I.T. Press, Massachusetts, 1974.

4^{th} *International Conference on Inverse Problems in Engineering*
Rio de Janeiro, Brazil, 2002

IDENTIFICATION OF THE GROUNDWATER TABLE LOCATION IN THE FOREST IMPACT PROBLEM

Anatoli Leontiev & Ronaldo da S. Busse
Instituto de Matemática,
Universidade Federal do Rio de Janeiro,
Rio de Janeiro-RJ, Brasil
anatoli@im.ufrj.br

Wilma Huacasi
LCMAT-CCT, Universidade Estadual
do Norte Fluminense,
Campos dos Goytacazes-RJ, Brasil
wilma@uenf.br

José Herskovits
PEM-COPPE,
Universidade Federal do Rio de Janeiro,
Rio de Janeiro-RJ, Brasil
jose@serv.com.ufrj.br

Cirstovão M. Mota Soares
Instituto de Engenharia Mecânica,
Instituto Superior Técnico,
Lisboa, Portugal
cmmsoares@alfa.ist.utl.pt

ABSTRACT

We propose here a mathematical model for the forest impact phenomenon. By this phenomenon it is mean the raising or the lowering of the groundwater table under the areas felled or recovered by the trees. Our formulation includes a boundary-value problem with contact and free boundary conditions. We offer a variational formulation of this problem, which is a quasi-variational inequality and prove its equivalence to the original problem. We describe a numerical algorithm for solution of the forest impact problem. Considering the free-contact boundary problem as a shape optimization problem we perform its boundary elements discretization. Taking the state variable and free boundary variable as independent variables, we treat the discretized problem as a nonlinear mathematical program and apply interior point algorithm to solve it. Numerical results for an illustrative 2D test problem are discussed.

INTRODUCTION

Various studies on the groundwater flow realized during last two centuries show increasing interest in the problem by virtue of the importance of water recurses management for the future of humanity. In the present paper we deal with the phenomenon of the forest impact on aquifers, the problem that in different forms appears in various fields of activity such as agriculture, civil engineering, etc.

By the forest impact on aquifers we meant the effect of raising or lowering of the groundwater table under the areas felled or recovered by the trees, see Fig. 1. From the hydromechanical point of view this is a problem of unconfined flow in porous media with possible fluid discharge through the water table owing to the tree roots suction. Mathematically, the water table can be considered as a free boundary, so this is a free boundary problem, see [1], [2]. The location of the water table under the forest suction effect, the flow characteristics as well as the region of the contact of the aquifer with the tree roots system are the unknowns of this problem, see [3].

To study the forest impact phenomenon the use of experimental methods and empirical formulas is common, see [3], [4], for example. The experiments consist in real time and real scaled monitoring of the water table response under a forest area and can take many years to obtain consistent details. To predict the groundwater level reduction, water balance models are applied, see [4].

Two dimensional model for the forest impact phenomenon proposed in this paper includes a boundary-value problem with a contact conditions which substitute for a part of the free boundary conditions.

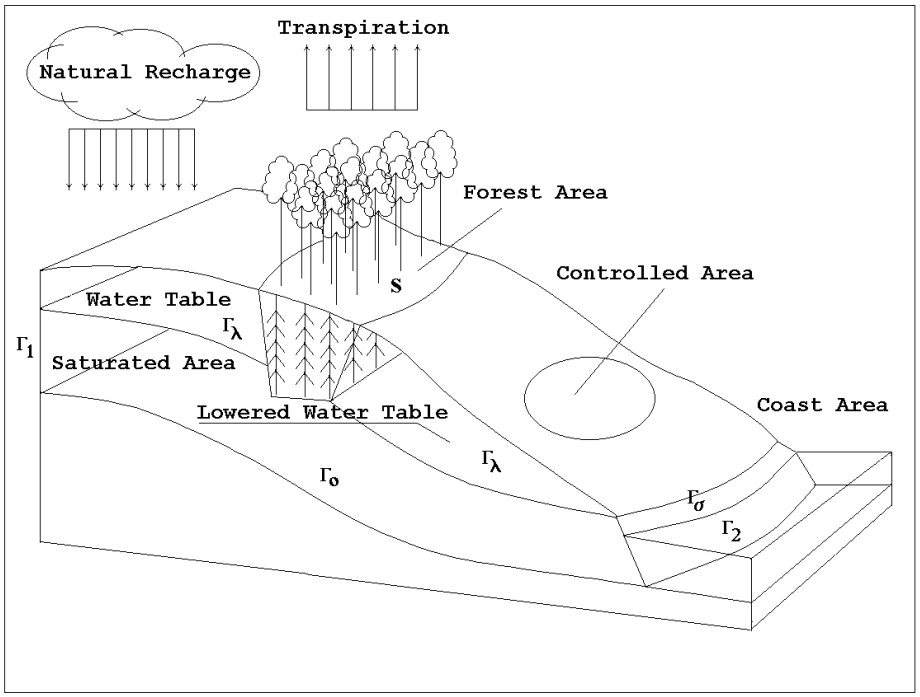


Fig. 1 Water Table and Tree Roots Sustem Interaction Scheme

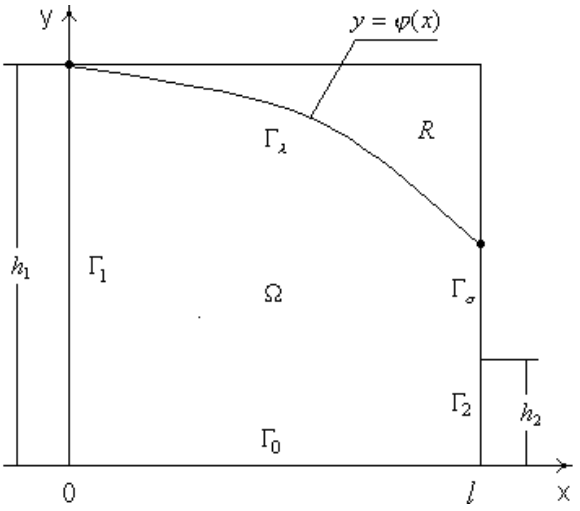


Fig. 2 Unconfined Fluid Flow. Classical Case

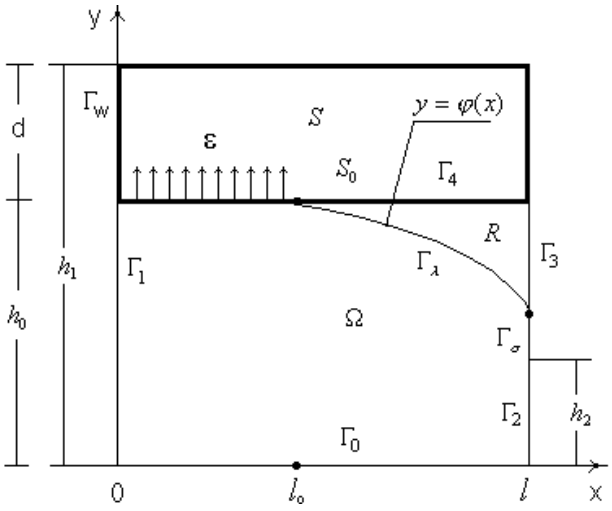


Fig. 3 Unconfined Fluid Flow with Suction

We use a Baiocchi-like transformation of the problem variables and obtain a quasivariational inequality associated to the boundary-value problem. We prove that this inequality is equivalent to the original boundary-value problem.

Our numerical technique to solve the forest impact problem is based on the shape optimization approach. We transform the free-contact boundary problem into a least squares like shape optimization problem. The objective functional contains one of the free boundary conditions, whereas the state equation together with the rest of the boundary conditions become the problem constraints. It is sought for the minimum of objective with respect to the shape of water table. This approach for classical seepage problem was used in [5] with finite elements descretization and in [6] using boundary elements method. The numerical example of the forest impact problem is shown and compared with different situations including the classical seepage problem.

THE CLASSICAL SEEPAGE PROBLEM

In the classical case of the unconfined flow through a porous media, the unknowns of the problem are the characteristics of the flow, like velocity potential $u(x, y)$, and the flow region (aquifer) Ω itself, see [1]. A part of the aquifer boundary Γ_λ, called the water table, is unknown *a priori* and has to be located, see Fig. 2.

In this paper we consider two dimensional steady flow through homogeneous and isotropic porous media with the permeability coefficient $k = 1$ and assume that the external pressure is equal to zero. Let R be an open and, for the sake of convenience, rectangular domain occupied by the porous media, h_1 and h_2 the fluid piezometric levels in the left and in the right sides of R respectively, $\Gamma_\circ$ the impermeable bottom and Γ_σ the seepage line. The classical case does not suppose any evaporation (or infiltration) effects on the water table.

Then, the classical problem of the unconfined flow through a porous media can be formulated as a free boundary problem:

PROBLEM 1. Find potential $u(x, y)$ and decreasing function $\varphi(x)$ that defines the location of the water table Γ_λ, satisfying

$$\begin{cases} \Delta u = 0 & \text{in} \quad \Omega, \\ u = h_1 & \text{on} \quad \Gamma_1, \\ u = h_2 & \text{on} \quad \Gamma_2, \\ u = y & \text{on} \quad \Gamma_\sigma \cup \Gamma_\lambda, \\ q = 0 & \text{on} \quad \Gamma_\circ \cup \Gamma_\lambda, \end{cases}$$

where $q \equiv \partial u/\partial n$ and n is the outward normal to $\Gamma_\circ \cup \Gamma_\lambda$.

In the unknown part Γ_λ of the boundary the function $u(x, y)$ has to fulfill two boundary conditions (free boundary conditions) $u = y$ and $q = 0$. Thus, the water table is considered as a free boundary. Problem 1 admits an unique solution pair $\{\varphi, u\}$, where $\varphi(x)$ is smooth and $u \in H^1(\Omega) \cap C^\circ(\overline{\Omega})$, see [7].

Performing the Baiocchi transformation [7], that is $w(x, y) = \int_y^{\varphi(x)} (u(x, t) - t)dt$, a variational inequality equivalent to Problem 1 can be obtained:

$$\int_R (w_x(v - w)_x + w_y(v - w)_y)dxdy \geq -\int_R (v - w)dxdy, \; w, \forall v \in K.$$

Here $K = \{v \in H^1(R) \mid v \geq 0 \text{ in } R, v = g \text{ on } \partial R\}$, the subscript x (or y) denotes the derivative with respect to x (or y) and function g is defined by using values of h_1, h_2 and l.

FOREST IMPACT PROBLEM

The difference between forest impact problem and classical seepage problem is in the possibility of the flow flux through the water table, which can appear when the aquifer attains the tree roots system. Let $\mathcal{R}$ be domain occupied by the porous media and $\mathcal{S}$ the tree roots system of the deepness $d > 0$, see *Fig. 3*. We suppose that at the part of the water table that reaches the tree roots system bottom $S_\circ$ there is the suction flux with given rate $\varepsilon(x)$. The left wall Γ_w of $\mathcal{S}$ is assumed impermeable. The contact area between aquifer and tree roots system is *a priori* unknown and can be defined together with the location of the rest of the water table Γ_λ, seepage Γ_σ and the velocity potential u in Ω. We suppose also that

the function $\varphi(x)$ that defines the portion $\Gamma_\lambda \setminus S_\circ$ of the water table is decreasing and denote $h_\circ \equiv h_1 - d$.

For the forest impact problem we define at the parts $\Gamma_1, \Gamma_2, \Gamma_\circ$ and Γ_σ of the boundary $\partial\Omega$ the same conditions as for the classical seepage problem. The part of the water table that does not contact $\mathcal{S}$ remains to be the free boundary and we put here conditions $u = y$ and $q = 0$. When $\Gamma_\lambda \cap S_\circ \neq \emptyset$ we have the flow with given rate $\varepsilon(x)$ through this part of the water table Γ_λ toward the interior of $\mathcal{S}$. Thus, we obtain the following mathematical formulation for the forest impact problem:

PROBLEM 2. Find potential $u(x, y)$ and decreasing function $\varphi(x)$ that defines the portion of the water table Γ_λ without contact with S, satisfying

$$\begin{cases} \Delta u = 0 & \text{in} \quad \Omega, \\ u = h_1 & \text{on} \quad \Gamma_1, \\ u = h_2 & \text{on} \quad \Gamma_2, \\ u = y & \text{on} \quad \Gamma_\sigma \cup (\Gamma_\lambda \setminus S_\circ), \\ q = 0 & \text{on} \quad \Gamma_\circ \cup (\Gamma_\lambda \setminus S_\circ), \\ q = -\varepsilon(x) & \text{on} \quad \Gamma_\lambda \cap S_\circ, \end{cases}$$

where $q \equiv \partial u/\partial n$ and n is the outward normal to $\Gamma_\circ \cup \Gamma_\lambda$.

At the water table we have conditions that take the form of free or contact boundary conditions. We call its "free-contact" boundary conditions. We obtain here a variational reformulation of this free-contact problem.

Let us consider in Ω the transformation:

$$w(x, y) = \int_y^{\psi(x)} (u(x, t) - t)dt + w_\circ(x), \quad (1)$$

where $\psi(x)$ a function that describe the whole water table Γ_λ and the function $w_\circ(x)$ is defined in the following form:

$$\begin{aligned} &w_\circ \in C^1[0, l], w_\circ(0) = d^2/2, w_\circ(l) = 0, \\ &w_\circ''(x) = -\varepsilon(x) \ \text{ on } [0, l_\circ), \\ &w_\circ''(x) = 0 \text{ on } (l_\circ, l]. \end{aligned} \quad (2)$$

Here the interval $[0, l_\circ)$ corresponds to the contact part of the water table and $(l_\circ, l]$ to the free one. Let $g(x, y)$ be a function of class $C^1(\overline{\mathcal{R}})$ such that $g = w$ on $\partial\mathcal{R}$ and $\mathcal{K}$ a nonempty, convex and closed subset of $H^1(\mathcal{R})$:

$$\begin{aligned} \mathcal{K} = \{v \in H^1(R) \ | v \geq w^\circ \text{ in } \mathcal{R} \\ \text{and } v = g \text{ on } \partial\mathcal{R}\}. \end{aligned} \quad (3)$$

Then, we have the following result:

THEOREM 1. Let $\{\varphi, u\}$ be a solution of Problem 2, $\varphi(x)$ is smooth, $u \in H^1(\Omega) \cap C^\circ(\overline{\Omega})$, w is given by formula (1), $w_\circ(x)$ is defined by conditions (2), $w^\circ(x, y) \equiv w_\circ(x)$ for $(x, y) \in \mathcal{R}$ and

$$w(x, y) = \begin{cases} w(x, y), & (x, y) \in \Omega, \\ w^\circ(x, y), & (x, y) \in \mathcal{R} \setminus \Omega. \end{cases}$$

Then w satisfies:

$$\begin{aligned} \textstyle\int_R (w_x(v - w)_x + w_y(v - w)_y)dxdy \geq \\ - \textstyle\int_R (v - w)dxdy, \ \ w, \forall v \in \mathcal{K}, \end{aligned} \quad (4)$$

where $\mathcal{K}$ is defined by (3). □

By the definition of function w°, the subset $\mathcal{K}$ depends implicitly on the flow through the contact part of Γ_λ. This part is unknown *a priori* and is defined by the function w. Hence, inequality (4) is a quasivariational one. The next theorem shows that if the solution w of quasivariational inequality (4) exists then the function $u = y - w_y$ together with the curve $\varphi(x)$ that separates two regions of $\mathcal{R}$ where $w = w^\circ$ and $w > w^\circ$, satisfy Problem 2.

Let be $u := y - w_y$ and the function $\varphi(x)$ is defined as

$$\begin{aligned} &\varphi(x) = \inf\{y | (x, y) \in \mathcal{R} \setminus \Omega\}, l_\circ < x < l, \\ &\varphi(l_\circ) = \lim_{x \to l_\circ^+} \varphi(x), \ \ \varphi(l) = \lim_{x \to l^-} \varphi(x). \end{aligned} \quad (5)$$

THEOREM 2. Let $w \in W^{2,p}(\mathcal{R}) \cap C^1(\overline{\mathcal{R}})$ with $1 \leq p < \infty$ be a solution of (4). Let be $\Omega = \{(x, y) \in \mathcal{R} \mid w(x, y) > w^\circ(x, y)\}$ and assume $\varepsilon'(x) \geq 0$. Let us consider $u := y - w_y$ in Ω and define $\varphi(x)$ by formula (5). Then the pair $\{u, \varphi\}$ is the solution of Problem 2. □

SHAPE OPTIMIZATION PROBLEM

An equivalent formulation of Problem 2 can be given in terms of shape optimization for the system governed by the Laplace equation. Let Φ be a set of all feasible shapes of the water table, formed by smooth curves. The optimization problem consists in finding $\psi \in \Phi$ and u such that:

$$\begin{cases} \min\limits_{\psi\in\Phi} (q)^2_{\Gamma_\lambda \setminus S_\circ} \\ \textit{where } q = \partial u/\partial n \textit{ and } u(x,y) \\ \textit{is a solution of problem:} \\ \begin{cases} \Delta u = 0 & \textit{in} \quad \Omega, \\ u = h_1 & \textit{on} \quad \Gamma_1, \\ u = h_2 & \textit{on} \quad \Gamma_2, \\ u = y & \textit{on} \quad \Gamma_\sigma \cup (\Gamma_\lambda \setminus S_\circ), \\ q = 0 & \textit{on} \quad \Gamma_\circ, \\ q = -\varepsilon(x) & \textit{on} \quad \Gamma_\lambda \cap S_\circ, \end{cases} \end{cases} \tag{6}$$

The objective functional contains the square of the flux along the free part of the water table. The choice of the optimal water table location forces the objective to be zero and vice versa.

In two-dimensional case for the problem governed by the Laplace equation the values of flux and potential verify on the frontier $\Gamma \equiv \partial\Omega$ the integral equation, [8]:

$$0.5u(\xi) + \textstyle\int_\Gamma q^*(\xi,\chi)u(\chi)d\Gamma = \int_\Gamma u^*(\xi,\chi)q(\chi)d\Gamma,$$

where $\chi \equiv (x,y) \in \Gamma$, $u^*(\xi,\chi)$ is the fundamental solution of the Laplace equation, $q^*(\xi,\chi)$ its normal derivative, and $\xi \in \Gamma$ is the collocation point.

In this way, to define the location of the water table we have the problem:

$$\begin{cases} \min\limits_{\psi\in\Phi} F(u,q), \\ \textit{where } q \textit{ and } u \textit{ verify at } \Gamma\textit{:} \\ 0.5u(\xi) + \displaystyle\int_\Gamma q^*(\xi,\chi)u(\chi)d\Gamma = \int_\Gamma u^*(\xi,\chi)q(\chi)d\Gamma, \end{cases} \tag{7}$$

where $F(u,q) = (q)^2_{\Gamma_\lambda\setminus S_\circ}$ and the boundary values are defined as in (6).

B.E.M. DISCRETIZATION

Formulation (7) furnishes an opportunity to apply the boundary elements discretization. We assume that the x-coordinates of the nodes at $\Gamma_\lambda \setminus S_\circ$ and the y-coordinates of the nodes at $\Gamma \cup S_\circ$ are fixed. Thus only the y-coordinates define the location of the nodes belonging to seepage and free part of the water table and x-coordinates define the nodes of the contact part of the water table.

For the discrete analog of (7) we consider as independent variables the flux at the boundary elements of Γ_1, the potential at the boundary elements of $\Gamma_\circ$, the flux at the boundary elements of Γ_2, Γ_σ and $\Gamma_\lambda \setminus S_\circ$, the the potential at the boundary elements of $\Gamma_\lambda \cup S_\circ$, y-coordinates of the seepage surface nodes, y- and x-coordinates of the water table nodes.

Performing this kind of discretization we obtain a nonlinear mathematical programming problem. To solve it we use Herskovits' interior point algorithm, [9]. We find the y-coordinates of free part of the water table and seepage surface nodes as well as x-coordinates of the contact part of the water table and values of potential and flux at the corresponding segments of the boundary.

NUMERICAL TESTS

For the test problem we choose: $h_1 = 6.3014$, $h_2 = 1.2359$, $\ell = 6.1592$ and $d = 1.3014$ ($h_\circ = 5.0$). This data is taken in order to compare the solution of the forest impact problem with the seepage one, considered in [6]. The suction flux is taken as $\varepsilon = 1$.

Our discretization includes 26 boundary elements, see *Fig. 4*. We are looking for the y-coordinates of ten nodes at the free part of the water table $W - M$ and the x-coordinates of three nodes at the contact part of the water table $B - W$. The position of the node 24 defines the location of the contact point of the water table (point W). The coordinates of the rest of the nodes are fixed. The water table initial position, used at the first iteration of the algorithm, is given by the line $B - W_\circ - M_\circ$ in *Fig.4*.

The mathematical program have 39 variables, 26 nonlinear equality constraints, 12

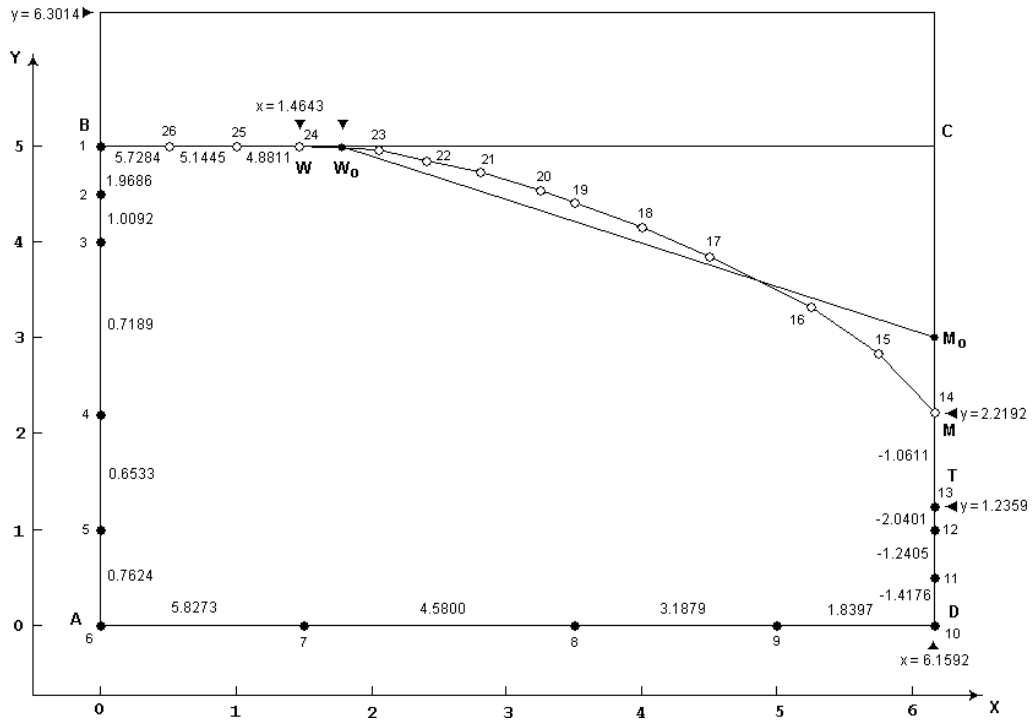


Fig. 4 B.E.M. Discretization

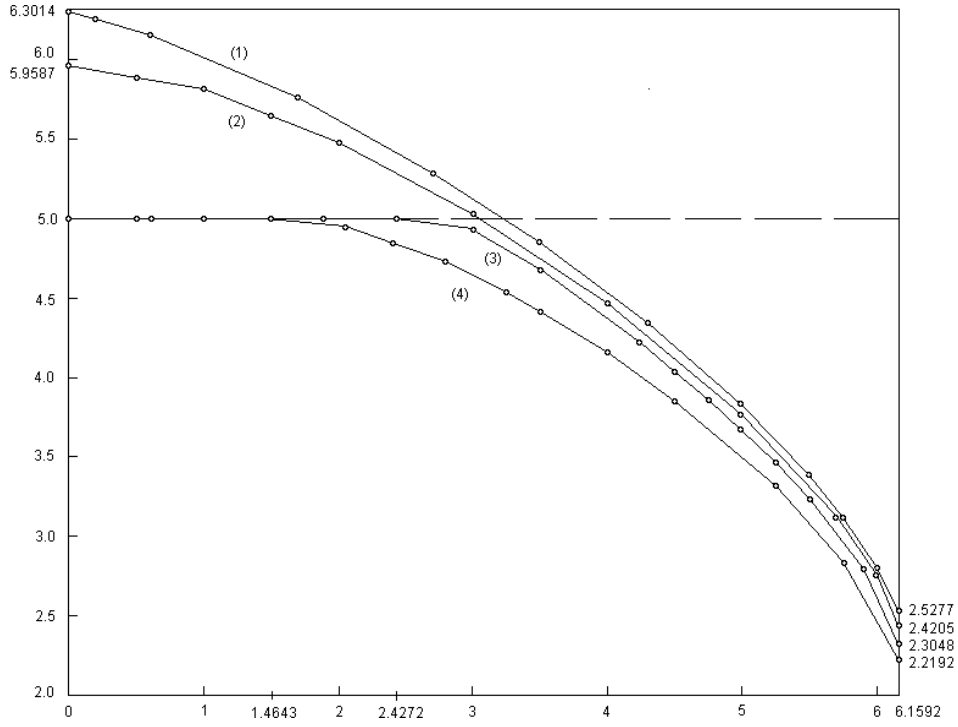


Fig. 5 Water Table Location. Numerical Results

"box" constraints and 2 linear inequality constrains. We adopt the algorithm stopping criterion with precision 10E−6 (see Herskovits [9] for details). With the different initial data, the convergence of the algorithm was obtained in no more than 20 iterations.

Fig. 4 shows the location of the water table (continuous line $B-W-M$) and corresponding nodes (`14-26`) positions calculated numerically as well as boundary data, i.e. flux at the segments $A-B$, $D-T$ and $T-M$ and potential for the segments $A-D$ and $B-W$.

We compare the location of the water table in the forest impact problem with the solution of another unconfined problems, considered for the same geometrical and piezometric parameters. The results are presented in *Fig. 5*. Here line (1) defines the location of the water table for the classical seepage problem, line (2) gives the location of the water table for the seepage problem with vertical impermeable wall Γ_w only (see *Fig. 3*), line (3) is the water table in the case of impermeable bottom $S_\circ$, line (4) is the solution of the forest impact problem with constant suction rate $\varepsilon = 1$.

CONCLUSIONS

We introduce the forest impact model at the form of "free-contact" boundary problem and obtain its equivalent variational formulation as a quasivariational inequality. This inequality seems for us to be more adequate that the free-contact formulation to study the properties of our model, such that existence and uniqueness of the solution and its regularity.

The numerical simulation shows that even for our model of forest impact, that takes into account only some principal characteristics of this phenomenon, the water table lowering owing to the forest suction is significative enough to be considered as an effective means for the control of groundwater.

ACKNOWLEDGEMENTS:

The authors gratefully acknowledge the support provided by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil), FAPERJ (Fundação da Amparo à Pesquisa do Estado do Rio de Janeiro), FUJB (Fundação Universitária José Bonifácio, Rio de Janeiro) and FCT-ICCTI(Portugal)/CNPq. The third author (W.H.) thanks the Institute of Mathematics of Federal University of Rio de Janeiro for his kindly invitation to stay as visiting professor during the Summer School-2002, while the principal part of the numerical results presented in this paper was obtained.

REFERENCES

1. P.Ya. Polubarinova-Kochina, *Theory of Ground Water Movement*, Princeton University Press, Princeton, 1962

2. A. Friedman, *Variational Principles and Free-Boundary Problems*, John Wiley & Sons, New York, 1982

3. G.E. Blight, *Lowering of the groundwater table by deep-rootes vegetation - The geotechnical effects of water table recovery*, In: Groundwater Effects in Geotechnical Engineering, Proceedings of the Ninth European Conference on Soil Mechanics and Foundation Engineering, Dublin. Eds. E.T.Hanrahan, T.L.L.Orr and T.F.Widdis, **1**, 285 (1987)

4. M.A. Bari and N.J. Schofield, *Lowering of a shallow, saline water table by extensive eucalipt reflorestation*, Journal of Hydrology, **133**, 273 (1992)

5. D. Begis and R. Glowinski, *Application de la méthode des éléments finis à l'approximation d'un problème de domaine optimal. Méthodes de résolution des problèmes approchés*, Applied Mathematics and Optimization, **2**, **n.2**, 130 (1975)

6. A. Leontiev and W. Huacasi, *Mathematical programming approach for unconfined seepage problem*, International Journal of Engineering Analysis with Boundary Elements, **25**, **n.1**, 49 (2001)

7. C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities. Applications to Free Boundary Problems*, John Wiley & Sons, Chichester, 1984

8. C.A. Brebbia, J.C.F. Telles and L.C. Wrobel, *Boundary Elements Techniques - Theory and Applications in Engineering*, Springer, Berlin, 1984

9. J. Herskovits, *Feasible direction interior-point technique for nonlinear optimization*, Journal of Optimization Theory and Applications, **99**, 121 (1998)

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ON THE INFLUX TO THE OUTFLUX MAPPING FOR THE TRANSPORT EQUATION

Nilson C. Roberty
Nuclear Engineering Program - COPPE
Universidade Federal do Rio de Janeiro,
P.O. Box 68509, 21945-970, Rio de Janeiro, RJ, Brazil
nilson@lmn.con.ufrj.br

ABSTRACT

The inverse problem for spatial reconstruction of the absorption and scattering coefficients is modeled in the context of the stationary one velocity linear transport boundary value problem. The first and the second order variational formulations for these problems are presented. The second order variational principle is used for the derivation of a second order form of the inverse transport equation adopted by the source detector methodology. The reconstruction problem is then stated as an study of the influx to the outflux mapping similar to the Dirichlet to Neumann mapping.

NOMENCLATURE

$\Re^3$	three dimensional euclidian space
$S^2 = \{ w \in \Re^3 \ such\ that\ \|w\| = 1 \}$	is the surface of the unit sphere.
Ω	convex domain.
$(\vec{s}, \vec{t})$	rotate coordinate.
$\Re^+$	non-negative real numbers.
T	unitary mapping
∇	gradient
div	divergent
g_i	trace on interfaces
s_a, s_t	absorption (total) coefficients (cross section)
s_s	scattering coefficients (cross section)
$L^\infty(\Omega)$	the vector space of all functions that are essentially bounded on Ω
$L^2(\Omega \times S^2)$	the vector spaces of all functions that are square integrable on $\Omega \times S^2$
$L^2(\Gamma^- \times S^2)$	same for $\Gamma^- \times S^2$ with weight $\|n_x.w\|$
$\Gamma^- = \{(x,w) \in \Omega \times S^2 : n_x w < 0\}$	is the influx, outflux surface
$f(x,w)$	the angular flux.
K	the scattering operator.
$G = s_t (I-K)$	the removal operator.
$R = \frac{1}{s_t} (I-K)^{-1}$	inverse of the removal operator
P_l	Legendre polynomials of grade *l*.
f	phase function.
L_w	the Leakage operator.
$\Gamma_w^{\mp} = \{x \in \partial\Omega : n_x.w < 0 (> 0)\}$	is in the influx (outflux) boundary for the direction *w*.
U	the operator for invertion of direction.
$P=(I+U)/2$	the projection operator.
f^{in}_{out}	influx (outflux) data.
Λ_s	influx to outflux mapping.
$L_{w,2}$	second order differential operator.
$f^{\pm}$	even (odd) parity angular flux.
$L^2_{\pm}(\Omega \times S^2)$	even (odd) parity subspace of $L^2(\Omega \times S^2)$.
$H^1_+(\Omega x S^2) = \{f \in L^2_+(\Omega x S^2)$: div(w f $) \in L^2_-(\Omega x S^2)$ $\}$	
F_1, F_2	first (second) order functional

∂F	variation of F.
$\Delta R^{ij}, \Delta G^{ij}$	difference between the operator for coefficients values j.
Q_s	minimum value for the second order functional.
Ψ_2	solution for the second order problem.
Sign(x)	=1 if x>0,=0 if x=0,=–1 if x<0.

INTRODUCTION

The radiative and particle transport can be modeled by a one group transport equation whose domain is the five-dimensional phase space $\Re^3 \times S^2$. The ray propagates thorough a medium which is the union of a finite number of contiguous sub-domains. $\grave{U} = \bigcup_{e=1}^{ne} \grave{U}_e$ with internal boundary $\bigcup_{e=1}^{ne} \Gamma_{ee'} = \overline{\Omega}_e \bigcap \overline{\Omega}_{e'}$ and an external boundary $\Gamma = \partial\overline{\Omega}$. The position of a photon or a particle such as a neutron and the direction in which it is propagating is characterized by the pair $(x,w) \in \Re^3 \times S^2$ where $\Re^3$ is the usual Euclidian space and $S^2 = \{w \in \Re^3, \text{ such that } \|w\| = 1\}$ is the surface of the unit sphere. As the ray propagates, its direction w introduces in the Euclidian plane an orthogonal parallel projection $\Pi_w : \Re^3 \to \Re^2$ of the domain Ω in a plane through the origin and perpendicular to w, $t = \Pi_w x$ and a line through the trace t that crosses the domain Ω following the direction $\Pi_{t,w} = \overline{\Omega} \bigcap \{t + sw, -\infty < s < \infty\}$. The ray direction and the respective plane induce in $\Re^3$ a rotated coordinate system that is the natural place for the formulation of the transport equation for all rays with the same direction. This rotated coordinate system is characterized by the mapping

$$T : \Re^3 \times S^2 \to \Re \times \Re^3 \times S^2$$
$$(x,w) \to T(x,w) = (s,t,w')$$

with $s = x.w$
$t = x-(x.w)w$
$w' = w$

which is one-to–one, continuous and continuous differentiable and has jacobian equal to one. Its inverse is $T^{-1}(s,t,w) = (x,w) = (t + w', w')$

Since the one group angular flux (photons or particles/area time)

$$f : \Re^3 \times S^2 \to \Re^+$$

is characterized by a function which has different regularities properties for spatial variations in directions parallel and perpendicular to the direction of ray propagation . This means that in a plane that crosses Ω and is perpendicular to w, we have a section of the angular intensity flux that we expect to be an L^2 function. In the direction of propagation of the radiation we have an attenuation or creation process which is proportional to directional derivatives, and so is an H^1 function. In order to write derivatives in the correct way we can use the mapping T to rotate the coordinate system, in this way:

$$\forall f : \Omega \times S^2 \to \Re^+$$
$$(T,f)(s,t,w) = f(T^{-1}(s,t,w)) = f(x,w) \quad (1)$$

and

$$w.\nabla f(x,w) = div(wf(x,w)) = \frac{\partial T f(s,t,w')}{\partial s} \quad (1b)$$

Since we also have internal boundaries $\Gamma_i = \Gamma_{ee'}$, the ray of direction w may have traces g_i in these interfaces, where these traces are spatial values for the rotated coordinate s, indexed by $0 \le i \le I(t,w)$, $g_0 < g_i < g_{I(t,w)}$ and obviously

$$\Pi_{t,w} = \bigcup_{i=1}^{I(t,w)} \{t + sw, g_{i-1} s \le g_i\}$$

Inside of each one of the sub domains Ω_e, the medium usually offers different properties for absorption and scattering, which are the process consider in this work. The functions absorption (or total) coefficients

$s_a, s_t : \Omega \to \Re^+$
and the scattering coefficient

$s_s : \Omega \times [-1,1] \to \Re^+$

are expected to be L^∞ functions, most frequently constants by parts Lebesgue simple functions.

THE DIRECT PROBLEM

The one velocity stationary linear transport problems can be stated as:

Find $f(x,w) \in L^2(\Omega \times S^2)$ such that

$$div(w f(x,w)) + s_t(x) f(x,w) - \int_{s^2} s_s(x, w.w') f(x,w') dw' = q(x,w) \quad \forall (x,w) \in \Omega \times S^2 \qquad (2a)$$

Subject to the boundary conditions

$$f(x,w) = f^{in}(x,w) \in L^2(\Gamma^- \times S^2) \text{ for } (x,w) \in \Gamma^- \qquad (2b)$$

$f(t + sw, w)$ continuous in Ω even for

$$(x,w) \in \Gamma_i \times S^2 \qquad (2c)$$

where the influx surface is

$$\Gamma^- = \{(x,w) \in \Omega \times S^2 : n_x w < 0\}$$

and the L^∞ removal and scattering cross sections, as well the $L^2(\Omega \times S^2)$ source term q(x,w) are given data.
In operator form we express this equation as

$$(L + s_t I - s_t K) f = q \qquad (3)$$

Where I is the identity operator, K is the scattering operator,

$$K : L^2(\Omega \times S^2) \to L^2(\Omega \times S^2)$$

$$\Psi(x,w) \mapsto (K\Psi)(x,w) = \int_{s^2} \frac{s_s(x, ww')}{s_t(x)} f(x,w') dw' \quad \forall (x,w) \in \Omega \times S^2$$

K is self-adjoint, $\|K\| \le b_0 < 1$. In this situation the removal $G = s_t(I - K)$ has a unique inverse operator $R = \frac{1}{s_t}(I - K)^{-1}$. This operator is given by

$$R : L^2(\Omega \times S^2) \to L^2(\Omega \times S^2)$$
$$(R\Psi)(x,w) = \Psi(x,w) + \int_{S^2} r(x, ww') \Psi(x,w') dw'$$

Where

$$r(x, ww') = \sum_{l=0}^{\infty} \frac{2l+1}{4p} \frac{s_s(x) f_e(x)}{s_t(x) - s_s(x) f_e(x)} P_l(w.w')$$

This is an absolutely and uniformly convergent series. Here P_l is the Legendre polynomial of order l and f_l is the l-coefficient of the Legendre expansion of the phase function f. We note that the phase function

$$f(x, ww') = \frac{s_s(x, ww')}{\int_{S^2} s_s(x, ww') dw'} = \frac{s_s(x, ww')}{s_s(x)}$$

is normalized to one.

The Leakage operator is an unbounded differential operator

$$L_w : L^2(\Omega \times S^2) \to L^2(\Omega \times S^2)$$
$$\Psi(x,w) \mapsto L_w \Psi(x,\Omega) = w \nabla \Psi(x,w) = div(w\Psi(x,w)) = T^{-1} \frac{\partial T\Psi}{\partial s}$$

which has as its domain the set of functions for which $T\Psi$ is absolutely continuous on every compact subset of the closure of $\Pi_{t,w}$, i.e.,

$T\Psi(\circ,t,w)\in C(\overline{\Pi}_{t,w})$ and $T^{-1}\dfrac{\partial Tf(\circ,t,w)}{\partial s}\in L^2(\Omega\times S^2)$ .We will call this set $H^1(\Omega\times S^2)$ and note that it is dense in $L^2(\Omega\times S^2)$ and has a range $R(L)\subset L^2(\Omega\times S^2)$. Since we have a complete set of directions in S^2, for every Leakage operator L_w given by a direction w, for which the influx boundary is to be prescribed, there is an operator L_{-w} with reverse direction $-w$, and for which the influx boundary is $\Gamma^-_{-w}=\Gamma^+_w$, where $\Gamma^+_w=\{x\in\partial\Omega : n_x.w>0\}$.

The set of unbounded leakage operators

$$\{Lw \ ; w\in S^2\}$$

forms a symmetric system which has self-adjoint properties.

The operator for inversion of direction used before

$$U:L^2(\Omega\times S^2)\to L^2(\Omega\times S^2)$$
$$\Psi(x,w)\mapsto(\cup\Psi)(x,w)=\Psi(x,-w)$$

is unitary, self adjoint and isometric. We can use this operator to form the parity projection operator

$$P:L^2(\Omega\times S^2)\to L^2(\Omega\times S^2)$$
$$\Psi\mapsto(P\Psi)(x,w)=\frac{1}{2}[\Psi(x,w)+\Psi(x,-w)]$$

That is, $P=[I+U]/2$. P is bounded, linear, idem potent, self-adjoint and decomposes $L^2(\Omega\times S^2)$ into complementary linear manifolds: one is the even parity manifold which is its range and the other is the odd parity manifold which is its null subspace. As a classical lemma, we have that every function in $L^2(\Omega\times S^2)$ has a unique representation with one component in each one of these manifolds.

THE INVERSE PROBLEM

The stationary one velocity linear transport boundary value problem for a prescribed removal operator and source–intensity is a well-solved problem which has a unique inverse [1]. If we choose spatial positions on the outgoing surface

$$\Gamma^+=\{(x,w)\in\partial\Omega\times S^2; w.n_x>0\}$$

and make experimental measures related with the outgoing radiation intensity

$$f(x,w)=f^{out}(x,w),\ (x,w)\in\Gamma^+$$

we can use this excess of information to make inference about the removal operator. Such is the situation occurring in the classical transmission tomography, in which the scattering part of the operator is neglected and only the total extinction cross section is spatially reconstructed, and in the emission tomography in which we reconstruct the source q . In the present work we are dealing with the problem of parameter identification related to the removal operator R. The complete set of parameters related to the scattering and absorption process is to be spatially reconstructed. We have a situation which is analogous to that investigated by Calderon, [11], Sylvester and Ulmann[7] and we ask for the definition of a function

$$\Lambda_s:\Gamma^-\to\Gamma^+$$
$$\Psi^{in}\mapsto\Psi^{out}=\Lambda_s(\Psi^{in}) \qquad (4)$$

which will be called the influx to the outflux mapping for the transport equation.

Heuristic Counting of degrees of freedom

Let us estimate the minimum dimension for the inverse transport problem by formulating the problem in an n-dimensional space and making the counting of degrees of freedom. Since in this case f^{in} is a function from $\Re^{n-1}\times\Re^{n-1}\to\Re^+$, the domain of f^{in} has 2(n-1) degrees of freedom. The same can be said about f^{out}. Then we have a total of *4(n-1)* degrees of freedom. In the most general problem, we are interested in the reconstruction of the function s_a (s_t) from $\Re^n\to\Re^+$, whose domain has n degree of freedoms and the function s_s from $\Re^n\times[-1,1]\to\Re^+$ whose domain has *n+1* degrees. Since the number of degrees of freedom in the data must exceed that of the variables, we

must have $4(n+1) \geq 2n+1$ and $n \geq 2.5$ is the minimal dimension. In general we will not expect good results for $n=2$, unless we have the special situation of isotropic scattering, in which we have $4(n+1) \geq 2n$ .and $n \geq 2$. For $n=3$ we will not have problem of this nature. Another way to see the problem is to compare it to the case of transmission tomography, in which only the non scattered part of a collimated ray injected into the influx boundary is collected by the detector in the antipode position on the outflux boundary. If we suppose that for the same set of data used in this transmission tomography we also collect rays in other directions, we see that we have plenty of measures to be used in the reconstruction. This has been numerically exploited in [4] and the main problem that we have to solve there was the difference in the magnitudes of the transmitted and of the scattered rays. A numerical experiment shows that it can be of at least three orders of magnitude. If we don't take the appropriate care in separating the transmitted from the scattered radiation, the last numerically disappears as noise in the first. With this reasoning, we are motivated for the spatial two–dimensional reconstruction of the scattering cross section, at least in the isotropic case.

SECOND ORDER FORMULATION OF THE DIRECT PROBLEM

Since we are interested in the study of the outflux to influx mapping, we will briefly introduce the second order formulation for this problem. The unbounded second order differential operator is

$$L_{w,2} = -L_w R L_w$$

and has domain inside the domain of the operator L_w with the additional restriction that $div(w div(w\Psi)) \in L^2(\Omega x S^2)$. With this we can formulate the second order problem as:
Find

$$\Psi \in L^2(\Omega x S^2) \qquad (5)$$

such that

$$\left(-LRL + s_t(I-K)\right)\Psi = q \quad \text{on} \quad (\Omega x S^2)$$

$$(I + sign(w \cdot n_x) RL)\,\Psi = \Psi_s \quad \text{on} \quad \Gamma x S^2 \qquad (5a)$$

where $\Psi_s = \begin{cases} \Psi^{in}(x,w) \; if \; w \cdot n_k < 0 \\ \Psi^{in}(x,-w) \; if \; w \cdot n_k > 0 \end{cases}$ (5.b)

$L^2(\Gamma x S^2)$ is the symmetrized influx boundary condition. In the internal interfaces Γ_i we also have that $\Gamma\Psi(\cdot,t,w)$ is essentially continuous and $\Gamma RL\Psi(\cdot,t,w)$ is naturally continuous.

VARIATIONAL FORMULATION OF THE DIRECT PROBLEM

We can use the projection operator to decompose $L^2(\Omega x S^2)$ in a complementary pair of manifolds

$$L^2_+(\Omega x S^2) = \{f \in L^2(\Omega x S^2) : P(f) = f\}$$
$$L^2_-(\Omega x S^2) = \{f \in L^2(\Omega x S^2) : P(f) = 0\}$$

Every function in $L^2(\Omega x S^2)$ has unique representation

$$f = f^+ + f^-$$

with $f^+ \in L^2_+$ and $f^- \in L^2_-$. The variational formulation is posed in the subspace of L^2_+ and L^2_-, the spaces

$$H^1_+(\Omega x S^2) = \{f \in L^2_+(\Omega x S^2) : \text{div}(w f) \in L^2_-(\Omega x S^2)$$

$$H^1_-(\Omega x S^2) = \{f \in L^2_-(\Omega x S^2) : \text{div}(w f) \in L^2_+(\Omega x S^2)\}$$

The first order variational formulation has been proposed by Pitkaranta in [3] and is given by the functional problem:
Find $f^+ \in H^1_+$ and $f^- \in H^1_-$ extremes for the functional

$$F_1[f^+, f^-] = \frac{1}{2}\int_{S^2}\int_{\Omega}\{f^+ div(w f^-) -$$

$$f^- div(wf^+) + f^+ R[f^+] - f^- R[f^-]\}$$
$$+\frac{1}{2}\int_{S^2}\int_{\partial\Omega}\{|w\cdot n_x|f^{+2} - (w\cdot n_x)f^+ f^-)$$
$$-2|w\cdot n_x|f^+ f_s\} \qquad (6)$$

The first variation with respect even (odd) parity radiation intensity is a minimum (maximum) and gives the parity form of (3) for q=0,

$$w.\nabla f^- + G[f^+] = 0$$
$$w.\nabla f^+ + G[f^-] = 0 \quad \forall (x,w)\in \Omega\times S^2$$

with the respective boundary condition

$$f^+ - sign(w.n_x)f^- = f_s \quad \forall (x,w)\in \Gamma\times S^2$$

The variational principle for the second order theory can be obtained from the first order formulation [12] and is written as:

Find $f^+ \in H_+^1$ that minimizes the functional

$$F_2[f^+] = \frac{1}{2}\int_{S^2}\int_{\Omega}\{RL[f^+]L[f^+] + f^+ G[f^+]\}$$
$$+\frac{1}{2}\int_{S^2}\int_{\partial\Omega}|w\cdot n_x|\{f^{+2} - 2f^+\Psi_2\} \qquad (7)$$

where

$$H_+^1 = \{\Psi\in L_+^2(\Omega xS^2) :$$
$$\|\Psi\|_1^2 = (L\Psi, L\Psi) + (\Psi,\Psi) < \infty\}$$

we note that $\|\ \|_1^2$ is equivalent to the energy norm of the second order self adjoint application $A = -LRL + G$.

SOME COMMENTS ON THE INFLUX TO THE OUTFLUX MAPPING

We have defined in (4) the operator Λ_s which maps the influx to the outflux. Given that the one-speed transport equation in fixed as a model, this mapping is characterized by the operator G. The related question here is:

(1) if the knowledge of the mapping Λ_s is sufficient to characterize the coefficients in the definition of operator G

(2) if the product of $L^2(\Omega xS^2)$ solutions to the direct problem is dense in $L^1(\Omega xS^2)$, and also, if the product of gradient of $H^1(\Omega xS^2)$ solution the direct problem is dense in $L^1(\Omega xS^2)$

These are important mathematical question for the scattering tomography formulated in the context of the transport theory. Based on the conjecture that these questions have a positive answer we established a methodology for the reconstruction of these coefficients. The fact is that the transport equation is a symmetric system in the sense o Friedrichs and has adjoint properties that permit us to deduce an inverse integral equation from both the first order and the second order variational formulation. This can be done also directly from the weak form of the direct problem formulated in (2).

To deduce these inverse transport equations for the determination of the coefficients we must choose variations with satisfy adjoint problems with prescribed absorption and scattering coefficients. The reference cross sections are expected to be as close as possible to the reconstructed one. Since the transport equations are not self-adjoint, the product of solutions to be considered in (2) is the solution for direct problems with streaming operator L_w and L_{-w}, respectively.

THE SOURCE-DETECTOR METHODOLOGY

Since we are more interested in the parallelism between the transport model for scattering tomography and the electrical impedance tomography, we will proceed to the deduction of the inverse transport equation in the context of the second order theory

Let *ns* source problems be given with data

$\Psi_i^{in}(x,w)\in L^2(\Gamma^- xS^2)$, *i=1,ns*

Let *nd* detectors problems be given with data

$\Psi_i^{in}(x,w)\in L^2(\Gamma^- xS^2)$, *i=1,nd*

The detectors problems are taken with

$\Psi_j(x,w) = \Psi_j^{out}(x,-w)$ on $\Gamma^- xS^1$

and with no internal source, that is, *q=0*.

The two types of problems have the same variational formulation:

Find $\Psi \in H^1(\Omega x S^2)$ that minimizes

$$F_2[\Psi] = (RL\Psi, L\Psi) + (s_t(I-K)\Psi, \Psi) + \int_{\Gamma \times S^2} |w \cdot n_x| \{\Psi - 2\Psi_s\} \Psi d\Gamma d\Omega \quad (8a)$$

The first variation of this functional is given by

$$dF(\Psi, \partial\Psi) = (RL\Psi, Ld\Psi) + (G\Psi, d\Psi) + \int_{\Gamma \times S^2} |w \cdot n_x| \{\Psi d\Psi - 2\Psi_s d\Psi\} \quad (8b)$$

We write this functional for the *ns* source problems and for the *nd* detector problems with prescribed reference cross sections.

Since the boundary conditions in this formulation are natural and all variations are in the same space $H^1(\Omega \times S^2)$, we can take

$$d\Psi^i(x,w) = \Psi^j(x,-w)$$
$$d\Psi^j(x,w) = \Psi^i(x,w)$$

Adopting a detector solution as variation for a source problem and a direct problem solution as the variation to the detector problem, after some manipulations we obtain:

$$(\Delta R^{ij}[L\Psi^i], L\Psi^j) + (\Delta G^{ij}[f^i], f^j = \int_{\Gamma \times S^2} |w \cdot n_x| \{\Psi_s^j \Psi^i - \Psi_s^i \Psi^j\} \quad (9a)$$

where

$$\Delta R^{ij}[\cdot] = \sum_{lodd} \frac{2l+1}{4p} \left[\frac{1}{s_t^i(x) - s_{sl}^i(x)} - \frac{1}{s_t^j(x) - s_{sl}^j(x)} \right] \int_{S^2} P_l(w.w')[\cdot] \quad (9b)$$

$$\Delta G^{ij}[\cdot] = \sum_{leven} \frac{2l+1}{4p} \left[\frac{1}{s_t^i(x) - s_{sl}^i(x)} - \frac{1}{s_t^j(x) - s_{sl}^j(x)} \right] \int_{S^2} P_l(w.w')[\cdot] \quad (9c)$$

As has already been pointed out before the coefficient values with index i are to represent estimated cross section values and those with index j are reference values used in the detector equations. We can note that the right hand side of (9.a) is just the difference due to violation of the reciprocity relation and is a consequence of the use of different values for the cross sections of source and detector problems.

So we may call this term a defect from the reciprocity. The systems (9) represent the second order inverse transport equation. There we can see that the odd Legendre expression coefficients are multiplied by the product of gradients (the current, in neutronic terminology) and the even coefficients by the angular flux, which are a magnitude greater than the current.

THE INVERSE SECOND ORDER TRANSPORT PROBLEM

It consists of the study of various properties of the map

$$\Phi$$
$$s \rightarrow \Lambda_s$$

that associates the set cross sections coefficients with the influx to outflux mapping. These properties include the continuity, injectivity and range. Following Calderon´s approach for the inverse conductivity problem [11], the functional (8a) is some kind of measure of the power necessary to maintain the flux (potential) on the boundary. The polarization of this quadratic form is the bilinear form (8a) that has been used to derive the inverse transport equation.

For

$$Q_s(\Psi, \Psi_{in}) = \inf_{\Psi \in H^1(\Omega \times S^2)} F_2(s_t, s_s, \Psi, \Psi_{in}) \quad (10a)$$

we obtain

$$Q_s(\Psi_2, \Psi_{in}) = \frac{1}{2} \int_{\Gamma \times S^2} |wn_x| \Psi_s \Psi_2 d\Gamma dw \quad (10b)$$

where $\Psi_s(x,w)$ is given by (5b) and Ψ_2 is solution that minimizes the functional.

Noting that on Γ

$$\Psi_2(x,w) = \Psi_s(x,w) - sign(w.n_x) RL\Psi_2(x,w)$$

we find that

$$Q_s = \frac{1}{2}\int_{\Gamma\times S^2} |wn_x|(-\Psi_s^2 + \Psi_s RL\Psi_2)d\Gamma dw \tag{10c}$$

Results (10a) and (10c) are similar to the equations (0.2) and (0.4) found in the reference [13]. The functional Q_s is a quantity which can be determined by measurements at the boundary of Ω and as a consequence of (10b) is that the unique self-adjoint operator associated to the quadratic form Q_s is the influx to outflux mapping

$$\Phi$$
$$s \rightarrow \Lambda_s$$

CONCLUSIONS

The inverse problem for the transport problem in which the scattering process is not negligible can be analyzed in a context that is analogous to that of the inverse conductivity problem. Since in the transport problem we have variables defined in the phase space, we face a problem with more degrees of freedom.

The theory can be formulated in two levels of variational formulation, that is, a first order theory which is comparable with the usual transmission tomography when the scattering is neglected and a second order theory, which as has been point before, is comparable to the inverse problem for elliptic system of differential equation. This research is now conducted in a numerical experimental level, with utilizes synthetic data, and in a mathematical analysis level. A new concept, parallel to the Dirichlet-to-Neumann mapping, (sometimes refereed to as the Liouville-Steklov mapping), the influx to the outflux mapping has been presented.

ACKNOWLEDGEMENTS

The author acknowledges the financial support provided by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico, FAPERJ – Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro and CAPES – Comissão de Aperfeiçoamento de Pessoal de Nível Superior.

REFERENCES

1. V.S. Vladimirov, Mathematical Problems in the One-velocity theory of Particle Transport, *Trans V.A. Steklov Math. Inst.* **61** (1961); translated by Atomic Energy of Canada, Ltd AECL-1661, 1993.

2 H. G. Kaper, G. K. Lead and A. J. Lindeman,. Formulation of a Ritz-Galerkin Type Procedure for the Approximate Solution of the Neutron Transport Equation, *Journal of Mathematical Analysis and Applications* **50**, 42-65 (1975)

3. J. Pitkaranta, A Non Self-adjoint Variational Procedure for Finite Element Approximation of the Transport Equation, *Transport Theory and Statistical Physics*, **4**(1), 1-24, 1975.

4. . R. F. Carita Montero, N. C. Roberty, and A. J. Silva Neto, Reconstruction Of Absorption And Scattering Coefficients With A Discrete Ordinates Method Consistent With The Source-Detector System, *4icipe* 2002

5. . A. T. Kauati , A. J. Silva Neto, and N. C. Roberty, The Source-Detector Methodology For Applications With Anisotropic Scattering, *4icipe* 2002

6 A. T. Kauati, A. J. Silva Neto and N. C. Roberty, The Source-Detector Methodology For The Construction And Solution Of The One Dimensional Inverse Transport Equation, *Inverse Problems in Engineering*, **9** , 45-66 (2001).

7. J. Sylvester and G. Uhmann, *The Dirichlet To Neumann Map And Application*, in Inverse problem and Partial Differential Equation Chapter8.

8. F. Natterer, *The Mathematics of Computerized Tomography*, John Wiley and Son, 1986.

9. K. M. Case and P.F. Zweilfel, *Linear Transport Theory*, Addison-Wesley , 1967.

10. Victor Isakov, *Inverse problem for Partial Differential Equation*, Applied Mathematical Sciences 127, Springer, 1998.

11. A. P. Calderon. On an inverse boundary Value Problem, seminar on Numerical Analysis and its Application to Continuum Physics, *Soc. Brasileira de Matematica*, Rio de Janeiro,65-73, 1980.

12. S. Kaplan and J. A. Davis, Canonical and Involutory Transformation of The Variational Problems of Transport Theory, *Nucl. Sci. Eng.*, **28**,(1967).166-176.

13. J. Sylvester and G. Uhlmann, A Global Uniqueness Theorem for an Inverse Boundary Value Problem, *Annals of Mathematics.*, 1**25**,(1987).153-169.

SOLID MECHANICS AND GEOPHYSICS

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

AN INVERSE VIBRATION PROBLEM FOR SIMULTANEOUSLY ESTIMATING THE TIME-DEPENDENT STIFFNESS COEFFICIENTS

Cheng-Hung Huang
Department of Naval Architecture and Marine Engineering
National Cheng Kung University
Tainan 701, Taiwan, R.O.C.
chhuang@mail.ncku.edu.tw

ABSTRACT

The Conjugate Gradient Method (CGM) with adjoint equations are applied to an inverse force vibration problem to estimate the unknown time-dependent stiffness coefficients (or spring constants) simultaneously in a multiple-degree-of-freedom system by using the simulated measured system displacement.

The numerical experiments are performed to show the validity of the present algorithm by using different types of stiffness coefficients and measurement errors. Results show that the excellent estimations on the time-dependent spring constants can be obtained simultaneously with any arbitrary initial guesses within a very short CPU time at Pentium III-500 MHz PC.

1. INTRODUCTION

The objective of the direct solutions for the force vibration problems is to calculate the system displacements, velocity and acceleration with time t when the initial conditions, external forces and time-dependent stiffness coefficients and damping coefficients are specified. In contrast, the inverse vibration problems that we are going to discussed here involve the estimation of the time-dependent stiffness coefficients simultaneously from the knowledge of the simulated measured system displacement at different time t.

The inverse heat conduction problems in estimating thermal properties for both linear and non-linear problems can be found in the literatures. For instant, Huang and Ozisik [1] have used the direct integration method together with the Levenberg-Marquardt method in estimating the temperature-dependent thermal conductivity and heat capacity. Huang et al. [2] used a very powerful inverse algorithm, i.e. Conjugate Gradient Method (CGM), to estimate the temperature-dependent thermal conductivity. Huang and Yan [3] estimated the temperature-dependent thermal conductivity and heat capacity simultaneously by using the CGM. Recently, Huang and Chin [4] extended the CGM to a two-dimensional inverse problem in estimating unknown thermal conductivity for the non-homogeneous material.

Many papers regarding the estimation of the damping and stiffness matrices in for the inverse vibration problems can also be found in the literatures. For example, Gladwell [5] has solved the inverse vibration problems in determining constant stiffness matrices for undamped system modeled by tridiagonal matrices. Lancaster and Maroulas [6] have solved the inverse vibration problem in estimating constant damping and stiffness matrices by means of the spectral theory of matrix polynomials.

In all the above references the system damping and stiffness matrices are all assumed constant and independent of time. Recently, Huang [7] used CGM as well as adjoint equation in the inverse force vibration problems in estimating the time-dependent stiffness coefficients for a single-degree-of-freedom problem and obtained good estimation.

The purpose of the present study is to extend the previous work by Huang [7] to a multiple-degree-of-freedom inverse vibration problem in estimating simultaneously the time-dependent stiffness coefficients. It is obvious that this should be more difficult than what have been done by Huang [7] previously.

2. THE DIRECT PROBLEM

The initial displacement and velocity conditions of the damped force vibration system

are $x_i(0)=0.0$ and $dx_i(0)/dt = y_i(0) = 0.0$, respectively. When $t > 0$, the time-dependent external forces $f_i(t)$ and time-dependent damping coefficients $C_i(t)$ are given, moreover the time-dependent stiffness coefficients $K_i(t)$ are also assumed known.

The system under consideration here is shown in Figure 1 and the mathematical formulation of this multiple-degree-of-freedom problem is given by:

$$\frac{d^2x_1(t)}{dt^2} = -\frac{[C_1(t)+C_2(t)]}{M_1}\frac{dx_1(t)}{dt} + \frac{C_2(t)}{M_1}\frac{dx_2(t)}{dt} - \frac{[K_1(t)+K_2(t)]}{M_1}x_1(t) + \frac{K_2(t)}{M_1}x_2(t) + \frac{f_1(t)}{M_1}$$
$$,t > 0 \qquad (1\text{-}1)$$

$$\frac{d^2x_i(t)}{dt^2} = \frac{C_i(t)}{M_i}\frac{dx_{i-1}(t)}{dt} - \frac{[C_i(t)+C_{i+1}(t)]}{M_i}\frac{dx_i(t)}{dt} + \frac{C_{i+1}(t)}{M_i}\frac{dx_{i+1}(t)}{dt} + \frac{K_i(t)}{M_i}x_{i-1}(t) - \frac{[K_i(t)+K_{i+1}(t)]}{M_i}x_i(t) + \frac{K_{i+1}(t)}{M_i}x_{i+1}(t) + \frac{f_i(t)}{M_i}$$
$$, t > 0 , i = 2 \text{ to } I\text{-}1 \qquad (1\text{-}i)$$

$$\frac{d^2x_I(t)}{dt^2} = \frac{C_I(t)}{M_I}\frac{dx_{I-1}(t)}{dt} - \frac{C_I(t)}{M_I}\frac{dx_I(t)}{dt} + \frac{K_I(t)}{M_I}x_{I-1}(t) - \frac{K_I(t)}{M_I}x_I(t) + \frac{f_I(t)}{M_I}$$
$$, t > 0 \qquad (1\text{-}I)$$

with the initial conditions

$$x_i(0) = 0.0 \text{ and } dx_i(0)/dt = y_i(0) = 0.0 , \; i = 1 \text{ to } I \qquad (2)$$

Here M_i represents the mass of the subsystem. There exists no exact solution for equation (1) for any arbitrary function of $K_i(t)$, $C_i(t)$ and $f_i(t)$. For this reason the numerical solution with the technique of the fourth-order Runge-Kutta method will be applied to solve equation (1) by reducing it into $(2 \times I)$ coupled first-order ordinary differential equations as shown below:

$$\frac{dx_1(t)}{dt} = y_1(t) , \quad t > 0 \qquad (3\text{-}1a)$$

$$\frac{dy_1(t)}{dt} = -\frac{[C_1(t)+C_2(t)]}{M_1}y_1(t) + \frac{C_2(t)}{M_1}y_2(t) - \frac{[K_1(t)+K_2(t)]}{M_1}x_1(t) + \frac{K_2(t)}{M_1}x_2(t) + \frac{f_1(t)}{M_1}$$
$$,t > 0 \qquad (3\text{-}1b)$$

$$\frac{dx_i(t)}{dt} = y_i(t) , \quad t > 0, \; i = 2 \text{ to } I\text{-}1 \qquad (3\text{-}ia)$$

$$\frac{dy_i(t)}{dt} = \frac{C_i(t)}{M_i}y_{i-1}(t) - \frac{[C_i(t)+C_{i+1}(t)]}{M_i}y_i(t) + \frac{C_{i+1}(t)}{M_i}y_{i+1}(t) + \frac{K_i(t)}{M_i}x_{i-1}(t) - \frac{[K_i(t)+K_{i+1}(t)]}{M_i}x_i(t) + \frac{K_{i+1}(t)}{M_i}x_{i+1}(t) + \frac{f_i(t)}{M_i}$$
$$, t > 0 , i = 2 \text{ to } I\text{-}1 \qquad (3\text{-}ib)$$

$$\frac{dx_I(t)}{dt} = y_I(t) , \quad t > 0 \qquad (3\text{-}Ia)$$

$$\frac{dy_I(t)}{dt} = \frac{C_I(t)}{M_I}y_{I-1}(t) - \frac{C_I(t)}{M_I}y_I(t) + \frac{K_I(t)}{M_I}x_{I-1}(t) - \frac{K_I(t)}{M_I}x_I(t) + \frac{f_I(t)}{M_I} \qquad ,t > 0 \qquad (3\text{-}Ib)$$

The direct problem considered here is concerned with the determination of the system displacement $x_i(t)$ and velocity $y_i(t)$ when the initial conditions, the time-dependent external forces $f_i(t)$, the damping coefficients $C_i(t)$ and stiffness coefficients $K_i(t)$ are all given.

Here the fourth-order Runge-Kutta method is used to solve the system of equation (3).

3. THE INVERSE PROBLEM

For the inverse problem, the time-dependent stiffness coefficients $K_i(t)$ are regarded as being unknown, but everything else in equation (3) is known. In addition, system displacements measured at some appropriate time are considered available.

Let the measured system displacement with time be denoted by $X_i(t)$, here $t = 0$ to t_f, and t_f represents the final time of the measurements. Then the inverse problem can be stated as follows: by utilizing the above mentioned measured system displacement data, $X_i(t)$, to estimate the unknown time-dependent stiffness coefficients $K_i(t)$.

In the present study, we haven't used real displacement measurements, rather, we used the exact time-dependent stiffness coefficients $K_i(t)$ to generate the simulated values of $X_i(t)$, then try to retrieve the time-dependent stiffness coefficients by using $X_i(t)$ and initial guesses of stiffness coefficients $K^0_i(t)$.

The solution of the present inverse vibration problem is to be obtained in such a way that the following functional is minimized:

$$J[\mathbf{K}(t)] = \int_{t=0}^{t_f} \sum_{i=1}^{I} [x_i(t) - X_i(t)]^2 \, dt \qquad (4)$$

here, x(t) are the estimated or computed displacements at time t. These quantities are determined from the solution of the direct problem given previously by using an estimated $\hat{K}_i(t)$ for the exact $K_i(t)$. Here the hat " $\wedge$ " denotes the estimated quantities.

4. CONJUGATE GRADIENT METHOD FOR MINIMIZATION

The following iterative process based on the conjugate gradient method is now used for the estimation of time-dependent stiffness coefficients $K_i(t)$ by minimizing the functional $J[\mathbf{K}(t)]$

$$\hat{K}_i^{n+1}(t) = \hat{K}_i^{n}(t) - \beta_i^{n} P_i^{n}(t) \quad \text{for } i = 1 \text{ to } I \text{ and } n = 0,1,2,\cdots\cdots \tag{5a}$$

or in vector form

$$\hat{\mathbf{K}}^{n+1}(t) = \hat{\mathbf{K}}^{n}(t) - \boldsymbol{\beta}^{n}\mathbf{P}^{n}(t) \tag{5b}$$

where β_i^n are the search step sizes in going from iteration n to iteration n+1, and $P_i^n(t)$ are the directions of descent (i.e. search directions) given by

$$P_i^{n}(t) = J_i'^{n}(t) + \gamma_i^{n} P_i^{n-1}(t) \tag{6a}$$

or in vector form

$$\mathbf{P}^{n+1}(t) = \mathbf{J}'^{n}(t) - \boldsymbol{\gamma}^{n}\mathbf{P}^{n-1}(t) \tag{6b}$$

which are a conjugation of the gradient directions $J_i'^{n}(t)$ at iteration n and the directions of descent $P_i^{n-1}(t)$ at iteration n-1. The conjugate coefficients are determined from

$$\gamma_i^{n} = \int_{t=0}^{t_J} (J_i'^{n})^2 dt \Big/ \int_{t=0}^{t_J} (J_i'^{n-1})^2 dt$$

$$\text{with } \gamma_i^{0} = 0 \text{ and } i = 1 \text{ to } I \tag{7}$$

We note that when $\gamma_i^n = 0$ for any n, in equation (7), the directions of descent $P_i^n(t)$ become the gradient direction, i.e. the "Steepest descent" method is obtained.

To perform the iterations according to equation (5), we need to compute the step sizes β_i^n and the gradient of the functional $J_i'^{n}(t)$. In order to develop expressions for the determination of these two quantities, a "sensitivity problem" and an "adjoint problem" are constructed as described below.

4-1.SENSITIVITY PROBLEM AND SEARCH STEP SIZE

Since the problem involves I unknown time-dependent stiffness coefficients $\mathbf{K}(t) = K_i(t_n) = \{K_1(t_n),\ldots, K_I(t_n)\}$, n = 1 to N. In order to derive the sensitivity problem for each unknown function, we should perturb one unknown stiffness coefficient at a time.

It is assumed that when $K_i(t)$ undergoes a variation $\Delta K_i(t)\delta(i-j)$, where $\delta(\bullet)$ is the Dirac-delta function and j = 1 to I, $x_i(t)$ and $y_i(t)$ are perturbed by $\Delta x_{i,j}(t)$ and $\Delta y_{i,j}(t)$. Then replacing in the direct problem $K_i(t)$ by $K_i(t)+\Delta K_i(t)\delta(i-j)$, $x_i(t)$ by $x_i(t)+\Delta x_{i,j}(t)$ and $y_i(t)$ by $y_i(t)+\Delta y_{i,j}(t)$, subtracting from the resulting expressions the direct problem and neglecting the second-order terms, we obtained the following I sensitivity problems, (i.e. j = 1 to I), for the sensitivity functions $\Delta x_{i,j}(t)$ and $\Delta y_{i,j}(t)$.

$$\frac{d\Delta x_{1,j}(t)}{dt} = \Delta y_{1,j}(t), \quad t > 0 \tag{8-1a}$$

$$\frac{d\Delta y_{1,j}(t)}{dt} = -\frac{[C_1(t)+C_2(t)]}{M_1}\Delta y_{1,j}(t) + \frac{C_2(t)}{M_1}\Delta y_{2,j}(t) - \frac{[K_1(t)+K_2(t)]}{M_1}\Delta x_{1,j}(t) + \frac{K_2(t)}{M_1}\Delta x_{2,j}(t) - \frac{\Delta K_1(t)\delta(1-j)}{M_1}x_1 - \frac{\Delta K_2(t)\delta(2-j)}{M_1}x_1 + \frac{\Delta K_2(t)\delta(2-j)}{M_1}x_2 \quad ,t > 0 \tag{8-1b}$$

$$\frac{d\Delta x_{i,j}(t)}{dt} = \Delta y_{i,j}(t), \quad t > 0 \text{ , } i = 2 \text{ to } I\text{-}1 \tag{8-ia}$$

$$\frac{d\Delta y_{i,j}(t)}{dt} = \frac{C_i(t)}{M_i}\Delta y_{i-1,j}(t) - \frac{[C_i(t)+C_{i+1}(t)]}{M_i}\Delta y_{i,j}(t) + \frac{C_{i+1}(t)}{M_i}\Delta y_{i+1,j}(t) + \frac{K_i(t)}{M_i}\Delta x_{i-1,j}(t) - \frac{[K_i(t)+K_{i+1}(t)]}{M_i}\Delta x_{i,j}(t) + \frac{K_{i+1}(t)}{M_i}\Delta x_{i+1,j}(t) + \frac{\Delta K_i(t)\delta(i-j)}{M_i}x_{i-1} - \frac{\Delta K_i(t)\delta(i-j)}{M_i}x_i - \frac{\Delta K_{i+1}(t)\delta(i+1-j)}{M_i}x_i + \frac{\Delta K_{i+1}(t)\delta(i+1-j)}{M_i}x_{i+1} \quad , t > 0 \text{ , } i = 2 \text{ to } I\text{-}1 \tag{8-ib}$$

$$\frac{d\Delta x_{I,j}(t)}{dt} = \Delta y_{I,j}(t)\,, \quad t > 0 \qquad (8\text{-Ia})$$

$$\frac{d\Delta y_{I,j}(t)}{dt} = \frac{C_I(t)}{M_I}\Delta y_{I-1,j}(t) - \frac{C_I(t)}{M_I}\Delta y_{I,j}(t) + \frac{K_I(t)}{M_I}\Delta x_{I-1,j}(t) - \frac{K_I(t)}{M_I}\Delta x_{I,j}(t) + \frac{\Delta K_I(t)\delta(I-j)}{M_I}x_{I-1} - \frac{\Delta K_I(t)\delta(I-j)}{M_I}x_I \,, \; t > 0 \qquad (8\text{-Ib})$$

with the initial conditions

$$\Delta x_{i,j}(0) = 0.0 \text{ and } \Delta y_{i,j}(0) = 0.0\,, \quad i = 1 \text{ to I and } j = 1 \text{ to I} \qquad (9)$$

The technique of fourth-order Runge-Kutta method is used to solve these sensitivity problems.

The functional $J(\hat{\mathbf{K}}^{n+1})$ for iteration n+1 is obtained by rewriting equation (4) as

$$J[\hat{\mathbf{K}}^{n+1}(t)] = \int_{t=0}^{t_f}\sum_{i=1}^{I}[x_i(\hat{\mathbf{K}}^n - \beta^n\mathbf{P}^n) - X_i(t)]^2 dt \,, \; i = 1 \text{ to I} \qquad (10)$$

where we replaced $\hat{\mathbf{K}}^{n+1}(t)$ by the expression given by equation (5). If estimated displacements x_i is linearized by a Taylor expansion, equation (10) takes the form

$$J[\hat{\mathbf{K}}^{n+1}(t)] = \int_{t=0}^{t_f}\sum_{i=1}^{I}[x_i(\hat{\mathbf{K}}^n) - \sum_{j=1}^{I}\beta_j{}^n\Delta x_{i,j}(P_j{}^n) - X_i(t)]^2 dt \qquad (11)$$

where $x_i(\hat{\mathbf{K}}^n)$ are the solutions of the direct problem by using estimate $\hat{\mathbf{K}}^n(t)$ for exact $\mathbf{K}(t)$ at time t. The sensitivity functions $\Delta x_{i,j}(P_j{}^n)$ are taken as the solutions of problem (8) at time t by letting $\Delta\mathbf{K}(t) = \mathbf{P}^n(t)$ in equation (8) [7].

Equation (11) is differentiated with respect to $\beta_j{}^n$ and equating them equal to zero. Finally I equations can be solved for I step sizes $\beta_j{}^n$.

4-2. ADJOINT PROBLEM AND GRADIENT EQUATION

To obtain the adjoint problems, equations (3-ia) and (3-ib) are multiplied by the Lagrange multipliers (or adjoint functions) $\lambda_{i,j}(t)$ and $\psi_{i,j}(t)$, respectively. The resulting expression is integrated over the correspondent time domain, then the result is added to the right hand side of equation (4) to yield the following expression for the functional $J[\mathbf{K}(t)]$:

$$J[\mathbf{K}(t)] = \int_{t=0}^{t_f}\sum_{i=1}^{I}[x_i(t) - X_i(t)]^2 dt + \int_{t=0}^{t_f}\{\lambda_{1,j}(t)\times[Eq(3-1a)] + \Psi_{1,j}(t)\times[Eq(3-1b)] + .. + \lambda_{i,j}(t)\times[Eq(3-ia)] + \Psi_{i,j}(t)\times[Eq(3-ib)] + .. + \lambda_{I,j}(t)\times[Eq(3-Ia)] + \Psi_{I,j}(t)\times[Eq(3-Ib)]\}dt \qquad (12)$$

It is assumed that when $K_i(t)$ undergoes a variation $\Delta K_i(t)\delta(i-j)$, i = 1 to I, where $\delta(\bullet)$ is the Dirac-delta function and j = 1 to I, $x_i(t)$ and $y_i(t)$ are perturbed by $\Delta x_{i,j}(t)$ and $\Delta y_{i,j}(t)$. Then replacing in the direct problem $K_i(t)$ by $K_i(t)+\Delta K_i(t)\delta(i-j)$, $x_i(t)$ by $x_i(t)+\Delta x_{i,j}(t)$ and $y_i(t)$ by $y_i(t)+\Delta y_{i,j}(t)$, subtracting from the resulting expressions the direct problem and neglecting the second-order terms. We thus find

$$\Delta J_j[\mathbf{K}(t)] = \int_{t=0}^{t_f}\sum_{i=1}^{I}2[x_i(t) - X_i(t)]\Delta x_{i,j}dt + \int_{t=0}^{t_f}\{\lambda_{1,j}(t)\times[Eq(8-1a)] + \Psi_{1,j}(t)\times[Eq(8-1b)] + .. + \lambda_{i,j}(t)\times[Eq(8-ia)] + \Psi_{i,j}(t)\times[Eq(8-ib)] + .. + \lambda_{I,j}(t)\times[Eq(8-Ia)] + \Psi_{I,j}(t)\times[Eq(8-Ib)]\}dt \qquad (13)$$

In equation (13), the integral terms containing first derivative of time are integrated by parts; the initial conditions of the sensitivity problem are utilized. Finally we found that the equations for adjoint problems are identical for j = 1 to I. For this reason the subscript j can be neglected and we obtained the following adjoint problems $\lambda_i(t)$ and $\psi_i(t)$:

$$-\frac{d\lambda_1(t)}{dt} = -\frac{[K_1(t)+K_2(t)]}{M_1}\Psi_1 + \frac{K_2}{M_2}\Psi_2 - 2(x_1 - X_1) \quad , t > 0, \qquad (14\text{-}1a)$$

$$-\frac{d\Psi_1(t)}{dt} = -\frac{[C_1(t)+C_2(t)]}{M_1}\Psi_1(t) + \frac{C_2(t)}{M_2}\Psi_2(t) + \lambda_1(t) \quad ,t > 0 \qquad (14\text{-}1b)$$

$$-\frac{d\lambda_i(t)}{dt} = \frac{K_i(t)}{M_{i-1}}\Psi_{i-1}(t) - \frac{[K_i(t)+K_{i+1}(t)]}{M_i}\Psi_i(t) + \frac{K_{i+1}(t)}{M_{i+1}}\Psi_{i+1}(t) - 2(x_i - X_i)$$

$$, t > 0 , i = 2 \text{ to } I\text{-}1 \qquad (14\text{-}ia)$$

$$-\frac{d\Psi_i(t)}{dt} = \frac{C_i(t)}{M_{i-1}}\Psi_{i-1}(t) - \frac{[C_i(t)+C_{i+1}(t)]}{M_i}\Psi_i(t) + \frac{C_{i+1}(t)}{M_{i+1}}\Psi_{i+1}(t) + \lambda_i(t)$$

$$, t > 0 , i = 2 \text{ to } I\text{-}1 \qquad (14\text{-}ib)$$

$$-\frac{d\lambda_I(t)}{dt} = \frac{K_I(t)}{M_{I-1}}\Psi_{I-1}(t) - \frac{K_I(t)}{M_I}\Psi_I(t) - 2(x_I - X_I) \qquad (14\text{-}Ia)$$

$$-\frac{d\Psi_I(t)}{dt} = \frac{C_I(t)}{M_{I-1}}\Psi_{I-1}(t) - \frac{C_I(t)}{M_I}\Psi_I(t) + \lambda_I \quad , t > 0 \qquad (14\text{-}Ib)$$

with the final conditions

$$\lambda_i(t_f) = 0.0 \text{ and } \psi_i(t_f) = 0.0 , \; i = 1 \text{ to } I \qquad (15)$$

The adjoint problems are different from the standard initial value problems in that the final time conditions at time $t = t_f$ is specified instead of the customary initial condition. However, this problem can be transformed to an initial value problem by the transformation of the time variables as $\tau = t_f - t$. Then the standard techniques of fourth-order Runge-Kutta method can be used to solve the above adjoint problems.

Finally, the following integral terms are left

$$\Delta J_1 = \int_{t=0}^{t_f} -\frac{\Psi_1 x_1}{M_1}\Delta K_1 dt \qquad (16a)$$

$$\Delta J_i = \int_{t=0}^{t_f} [\frac{\Psi_{i-1}x_{i-1}}{M_{i-1}} - \frac{\Psi_{i-1}x_i}{M_{i-1}} - \frac{\Psi_i x_{i-1}}{M_i} + \frac{\Psi_i x_i}{M_i}]\Delta K_i dt \; ; \; i = 2 \text{ to } I \qquad (16b)$$

From definition [13], the functional increment can be presented as

$$\Delta J_i = \int_{t=0}^{t_f} (J_i' \Delta f_i)\, dt \quad ; i = 1 \text{ to } I \qquad (17)$$

A comparison of equations (16) and (17) leads to the following expression for the gradient of functional J_i' :

$$J_1'[K(t)] = \frac{\Psi_1 x_1}{M_1} \qquad (18a)$$

$$J_i'[K(t)] = (\frac{\Psi_{i-1}x_{i-1}}{M_{i-1}} - \frac{\Psi_{i-1}x_i}{M_{i-1}} - \frac{\Psi_i x_{i-1}}{M_i} + \frac{\Psi_i x_i}{M_i}) \; ; i = 2 \text{ to } I \qquad (18b)$$

We note that $J_i'[K(t)]$ is always equal to zero at $t = 0$ and t_f since $x_i(0) = 0.0$ and $\psi_i(t_f) = 0.0$, therefore the estimated values of $K_i(t)$ will deviate from exact values near both initial and final time. For this reason some estimated values of $K_i(t)$ near $t = 0$ and t_f should be discarded.

4-3. STOPPING CRITERION

If the problem contains no measurement errors, the traditional check condition is specified as

$$J[\hat{\mathbf{K}}^{n+1}(t)] < \varepsilon \qquad (19)$$

where ε is a small-specified number. However, the measured displacements may contain measurement errors. Therefore, we do not expect the functional equation (4) to be equal to zero at the final iteration step. Here we use the discrepancy principle as the stopping criterion, i.e. we assume that the residuals for the displacement and velocity may be approximated by

$$x_i(t) - X_i(t) \approx \sigma_i \; , i = 1 \text{ to } I \qquad (20)$$

where σ_i are the standard deviation of the displacement measurements, which are assumed to be a constant. Substituting equation (20) into equation (4), the following expression is obtained for stopping criteria ε:

$$\varepsilon = \sum_{i=1}^{I} (\sigma_i^2) t_f \qquad (21)$$

Then, the stopping criterion is given by equation (19) with ε determined from equation (21).

5. RESULTS AND DISCUSSIONS

The objective of this work is to show the validity of the CGM in estimating simultaneously the stiffness coefficients $K_i(t)$ in the inverse force vibration problems with no prior information on the functional form of the unknown quantities.

To illustrate the accuracy of the conjugate gradient method in predicting stiffness coefficients $K_i(t)$ in a damped vibration problem from the knowledge of transient displacement recordings, one specific example having different form of stiffness coefficients are considered here.

In order to compare the results for situations involving random measurement errors, we assume normally distributed uncorrelated errors with zero mean and constant standard deviation. The simulated inexact measurement displacement data $X_i(t)$ can be expressed as

$$X_i(t) = X_{i,exact}(t) + \omega_i(t)\sigma_I \qquad (22)$$

where $X_{i,exact}(t)$ are the solution of the direct vibration problem with an exact stiffness coefficients $K_i(t)$; σ_i are the standard deviation of the measured displacements and $\omega_i(t)$ are the random variables that are generated by subroutine DRNNOR of the IMSL [8] and will be within -2.576 to 2.576 for a 99% confidence bound.

One of the advantages of using the conjugate gradient method to solve the inverse problems is that the initial guesses of the unknown quantities can be chosen arbitrarily. In all the test cases considered here the initial guesses of $\hat{K}_i(t)$ is taken as $\hat{K}_{i\,initial}(t) = 0.0$.

We now present below the numerical experiments in determining $K_i(t)$ simultaneously by the inverse analysis using the CGM in a two-degree-of-freedom problem, i.e. I = 2. The initial conditions for displacement and velocity are both assumed zero, i.e. $x_i(0) = 0$ and $y_i(0) = 0$. Moreover, due to the singularity at t = 0 and t_f that was discussed previously, we thus neglect the first and last ten estimated values of stiffness coefficients in the present study

The parameters that used in the present test case are taken as:
$M_1 = 1.0$, $M_2 = 3.0$, $f_1(t) = 50.0$, $f_2(t) = 60.0$, $C_1(t) = 8.0$ and $C_2(t) = 5.0$.

Time interval is chosen as 36 and a time step $\Delta t = 0.3$ is used, therefore a total of 240 unknown discretized stiffness coefficients are to be determined in the present study. However we have discarded the first and last ten estimated values for $K_1(t)$ and $K_2(t)$, respectively, thus only 200 estimated values are reported here. The number of measured displacements for system 1 and 2 are both 120.

The unknown transient stiffness coefficients $K_1(t)$ and $K_2(t)$ are assumed as:

$$\begin{cases} K_1(t) = 18. - 5.0 \times \mathrm{SIN}(\dfrac{2\pi t}{36}) \\ K_2(t) = 8. + 3.0 \times \mathrm{COS}(\dfrac{2\pi t}{36}) \end{cases} ;\ \text{for } 0 < t \le 36 \qquad (23)$$

The inverse analysis is first performed by using the exact displacement measurements, i.e. assuming no measurement errors ($\sigma_1 = \sigma_2 = 0.0$). When the stopping criteria is set as $\varepsilon = 0.4$, after 78 iterations the inverse solutions are converged, J is calculated as 0.39 and CPU time at Pentium III-500 MHz PC is about 8 seconds. The exact and estimated stiffness coefficients $K_1(t)$ and $K_2(t)$ are shown in Figure 2 while Figure 3 shows the measured and estimated displacement, $X_i(t)$ and $x_i(t)$.

The average errors for the estimated stiffness coefficients and displacements are ERR1 = 1.01 % and ERR2 = 0.59 %, respectively, where the definition for ERR1 and ERR2 is given as

$$\mathrm{ERR1}\,\% = \left[\sum_{i=1}^{I} \sum_{n=10}^{N-10} \left| \frac{K_i(t_n) - \hat{K}_i(t_n)}{K_i(t_n)} \right| \right] \div [I \times (N-20)] \times 100\% \qquad (24a)$$

$$\mathrm{ERR2}\,\% = \left[\sum_{i=1}^{I} \sum_{n=10}^{N-10} \left| \frac{X_i(t_n) - x_i(t_n)}{X_i(t_n)} \right| \right] \div [I \times (N-20)] \times 100\% \qquad (24b)$$

Where n is the index for time and N represents the total number of discreted time. From those figures we concluded that the present algorithm has been applied successfully in the inverse vibration problem in estimating stiffness coefficients $K_i(t)$ since the estimated results are very accurate.

Next, let us discuss the influence of the measurement errors on the inverse solutions. When the measurement error for the displacements measured by sensors for subsystem

1 and 2 are taken as $\sigma_1 = 0.06$ (about 1 % of the average measured displacement for subsystem 1) and $\sigma_2 = 0.14$ (about 1 % of the average measured displacement for subsystem 2), then the stopping criteria ε can be calculated from equation (21). After 68 iterations (CPU time is about 7 seconds), the inverse solutions can be obtained and plotted in Figure 4 for the exact and estimated stiffness coefficients and in Figure 5 for the measured and estimated displacements. The average errors for the estimated stiffness coefficients and displacements are ERR1 = 2.73 % and ERR2 = 0.88 %, respectively.

Then, the measurement error is increased to $\sigma_1 = 0.3$ (about 5 % of the average measured displacement for subsystem 1) and $\sigma_2 = 0.7$ (about 5 % of the average measured displacement for subsystem 2). After 14 iterations (CPU time is about 2 seconds), the inverse solutions can be obtained. In Figure 6 the exact and estimated stiffness coefficients are shown. The average errors for the estimated external forces and displacements are ERR1 = 5.09 % and ERR2 = 4.61 %, respectively.

From the above Figures and data we learned that reliable inverse solutions can still be obtained when the large measurement errors are considered

6. CONCLUSIONS

The Conjugate Gradient Method (CGM) was successfully applied for the solution of the inverse force vibration problem in a multiple-degree-of-freedom system to determine simultaneously the unknown transient stiffness coefficients by utilizing simulated displacement measurements. Several test cases involving different system parameters, measurement errors and stiffness coefficients were considered. The results show that the inverse solutions obtained by CGM are still reliable as the measurement errors are increased. Moreover the CPU time needed in the inverse calculations is very short and the initial guesses for external forces can be arbitrarily chosen as zero.

ACKNOWLEDGMENT

This work was supported in part through the National Science Council, R. O. C., Grant number, NSC-89-2611-E-006-058.

REFERENCES

1. C. H. Huang and M. N. Ozisik, A Direct Integration Approach for Simultaneously Estimating Temperature Dependent Thermal Conductivity and Heat Capacity, Numerical Heat Transfer, Part A, Vol. 20, No. 1, pp.95-110, (1991).
2. C. H. Huang , J. Y. Yan and H. T. Chen, The Function Estimation in Predicting Temperature Dependent Thermal Conductivity without Internal Measurements, AIAA, J. Thermophysics and Heat Transfer, Vol. 9, No. 4, pp. 667-673, October-December, (1995).
3. C. H. Huang and J. Y. Yan, An Inverse Problem in Simultaneously Measuring Temperature Dependent Thermal Conductivity and Heat Capacity, Int. J. Heat and Mass Transfer, Vol. 38. No. 18, pp. 3433-3441, (1995).
4. C. H. Huang and S. C. Chin, A two-dimensional inverse problem in imaging the thermal conductivity of a non-homogeneous medium, Int. J. Heat and Mass Transfer, Vol. 43, No. 22, pp. 4061-4071, (2000).
5. G. M. L. Gladwell, Inverse Problem in Vibration, The Netherlands: Kluwer Academic Publishers, (1986).
6. P. Lancaster and J. Maroulas, Inverse Eigenvalue Problems for Damped Vibrating Systems, Journal of Mathematical Analysis and Applications, Vol. 123, No. 1, pp. 238-261, (1987).
7. C. H Huang, A nonlinear inverse vibration problem of estimating the time-dependent stiffness coefficients by conjugate gradient method, Int. J. Numerical Methods in Engineering, Vol. 50, pp. 1545-1558, (2001).
8. IMSL Library Edition 10.0, User's Manual: Math Library Version 1.0, IMSL, Houston, TX, (1987).

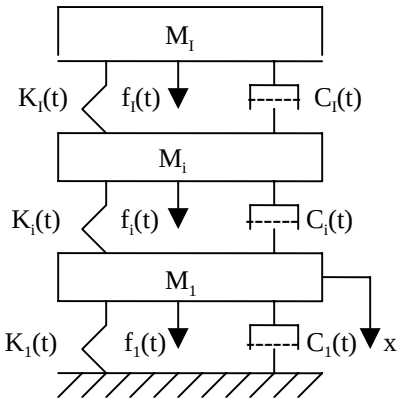


Figure 1. A multiple-degree-of-freedom nonlinear force vibration system for the present study.

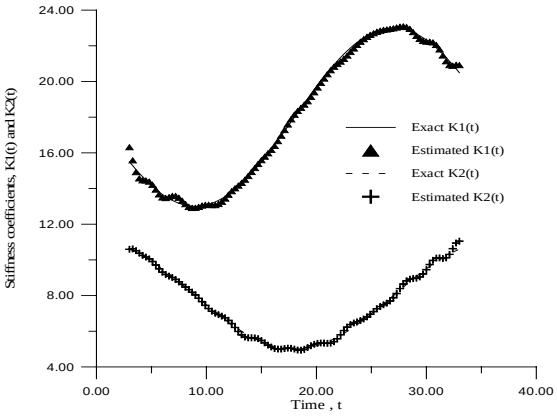


Figure 2. The exact and estimated stiffness coefficients using displacement measurements with $\sigma_1 = \sigma_2 = 0.0$.

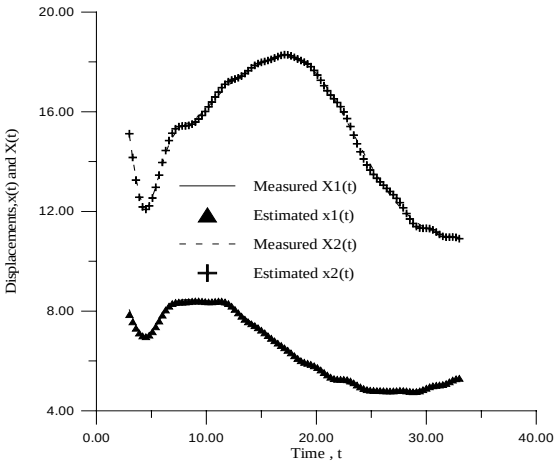


Figure 3. The measured and estimated displacements with $\sigma_1 = \sigma_2 = 0.0$

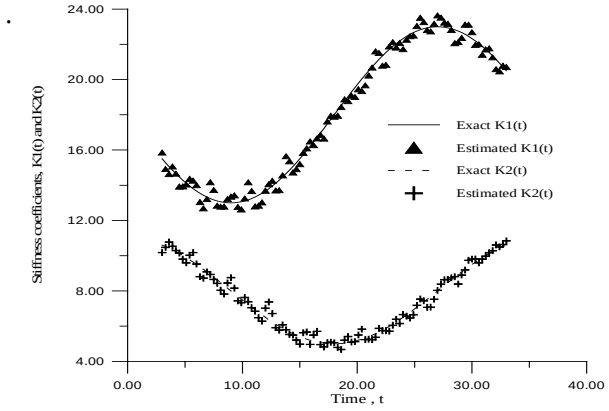


Figure 4. The exact and estimated stiffness coefficients using displacement measurements with $\sigma_1 = 0.06$ and $\sigma_2 = 0.14$.

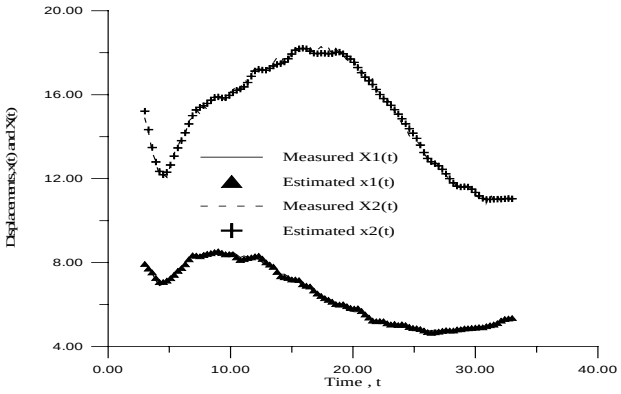


Figure 5. The measured and estimated displacements with $\sigma_1 = 0.06$ and $\sigma_2 = 0.14$.

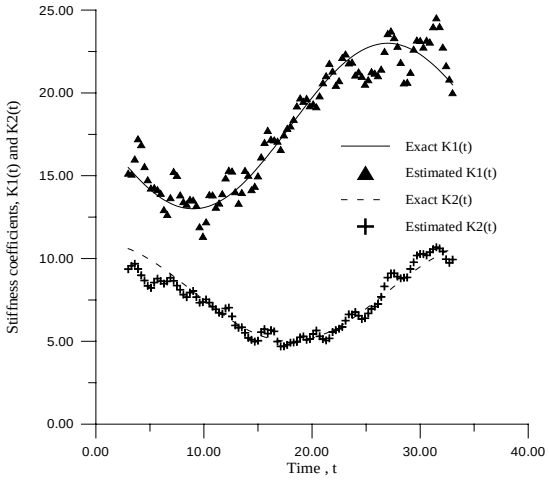


Figure 6. The exact and estimated stiffness coefficients using displacement measurements with $\sigma_1 = 0.3$ and $\sigma_2 = 0.7$.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

RESPONSE SURFACE METHOD FOR SOLUTION OF STRUCTURAL IDENTIFICATION PROBLEMS

Rolands Rikards and Janis Auzins
Institute of Materials and Structures
Riga Technical University, Riga, Latvia
rikards@latnet.lv

ABSTRACT

The paper is focused on the application of the response surface method (RSM) for the solution of structural identification problems. The approximating functions are obtained from the data of deterministic numerical experiment. The numerical experiment is performed in the sample points of experimental design. A minimal mean squared distance Latin hypercube (MMSDLH) design is used in the present paper. A local approximation method is employed for building the response surfaces. An example of the application of the response surface method and experimental design for the identification of elastic properties of a laminated composite material is discussed. The elastic properties of carbon/epoxy laminate are determined employing experimentally measured eigenfrequencies of composite plates. The identification functional represents differences between experimentally measured and numerically calculated frequencies, which are dependent on variables to be identified. The parameters to be identified are the five elastic constants of the material. The elastic constants identified from the vibration test have been compared with the values obtained from an independent static test. A good agreement of the results is observed.

INTRODUCTION

In structural optimization and identification, some problems require too much computational time when conventional methods of minimization are used. For example, it takes several hours of computer time for one variant of the finite element solution to be calculated. For complex optimum design problems it is necessary to perform calculations of several thousand variants. Similarly, the solution of some identification problems can also require large computational efforts. In order to reduce computational efforts, methods based on approximation concepts can be used. Nowadays these methods take a dominant position in structural optimization [1]. Approximation methods also are employed to solve identification problems [2]. The development of approximation functions has become a separate problem in optimum structural design. The approximating models can be built in different ways. Empirical model building theory is discussed in [3]. To construct a more general model of the original function, the method of experimental design [4,5] can be employed together with approximate model building [6-8]. A simplified model, called “metamodel”, is built using the results of a numerical experiment in the points of experimental design. Response analysis using the simplified model is computationally much less expensive than a solution using the original model. Although there is a wide literature about experimental designs and the building of approximating functions, it should be noted that there are some special features present in the experimental design that are not present in the physical experiment. The main features are as follows.

1) The results obtained in the numerical experiment are deterministic and without statistical errors. Repetition of the results is 100%. This means that there is no statistical dispersion of the model parameters. However, computer models produce numerical noise as a result of the incomplete convergence of iterative processes, round-off errors, and the discrete representation of continuous physical phenomena when a different number of calculation steps or a different finite element grid is generated [9]. In deterministic computer experiments, replication at a sample point is meaningless, therefore the points should be chosen to fill the design space.

2) The mathematical model of the object is unknown, i.e., the form of the regression equation is not known. Therefore, well-known criteria for experimental design optimality, for example, *D*-optimality, cannot be used. Such criteria can be used only in the case when the form of the regression equation is known.

There is a wide literature about the different methods of experimental design. Among the methods, the space filling designs can specifically be emphasized. The first space filling design for a computer experiment was proposed in [4]. In this work, the designs in which the number of levels for each variable is equal to the total number of runs were first proposed. In [4], the space filling criterion based on a function similar to potential energy of gravity was first used. Later, the same kind of experimental designs was proposed as a Monte Carlo integration technique by McKay et al. [10], and the name "Latin hypercube samplings" was introduced. Numerous space filling experimental designs have been developed in an effort to provide more efficient and effective means for sampling deterministic computer experiments based on Latin hypercubes. Different space filling criteria for Latin hypercube designs was proposed by many authors: Maximin Latin hypercubes [11], Minimal Integrated Mean Square Error designs [12], Orthogonal array-based Latin hypercube designs [13], Orthogonal Latin hypercubes [14], Integrated Mean Square Error (IMSE) optimal Latin hypercubes [15].

Employing the approach of experimental design and approximation proposed by Eglais [4,6], good results for the problems based on numerical experiment can be obtained. This approach based on global approximation was used in [2] for solution of optimal design and identification problems. However, sometimes the results of the approximation are not satisfactory. Therefore, in the present paper a minimal Mean Square Distance Latin hypercube (MMSDLH) design and local approximation method are employed to solve an identification problem similar to that described in [2]. Thus the accuracy of the solution can be improved.

In the past few years, the so-called non-parametric approximation methods have been widely used for the design and analysis of computer experiments: local polynomial approximation [16,17], Kriging [18]. Finally, other statistical techniques such as Multivariate Adaptive Regression Splines [19] and Radial Basis Functions [20-22] are beginning to draw the attention of many researchers. However, these methods are computationally expensive not only for metamodel building, but also in the case of using the metamodels for prediction. In the present paper, a local approximation method with weight functions is employed for the solution of the identification problem considered.

EXPERIMENTAL FREQUENCIES

Experiments have been performed on unidirectional carbon/epoxy laminate (see Figure 1). Plates were tested for vibrations in order to measure eigenfrequencies and corresponding modes. Experiments were performed with free-free boundary conditions on all edges of the plate, in order to exclude the influence of boundary conditions on the results of the identification. The plate dimensions are as follows: a=b=207.5 mm; h=2.0 mm. Density of the material ρ=1535 kg/m^3. Experimental eigenfrequencies $f_i^{\exp}$ are presented in the third column of Table 1. Since not all of the frequencies were observed experimentally, frequencies were ranged according to the finite element solution. In the second column, the frequencies f_i^{FEM} obtained by FEM using the identified elastic constants (see section Results and verification) are presented. Other quantities presented in Table 1 are explained in the section Results and verification (see below).

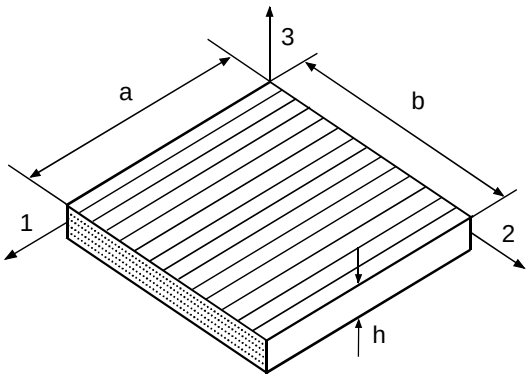


FIGURE 1 Laminated composite plate.

It can be seen that in the range of the first 17 numerical frequencies only 12 experimental frequencies were observed. It should be noted that frequencies are identified through mode shapes and for numerical and experimental frequencies

the same modes were observed. These experimental frequencies were employed for identification.

TABLE 1 Experimental frequencies.

No.	f_i^{FEM}, Hz	$f_i^{\exp}$, Hz	Δ_i, %	Δf_i, %
1	97.9	97	+0.92	11.15
2	124.3	123	+1.06	6.94
3	235.7	237	-0.55	9.39
4	342.3	341	+0.38	6.89
5	455.3	458	-0.79	8.12
6	502.2	502	+0.04	1.43
7	539.4	541	-0.29	2.69
8	651.1	653	-0.29	5.07
9	676.5	-	-	-
10	779.7	-	-	-
11	860.8	-	-	-
12	1113	-	-	-
13	1169	1168	+0.08	7.96
14	1215	-	-	-
15	1376	1381	-0.36	1.52
16	1407	1413	-0.42	2.05
17	1503	1512	-0.59	3.04

IDENTIFICATION FUNCTIONAL AND APPROXIMATION

The parameters x to be identified are five elastic constants of the transversally isotropic material of the plate

$$x = (x_1, x_2, ..., x_5) = (E_1, E_2, G_{12}, G_{23}, \nu_{12}) \quad (1)$$

Here E_1 and E_2 are Young's modulus in the fiber and transverse direction, respectively, G_{12} is the in-plane shear modulus, G_{23} is the transverse shear modulus and ν_{12} is Poisson's ratio. Directions of the material axes, which are also the plate axes, are denoted 1-2-3, where 1 is the fiber direction and 2, 3 are the transverse directions.

In [2] it was assumed that the functional to be minimized describes the deviation between the experimentally measured $f_i^{\exp}$ and the numerically calculated $f_i(x)$ frequencies

$$\Phi(x) = \sum_{i=1}^{I} k_i \frac{[f_i^{\exp} - f_i(x)]^2}{(f_i^{\exp})^2} = \sum_i k_i \varepsilon_i^2 \quad (2)$$

Here ε_i is the relative discrepancy or residual and k_i are the weighting coefficients for the selected frequencies. In (2) the integer I is the number of all frequencies used in the analysis. It is possible to assign non-negative weights to each residual. For simplicity, only unity values are used. The estimation can be based on any set of frequencies by assigning weights of zeros and ones as appropriate.

The numerical frequencies $f_i(x)$ are functions of elastic constants. These functions are obtained as approximation of the finite element solution, which is performed in the sample points of the experimental design. The frequencies and corresponding vibration modes (eigenvectors) are obtained by solving an eigenvalue equation

$$\mathbf{K}(x)\mathbf{U} - \lambda_i(x)\mathbf{M}\mathbf{U} = 0 \quad (3)$$

Here $\mathbf{K}$ is the plate stiffness matrix, which depends on x, $\mathbf{M}$ is the mass matrix, $\mathbf{U}$ is the displacement vector (eigenvector) and $\lambda_i = \omega_i^2$ is the eigenvalue and $\omega_i = 2\pi f_i$ is the circular frequency (rad/s).

For identification, the functional (2) can be used, but it is more appropriate to employ the eigenvalues instead of frequencies

$$\Phi(x) = \sum_{i=1}^{I} k_i \frac{[(2\pi f_i^{\exp})^2 - \lambda_i(x)]^2}{[(2\pi f_i^{\exp})^2]^2} \quad (4)$$

The functionals (2) and (4) were employed for identification in [2], where, instead of the original functions $\lambda_i(x)$, the approximating functions $\hat{\lambda}_i(x)$ were used. Thus, in [2] the approximations were performed for each frequency. Employing the same functional (4) procedure of identification can be modified so that the approximation is performed not for each frequency but for the whole functional $\Phi(x)$. Thus, the function to be approximated and minimized is as follows.

$$\Phi(x^j)=\left(\sum_{i=1}^{I}k_i\left(\frac{\left(2\pi f_i^{\exp}\right)^2-\lambda_i(x^j)}{\left(2\pi f_i^{\exp}\right)^2}\right)^2\right)^p \qquad (5)$$

Here $\lambda_i(x^j)$ is the *i*-th eigenvalue calculated by the finite element method in a sample point $x^j=(E_1^j,E_2^j,G_{13}^j,G_{23}^j,\nu_{12}^j)$ of a 5-dimensional space of identification parameters, *j* is the number of the sample point (run) in the experimental design (*j*=1,2,...*N*), *N* is the total number of sample points (number of runs) in the experimental design (see below), *p*=1, 1/2,1/4 or 1/8. The value of *p* is chosen to improve the quality of approximation. The best results were obtained (see below) with *p*=1/2 and 1. Note that hereafter the upper index for the variable *x* denotes the number of the point in the experimental design, but the lower index denotes the component of variable *x*.

The functional Φ is minimized employing local approximation:

$$\hat{\Phi}(x)=\beta_0+\sum_{i=1}^{5}\beta_i x_i+\sum_{i=1}^{5}\sum_{k=i}^{5}\beta_{ik}x_i x_k \qquad (6)$$

Here the lower index is used for the component of variable *x* (*i*=1,2,.., 5), but the upper index (see expression (5)) is employed to indicate the number of sample point in the experimental design (*j*=1,2,...*N*). In approximation (6) coefficients are calculated by

$$\beta=\arg\min_{\beta}\sum_{j\in N_X}w(x-x^j)\times\left(\frac{\Phi_j-\beta_0-\sum_{i=1}^{5}\beta_i x_{ji}-\sum_{i=1}^{5}\sum_{k=i}^{5}\beta_{ik}x_i^j x_k^j}{\Phi_j}\right)^2 \qquad (7)$$

where $\beta_0,\beta_i,\beta_{ik}$ are coefficients of the local quadratic approximation (dependent on *x*), N_X is the set of numbers of the nearest neighbors of the point *x*. In the case when the Gaussian weight function $w(x-x^j)=\exp(-G\|x-x^j\|^2)$ is used in (7), all points of experimental design are considered as neighbors $N_X=\{1,2,...,N\}$. Here $\|x-x^j\|$ is the Euclidean distance between *x* and x^j, *G* is the coefficient of the Gaussian function. If *G*=0, then the conventional least squares method is obtained (without weighting coefficients and without division by Φ_j in (7)). Usually *G*=0.75 was used.

MINIMAL MEAN SQUARE DISTANCE DESIGN

For the computer experiment, the Minimal Mean Squared Distance (MMSD) experimental designs were employed. These designs were proposed in [23]. The MMSD designs are space filling designs that give minimal Mean Squared Distance (MSD) between the mesh points in design space R^m and the nearest point from experimental design *D*

$$MSD=\sqrt{\left(\frac{1}{n}\right)\sum_{v=1}^{n}\min_{u=1,...,N}\left[\sum_{i=1}^{m}\left(w_i^v-x_i^u\right)^2\right]} \qquad (8)$$

where w_v are points from a large sample in design space R^m (*v*=1,..., *n*), *N* is the number of points of the experimental design and *n* is the number of mesh points. Approximately *n*=1000000 equidistant mesh points for low dimensions (*m*=2,3) are employed and a 100000-point Latin hypercube sample for large-scale designs (*m*>3) is used. These designs give points uniformly distributed in the design space and tend to minimize the expected mean squared error of the local quadratic approximation [23]. Fang and Wang [24] introduced a similar criterion, named Mean Squared Error. In [23], a quick search algorithm for the minimization of the MSD criterion for Latin hypercube designs in the unit cube $[-1,1]^m$ as well as for designs with unconstrained level values and numbers in unit cubes or *m*-dimensional spherical regions was proposed.

For the purpose of comparing with other designs, the distances and other characteristics of experimental designs are computed after the designs are scaled into the unit cube $[0, 1]^m$, although the designs are mostly constructed in an *m*-dimensional cube $[-1, 1]^m$.

For comparing with other space filling designs, four additional criterions have been used.

1. Eglais' criterion [4], later proposed also by Morris and Mitchell in a more general form [25]

$$\Phi_2 = \left[\sum_{u=1}^{N-1} \sum_{v=u+1}^{N} \frac{1}{\sum_{i=1}^{m} \left(x_i^u - x_i^v \right)^2} \right]^{1/2} \qquad (9)$$

2. The MINDIST criterion, which seeks to maximize the minimum distance between any pair of points in the data collection plan [11]

$$\text{MINDIST} = \min_{u,v=1,\ldots,N} \sum_{i=1}^{m} \left(x_i^u - x_i^v \right)^2 \qquad (10)$$

3. The entropy criterion first proposed by Shewry and Wynn (1987) [26] and then adopted by Currin et al. (1991) [27]. The entropy criterion for designs in unit cube $[0,1]^m$ is equivalent to the minimization of $-\log|\mathbf{C}|$, where $\mathbf{C}$ is the $N \times N$ covariance matrix of the design with elements

$$c_{ij} = \exp\left\{ -\Theta \sum_{k=1}^{m} \left| x_k^i - x_k^j \right|^q \right\}, \; 0 < q \le 2 \quad (11)$$

where $i, j=1,\ldots,N$. Throughout this paper the value $q = 2$ is selected thus that the correlation between two points is a function of their Euclidean distance L_2, and Θ is set equal to 2.

4. The discrepancy criterion, which averages the squared difference in the cumulative density function [28]

$$(D_C)^2 = \left(\frac{13}{12} \right)^m - \frac{2}{N} \sum_{u=1}^{N} \prod_{i=1}^{m} \left[1 + 0.5 \left| x_i^u - 0.5 \right| - 0.5 \left| x_i^u - 0.5 \right|^2 \right] + \frac{1}{N^2} \sum_{u=1}^{N} \sum_{v=1}^{N} \prod_{i=1}^{m} \left[1 + 0.5 \left(\left| x_i^u - 0.5 \right| + \left| x_i^v - 0.5 \right| - \left| x_i^u - x_i^v \right| \right) \right]$$

Table 2 shows a comparison of three 16-run designs of 7 variables for all five criterions. MMSDLH stands for Minimal MSD Latin hypercube design; ULH stands for Uniform Design Based On Centered L_2 Discrepancy U_n (n^5) [29] and MBLH is a Minimum Bias Latin hypercube [28]. It can be seen that the MMSDLH plan performs better than the others according to all five criterions.

For the identification problem formulated in the present paper, an MMSDLH-type design with 101 runs and 5 factors is employed. For this design the values of criterions are as follows: MSD=0.2051, Φ_2 =89.0740, MINDIST=0.3808, Entropy=69.7806, D_C =0.0453.

TABLE 2 Comparison of 16-point designs for 7 variables

Design	MMSDLH	ULH	MBLH
MSD	0.3942	0.4006	0.3947
Φ_2	9.5196	9.5449	9.5281
MINDIST	0.8869	0.8353	0.8000
Entropy	0.2900	0.3123	0.3147
D_C	0.2464	0.2289	0.2468

MINIMIZATION

Unlike the parametric quadratic approximation commonly used in the response surface method, the minimization of a locally approximated function is more difficult. Generally, any method of non-linear programming can be used. However, using the derivatives is not appropriate because the approximating function cannot be smooth enough and may have a lot of local extremes. Two methods are employed in order to obtain a global minimum of the locally approximated function of interest.

The first method is iterations. A randomly selected point in the design space is taken as a starting point. Subsequently, a local approximation is built in this point and the coefficients β are found according to (7). Then the minimum point of the approximating function is calculated with fixed values of coefficients β. This is a simple problem, which requires the solution of a system of only five linear algebraic equations. Afterwards in this new point a local approximation is built and the search is continued. We should be convinced that the true minimum

was found. In the case when the process converges (and converges to the same point from all starting points), there is a high probability that the actual minimum was found. From experience it can be concluded that in this case the physical parameters of the plate are correctly identified.

Unfortunately, two alternative cases are found to be more common. First, the process can converge to a point outside of the region in which the experimental design was planned. In this case the center of the experimental design (for FEM calculations) should be moved or the bounds should be shifted.

In the worst case, when the iterative procedure diverges or gives a lot of local extremes, the second method, a global search, is used [30]. Approximately 100000 points from the randomly selected Latin hypercube type sample are tested and the best point is selected. Then the search domain is reduced around this point and a new random search is performed until an acceptable accuracy of the extreme values is obtained.

This is a computationally more expensive way than the iterative search, since one calculation of the approximating function needs to solve the system of 21 linear equations (the Cholesky decomposition method has been used). The entire process requires about one to two minutes of calculation time on a Pentium 800 MHz processor, but compared with the time of the FEM simulation this time is negligible.

After the minimum of the approximating function of interest is found, N_a confirmation points near the optimal values should be calculated to verify the accuracy of identification. These points can be used as additional points and optimization may be recalculated employing $N+N_a$ design points in order to improve the accuracy of the solution. Note that when the optimum of the locally approximated function is found, the true value of the function is verified by FEM in any case.

RESULTS AND VERIFICATION

To build the local approximations, an MMSDLH type design with N=101 sample points in five dimensions is used. These sample points are distributed in the domain of interest, which is formed by the lower and upper bounds of variables. The initial guess values of these bounds can be chosen employing the elastic constants of a similar material. If the identified values are outside of the region, the bounds should be shifted and the procedure of identification should be repeated. Thus, the domain of interest is corrected in few stages of identification. For the present example, in the first stage the domain of interest was chosen as follows

$$\begin{aligned} 168 &\le E_1 \le 174 \\ 9.5 &\le E_2 \le 11.5 \\ 5.2 &\le G_{12} \le 7.2 \\ 4 &\le G_{23} \le 8 \\ 0.2 &\le \nu_{12} \le 0.45 \end{aligned} \tag{12}$$

Here the Young's and shear modulus are given in GPa, but Poisson's ratio is a non-dimensional quantity. In sample points (j=1,2,…,N), the equation (3) was solved and eigenvalues $\lambda_i(x^j)$ were obtained. These eigenvalues are treated as original function. Approximations $\hat{\Phi}(x)$ of the original function (5) were obtained using the local approximation method described above.

Twelve experimentally measured frequencies, which are presented in Table 1, can be used in identification in any combination. First, all 12 experimental frequencies were used in identification by minimizing the functional (5). Then, only the first six frequencies were employed in identification. In this case of minimizing the functional (5), the following elastic constants were obtained

$$x^* = (170.7, 10.4, 6.2, 5.6, 0.34) \tag{13}$$

Practically the same results were obtained employing all 12 experimental frequencies in minimization of the functional (5). It should be noted that, employing the global approximations, the results (13) for the first four constants are approximately the same. The exception is Poisson's ratio, which can be reliably determined only by using the local approximations.

The verification of the results was performed by calculating with FEM the original function in the point of optimum (13). Then the numerical values were compared with the experimental frequencies. Residuals were calculated by the expression

$$\Delta_i = \frac{f_i^{FEM}(x^*) - f_i^{\exp}}{f_i^{\exp}} 100 \tag{14}$$

The results are shown in Table 1. It can be seen that the differences between the experimental and numerical frequencies are very small. Mostly the residuals do not exceed 1%, even for those six frequencies which were not used in identification (frequencies i=7,8,13,15,16,17). However, since higher frequencies are less sensitive to elastic constants that lower frequencies, in addition to the residuals Δ_i the range of each frequency in the space of the experimental design should be compared. In the last column of Table 1 the relative amplitude of each frequency Δf_i in the experimental design space is presented. The relative amplitude is calculated by the expression

$$\Delta f_i = \frac{f_i^{\max} - f_i^{\min}}{f_i^{\exp}} 100 \qquad (15)$$

Here

$$f_i^{\max} = \max_{x^j} f_i(x^j), \quad f_i^{\min} = \min_{x^j} f_i(x^j)$$

In Table 1 it can be seen that lower frequencies are more sensitive to the elastic constants. However, the amplitude in the design space for the 8th and 13th frequency is also considerable.

TABLE 3 Comparison of elastic constants obtained from vibration and static tests.

Elastic constant	Vibration test	Static test		
		RTU	IAI	DLR
E_1, GPa	170.7	176 (143)	165 (175)	192 (147)
E_2, GPa	10.4	8.9 (9.6)	9.2 (11.8)	10.6 (9.7)
G_{12}, GPa	6.2	5.2	5.4	6.1
G_{23}, GPa	5.6	-	-	-
ν_{12}, GPa	0.34	0.34	-	0.31 (0.34)

In order to validate results, it is necessary to compare the properties obtained from the vibration tests through identification with those obtained from an independent test. Conventional static test was selected as the independent test. Static tests were performed according to ASTM guidelines (RTU-Riga Technical University and IAI-Israel Aircraft Industry LTD) and DIN standards (DLR-German Aerospace Center). Results are presented in Table 3. The values obtained by the compression test are given in parenthesis. Generally, a good agreement of the results is observed.

CONCLUSIONS

The elastic constants of an unidirectionally reinforced laminate have been determined employing the identification procedure based on the experimental design and response surface method. For this, minimal Mean Square Distance Latin Hypercube (MMSDLH) designs and local approximations were used. It was shown that the elastic constants obtained from vibration tests through identification are in good agreement with the values obtained by conventional static tests.

ACKNOWLEDGEMENTS

Investigations concerning the development of identification methods for laminated composites are sponsored through Contract No. GRD-CT-1999-00103 by the Commission of European Union. We would also like to thank Dr. H. Abramovich (Israel Institute of Technology) for providing the vibration test data, Dr. Green (IAI - Israel Aircraft Industry LTD) and Dr. R. Zimmermann (DLR - German Aerospace Center) for providing the static test data.

REFERENCES

1. J.-F. M. Barthelemy and R. T. Haftka, Approximation concepts for optimum structural design – a review, *Structural Optimization*, **5**, 129, (1993).

2. R. Rikards, A. Chate and G. Gailis, Identification of elastic properties of laminates based on experiment design, *Int. J. Solids and Structures*, **38**, 5097, (2001).

3. G. E. P. Box and N. R. Draper, *Empirical model-building and response surfaces*, John Wiley & Sons, New York, 1987.

4. P. Audze and V. Eglais, New approach to planning out of experiments, In: *Problems of Dynamics and Strength* vol. 35 (Ed. E. Lavendel), Zinatne, Riga, 1977, pp. 104-107 (in Russian).

5. R. T. Haftka, E. P. Scott and J. R. Cruz, Optimization and experiments: A survey. *Appl. Mech. Rev.*, **51** (**7**), 435, (1998).

6. V. Eglais, Approximation of data by multi-dimensional equation of regression, In: *Problems of Dynamics and Strength* vol. 39 (Ed. E. Lavendel), Zinatne, Riga, 1981, pp. 120-125 (in Russian).

7. R. D. Meyer, D. M. Steinberg and G. Box, Follow-Up Designs to Resolve Confounding in Miltifactor Experiments, *Technometrics*, **38**, 303, (1995).

8. A. I. Khuri and J. A. Cornell, *Response Surfaces: Designs and Analyses*, Marcel Dekker, New York, 1996.

9. A. A. Giunta, J. M. Dudley, R. Narducci, B. Grossman, R. T. Haftka, W. H. Mason, and L. T. Watson, Noisy Aerodynamic Response and Smooth Approximations in HSCT Design, In: *Proceedings of the 5th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Panama City Beach, FL, Sept. 1994, pp.1117–1128.

10. M. D. McKay, W. J. Conover and R. J. Beckman, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21**(2), 239, (1979).

11. M. E. Johnson, L. M. Moore and D. Ylvisaker, Minimax and Maximin Distance Designs. *Journal of Statistical Planning and Inference*, **26** (2), 131, (1990).

12. J. Sacks, W. J. Welch, T. J. Mitchell and H. P. Wynn, Design and analysis of computer experiments, *Statistical Science*, **4** (4), 409, (1989).

13. B. Tang, Orthogonal Array-Based Latin Hypercubes, *Journal of the American Statistical Association*, **88** (424), 1392, (1993).

14. K. Q. Ye, Column orthogonal Latin hypercubes and their application in computer experiments, *Journal of American Statistical Association*, **93**, 1430, (1998).

15. J.-S. Park, Optimal Latin-Hypercube Designs for Computer Experiments, *J. Statistical Planning and Inference*, **39** (1), 95, (1994).

16. W. S. Cleveland and S. J. Devlin, Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, **83**, 596, (1988).

17. J. R. Koehler and A. B. Owen, Computer Experiments, In: *Handbook of Statistics* (Eds. Ghosh, S. and Rao, C. R.), Elsevier Science, New York, 1996, pp. 261-308.

18. A. J. Booker, J. E. Dennis Jr., P. D. Frank, D. B. Serafini, V. Torczon and M. W. Trosset, A Rigorous Framework for Optimization of Expensive Functions by Surrogates, *Structural Optimization*, **17** (1), 1, (1999).

19. J. H. Friedman, Multivariate Adaptive Regression Splines, *The Annals of Statistics*, **19** (1) 1, (1991).

20. R. L. Hardy, Multiquadratic Equations of Topography and Other Irregular Surfaces, *J. Geophys. Res*., **76**, 1905, (1971).

21. N. Dyn, D. Levin and S. Rippa, Numerical Procedures for Surface Fitting of Scattered Data by Radial Basis Functions, *SIAM Journal of Scientific and Statistical Computing*, **7** (2), 639, (1986).

22. M. J. D. Powell, Radial Basis Functions for Multivariable Interpolation: A Review, In: *Algorithms for Approximation* (Eds. Mason, J. C. and Cox, M. G.), Oxford University Press, London, 1987, pp. 143-167.

23. J. Auzins, New Experimental Designs for Computer Experiments, In: *Scientific Proceedings of RTU: Transport and Mechanical Engineering*, series 1, vol. 53, Riga, Riga Technical University, p. 8, (2001).

24. K.-T. Fang and Y. Wang, *Number-Theoretic Methods in Statistics*, Chapman & Hall, London, 1994.

25. M. D. Morris and T.J. Mitchell, Exploratory designs for computer experiments, *Journal of Statistical Planning and Inference*, **43**, 381, (1995).

26. M. Shewry and H. Wynn, Maximum entropy design, *Journal of Applied Statistics*, **14** (2), 165, (1987).

27. C. Currin, T. Mitchell, D. Morris and D. Ylvisaker, Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments, *Journal of the American Statistical Association*, **86**, 953, (1991).

28. K. Palmer and K.-L. Tsui, A Minimum Bias Latin Hypercube Design, *IIE Transactions*, **33**(9), 793 (2001).

29. K. T. Fang, W.C. Shiu, W.C. and J.X. Pan, Uniform designs based on Latin squares, *Statistica Sinica*, **9**, 905, (1999). <www http://www.math.hkbu.edu.hk/UniformDesign/ >

30. J. Auzins, A. Janushevskis and O. Onzevs, Optimization of Multibody Vibration Response by Global Search Procedure, In: *Proceedings of ECCM 99*, European Conference on Computational Mechanics, CD-ROM, Munich, Germany, p. 17 (1999).

Identification of weak non-linear damping forces from system response.

C. Meskell
Department of Mechanical and Manufacturing Engineering,
Trinity College,
Dublin, Ireland
cmeskell@tcd.ie

ABSTRACT

Non-linear damping forces can be quantified using the force state mapping technique and this method of non-linear identification has been successfully applied to a fluidelastic system. However, when applied to a system with low damping this method becomes unacceptably sensitive to small phase distortions which may be caused by either the instrumentation or numerical differentiation. It has been shown previously that a simple optimization technique (downhill simplex search) can be used to accurately identify light damping in a single degree of freedom linear system, and this technique has been used to obtain equivalent linearized damping parameters in a fluidelastic system where the damping mechanisms are inherently non-linear. This approach has the advantage that it is insensitive to phase errors as only one response measurement is required as input, however it is currently restricted to linear systems. This paper extends the technique to explicitly account for non-linear damping terms. Response data for a single degree of freedom system with a small cubic damping term are obtain by numerical simulation and are used to demonstrate the enhanced procedure. The technique is shown to be robust in the presence of line noise and the effect of initial parameter estimates is explored.

INTRODUCTION

When considering the dynamic response of a system, the damping mechanisms are often of most interest, since it is the damping which will govern the amplitude of motion, in the case of forced vibration, and the longevity of vibration in transient response. However, these damping forces are often orders of magnitude smaller than the stiffness forces. This, coupled with the fact that dissipation is often due to non-linear mechanisms (e.g. Coulomb damping; fluidelastic damping), makes quantification of the parameters associated with the damping forces difficult.

This situation is typified by fluidelastic systems in which the fluid dynamics is strongly coupled with structural dynamics. The resulting vibration may be self excited, large amplitude and self limiting, a phenomenon referred to as fluidelastic instability. One example of fluidelastic instability can be found in heat exchangers subjected to cross flow. An excellent review of fluidelastic instability in tube bundles (i.e. heat exchangers) has been published by Price [1]. It can be concluded from this review that the self exciting nature of fluidelastic instability may be described well with a linear model: a negative linear damping, which is caused by the coupling of the fluid and structural system, increases with flow velocity. The onset of instability (the "critical velocity") is then predicted when the total damping becomes negative. Such a model predicts a dynamic divergence. However, in the physical system, limit cycle behaviour is often observed. Prediction of such behaviour requires a non-linear model of the fluid force, particularly the damping. While it is true that structural non-linearities, such as impacting will be more substantial, Price noted that a non-linear force model is still desirable, since it will determine the energy available in the system. The experimental identification of a non-linear model is problematic, since even in a post-critical regime, the non-linear forces are extremely weak. These general observations can be equally well applied to other fluidelatic systems such as aerofoil/hydrofoil flutter or galloping of bluff bodies [2, 3].

Although fluidelastic behaviour was the motivation for this study, the techniques described below are not restricted to such systems. The central issue here is the *parametric* identification of small non-linear damping forces in an otherwise linear system.

PARAMETER ESTIMATION PROCEDURES

Marsi & Caughey [4] described a parametric estimation procedure, referred to as force state mapping, which has the advantage that the identification problem is linear, even for a non-linear model, but it does require that the structure's state variables (displacement and velocity) be measured simultaneously with the total excitation force (including the system acceleretion). As well as having been applied to systems which include strong non-linear ele-

ments [4, 5, 6], it has also been used successfully to identify a weakly non-linear model for fluidelastic instability [7]. However, it has been noted that for a system with light damping, the estimates of the damping parameters are very sensitive to small phase distortions of the measured signals, which can easily result from the instrumentation. This will be true for any time domain technique which requires several synchronous measurements. For this reason, a technique which requires only a single measured response is desirable.

Mottershead & Stanway[9] proposed such a scheme to directly estimate the parameters of a non-linear single degree of freedom system from only a single noisy observation channel (e.g. acceleration). Unlike the force state mapping technique, this procedure is iterative, and requires an inital estimate of the parameters. The authors applied the technique to identify the parameters of a system with nth-power velocity damping. Similar techniques have been applied by Yar & Hammond[10] and Stanway *et al.*[11] to identify non-linear hysteretic and damping behaviour in systems with a single degree of freedom. In all three studies, a random force was used to excite the system the system.

The iterative technique is started with an initial estimate for the parameter set to be identified. This parameter vector includes the initial conditions of the system. The excitation force is assumed to have been measured exactly, which in practice is probably not the case. The measured force is used as input to a simulation of the system with the parameter values set by the initial estimate. The simulation is based on a Runge-Kutta integration of equation of motion of the system. The total error between the resulting simulated response and the experimentally measured response is then calculated. For example, if acceleration was measured

$$\varepsilon = \int_0^T \left[\ddot{y}(t) - \ddot{x}(t)\right]^2 dt \qquad (1)$$

where $\ddot{y}(t)$ is the measured signal, $\ddot{x}(t)$ is the simulated signal and T is the observation time. The acceleration response is used here as it is arguably the easiest to obtain experimentally. The identification then proceeds by minimizing ε with respect to the parameter set. The minimization is achieved by using the Gauss-Newton procedure. This is a first order minimization routine and so at each iteration the first derivative of the cost function with respect to each parameter is required. These derivatives are calculated as part of the Runge-Kutta scheme, however, this can introduce a phase error into the simulated time records which may translate into an error in the final parameter estimates.

Meskell & Fitzpatrick [8] employed a similar method to identify the equivalent linearized parameters in a fluidelastic system. The approach used by the authors is broadly similar to the one discribed above, but with three important differences:

- The downhill simplex search developed by Nelder & Mead [12] was used to minimize the cost function. This algorithm has the advantage that it does not require any derivatives with respect to the parameter set and it is straight forward to implement. This alleviates the additional source of error due to numerical estimation of derivatives. Although it is true that the minimum is approached more slowly than in the Gauss-Newton method, the modern desktop computer capacity means that this is not such a critical issue. The Downhill simplex scheme also has the advantage that it is quite robust.

- In the previous work [8] the measured response data $\ddot{y}$ was obtained from a free decay test with no additional excitation. As will be shown below this offers considerable advantages over forced response, particularly in terms of the reliability of the parameter estimates obtained.

- Since the model that was identified previously was linear, at each iteration the response $\ddot{x}$ could be calculated with an analytical solution rather than with a Runge-Kutta integration. This meant that the identification procedure was very inexpensive computationally. While it is possible to develop analytical solutions for the free response of some weakly non-linear systems, this is not generally the case. Therefore, in this paper a fourth order Runge-Kutta integration scheme is used to obtain the response for both free and forced situations.

This paper will extends framework employed by Meskell & Fitzpatrick to include weak non-linear damping forces and explores the effect of measurement noise and initial parameter estimates on the final identified values.

EFFECT OF EXCITATION IN THE PRESENCE OF LINE NOISE

The studies discussed above [9, 10, 11] have shown that the estimates of parameters may become unreliable in the presence of line noise on the measured response and as would be expected weak forces are more prone to error. This problem can be alleviated somewhat by the addition of a noise model [11]. Alternatively, if the measured response is transient (i.e. a free decay test), the technique becomes more robust to line noise, even for the weak non-linear forces of interest here. An additional advantage is that an identification technique which depends only on free response data will not be prone to errors due to noise or relative phase distortion in the excitation force measurement.

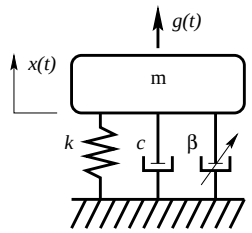


Figure 1: Schematic of weakly non-linear system.

However, the discussion here will be confined to the effect excitation on the parameter estimates in the presence of line noise without phase distortion of the underlying signal.

In order to illustrate the difference between parameter estimates obtained from forced and free response, consider the specific single degree of freedom system shown in Figure . This is basically a linear system with a weak non-linear cubic damping element.

The equation of motion for the system is:

$$\ddot{x} + \frac{k}{x} + c\dot{x} + \beta\dot{x}^3 = g(t) \tag{2}$$

Note that the parameters k, c and β are mass normalized, as is the excitation force $g(t)$. A fourth order Runge-Kutta scheme was used to generate 4 seconds of data at a sample rate of 2048Hz with the following system paremeter values:

Parameter	Value
$k = \omega^2$	$1600s^{-2}$
c	$1.0s^{-1}$
β	$2.0sm^{-2}$

Table 1: Exact system parameter values.

For the random excitation forced response simulation the excitation, $g(t)$ was band limited white noise (0-100Hz) with a variance of $9.0\mathrm{N}^2\mathrm{kg}^{-2}$. The system was started from equilibrium i.e. $x(0) = 0, \dot{x}(0) = 0$.

For the simulation with deterministic excitation, $g(t)$ was a Swept Sine with a frequency range of 0-100Hz. The period of the Swept SIne was 4 seconds. The amplitude on the excitation was $3.8\mathrm{ms}^{-2}$. The simulation was initially started from equilibriuim, but was allowed to run for 10 periods of the excitation to allow the system to establish a periodic response. Once this was achieved the response signals for one period (4 s) were recorded. The state variables at the start of this final period were also noted ($x(0) = -2.5^{-4}$, $\dot{x}(0) = -5.3 \times 10^{-2}$). In this way the initiatial conditions of the system are known.

For the transient response, $g(t)$ was zero, but the system was released from rest with a non-zero displacement: $x(0) = 0.01\mathrm{m}$, $\dot{x}(0) = 0$.

This choice of initial conditions and amplitudes yields the comparable standard deviation in the acceleration responses associated with the three simulations. As an indication of the relative contributions of the force terms in equation 2, the standard deviation of each term is listed in table 2 below for the three different types of excitation. It

	Standard deviation (ms^{-2})		
Force term	Free	Random	Swept Sine
Excitation	-	3.0	2.7
Acceleration	5.3	5.3	5.3
Stiffness	5.3	4.5	4.5
Linear damping	0.13	0.11	0.11
Cubic damping	0.011	0.009	0.013

Table 2: Response statistics

can be seen from this that the magnitude of the acceleration signal, which will be used as input for the identification procedure, is comparable in all three cases. It is also worth noting that the cubic damping term, which is of most interest in this study, has the lowest value and is nearly three orders of magnitude smaller than the stiffness.

The data obtained was used as input to the algorithm described above: a Runke-Kutta integration based on the current parameter estimates at each step of a downhill simplex search. Line noise was simulated by adding exactly the same random Gaussian signal to the transient and forced responses. Six levels of noise were examined in the range 0%-10% where the percentage indicates the standard deviation of the noise as a fraction of the standard deviation of the response. Figure 2 shows an example of noise corrupt free response. In this instance the noise is 10% with the noise free signal superimposed for comparison.

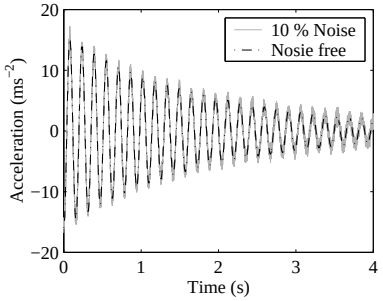


Figure 2: Noise corrupt free response.

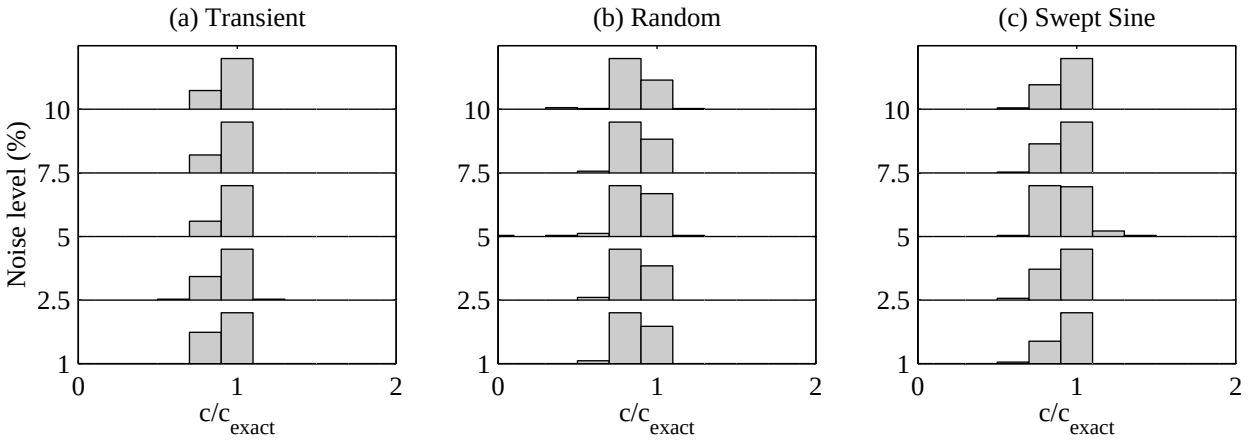


Figure 3 Distribution of estimates of c from different initial estimates at various levels of noise

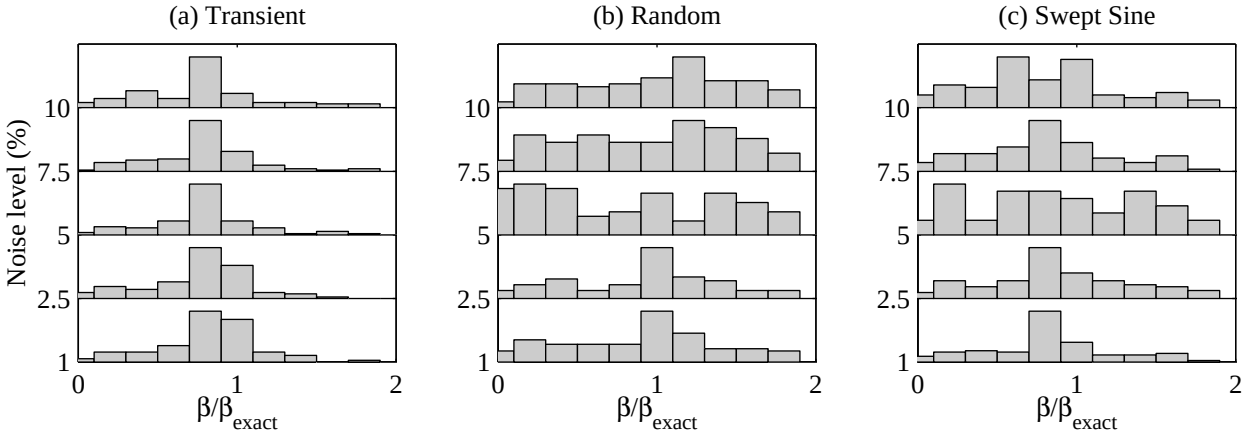


Figure 4 Distribution of estimates of β from different initial estimates at various levels of noise

There are five parameters to be identified: the initial displacement and velocity; stiffness term ; the linear damping; and the cubic damping. So the parameter set is

$$\mathbf{P} = \begin{bmatrix} x(0) & \dot{x}(0) & k & c & \beta \end{bmatrix} \tag{3}$$

In practice it is the last two parameters (the damping parameters) which are of primary interest and so attention will focus on these.

In order to isoloate the effect of the excitation signal on the final parameter estimates from the effect of the initial parameter estimates, the identifcation procedure was applied to each of the three response records using 100 different sets of initial parameter estimates. The initial conditions for the system were assumed to be known exactly,as was the stiffness. Thus, the initial parameter estimate set is given by

$$\mathbf{P}_0 = \begin{bmatrix} x_0 & \dot{x}_0 & 1600 & c_0 & \beta_0 \end{bmatrix} \tag{4}$$

with $x_0 = 0.01, 0$ or -2.5×10^{-4} and $\dot{x}_0 = 0, 0$ or -5.3×10^{-2} for free, random and swept sine response, respectively. Both c_0 and β_0 were systematically assigned values upto 100% over- and underestimating the exact values.

As with any iterative optimization technique the downhill simplex method requires a termination criterion in order to determine that acceptable convergence on a minimum has been achieved. In this case, the optimization of the parameter set was terminated once the normalized size of the simplex bracketing the solution fell below a certain value. This metric is an indication of how rapidly the solution set is changing, and so is an indirect measure of the gradient of the cost function in the region of the current estimate of the parameter set. A secondary termination criterion specifying a maximum number of iterations was also set. It was found that 1000 iterations of the scheme was sufficient, with the cost function (equation 1) often minimized in less than 300 iterations.

Figures 3 and 4 show the distribution of the final estimates of c and β (normalized with the exact values) obtained from the 100 different initial estimate sets for the

three types of response at 5 levels of line noise. Results for zero noise is not shown here as all three responses identified the damping parameters to within 1% of the exact values regardless of the initial estimates. It should also be noted that the final estimate of stiffness, k, and initial conditions, $(x_0, \dot{x}_0)$, are always within 1% of the correct values. This is significant in that it demonstrates that, the optimization procedure does not diverge, at least in these three parameters, as the initial estimates were exact.

The distributions in Figure 3 show that all three response types provide good estimates for the linear damping c even at high levels of noise. The transient and swept sine response data perform comparably well, with the majority of estimates being within 10% of the exact value. However, the random excitation data is more likely to underestimate the value of c.

The estimates of the cubic damping show a different behaviour. Firstly, it is apparent that there is a larger spread in the estimates when compared to those for c for all three response types. At low levels of line noise (up to 2.5%) the three types of response data again provide reasonable reliability, with the majority of estimates within 30% of the exact value of β. However, even at these low levels of noise it is apparent that the transient response offers superior reliability when compared to the random response data. The swept sine estimates are marginally less reliable at these low noise levels, as indicated by the slightly more significant spread in the distribution. At higher levels of noise, it is obvious that the transient response data is more likely to offer an accurate estimate of β. Indeed, there is no significant reduction in the likelihood of an acceptable level of accuracy with increased noise for estimates based on the free response data. For both the swept sine and the random response data, the estimates of β deteriorate significantly at higher levels of noise although the estimates from the random response data are arguably worse than those from the deterministic excitation.

For the random response data it seems that, in the presence of noise, the final estimate of β is close to the initial estimate. In other words, the cost function (equation 1) is relatively insensitive to the value of β. Similar behaviour can be seen in the results of Yar & Hammond [10]. It ist empting to argue that the reason for the poor prediction of β is due to its small contribution to the total force. However, the free response data has very similar relative contrubutions as can be seen in Table 2.

These observations can be summarized by considering the absolute error in the parameter estimates averaged over all the initial parameter sets. Figure 5 shows the variation in the average error in the estimate of the linear dampingf c as noise level increases. As discussed above, the transient response offers marginally more reliable results, but all three response types offer reasonable estimates. In figure 6, the variation of the mean error in the final estimates of β is seen to be similar in trend to that in figure 5. However, the size of the error far more significant. In this case the transient response offers far superior results.

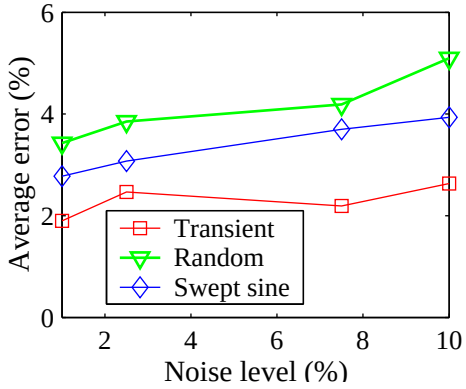


Figure 5 Mean % error in estimate of c.

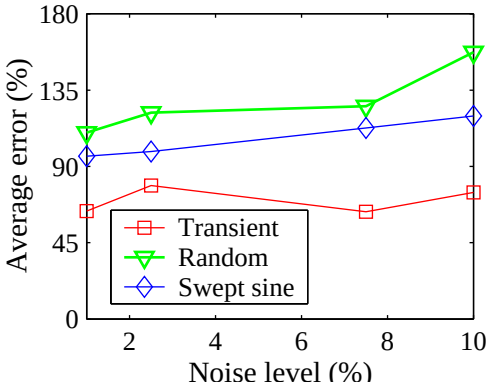


Figure 6 Mean % error in estimate of β.

EFFECT OF INITIAL ESTIMATES

The performance of the identifcation procedure will depend on the initial estimate of the parameters. To assess this, the error associated with the final parameter estimates will be considered when there is 10% line noise. As before,the first three parameters of the initial parameter set (equation 4) are given the appropriate exact values. The values of c_0 and β_0 are again varied systematically in the range $0 \leq c_e xact \geq 2 \times c_e xact$ and $0 \leq \beta_e xact \geq 2 \times \beta_e xact$ respectively. For transient response it was found that for $c_0 = 0$ the procedure diverged, but even a small non-zero value produced convergence.

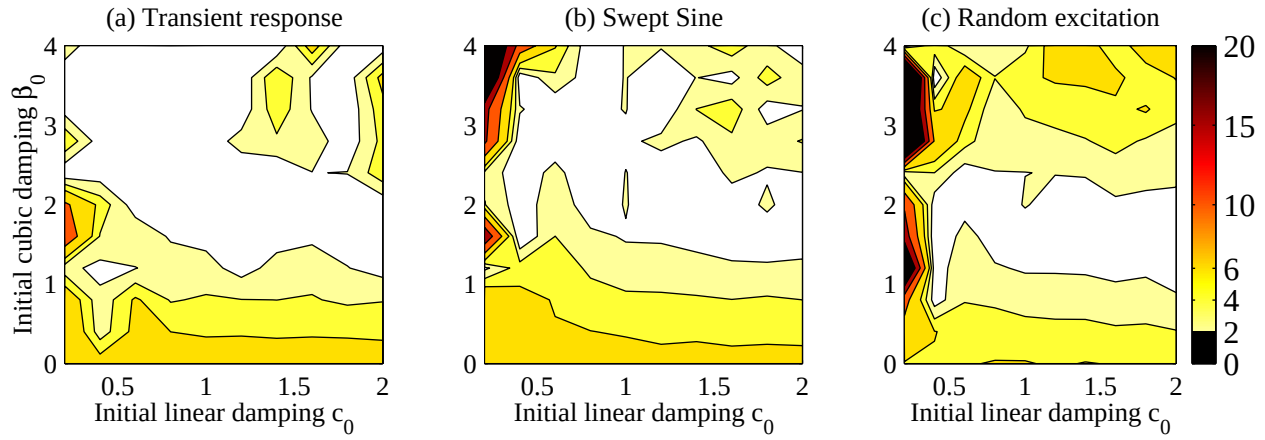


Figure 7 Variation of percentage error in estimate of c for various initial parameter sets. Noise level 10%.

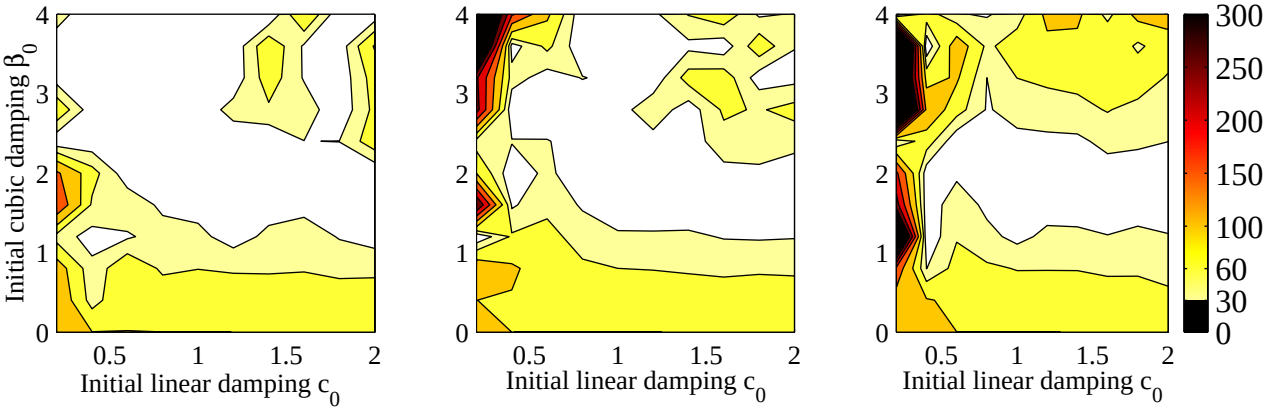


Figure 8 Variation of percentage error in estimate of β for various initial parameter sets. Noise level 10%.

The absolute error in the final estimates of c and β for the range of starting estimates is shown in figures 7 and 8, respectively. Note that these errors are shown as percentages.

Agian it is apparent that the transient response data will, generally yield better values for the system parameters. This is idicated by the much larger area of low error (white region) in figure 7(a) compared to figure 7(b) and (c). The same is true of figure 8.

In figure 7(a), the identified value of linear damping, c, is within 10% of the actual value of $1.0s^{-1}$ except for low values of c_0, and even then the final estimate is reasonable. Although the maximum error associated with the cubic damping β is more than an order of magnitude greater (see figure 8), the error is less than 100% as long as c_0 is not very small. For transient response the estimate is better than 20% for about half the range of initial estimates evaluated.

In contrast, figure 7(c) and figure 8(c) demonstrate that the random excitation will yield accurate results only if the initial estimates are close to the exact values.

Examining the figure 7(a) and figure 8(a) together suggests that the best strategy for choosing initial estimates is to slightly under-estimate the linear damping (i.e. $c_0 < c_{exact}$) and over-estimate the cubic damping (i.e. $\beta_0 > \beta_{exact}$). In practice, the exact values are unknown; this is afterall why an identification technique is needed. However, an appropriate value for c_0 could be achieved by first estimating the equivalent linear damping (i.e. ignore the non-linear damping [8]) which could be scaled by a value close to unity.

Yar & Hammond [10] recommended that the damping parameters should be initially over-estimated. The authors of that study were examining a system with a hysteretic restoring force subject to random excitation. Further investigation is needed to clarify whether the appropriate stategy for choosing the initial parameter values depends on the system model or on the excitation or both.

Conclusions

An iterative time domain technique has been described for the parametric identification of a single degree of freedom model with weak non-linear damping. Although it is an iterative paramteric identification technique, only an initial estimate of the parameter set is required, not the derivatives of the cost function. The specific case of a system with cubic damping was considered in which the contribution of the non-linear damping force was 2-3 orders of magnitude smaller than that of the stiffness force.

It has been shown that parameter estimates based on free response data are considerably more reliable in the presence of line noise when compared to those values obtained from forced response data. Both random and deterministic excitation signals have been examined and in both cases the transient response offered superior parameter estimates.

Using the procedure, the effect of initial parameter estimates has been explored for the test system and strategy for choosing these values is discussed. However, it is likely that this strategy for choosing the initial estimates is not general but rather it is specific to the free response of the particular system under investigation. Indeed, further work is needed to establish if the overall approach is applicable to other systems and if it is robust in the presence of unknown excitation, such as would be the case in fluidelastic systems where turbulence will always be present.

In principle this method can be easily extended to multi-degree of freedom systems which have weakly non-linear components. However, care should be taken as even with a single degree of freedom system the dimension of the parameter space was 5. (the likelihood of success with any iterative optimization procedure will reduce as the number of parameters to be found increases). One possible strategy for overcoming this issue and reducing the size of the parameter space would be to use the parameter estimates of an equivalent linearized system which can be obtained with a one step procedure (such as least squares) as the initial values. This type of approach has been used previously with some success [11].

Not withstanding these comments, the basic method has been shown to be a likely candidate for anaylsis of experiemntal data. It is widely known that for linear systems estimating the damping can most easily be done from free response data. This paper suggests that the sam is true for non-linear systems. Thus, for a non-linear system when non-linear damping parameters are of primary interest, the transient response offers more accurate results.

References

[1] S. J. Price. A review of theoretical models for fluidelastic instability of cylinder arrays in cross-flow. *Journal of Fluids and Structures*, 9:463–518, 1995.

[2] J. Horacek and Zolotarev I., editors. *Proceedings of the 3rd International conference, Engineering aero-hydroelastciity*. Institue of thermodynaimcs, 1999.

[3] Paidoussis, editor. *Fluid-structure interactions, aeroeleasticity, flow-induced vibration and noise*, volume 1. ASME, 1997.

[4] S. F. Masri and T. K. Caughey. A non-parametric identification technique for nonlinear dynamic problems. *Journal of applied mechanics*, 46:433–447, 1979. Original ref to FSM for a sdof problem.

[5] K. Q. Xu and H. J. Rice. On an innovative method of modeling general nonlinear mechanical systems. part 1: Theory and numerical simulations. part 2: Experiments. *Journal of Vibration and Acoustics*, 120:125–137, 1998.

[6] K. Worden. Data processing and experiment design for the restoring force surface method. part (i): Integration and differentiation of measured time data. part (ii): Choice of excitation signal. *Mechanical systems and signal processing*, 4(4):295–344, 1990.

[7] C. Meskell, J.A. Fitzpatrck and H.J. Rice. Application of force-state mappinbg to a non-linear fluid-elastic system. *Mechanical systems and signal processing*, 15(1):75–85, 2001.

[8] C. Meskell and J.A. Fitzpatrick. Identification of linearised parameters for fluidelastic instability. In Paidoussis et al., editor, *Fluid-structure interactions, aeroeleasticity, flow-induced vibration and noise*, volume 1, pages 319–324. ASME, 1997.

[9] J.E. Mottershead and R. Stanway. Identification of nth-power velocity damping. *Journal of sound and vibration*, 105(2):309–319, 1986.

[10] M. Yar and J.K. Hammond. Parameter estimation for hysteretic systems. *Journal of sound and vibration*, 117(1):161–172, 1987.

[11] R. Stanway, J. Sproston, and R. Firoozian. Identification of the damping law of an electro-rheologoical fluid: a sequential filtering approach. *Journal of dynamic systems, measurement and control*, 111:91–96, 1989.

[12] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

4^{th} *International Conference on Inverse Problems in Engineering*
Rio de Janeiro, Brazil, 2002

ON THE IDENTIFICATION OF CONSTITUTIVE PARAMETERS OF VISCOELASTIC MATERIALS BY MEANS OF A TIME DOMAIN TECHNIQUE

Daniel Alves Castello and Fernando Alves Rochinha
Solid Mechanics Laboratory
Mechanical Engineering Department, Federal University of Rio de Janeiro, UFRJ
Rio de Janeiro, RJ, Brazil
castello@mecsol.ufrj.br and faro@serv.com.ufrj.br

ABSTRACT

The present work approaches the problem of identification of the elastic and damping fields of a medium by means of a time domain technique. This technique is within the inverse problems scope, i.e., the solution of the problem is sought by means of the minimization of a suitable error function which includes data from both the system model and the experiment setup for the same input excitation. In order to assess the effectiveness of the proposed method, simulations on a bar-like structure have been performed under impact loading and considering the corrupting effects of noise.

NOMENCLATURE

Matrices

$\mathbf{C}$	Observability matrix.
$\mathbf{D}$	System damping matrix.
$\mathbf{K}$	System stiffness matrix.
$\mathbf{M}$	System mass matrix.

Vectors

$\mathbf{d}$	Direction of descent.
$\mathbf{f}$	External force.
$\mathbf{p}$	Parameter vector.
$\mathbf{x}$	System displacement field.
$\mathbf{y}$	System observable variable.
$\mathbf{y}^E$	System measured data.
$\boldsymbol{\lambda}$	Lagrange multiplier.

Scalars

E	Elastic field.
G	Damping field.
n	Number of degrees of freedom.
β	Search step size.
A	Bar cross section area.
Φ	Classical Lagrangian piecewise linear shape functions.

INTRODUCTION

Aiming at taking advantage of the dynamical properties of each material in a system design, it is required the fully understanding of the mechanical behavior of these materials. This behavior can be described by different models such that the designer has some freedom to choose the most suitable one for a certain type of application that the material will be part of. Once one has in hands the chosen model that will be used to describe the mechanical behavior of the material under study, the next step usually consists in determining the set of parameters that characterizes this model. The identification of these parameters provides a mathematical model which enables one to simulate and predict the response of the material when it is subjected to a certain excitation. In particular, the mechanical behavior of viscoelastic materials is of great interest in engineering sciences such as mechanical, civil, aerospace and biomechanical. The technical literature concerned with the identification of viscoelastic materials is very extensive and it presents different approaches to the problem [1], [2], [3], [4], [5], [6], [7] and [8] can be cited as the most recent ones. The present work is built on the use of a constitutive equation for viscoelastic materials parameterized by a set of unknown constitutive parameters and makes use of a time do-

main technique to identify this set of parameters. The solution technique is within the inverse problems scope, i.e., the solution is sought by means of the minimization of a suitable error function which includes data from both the system model and the experiment. The technique takes into account the constraint associated to the system evolution equation as being part of an extended error function what naturally gives rise to the Lagrange multiplier variables which are obtained via solution of an adjoint problem [9], [11] . The effectiveness of the technique is assessed on simulations performed on a bar-like structure, where strains or displacements are measured at a subset of the system degrees of freedom. The simulated experiment consists on a bar under dynamic loading excitation and in order to furnish realism to the simulations, it is considered the corrupting effects noise.

DIRECT PROBLEM

Consider an $n-DOF$ linear dynamic system such that its discretized evolution equation is given by

$$\begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{D}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{f}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \\ \mathbf{x}(0) = \mathbf{x}_0 \\ \dot{\mathbf{x}}(0) = \dot{\mathbf{x}}_0 \end{cases} \tag{1}$$

where $\mathbf{M}$, $\mathbf{D}$ and $\mathbf{K}$ are nxn matrices describing the mass, damping and stiffness properties respectively and the n dimensional vectors $\mathbf{x}$ and $\mathbf{f}$ correspond to the system displacement field and to the external loading applied to the system. The matrix $\mathbf{C}$ associates the system DOF to the measured observable variables $\mathbf{y}$, which in turn, can be displacements or strains. The direct problem consists basically in determining the transient displacement field $\mathbf{x}(t)$ when the external load is known. It should be emphasized that the direct analysis assumes a priori that the material behavior is known, fact that, for the present problem, means that one has in hands the constitutive equation between stress and strain for the material under study and moreover, the actual value of the parameters of this constitutive equation is available. In equation (1) it is implicit that the property matrices $\mathbf{D}$ and $\mathbf{K}$ are in some way functions of the parameters that characterize the material constitutive equation, viz.

$$\mathbf{K} = \mathbf{K}(\mathbf{p}) \qquad \text{and} \qquad \mathbf{D} = \mathbf{D}(\mathbf{p}) \tag{2}$$

where the vector $\mathbf{p}$ contains both elastic and damping parameters upon which the material constitutive equation is defined.

INVERSE PROBLEM

For the inverse problem, the elastic and damping parameters $\mathbf{p}$ are considered to be unknown. It is also assumed that there is set of experimental data available $\mathbf{y}^E(t)$, $t \in [0, t_f]$, which can be used as the additional information for the estimation of the parameters $\mathbf{p}$ and consequently the matrices $\mathbf{K}$ and $\mathbf{D}$. The idea is to minimize a suitable error function which consists basically of the norm of the difference between the measured data $\mathbf{y}^E(t)$ and the data obtained from the system model $\mathbf{y}(t)$ for the same input excitation. The error function $\hat{J}_1(\mathbf{p})$ is defined as follows

$$\hat{J}_1(\mathbf{p}) = \int_0^{t_f} [\mathbf{y} - \mathbf{y}^E]^T [\mathbf{y} - \mathbf{y}^E] dt \tag{3}$$

Therefore, the goal of the inverse problem step is to estimate the N_p-dimensional vector of unknown parameters $\mathbf{p}$ through the minimization of $\hat{J}_1(\mathbf{p})$. The search step size determination will be presented later.

Parameter Estimation

The technique used for parameter estimation is the Conjugate Gradient Method, which is a powerful iterative technique for solving linear and nonlinear inverse problems of parameter estimation [9]. In the iterative procedure of the conjugate gradient method, at each iteration a suitable step size is taken along a direction of a descent in order to minimize the error function as follows

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - \beta^{(k)} \mathbf{d}^{(k)} \tag{4}$$

where k indicates the current iteration, $\beta^{(k)}$ is the search step size, $\mathbf{d}^{(k)}$ is the direction of descent which is defined as follows

$$\mathbf{d}^{(k)} = \nabla \hat{J}^{(k)} + \gamma^{(k)} \mathbf{d}^{(k-1)} \tag{5}$$

For the conjugation coefficient $\gamma^{(k)}$, among some possibilities, one has chosen

$$\gamma^{(k)} = \frac{\nabla \hat{J}^{(k)} \cdot [\nabla \hat{J}^{(k)} - \nabla \hat{J}^{(k-1)}]}{\nabla \hat{J}^{(k-1)} \cdot \nabla \hat{J}^{(k-1)}} \tag{6}$$

Further details about the previous choice may be found in [10].

Sensitivity Problem

The sensitivity function $\Delta\mathbf{x}(t)$, which is the solution of the sensitivity problem, is defined as the directional derivative of the displacement field $\mathbf{x}(t)$ in the direction of the perturbation of the unknown parameter vector $\mathbf{p}$ [9]. The presentation of the sensitivity problem is required in order to obtain the search step size $\beta^{(k)}$. Aiming at obtaining the sensitivity problem one assumes that the displacement field $\mathbf{x}(t)$ is perturbed by an amount $\Delta\mathbf{x}(t)$ when the unknown vector of parameters $\mathbf{p}$ is perturbed by $\Delta\mathbf{p}$ such that

$$\mathbf{x}(t, \mathbf{p} + \Delta\mathbf{p}) = \mathbf{x}(t, \mathbf{p}) + \Delta\mathbf{x}(t, \mathbf{p}) \tag{7}$$

$$\mathbf{D}(\mathbf{p} + \Delta\mathbf{p}) = \mathbf{D}(\mathbf{p}) + \Delta\mathbf{D}(\mathbf{p}) \tag{8}$$

$$\mathbf{K}(\mathbf{p} + \Delta\mathbf{p}) = \mathbf{K}(\mathbf{p}) + \Delta\mathbf{K}(\mathbf{p}) \tag{9}$$

The evolution equation for the system under this new set of parameters casts as follows

$$\mathbf{M}[\ddot{\mathbf{x}} + \Delta\ddot{\mathbf{x}}] + [\mathbf{D} + \Delta\mathbf{D}][\dot{\mathbf{x}} + \Delta\dot{\mathbf{x}}] + \\ +[\mathbf{K} + \Delta\mathbf{K}][\mathbf{x} + \Delta\mathbf{x}] = \mathbf{f} \tag{10}$$

and

$$\mathbf{y} + \Delta\mathbf{y} = \mathbf{C}[\mathbf{x} + \Delta\mathbf{x}] \tag{11}$$

Where $\mathbf{x} = \mathbf{x}(t, \mathbf{p})$ e $\mathbf{y} = \mathbf{y}(t, \mathbf{p})$. Applying the initial conditions to the new solution $\mathbf{x}(t, \mathbf{p} + \Delta\mathbf{p})$ and considering that the equation (7) must hold leads to the following initial conditions for the sensitivity problem

$$\Delta\mathbf{x}(0, \mathbf{p}) = \mathbf{0} \qquad \text{and} \qquad \Delta\dot{\mathbf{x}}(0, \mathbf{p}) = \mathbf{0} \tag{12}$$

Hence, disregarding the second order terms of equation (10) and considering that the terms associated to the evolution equation of the system, which are present in equations (10) and (11), are automatically satisfied, enables one to state the sensitivity problem as follows

$$\begin{cases} \mathbf{M}\Delta\ddot{\mathbf{x}} + \mathbf{D}\Delta\dot{\mathbf{x}} + \mathbf{K}\Delta\mathbf{x} = \\ \Delta\mathbf{D}\dot{\mathbf{x}} + \Delta\mathbf{K}\mathbf{x} \\ \Delta\mathbf{y} = \mathbf{C}\Delta\mathbf{x} \\ \Delta\mathbf{x}(0) = \mathbf{0} \qquad \Delta\dot{\mathbf{x}}(0) = \mathbf{0} \end{cases} \tag{13}$$

Adjoint Problem

The adjoint problem naturally appears when one considers that the displacement field $\mathbf{x}(t)$ needs to satisfy the evolution equation described in (1), which is the solution of the direct problem. Therefore, instead of considering the evolution equation as an additional constraint of the minimization problem, one may consider it naturally inherent to the own functional to be minimized. The price that has to be paid is the inclusion of a new set of variables into the problem under study, which are simply the well known Lagrange Multipliers $\boldsymbol{\lambda}(t)$. The Lagrange multipliers $\boldsymbol{\lambda}$ here belong to the n-dimensional vector space. So, the new functional that has to be minimized $\hat{J}(\mathbf{p})$ encompasses the one defined in (3) and a new one $\hat{J}_2(\mathbf{p})$ which is defined as follows

$$\hat{J}_2(\mathbf{p}) = \int_0^{t_f} \boldsymbol{\lambda}^T[\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\dot{\mathbf{x}} - \mathbf{f}]\mathbf{dt} \tag{14}$$

Therefore the identification problem becomes the minimization of the functional $\hat{J}(\mathbf{p})$ which casts as

$$\hat{J}(\mathbf{p}) = \hat{J}_1(\mathbf{p}) + \hat{J}_2(\mathbf{p}) \tag{15}$$

The concrete definition and presentation of the adjoint problem will be possible only after the determination of the functional variation which is addressed in the next subsection.

Functional Variation

In order to perform the iterative process of parameter updating described in (4) it is clear that one has to determine the gradient of the functional $\nabla\hat{J}(\mathbf{p})$ at each iteration. The point is that the gradient determination is not an easy task since the functional depends on the system response $\mathbf{y}(t)$ which, in general, does not possess an analytic expression as a function of time t and the parameters $\mathbf{p}$. Aiming at overcoming this drawback one may determine the variation of functional $\Delta\hat{J}(\mathbf{p})$ when the parameter vector $\mathbf{p}$ suffers a variation of $\Delta\mathbf{p}$ and based on some suitable assumptions, try to extract, if it is feasible, the gradient out of this functional variation. The functional variation demands the calculation of the functionals $\hat{J}_1$ and $\hat{J}_2$ evaluated at $\mathbf{p} + \Delta\mathbf{p}$. For the functional $\hat{J}_1$ one has

$$\hat{J}_1(\mathbf{p} + \Delta\mathbf{p}) = \\ = \int_0^{t_f} [\mathbf{y} + \Delta\mathbf{y} - \mathbf{y}^E]^T[\mathbf{y} + \Delta\mathbf{y} - \mathbf{y}^E]dt = \\ \int_0^{t_f} [\mathbf{y} - \mathbf{y}^E]^T[\mathbf{y} - \mathbf{y}^E]dt +$$

$$2\int_0^{t_f}[\mathbf{y}-\mathbf{y}^E]\Delta\mathbf{y}dt \tag{16}$$

where the second order terms have been disregarded. Performing similar steps for the second functional $\hat{J}_2$ one has

$$\begin{aligned}&\hat{J}_2(\mathbf{p}+\Delta\mathbf{p}) = \\ &\int_0^{t_f}\boldsymbol{\lambda}^T[\mathbf{M}\ddot{\mathbf{x}}+\mathbf{D}\dot{\mathbf{x}}+\mathbf{K}\mathbf{x}-\mathbf{f}]\,dt + \\ &\int_0^{t_f}\boldsymbol{\lambda}^T[\mathbf{M}\Delta\ddot{\mathbf{x}}+\mathbf{D}\Delta\dot{\mathbf{x}}+\mathbf{K}\Delta\mathbf{x}]\,dt + \\ &\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{D}\dot{\mathbf{x}}\,dt+\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{K}\mathbf{x}\,dt\end{aligned} \tag{17}$$

Subtracting $\hat{J}(\mathbf{p})$ from $\hat{J}(\mathbf{p}+\Delta\mathbf{p})$ and integrating by parts the terms containing the time derivatives of the variation $\Delta\mathbf{x}$,one reaches to the variational of the functional $\hat{J}$

$$\begin{aligned}&\Delta\hat{J}(\mathbf{p}) = \\ &\mathring{A}+\int_0^{t_f}[\mathbf{M}\ddot{\boldsymbol{\lambda}}+\mathbf{D}\dot{\boldsymbol{\lambda}}+\mathbf{K}\boldsymbol{\lambda}]\Delta\mathbf{x}\,dt + \\ &\int_0^{t_f}2(\mathbf{y}-\mathbf{y}^E)^T\Delta\mathbf{y}\,dt + \\ &\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{D}\dot{\mathbf{x}}\,dt+\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{K}\mathbf{x}\,dt\end{aligned} \tag{18}$$

where $\mathring{A}$ corresponds to

$$\begin{aligned}&\mathring{A} = \boldsymbol{\lambda}(t)^T\mathbf{M}\Delta\dot{\mathbf{x}}(t)-\dot{\boldsymbol{\lambda}}^T(t)\mathbf{M}\Delta\mathbf{x}(t) + \\ &\boldsymbol{\lambda}(t)^T\mathbf{D}\Delta\mathbf{x}(t)\Big|_{t=0}^{t=t_f}\end{aligned} \tag{19}$$

It is clear that the term of $\mathring{A}$ associated to $t=0$ is null due to the initial conditions of the sensitivity problem and one may choose the Lagrange Multipliers such that it is null at $t=t_f$ as well its first time derivative inasmuch as the user has this freedom in hands. Hence the term $\mathring{A}$ containing data at the final and initial instants of time disappears from equation (18).

Considering that the variation of the output $\Delta\mathbf{y}$ has a straightforward relation with the variation of the displacement vector $\Delta\mathbf{x}$ as shown in equation (13) one may write rewrite equation (18) as follows

$$\begin{aligned}&\Delta\hat{J}(\mathbf{p}) = \\ &\int_0^{t_f}[\mathbf{M}\ddot{\boldsymbol{\lambda}}-\mathbf{D}\dot{\boldsymbol{\lambda}}+\mathbf{K}\boldsymbol{\lambda}+2\mathbf{C}^T(\mathbf{y}-\mathbf{y}^E)]\Delta\mathbf{x}\,dt + \\ &\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{D}\dot{\mathbf{x}}\,dt+\int_0^{t_f}\boldsymbol{\lambda}^T\Delta\mathbf{K}\mathbf{x}\,dt\end{aligned} \tag{20}$$

As it has already been mentioned, the goal is to obtain an expression for $\Delta\hat{J}(\mathbf{p})$ as a straightforward function of the parameter variation $\Delta\mathbf{p}$ and it is clear that it cannot be achieved in equation (20) since there is one term containing the variation $\Delta\mathbf{x}$ which is likely to have a very complicated relation with $\Delta\mathbf{p}$. In order to obtain a simpler relation between $\Delta\hat{J}(\mathbf{p})$ and $\Delta\mathbf{p}$ one can make use of an adjoint problem defined as follows

$$\mathbf{M}\ddot{\boldsymbol{\lambda}}-\mathbf{D}\dot{\boldsymbol{\lambda}}+\mathbf{K}\boldsymbol{\lambda} = 2\mathbf{C}^T(\mathbf{y}^E-\mathbf{y}) \tag{21}$$

under the following conditions

$$\boldsymbol{\lambda}(t_f)=\mathbf{0} \qquad \dot{\boldsymbol{\lambda}}(t_f)=\mathbf{0} \tag{22}$$

It should be emphasized that the problem stated by equations (21) and (22) can be changed to a problem with initial conditions rather than with final conditions with a suitable change of variables.

Gradient Equation

To obtain the gradient of the functional $\hat{J}(\mathbf{p})$ it is necessary to obtain the matrices $\Delta\mathbf{D}$ and $\Delta\mathbf{K}$ as functions of the variations of the parameters $\Delta\mathbf{p}$. This task is accomplished by expressing the damping and stiffness matrices as functions of the parameters and then evaluating their variations as follows

$$\Delta\mathbf{K}(\mathbf{p}) = \sum_{j=1}^{j=N_p}\frac{\partial\mathbf{K}}{\partial p_j}\Delta p_j \tag{23}$$

and

$$\Delta\mathbf{D}(\mathbf{p}) = \sum_{j=1}^{j=N_p}\frac{\partial\mathbf{D}}{\partial p_j}\Delta p_j \tag{24}$$

Hence, the variation of the functional $\hat{J}$ casts as

$$\begin{aligned}&\Delta\hat{J}(\mathbf{p}) = \\ &\sum_{j=1}^{j=N_p}\Delta p_j\int_0^{t_f}\boldsymbol{\lambda}^T(t)\Big[\frac{\partial\mathbf{K}}{\partial p_j}\,\mathbf{x}(t)+\frac{\partial\mathbf{D}}{\partial p_j}\,\dot{\mathbf{x}}(t)\Big]\,dt = \\ &\Delta\mathbf{p}^T\,\nabla\hat{J}(\mathbf{p})\end{aligned} \tag{25}$$

where each component of the gradient vector is given by

$$[\nabla \hat{J}(\mathbf{p})]_j = \int_0^{t_f} \boldsymbol{\lambda}^T(t) \Big[\frac{\partial \mathbf{K}}{\partial p_j} \mathbf{x}(t) + \frac{\partial \mathbf{D}}{\partial p_j} \dot{\mathbf{x}}(t) \Big] \, dt \qquad (26)$$

where $j \in \{1, 2, ..., N_p\}$ and N_p is the number of parameters that characterize the constitutive equation of the material.

Search Step Size

The search step size $\beta^{(k)}$ that appears in equation (4) is obtained through the minimization of the functional $\hat{J}_1$ at the iteration $k+1$. Accomplishing the corresponding minimization leads to

$$\beta^{(k)} = \frac{\int_0^{t_f} \Delta \mathbf{y}^T(t, \mathbf{p}^{(k)}) [\mathbf{y}(t, \mathbf{p}^{(k)}) - \mathbf{y}^E(t)] \, dt}{\int_0^{t_f} \Delta \mathbf{y}^T(t, \mathbf{p}^{(k)}) \Delta \mathbf{y}(t, \mathbf{p}^{(k)}) \, dt} \qquad (27)$$

Further details about this choice may be found in [9].

CONSTITUTIVE EQUATION

It should be emphasized that the starting point of the present technique is the constitutive equation of the material, i.e., it is out of it that one is able to define the matrices $\mathbf{K}$ and $\mathbf{D}$ as being functions of the parameters $\mathbf{p}$ that characterize the constitutive equation. For the first trial one may consider a material which possesses a simple one-dimensional localized constitutive relation between stress σ and strain ϵ that is given by

$$\sigma(x,t) = E(x)\epsilon(x,t) + G(x)\dot{\epsilon}(x,t) \qquad (28)$$

where $E(x)$ and $G(x)$ represent the elastic and the damping fields over the entire body respectively. Although, at first sight, one may consider this model quite simple, it can be used as a simple approach to characterize Functionally Graded Materials with slight viscoelastic behavior [12].

NUMERICAL ILLUSTRATIONS

Noise

In order to introduce a more realistic scenario to the simulation one may introduce some Gaussian noise to the experimental data. The level of noise in the analyzed signal can be quantified by means of the signal-to-noise ratio, which is defined as follows

$$SNR = 10 \log \frac{{\sigma_s}^2}{{\sigma_n}^2} \qquad (29)$$

where σ_s and σ_n are the variances of the signal and the noise respectively.

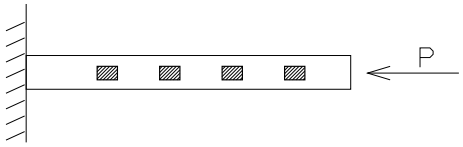


Figure 1: Virtual experiment sketch and its 4 strain sensor locations.

Examples

In order to assess the effectiveness of the proposed approach to identify mechanical system properties from a certain set of experimental data, a bar-like structure will be considered. The virtual experiment consists basically of a bar instrumented with four strain sensors along its length and which is subjected to a dynamic loading such as an impact.

A brief sketch of the virtual experiment is depicted in figure (1) and it has been chosen four equally spaced positions at which strain measures will be taken during the experiment. The properties of the bar have been chosen as follows: cross-section area $A = 2.84\text{x}10^{-4}\ m^2$, length $L = 2.03\ m$, specific mass $\rho = 4408.2\ kg/m^3$. The simulation data have been obtained from a finite element element model of the bar. The one-dimensional finite element model has 82 elements and it has been considered that a compressive force $P(t)$ has been applied at the boundary $x = 2.03\ m$ as shown in picture (2). The force $P(t)$, in Newtons, is defined as follows

$$P(t) = 125\,[1 - \cos(\Omega t)] \qquad t \in [0, T_{imp}]; \qquad (30)$$

where $\Omega = 2.52\text{x}10^5\ rad/s$ and $T_{imp} = 2.50\text{x}10^{-5}\,s$ and the impact force is zero for $t \in (T_{imp}, t_f]$. The same definition for the impact force has been used by Rusovici in [3].

All the experimental data possess 8192 points and the sampling frequency was adopted equal to $4MHz$. Here, the components of the stiffness and damping matrices are represented as follows

$$\mathbf{K}_{i,j} = \int_0^L E(x) A \frac{\partial \Phi_i}{\partial x} \frac{\partial \Phi_j}{\partial x} \, dx \qquad (31)$$

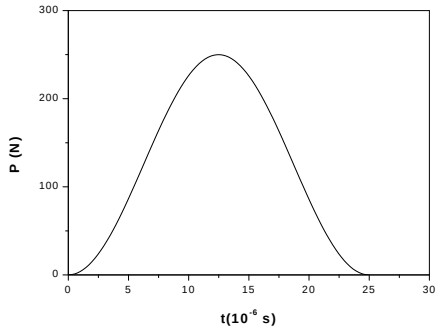


Figure 2: Impact force applied to the bar.

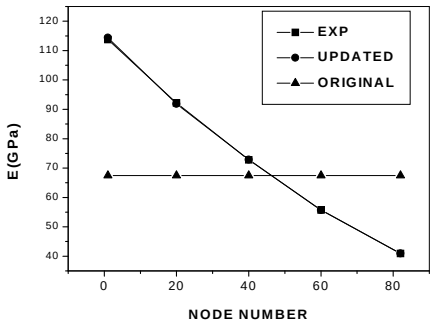


Figure 3: Elastic field for case S1.

$$\mathbf{D}_{i,j} = \int_0^L G(x) A \frac{\partial \Phi_i}{\partial x} \frac{\partial \Phi_j}{\partial x}\, dx \qquad (32)$$

where Φ corresponds to the classical Lagrangian piecewise linear shape functions and $i, j \in \{1, \ldots, N\}$ and N is the number of nodes of the finite element mesh.

For the first example (S1) it has been considered that the elastic E field is a linear distribution defined by its values at the nodes 1 (0.00 m), 20 (0.48 m), 40 (0.98 m), 60 (1.48 m) and 82 (2.03 m), which were set to be 113.8, 92.2, 72.8, 55.7 and 40.9 respectively, in GPa. The damping field G is defined similar to the elastic field and at the same nodes such that its nodal values were set to be 3.71x10^5, 3.18x10^5, 2.65x10^5, 2.12x10^5 and 1.59 10^5 in Ns/m^2. The strain sensors are located at the nodes 20 (0.45 m), 40 (0.95 m), 60 (1.45 m) and 80 (1.95 m).

It is assumed for the iteration process that the initial damping field is null and that the elastic field is uniform over the bar and its value is equal to a characteristic value that is assumed to be obtained by means of a static test on the bar. The signal-to-noise ratio adopted here is 30 dB. The result obtained for the elastic and damping fields are depicted in Fig.(3) and in Fig.(4) respectively and the term "original" refers to the original finite element model of the system.
It is clear from Fig.(3) that the elastic field has been determined quite accurately and that although the obtained damping field has some oscillations it is also an effective result. The number of iterations for this case is 75.

For the second case to be analyzed (S2) everything has been maintained equal to the first case (S1) but the damping field. The damping field has been defined at the same set of nodes as the previous example but its nodal values have been changed to: 1.59x10^5, 2.12x10^5, 2.65x10^5, 3.18x10^5 and 3.71x10^5 in Ns/m^2. The result obtained for the elastic field is graphed in Fig.(5) and the obtained damping field is graphed Fig.(6). As in the previous example the obtained elastic field has been perfectly determined and the obtained damping field has also been effective. The number of iterations for this case is 56. It should be remarked that the results presented for the two examples have been determined taking into account real-like limitations such as few measurement sensors and measured signals polluted with noise.

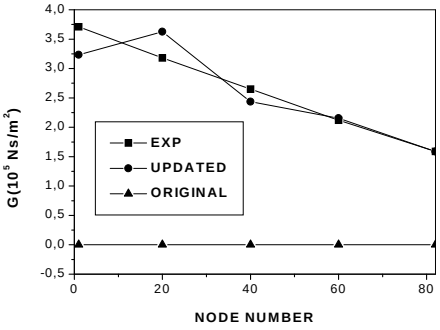


Figure 4: Damping field for case S1.

Concluding Remarks

A time domain technique aiming at identifying the unknown parameters that characterize a viscoelastic model for a certain material has been presented. In order to assess the effectiveness of the approach some simulations have been

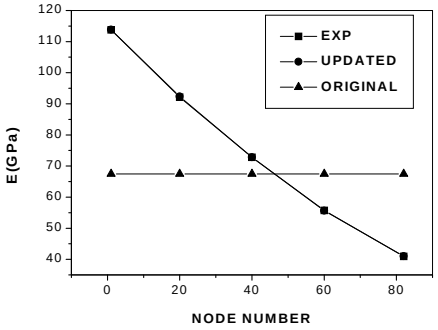


Figure 5: Elasic field for case S2.

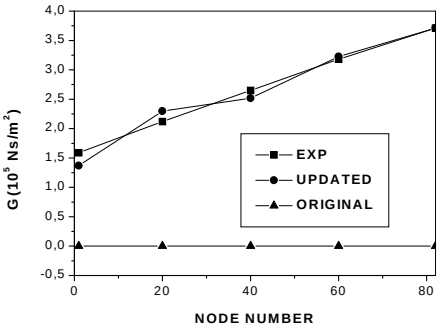


Figure 6: Damping field for case S2.

performed on a bar-like structure subjected to an impact loading. It was considered that the constitutive law of the material for stress and strain is characterized by distributed elastic and damping fields. The measured signals have been polluted with white noise to furnish more realism to the simulations and the results provided by the present approach has shown to be effective for the examples that have been analysed.

ACKNOWLEDGEMENTS

The authors are grateful to Mr. Alexandre Santos Hansen and to Mr. Rodrigo Penha Andrade Rocha for their help with WinEdt program.

REFERENCES

1.Gavrus, A., Massoni, E. and Chenot, J.L., An Inverse Analysis Using a Finite Element Model for the Identification of Rheological Parameters, *Journal of Materials Processing Technology*, vol. 60, pp. 447-454,(1996).

2.Dietrich, L., Lekszycki, T. and Turski K., 'Problems of Identification of Mechanical Characteristics of Viscoelastic Composites, *Acta Mechanica*, Vol. 126, pp. 153-167.,(1998).

3.Rusovici, R., *Modelling of Shock Wave Propagation and Attenuation in Viscoelastic Structures*, Ph.D. Dissertation, Virginia Polytechnic Institute and State University,(1999).

4.Haupt, P., Lion, A. and Backhaus, E., On the Dynamic Behaviour of Polymers under Finite Strains: Constitutive Modelling and Identification of Parameters, *International Journal of Solid and Structures*, Vol. 37, pp. 3633-3646, (2000).

5.Mossberg, M., *Identification of Viscoelastic Materials and Continuous-Time Stochastic Systems*, Ph.D. Dissertation, Uppsala University, (2000).

6.Sarron, J.C., Blondeau, C., Guillaume, A. and Osmont, D., Identification of Linear Viscoelastic Constitutive Models, *Journal of Biomechanics*, Vol. 33, pp. 685-693, (2000).

7.Janno, J. and von Wolfersdorf, L., An Inverse Problem for Identification of a Time- and Space-Dependent Memory Kernel in Viscoelasticity,*Inverse Problems*, Vol.17, pp.13-24, (2001).

8.Mossberg, M., Hillström, L. and Söderström T., Non-Parametric Identification of Viscoelastic Materials from Wave Propagation Experiments, *Automatica*, Vol. 37, pp. 511-521, (2001).

9.Özisic, M.N. and Orlande, H.R.B, *Inverse Heat Transfer: Fundamentals and Applications*, Taylor and Francis, (2000).

10.Daniel, J.W., *The Approximate Minimization of Functionals*, Prentice Hall Inc., (1971).

11.Huang, C.-H., *An Inverse Non-Linear Force Vibration Problem of Estimating the External Forces in a Damped System with Time-Dependent System Parameters*, Journal of Sound and Vibration , 242(5), 749-765, 2001.

12.Paulino, G.H. and Jin, Z.-H., *Viscoelastic Functionally Graded Materials Subjected to Antiplane Shear Fracture*, Transactions of the ASME, 68, 284-293, 2001.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ON SOME APPLICATIONS OF THE REGULARIZED MOORE-PENROSE PSEUDOINVERSION METHOD IN APPLIED GEOPHYSICS

Vesselina Iv. Dimova
Department of Geophysics, Utrecht University,
Utrecht, The Netherlands,
vessa@geo.uu.nl

ABSTRACT

The classical inverse problems, arising in gravity, magnetic and electrical prospecting are considered. The main task is reduced to seeking a solution of Fredholm's integral equation from the first kind. After appropriate discretization this integral equation is transformed into overdetermined system of algebraic equations. This is an incorrectly posed problem since the requirements for the existence, uniqueness and the stability of the solution do not hold. In this case we can not solve the problem neither by the classical mathematical methods, such as Gauss, Jordan, etc., nor by the very often used in applied geophysics method like Least Square Method, Singular Value Decomposition, Moore-Penrose pseudoinversion, etc. The ill-posedness requires the application of the classical Tikhonov's regularization, or the regularized pseudoinversion method of Moore-Penrose. The way of stating and solving the problems arose in potential-field geophysical methods is shown. An interpretation of a gravity profile over Raguba Field, Sirte Basin, Libya is performed. The comparison between the result obtained from this interpretation and the results received by the use of the most common methods used in applied geophysics is shown and appears to be interesting.

Key words: Applied geophysics, Tikhonov's regularization, Moore-Penrose pseudoinversion.

NOMENCLATURE

A - linear bounded operator; the law governing the phenomena;
Ah - operator given with error h;
A^T - transposed matrix;
$\overline{A}$ - matrix with exactly prescibed coefficients;
$\overline{A}^+$ - pseudoinverse matrix of Moore-Penrose;
h - error in setting the operator;
M - point belonging to the Earth's surface;
$M^\alpha[z]$ - Tikhonov's regularizing functional;
$\boldsymbol{p}$ - magnetization;
q - density of the electrical charges;
R^n - n- dimensional space;
U - space of the results;
u - result (vector); right-hand side of the operator equation;
u_δ - result; right-hand side of the operator equation, given with error δ ;
V - gravitational potential;
W - magnetic potential;
Z - space of the causes; space of the sought solutions;
z - causes; vector sought;
z^* - pseudosolution;
Z^* - space of the pseudosolutions;
$\bar{z}$ - normal pseudosolution;
α - regularization coefficient;
δ - error in setting the result, the right-hand side of the operator equation;
η - function relating h with δ;
φ - electrical potential;
Λ - incompatibility measure;
ρ - density of the gravitational masses;
τ - body under investigation;

INTRODUCTION

In classical mechanics and physics two paradigms dominate:

The concept of exactness, which presumes that all the quantities are prescribed accurately and that all mathematical operations are exact (Aristotle).

The concept of the determinism, according to which known causes evolve continuously into uniquely determined effects (Laplace).

In order to comment the above-mentioned paradigms, let first remind that in the exact natural sciences, including in geophysics, each phenomena is described by a cause-result relation. This relation is represented in the form of equation, or system of equations. These equations can be of any type: algebraic, differential, integral, operator equations, etc. For example let is known the body τ, in which are displayed either gravitational masses with density $\rho(x,y)$, or magnetic masses with magnetization $p(x,y,z)$, or electrical charges with density $q(x,y,z)$ (Fig. 1). Then the gravitational V, the magnetic W and the electrical Φ potentials can be represented by the expressions:

$$\int_{\tau} \frac{\rho(M)d\tau_M}{R(Q,M)} = V(Q) \tag{1}$$

$$\int_{\tau} p(M)grad_M \frac{1}{R(Q,M)} d\tau_M = W(Q) \tag{2}$$

$$\int_{\tau} q(M)\frac{d\tau_M}{R(Q,M)} = \Phi(Q) \tag{3}$$

where:

- Q is the observation point,
- M is the point which belongs to the body τ,
- $R(Q,M)$ is the distance between the points Q and M.

If we linearize (1)-(3) and generalize the problem, we can wire them in the form of operator equation

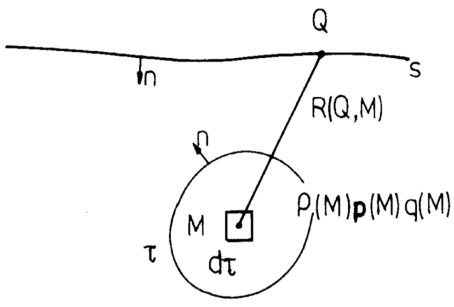


Fig. 1. Basic scheme of the potential field geophysical problem

$$Az = u \quad , \quad z \in Z \quad , \quad u \in U \tag{4}$$

where Z is the causes space, U is the results space. Both spaces are Hilbert spaces. A is the low governing the phenomena. It is a linear bounded operator, which acts from Z into U , i.e. to a given reason $z \in Z$ it prescribes a corresponding result $u \in U$.

According to the relation (4), two main problems arise:

1. Direct problem – given the reason z and low governing the phenomena A, determine the result u.

2. Inverse problem, which is also the main problem in the geophysics – given both the result u and the low governing the phenomena A, determine the cause z.

Now we can comment the above-mentioned paradigms.

- The right hand side of (4) is a result of a measurements, thus it inevitably carries measuring errors. The left hand side is also prescribed with errors, due to the rounding errors in the computation.

- One result can be caused my numerous different causes.

In the light of the above-mentioned peculiarities it arise the question: What type and which relations from the class (4) can describe real phenomena?

The answer of the above questions has been given by J. Hadamard in 1932 year. [2], [3] in the form of the following conditions:

1. The solution of the posed problem should exist;
2. The solution should be unique;
3. The solution should be stable, i.e. small changes in the data should lead to small changes in the result.

If Hadamard's requirements do hold, then the problem is correctly posed. If even one of them is broken, then the problem is incorrectly posed.

MATHEMATICAL BACKGROUND

Taking into account the aforesaid we can conclude that the inverse problem in geophysics is incorrectly posed. To comprehend the essence of the problem and to hold out a way for its solution let us suppose we have lienarized the relation (1)–(3) and that we have discretized it. Then in (4) A is a matrix and z and u are vectors. Thus we can write the system of linear algebraic equation in the form [4]:

$$\overline{A}z = \overline{u} \quad , \quad z \in R^n \quad , \quad u \in R^m \tag{5}$$

where $\overline{A}$ is a nonzero real $m \times n$ matrix. This system could not have a solution in the classical sense. However always there exist a nonempty set Z^* of the pseudosolutions, i.e. such a vectors $z^* \in R^m$ for which:

$$\left\|\overline{A}z^* - \overline{u}\right\| = \inf\left\{\left\|\overline{A} - \overline{u}\right\| : z \in R^n\right\} \tag{6}$$

If in addition the following equality holds

$$\left\|\overline{z}\right\| = \inf\left\{\left\|z\right\| : z \in Z^*\right\} \tag{7}$$

then the pseudosolution $\overline{z}$ is called normal pseudosolution, received by the Least Square Method (LSM). This is the situation when we have exact data.

Taking into account that the data are approximately prescribed, i.e. $\left\|A_h - \overline{A}\right\| \le h, \left\|u_\delta - \overline{u}\right\| \le \delta$, where h is the error in setting the operator and δ is the error in setting u, we can pose the following problem: Given the approximate data A_h, u_δ, h, δ, construct a stable in R^n approximation $z_\eta\left(\eta \equiv (h, \delta)\right)$ to the normal pseudosolution $\overline{z}$ of the system (5) : $\left\|z_\eta - \overline{z}\right\| \to 0$ when $\eta \to 0$.

The solution of the posed problem, i.e. the unique element z^α in respect to $\alpha > 0$ for given η and α is received as a solution of the variational problem

$$M^\alpha\left[z^\alpha\right] = \inf\left\{M^\alpha[z] : z \in R^n\right\} \tag{8}$$

where

$$M^\alpha[z] = \alpha\left\|z\right\|^2 + \left\|A_h z - u_\delta\right\|^2 \quad , \quad z \in R^n \quad , \quad \alpha > 0 \tag{9}$$

It is known [4] that the problem (8)–(9) is equivalent to seeking the solution of the Euler's equation

$$A_h^t A_h z^\alpha + \alpha z^\alpha = A_h^T u_\delta \tag{10}$$

which leads to

$$z_\eta^\alpha = \left(A_h^T A_h + \alpha E\right)^{-1} A_h^T u_\delta \tag{11}$$

The element (11) is a solution of the posed problem if we substitute α with the root of the equation

$$\rho(\alpha) = \left\|A_h z^\alpha - u_\delta\right\|^2 + \left(\hat{\mu}_\eta + \delta + h\left\|z^\alpha\right\|\right)^2 \quad , \quad \alpha > 0 \tag{12}$$

where

$$\hat{\mu}_\eta = \inf\left\{\left\|A_h z - u_\delta\right\| + \delta + h\left\|z\right\| : z \in R^n\right\} \tag{13}$$

Now in the light of the above outlined we will proceed with the Moore-Penrose method [5],[6],[4].

We introduce the normed spaces of the matrixes U, U^*, U_m, U_n with dimensions $m \times n, n \times m, m \times m, n \times n$ and with Euclidian norms $\|\cdot\|, \|\cdot\|_*, \|\cdot\|_m, \|\cdot\|_n$. The normal solution $\overline{z}$ of (3) for exact data $(\overline{A}, \overline{u})$ has the form $\overline{z} = \overline{A}^+ \overline{u}$. Here $\overline{A}^+$ is the pseudoinverse matrix, which is a solution of the following extreme problem: determine such a matrix $\widetilde{Z} \in U^*$, for which

$$\left\|\overline{A}\widetilde{Z} - E\right\|_m = \inf\left\{\left\|\overline{A}Z - E\right\|_m : Z \in U^*\right\} \tag{14}$$

where $E \in U_m$ is identity matrix. The posed problem can have a non-unique solution. Its unique normal solution is

$$\left\|\overline{A}^+\right\|_* = \inf\left\{\left\|\widetilde{Z}\right\|_* : \widetilde{Z} \in U_0^*\right\} \tag{15}$$

where U_0^* is the set of the solutions to the problem (14). For example, if the matrix $\overline{A}$ with dimensions $m \times n$ is a matrix with full rank, then the matrix $\overline{A}^+$ has the form
[10]:

$$\overline{A}^+ = \begin{cases} \left(\overline{A}^T \overline{A}\right)^{-1} & , \quad m \ge n \\ \overline{A}^T \left(\overline{A}\overline{A}^T\right)^{-1} & , \quad m \le n \end{cases} \tag{16}$$

Many researchers consider Moore-Penrose pseudoinversion method as an appropriate

approach for solving the main problem in the inverse geophysical theory. However the method is not fit for solving inverse geophysical problems. This is due to the fact that Moore-Penrose method is stable in respect to the errors in the right hand side in (5), but is not stable in respect to the errors in setting the matrix [8].

We will see first that the method is stable in respect to the errors in setting the vector $\bar{u}$. Let instead of $\bar{u}$ it is given the vector $\bar{u}_\delta$ such that $\|\bar{u}_\delta - u\| \le \delta, \delta > 0$. Constructing the pseudomatrix $\bar{A}^+$ we can determine the approximate normal solution of (5): $z_\delta = \bar{A}^+ \bar{u}_\delta$. Since any linear operator acting in a finite space is continuous, then $z_\delta = \bar{A}^+ \bar{u}_\delta \to \bar{z} = \bar{A}^+ \bar{u}$, i.e. the problem for finding the normal pseudosolution of the system (5) is unstable in respect to the error in setting the vector $\bar{u}_\delta$. We can determine the error in receiving $\bar{u}_\delta$:

$$\|z_\delta - \bar{z}\| \le \|\bar{A}^+\| \|\bar{u}_\delta - \bar{u}\| \le \|\bar{A}^+\| \delta \tag{17}$$

where the norm of the linear operator is determined as

$$\|\bar{A}^+\| = \sup_{x \ne 0} \frac{\|\bar{A}^+ x\|}{\|x\|} = \sup \|\bar{A}^+ x\| \tag{18}$$

Let now instead of the matrix $\bar{A}$ in (5) it is given the matrix $A_h = \{a_{ij} + h_{ij}\}$ where h_{ij} is the error in setting $\bar{A}$. The matrix A_h being a result of measuring or calculations carries inevitable errors. Let us consider the system

$$\begin{aligned} &\bar{z}_1 + \bar{z}_2 = 1 \\ &\bar{z}_1 + \bar{z}_2 = 1 \end{aligned} \tag{19}$$

and let for example

$$A_h = \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \end{pmatrix} \tag{20}$$

i.e. let us consider the system

$$\begin{aligned} &\bar{z}_1 + \bar{z}_2 = 1 \\ &(1+\varepsilon)\bar{z}_1 + \bar{z}_2 = 1 \end{aligned} \tag{21}$$

For any $\varepsilon \ne 0$ the system (21) has unique solution, which we can receive by the pseudoinversion method $\bar{z}_h = \{o, \bar{z}\}^T$. This solution tends to the exact normal pseudosolution $\bar{z}_e = \{1/2, 1/2\}^T$ when $\varepsilon \to 0$. If we estimate the error in receiving the "approximate solutions" as prescribing the error h to the matrix $A_h \left(h > 0, \|\bar{A} - A_h \le h\|\right)$ and we write all possible solutions $\bar{z}_h$, then taking the least upper bound of the deviation $\bar{z}_h$ from $\bar{z}$, we receive

$$\sup \|\bar{z}_h - \bar{z}\| = \infty \tag{22}$$

Thus it terns out that the problem for obtaining the normal pseudosolution of the system (5) is unstable in respect to the errors in the matrix, i.e. this problem is incorrectly posed. Therefore it requires regularization.

For achieving the goal we will use a technique similar to the above described, and we will be based on the smoothing functional [4]:

$$\begin{aligned} &M^\alpha[Z] = \alpha \|Z\|_*^2 + \|A_h Z - E\|_m^2 \quad , \\ &\alpha > 0 \quad , \quad Z \in U^* \end{aligned} \tag{23}$$

For any $\alpha > 0$ this functional has unique extremum

$$Z^\alpha = \left(aI + A_h^T A_h\right)^{-1} A_h^T \tag{24}$$

which realize its infimum U^*.

The coefficient $\alpha(h) > 0$ in (24) is determined as a solution of the equation (generalized discrepancy)

$$\rho(\alpha) = \|A_h Z^\alpha - E\|_m^2 - \left(\Lambda_h + h\|Z^\alpha\|_*\right)^2 \tag{25}$$

where

$$\Lambda_h = \inf \left\{ \|A_h Z - E\|_m + h\|Z\|_* : Z \in U^* \right\} \tag{26}$$

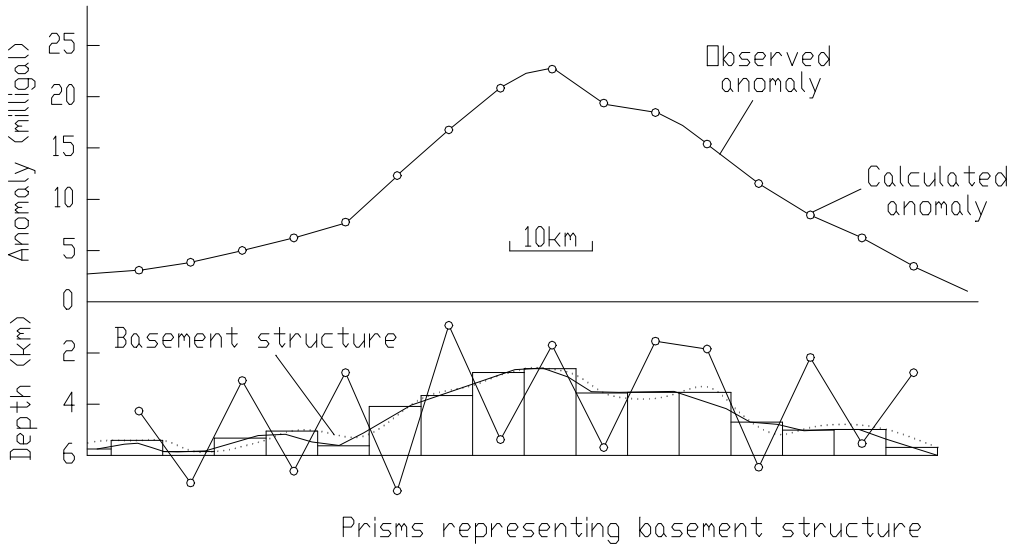


Fig. 2. Interpretation of the gravity profile over Raguba Field, Libya.

We can prove [4], that $Z^{\alpha(h)} \to \overline{A}^{+}$ when $h \to 0$.

Now the regularized approximation $Z^{\alpha(h)}$ can be used for constructing approximations to $\bar{z}$:

$$z_\delta = Z^{\alpha(h)} u_\delta \tag{27}$$

From the inequality

$$\begin{aligned} \left\| z_\delta - \bar{z} \right\| \le \left\| Z^{\alpha(h)} - \overline{A}^{+} \right\|_* \left\| u_\delta \right\| + \\ + \left\| \overline{A}^{+} \right\|_* \left\| u_\delta - \bar{u} \right\| \end{aligned} \tag{28}$$

follows the convergence

$$z_\delta \to \bar{z} \quad \text{when} \quad \eta = (h, \delta) \to 0 \tag{29}$$

EXAMPLE AND DISCUSSION

In [9] Murthy and Rao have interpreted a gravity profile over Raguba Field, Sitre Basin, Libya. The same problem was solved by using the program TANGRA, which realize the method described in this article. The level of the errors is: $h = 10^{-3}, \delta = 10^{-3}$. For the data of [9]: $H = 5.7km, \rho_2 - \rho_1 = 0.2g/cm^2$, we determined the contact surface by using the nonregularized Moore-Penrose method – the crocked line, and by the regularized Moore-Penrose method – the dotted line. The comparison of the results received in [9] - the thick line with our results (Figure 2), show that: a) The results received in [9] and the results received by the regularized Moore-Penrose method are apparently qualitatively close, but from quantitative aspect they differ significantly. In some points the differences rich up to 100-150 meters. This can be seen on Figure 2; b) The results received by the nonregularised Moore-Penrose method are so instable, that they are not acceptable for the needs of the Applied Geophysics. The reason for the malfunction of the nonregularized Moore-Penrose method and the method proposed in [9] is the fact that in these methods the error in setting the data are not prescribed. According to Leonov-Yagola's theorem without setting the errors in the data we can not solve ill-posed problems [11].

CONCLUSION

The lienarized inverse problem arisen in the gravity, magnetic and resistivity prospecting is reduced to a system of linear algebraic equations. This system is usually overdetermined. It is proposed to use the regularized Moore-Penrose method in seeking the solution to the main geophysical problem. An example about the contact problem in gravity prospecting is considered. Attention is drawn to the fact that the most common methods used in solving inverse problems in Applied Geophysics like the classical

LMS, classical pseudoinversion method of Moore-Penrose, and many other do not use the errors in the data. Due to Leonov-Yagola's theorem the results obtained by them are not reliable.

REFERENCES

1. D. Zhidarov, *Inverse gravimetric problems in gravity prospecting and geodesy*, Elsevier, Amsterdam, 1990, p. 284.

2. J. Hadamard, *Lectures on Cauchy problem in linear partial differential equation*, Yale University, New Haven, 1923, p. 351.

3. S. Banach, *Théorie des opérations linéares, Monografje matematyczne*, t.1, Warszawa, 1932, p. 254.

4. A.N. Tikhonov, A.S. Leonov and A.G. Yagola, *Nonlinear ill-posed problems*, Moscow, Nauka, 1995, p. 311 (in Russian).

5. E.N. Moore, On the reciprocal of the general algebraic matrix, *Bull. Amer. Math. Soc*., vol. **26**, p. 394-395, (1920)

6. R.A. Penrose, A generalized inverse for matrix, *Proc. Camb. Phil. Soc*.,vol. **51**, No 3, p. 406-413 (1955)

7. T.B. Yanovskaya and L.N. Porokhova, *Inverse problems in Geophysics*, Leningrad State University press, 1983, p. 210 (in Russian).

8. I.V. Kochikov, G.M. Komarshina and A. G. Yagola, *Numerical methods in oscillating spectroscopy*, Series "Matehmatics and kibernetics", **1**, Znanie, Moscow, 1989, p. 47.

9. I.V.R. Murthy and S.J. Rao, A FORTRAN program for solving gravity anomalies of two – dimensional basement structures, *Computers & Geosciences*, vol. **15**, No 7, pp. 1149-1156, (1989)

10. V.V. Voevodin and Y. A. Kuznetsov, *Matrix and calculations*, Nauka, Moscow, 1984, p. 320 (in Russian).

11. A.S. Leonov and A.G. Yagola, Can an ill-posed problem be solved if the data error is unknown?, Moscow University, Physics, vol. 50, No 4, p. 25-28.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ELECTRIC AND SEISMIC INVERSION IN ANISOTROPIC INHOMOGENEOUS MEDIA

Jörg V. Herwanger, Christopher C. Pain and Cassiano R.E. de Oliveira
Department of Earth Science and Engineering
Imperial College of Science, Technology and Medicine
London, UK
jorg@btl.net; c.pain@ic.ac.uk; c.oliveira@ic.ac.uk

ABSTRACT

We present the development, implementation and application of a multidimensional anisotropic resistivity inversion technique. We use finite elements to discretise the anisotropic Laplace equation governing the forward problem. The inverse problem is posed as an optimisation problem and is solved using a variant of the popular Marquardt-Levenberg algorithm. Inversion for the anisotropic conductivity distribution increases the number of model parameters by a factor of six and therefore the ill-posedness of the electrical inversion problem is increased. We counter this ill-posedness by introducing terms for smoothness, structural and anisotropy constraints in the error-functional.

We apply the inversion algorithm to survey data from an electric tomographic study between two boreholes at a hydrological test–site. The resulting electric inversion images are compared to the results anisotropic seismic images from same study area. Both the electric and the seismic experiment scan a depth interval of 20–115 meters between two wells spaced at 25 meters. The number of data is approximately 8000 for each survey and the subsurface in the inter-well region is discretised in elements of approximately 1.5 meters in both x– and z–directions.

A comparison of anisotropic seismic velocity distribution and electric conductivity distribution shows an amazing correlation between the two tomograms. Both methods clearly delineate an anisotropic body of highly layered and fractured siltstones underlain by an isotropic sandstone body. Zones of fractured rock and zones of highly layered sedimentary rock both result in electric and seismic anisotropy.

INTRODUCTION

Earth materials are known to exhibit anisotropic behaviour for both electric current and seismic waves [1]. Anisotropy can be caused by fine layering of sediments, aligned fractures, preferential stress direction or aligned crystals.

Electric tomograms give an image of the distribution of electrical conductivity (or resistivity) in the surveyed region and seismic tomograms give an image of seismic velocity. These images can subsequently be interpreted in terms of parameters of direct interest to engineers, geologist or hydrologists, such as shear and bulk modulus, lithology or hydraulic conductivity. Especially electric tomography has created an interest for application in hydrology due to its potential to image hydraulic flow, since pathways of hydraulic and electric flow are similar [2], and a number of papers have been published on application of resistance tomography to hydrological problems [3,4,5].

Despite the knowledge of the presence of electrical anisotropy in Earth materials since early days of exploration [6], to our knowledge no anisotropic multi-dimensional electrical inversion algorithm has been published or is commercially available. In a previous paper we have demonstrated the need for such an algorithm by showing that field data from an anisotropic test–site can not be adequately explained by isotropic models [7]. In this paper we describe the development of an algorithm that is capable of inverting data from state-of-the-art geo–electrical surveys comprised of up to 10000 datapoints and apply the newly developed algorithm to field–data from a survey at a hydro–geological test site.

In a first section we describe the geophysical crosswell experiments used to acquire data to test our new electric inversion algorithm and to benchmark its results by a seismic crosswell

tomographic study. In the next section, the development of the anisotropic electric inversion algorithm is described. Due to the increased number of inversion parameters in anisotropic inversion, the ill-possedness of the inverse problem increases and thus the inversion models become more ambiguous. Therefore special attention is given to the application of structural constraints and anisotropy constraints in conditioning the inverse problem. Finally, field data inversion models for both anisotropic electric and seismic crosswell inversions are presented.

GEOPHYSICAL CROSSWELL EXPERIMENT

This section describes an electrical and a seismic field experiment. The two experiments were carried out over the same depth interval of approximately 20–115 *m* and the two wells used in the experiments are spaced at 25 *m*. The size of the scanned region is approximately equal in size to half a football pitch. Scanning the same region with different methods allows the resulting inversion images to be compared as a method of quality control of the inversion images.

We first describe the field site and then give a detailed description of the electric and the seismic experiment.

Field Site

Electric and seismic crosswell experiments were carried out at the Reskajeage hydrological test site in Cornwall, UK. A large amount of hydrological and geological data has been previously acquired [8] and can be used to compare results of geophysical inversion models with geological and hydrological data. A series of unlined boreholes penetrate up to 300 *m* into the Mylor slates, a series of marine metasediments of Devonian age. The bedding planes of the sedimentary rock are inclined at an angle of 10-15 degrees in the direction between the two wells used in this study. There is also evidence of abundant open fractures. The highly fractured intervals correspond to hydraulically transmissive zones.

The vertical plane through the two wells used in this study is at right angles to the strike direction of both the bedding planes and the dominant fracture set, which is some justification for obtaining two dimensional crosswell resistivity and seismic data.

Electric Experiment

In geophysical DC-electric experiments current is injected into the ground by means of two electrodes located either on the Earth surface or in boreholes. The resulting electrical field is a function of the conductivity distribution in the Earth and is monitored by measuring voltages between two further electrodes. For one datum one needs to know (i) the location of the two current electrodes, (ii) the strength of the injected current, (iii) the location of the two potential electrodes and (iv) the potential difference (voltage) between the two potential electrodes.

A sketch of an electric crosswell experiment is shown in figure 1. Current is injected between an electrode in the left borehole (labelled C_1) and a remote electrode (labelled C_∞). This acquisition geometry is referred to as Pole–Pole geometry. Dashed lines schematically show the resulting electrical potential. Note that lines of constant potential are perpendicular to the Earth's surface at the point of intersection as a result of Neumann boundary conditions at the free surface.

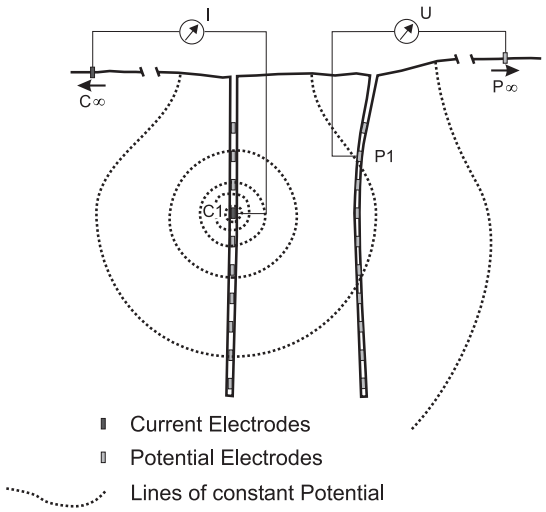


Figure 1: Sketch of electrical crosswell experiment in Pole-Pole geometry.

The instrumentation for the field experiment at Reskajeage Quarry Borehole testsite consists of two custom-built downhole electrode strings with 32 electrodes per string at 1 *m* spacing interfaced with a 64–channel fully automated digital resistivity meter.

For the electric experiment current was injected at 88 electrode locations in one borehole and the resulting potential field monitored at 87

locations in the other borehole covering a depth interval of approximately 20–105 *m* below the Earth surface. Slowly varying alternating current (using a boxcar-shaped time function) at 1 *Hz* is used. Using direct current, polarisation at the electrodes would result, an effect that is hard to quantify or correct for. Using alternating current at higher frequencies, faster acquisition would be possible, but polarisation effects (skin effect) cannot be neglected and the governing equation changes from Laplace's equation to a diffusion equation. The choice of acquisition parameters is a trade-off between accuracy and speed. Our dataset consists of 7656 datapoints and the acquisition time was 2 hours.

Seismic Experiment

In seismic crosswell experiments a mechanical pulse is generated in one well by either an acoustical, air-pulse, piezo-electric or mechanical source. The resulting disturbance travels as a seismic wave. The wavefield is monitored at discrete receiver location in a further well using hydrophones or geophones, measuring the pressure or velocity-field, respectively.

The Reskajeage seismic crosswell experiment was carried out using a 48–channel hydrophone string (using piezo–electric receivers) at half–metre spacing and a sparker source (using the "bang" generated by an electromagnetic discharge across a spark–gap as signal) with a centre frequency of 2 *kHz*, fired at one–metre depth intervals. The signals are monitored using a 48 channel digital seismograph with a 24–bit dynamic range and a sampling interval of 1/32 *ms*.

A total of 95 source positions over a depth interval from 17.5–112.5 *m* were recorded at more than 200 receiver locations ranging from 13.5–119 *m* in the opposite well, resulting in more than 20000 individual traces. The first arrival times of the signals are extracted from the seismic recordings and form the data for the inversion. Times were picked manually, using interactive computer software, in a variety of different gathers (common shot, common ray angle and common receiver gathers) and using different gain settings in order to achieve high picking accuracy and reliability.

ANISOTROPIC ELECTRIC INVERSION

This section describes the development of a novel multi–dimensional anisotropic resistivity inversion method. Electrical anisotropy on a macroscopic scale can be caused by fractures in the earth, layered strata or fibrous materials, and the conductivity distribution needs to be expressed as a spatially varying tensor. The aim of this work is to present an algorithm that can invert for this tensor. The extra non-uniqueness of the inverse solution over and above isotropic solutions is handled with spatial regularization and flexible anisotropy penalisation (model covariance). The algorithm is implemented in a robust and efficient finite element framework and uses a least squares procedure, which treats the model covariance implicitly.

Forward Problem

DC-electrical experiments, where current of strength I_s is injected at a source location r_s into a body with conductivity distribution $\underline{\underline{\sigma}}(r)$, are governed by Laplace equation:

$$\nabla \bullet \underline{\underline{\sigma}}(r) \nabla \Phi_s = I_s \delta(r - r_s) . \qquad (1)$$

The resulting potential field resulting from this source-experiment is given by Φ_s. Note, that the conductivity distribution $\underline{\underline{\sigma}}(r)$ is both inhomogeneous and anisotropic, i.e. $\underline{\underline{\sigma}}(r)$ is a function of the location *r* and at a given location dependent on the direction of observation.

In all physical experiment current can only flow in a closed circuit and thus at least two current electrodes are needed. The solution of a problem with multiple sources can be found by superposition. To make the solution of Laplace equation unique, appropriate boundary equations need to be defined. In the presented experiment, Neumann boundary conditions are applied at the Earth's surface (normal derivative of potential equals zero) and at the other boundaries of the modelling domain Dirichlet boundary conditions (potential equals a constant) are applied.

We solve Laplace equation by discretisation using finite elements with unstructured elements using 8-node hexahedral elements with tri-linear basis functions. The use of unstructured meshes allows for element–sizes to vary over the modelling domain. For example, we use a fine mesh in the vicinity of the sources where the solution (the electrical potential) varies quickly with space and a good resolution is needed. Near the border of the modelling domain, where the solution is slowly varying in space, large elements are used. This decreases the size of the employed FE–mesh and thus is computationally efficient.

The set of linear equations resulting from disretisation is solved using a preconditioned conjugate gradient solver. This allows solving very large problems with up to a million nodepoints without excessive memory requirements.

Material Properties

In this section we describe the material properties we invert for and their relationship to the conductivity tensor.

The conductivity tensor $\underline{\underline{\sigma}}$ can be described by its eigenvalues $\hat{\sigma}^1$, $\hat{\sigma}^2$ and $\hat{\sigma}^3$ and the Euler angles α, β and γ. The conductivity in terms of eigenvalues and rotations can then be written as:

$$\underline{\underline{\sigma}} = R^T \hat{\underline{\underline{\sigma}}} R \,. \tag{2}$$

The diagonal matrix $\hat{\underline{\underline{\sigma}}}$ contains the Eigenvalues $\hat{\sigma}^1$, $\hat{\sigma}^2$ and $\hat{\sigma}^3$ of the conductivity tensor as diagonal elements and the rotation matrix R contains the projections of the Eigenvectors onto the *x*-, *y*- and *z*-axis of the coordinate system. In our description of the Euler angles we follow the definitions given in [9]. The description of the conductivity tensor using Eigenvalues and rotations is instrumental in applying an anisotropy penaltiy in our inversion framework.

The numerical values found for conductivity in materials occurring naturally in the Earth vary over orders of magnitude. This could pose a problem in the inversion process, where it is advisable to use model parameters of similar magnitudes. For this reason we use the logarithm of the conductivity and the Euler angles as model parameters. We thus invert for six material properties at each node point of the FE–mesh, namely:

$$\begin{aligned} &m^1 = \ln \hat{\sigma}^1, \; m^2 = \ln \hat{\sigma}^2, \; m^3 = \ln \hat{\sigma}^3 \\ &m^4 = \alpha, \; m^5 = \beta \;\text{ and }\; m^6 = \gamma. \end{aligned} \tag{3}$$

Using the logarithm of conductivity has the added advantage of introducing positivity constraints, which is physically an entirely reasonable constraint. Using a node-based finite element description, each of the material properties $m^1, \ldots, m^6$, is sampled at the node-points and the discretized model vector **m** is of length 6 × number of nodes.

Error Functional

In order to solve the inverse problem of reconstructing electrical conductivities from observed electrical potentials we minimize the error–functional:

$$F = F_d + F_r \,. \tag{4}$$

F_d is a measure of data misfit (i.e. how good the predicted data from an inversion model matches the observed field data) and F_r forms the regularisation contribution to the error–functional.

The data misfit is calculated by summing the squared differences between observed data and the data predicted from solving the forward problem:

$$F_d = \sum_{i=1}^{NData} w_i (d_i^{obs} - d_i^{pre})^2 \tag{5}$$

The contribution of each datum *i* to this functional is additionally weighted according to the error w_i associated with this datum.

The regularisation part of the error–functional consists of three parts, penalizing structure F_r^s, anisotropy F_r^a and deviation from a desired starting model. The implementation of these penalties is discussed in the next section.

Use of Model Covariance in Error Functional

The success of the presented inversion methodology relies on the use of model covariance information. For most practical geophysical inverse problems the available data cannot uniquely determine an inversion model and the inverse problem is ill-posed. However, by including prior information or allowing only certain classes of models, unique and meaningful solutions can be found.

Proposed measures for the desired model covariance information cited in the geophysical literature include requirements on the roughness [10] of the inversion model or previous knowledge about stochastic properties of the model [11,12]. In the following, three ways of using model covariance information that are implemented in our code are discussed. Since we solve for anisotropic material properties, we introduce a penalty function that limits the amount of anisotropy allowed in the inversion model.

Structure Penalty. In order to impose structural constraints we have designed the following functional:

$$F_r^s = \frac{1}{2} \sum_{\mu=1}^{6} \lambda_s^\mu \int \nabla^T m^\mu \, \underline{\underline{k}} \nabla m^\mu \, d\Omega \,. \tag{6}$$

For each of the 6 material properties $m^1, \ldots, m^6$ a scalar product $\nabla^T m^\mu \underline{\underline{k}} \nabla^T m^\mu$ of the gradient:

$$\nabla m^\mu = (\tfrac{\partial}{\partial x} m^\mu, \tfrac{\partial}{\partial y} m^\mu, \tfrac{\partial}{\partial z} m^\mu)^T \qquad (7)$$

is calculated and integrated over the whole domain Ω. The resulting number, measuring the amount of "structure" contained in the model, is weighted by a structure penalty level (also known as Lagrange multiplier) λ_s^μ. The tensor $\underline{\underline{k}}$ needs to be positive definite in order to define a scalar product. In the simplest case, $\underline{\underline{k}}$ is an identity matrix and the scalar product reduces to the squared gradient $(\tfrac{\partial}{\partial x} m^\mu)^2 + (\tfrac{\partial}{\partial y} m^\mu)^2 + (\tfrac{\partial}{\partial z} m^\mu)^2$. However, in order to allow the gradient to be spatially and directionally variable we use a positive definite tensor function $\underline{\underline{k}}(x,y,z)$. For example, if inhomogeneous but isotropic smoothing is desired, the tensor is a diagonal, with the three diagonal elements being spatially varying, i.e. $k_{11}=k_{22}=k_{33}=f(x,y,z)$. In the most general case spatially and directionally varying smoothness constraints can be forced by this formulation. A similar functional has been proposed in [13].

Using the finite element formulation the equation for structural constraints is written in matrix form as:

$$F_r^s = \frac{1}{2}\mathbf{m}^T \mathbf{K}\mathbf{m} . \qquad (8)$$

Anisotropy Penalty. It can be useful to remove some of the ambiguity associated with electrical inversion and help the inversion algorithm to find a "good" local minimum by penalizing anisotropy. The contribution to the error–functional that achieves this takes the form:

$$F_r^a = \frac{1}{2}\lambda_a \int (m^1 \quad m^2 \quad m^3)\, \underline{\underline{a}} \begin{pmatrix} m^1 \\ m^2 \\ m^3 \end{pmatrix} d\Omega . \qquad (9)$$

The material properties m^1, m^2 and m^3, were defined in equation (3). The matrix $\underline{\underline{a}}$ has the form of a discretised Laplacian and a typical form would therefore be:

$$\underline{\underline{a}} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} . \qquad (10)$$

The anisotropy penalty level λ_a influences the degree to which anisotropy in the inversion model is penalized. For large values of λ_a the resulting inversion model is isotropic and for small values of λ_a the inversion model can be very anisotropic. Using the finite element approximation the expression for the anisotropy penalty becomes:

$$F_r^a = \frac{1}{2}\mathbf{m}^T \mathbf{A}\mathbf{m} . \qquad (11)$$

Step Length Damping. Additionally the structure **m** of an inversion model with respect to a known (or desired) structure $\mathbf{m_0}$ can be used for regularisation. If a good guess of a starting model is available, for example from detailed knowledge of the geology in the survey area, a penalty for deviation from a starting model is required. The functional that achieves this is given by:

$$F_r^l = \frac{1}{2}\sum_{\mu=1}^{6} \lambda_l^\mu \int \left(m^\mu - m_0^\mu\right)^2 d\Omega . \qquad (12)$$

The steplength penalties λ_l^μ in the directions of each of the six materials can be chosen individually. In discretised form the steplength damping becomes:

$$F_r^l = \left(\mathbf{m}-\mathbf{m_0}\right)^T \mathbf{M}\left(\mathbf{m}-\mathbf{m_0}\right), \qquad (13)$$

in which the matrix **M** is the mass-matrix of the finite element system.

Least–Squares Inversion

In our inversion program, an initial user supplied starting model is iteratively updated. The model updates are calculated using a Marquardt-Levenberg type method with additional terms for the model-covariance information. The equations for obtaining model updates **Δm** solved in each iteration step are given by:

$$\left(\mathbf{J}^T\mathbf{W}\mathbf{J} + \mathbf{C}^{-1} + \nu\mathbf{M}\right)\Delta\mathbf{m} = -\mathbf{J}^T\mathbf{W}(\mathbf{d}^{obs} - \mathbf{d}^{pre}(\mathbf{m}_{old})) - \mathbf{C}^{-1}\mathbf{m}_{old} \qquad (14)$$

In this equation **J** is the Jacobian, **W** is the data covariance matrix, **M** is the mass matrix controlling the steplength damping and **C** is the model covariance matrix. Note that $\mathbf{C}^{-1}=\mathbf{K}+\mathbf{A}$ with **K** and **A** defined in equations (8) and (11), containing the disretized structure and anisotropy information used for regularization.

The steplength damping factor ν is adjusted automatically at each iteration: ν is increased by a factor of 10 (i.e. the steplength is decreased) if either the conjugate gradient solver is not converging well, or the updated model performs worse in terms of data-misfit than the old model. If the updated model results in a smaller data-misfit and the conjugate gradient algorithm converges well, ν is decreased by a factor of 10, resulting in a larger steplength.

Matrix equation (14) is solved using preconditioned conjugate gradients (using the same solver as for the forward problem). This has the advantage, that the matrices are never explicitly formed and stored in memory. This makes the algorithm very memory efficient, allowing the solution of large-scale inversion problems (e.g. several 100000 nodepoints). The elements of the matrix are assembled when they are needed.

A full description of the algorithm, including efficient calculation of the Jacobian and computational issues is given in [14,15].

Choice of penalty levels: To find appropriate values for penalty levels λ_s^{μ} and λ_{α} for structural constraints and anisotropy constraints, respectively, we advocate running a series of inversion with penalty levels varying on a logarithmic scale. For each inversion we plot residual maps, i.e. colourcoded data–residuals as function of source and receiver position. Large values for the penalty levels clearly create correlated residuals. As the penalty levels are decreased the residual maps become less correlated. The displayed inversion images picture the smoothest and least anisotropic model for which the residual map shows uncorrelated residuals.

ANISOTROPIC SEISMIC TRAVELTIME INVERSION

The seismic inversion for anisotropic velocity models uses first arrival travel times as input data. The subsurface is parameterised with a piece-wise homogeneous medium, and 6 model parameters describe the stiffness tensor in each homogeneous region.

We have used code developed by R.G. Pratt to invert the seismic travel times into a distribution of anisotropic velocities. Details of the inversion algorithm and methodology are not included since they are well documented in a number of papers including [16,17,18]. In this study, we use the seismic inversion models as a benchmark for inversion models from our newly developed electric inversion algorithm. In [19] a detailed description of the seismic tomograms at the test-site including choice of inversion parameters is given.

Assuming a transversely isotropic (TI) medium, the 6 model parameters used to describe the stiffness tensor in each region can be mapped to (i) velocity along the axis of symmetry, (ii) the Thomsen (anisotropy) parameters ε and δ and (iii) the tilt angle of the symmetry axis with respect to the vertical. The anisotropy parameter ε measures the fractional difference between the P-wave velocities perpendicular $v_{\perp}$ and parallel $v_{\parallel}$ to the symmetry-axis

$$\varepsilon = \frac{v_{\perp} - v_{\parallel}}{v_{\parallel}}. \qquad (15)$$

The anisotropy parameter δ can be thought of as describing the shape of the wavefront of a compressional wave in a TI-medium. For ε = δ the wavefront is elliptical, whereas for ε ≠ δ the wavefront can deviate markedly from elliptical. The Thomsen parameters were introduced and are fully explained in [20] and are widely used in exploration seismology.

ANISOTROPIC GEOPHYSICAL IMAGES

In this section we present anisotropic inversion images from the Reskajeage test–site. The ability to compare the inversion images derived from anisotropic electrical inversion with (independently) calculated anisotropic traveltime tomograms confirms the quality of the electric inversion images.

Anisotropic Electric Inversion Image

Figure 2 shows anisotropic resistivity inversion images calculated using the newly developed algorithm. The left image shows average resistivities and the right image shows the reconstructed level of anisotropy in percent.

A total of 7656 datapoints, are used to invert for 6 material properties at each of the approximately 60000 nodepoints of the three-dimensional FE–mesh. Clearly, this inverse problem is severely underdetermined. In order to solve this problem we use the structural constraints described above. For example, since the acquisition geometry is essentially two dimensional (given by the x–z plane), it is reasonable not to allow conductivities to vary in y–direction. This is easily accomplished by setting k_{22} in equation (6) to a large number (here $k_{22}=10^6$), while imposing reasonable values for the smoothness in x– and z–directions (here $k_{11}=k_{33}=1$).

The nodespacing in the region between the boreholes, the region displayed in figure 2, is 1.5 *m*. At each of the nodes the arithmetic mean

$$\widetilde{\sigma} = \left(\frac{\hat{\sigma}^1 + \hat{\sigma}^2 + \hat{\sigma}^3}{3}\right) \qquad (16)$$

of the three conductivity eigenvalues is calculated. It has become habitual in the geophysical literature to plot resistivities rather than conductivities. We follow this convention and display resistivities $\rho = 1/\tilde{\sigma}$ on a logarithmic scale. The inverted resistivities range from 300 to 1000 Ωm, values that are typical for sedimentary rocks. The percentage anisotropy is calculated by evaluating:

$$\left(\hat{\sigma}^1 \quad \hat{\sigma}^2 \quad \hat{\sigma}^3\right) \underline{\underline{a}} \left(\hat{\sigma}^1 \quad \hat{\sigma}^2 \quad \hat{\sigma}^3\right)^T \cdot \tilde{\sigma}^{-1} \cdot 100\% \quad (17)$$

at each nodepoint.

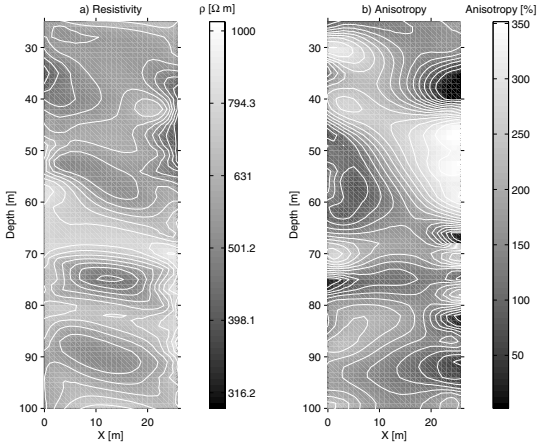


Figure 2: Anisotropic resistivity tomogram. In the left image average resistivity is displayed and on the right electric anisotropy is shown.

The top part of the tomogram shows a conductive, highly anisotropic band, dipping from the left towards the right. This observation is in accordance with the geological log showing finely layered and highly fractured siltstones. At a depth of around 65 *m* a region of high resistivity and low anisotropy is encountered. This region coincides with a well–cemented sandstone body. From 70 *m* to 80 *m* a horizontal zone of low resistivity and high anisotropy images one of the most hydraulically active fracture zones in the boreholes. Below 80 *m* the geological units become smaller. This is reflected by a complex pattern in both the resistivity and the anisotropy tomograms.

Anisotropic Seismic Inversion Image

Seismic anisotropic velocity images (Fig. 3) are used to assess the quality of the anisotropic electric images.

A total of 7440 traveltime picks served as input to the seismic inversion. The existence of significant anisotropy was clearly observed in the raw travel-time data. Steeply dipping rays arrive anomalously late and horizontal rays arrive anomalously early. The subsurface was parameterised in blocks of 1.5 × 1.5 m. Within each block, we invert for 6 parameters, resulting in a total of 8640 model parameters to be estimated.

The left image shows directionally averaged seismic velocity in each of cells in the region between the two wells. On the right seismic anisotropy ε is displayed.

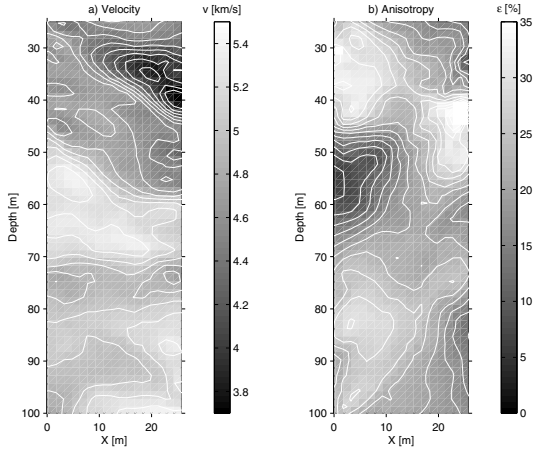


Figure 3: Anisotropic velocity tomogram. In the left image average seismic velocity is shown and the right image displays seismic anisotropy ε.

The correlation between seismic and electric average velocity and resistivity on the one hand and between seismic and electric anisotropy on the other hand is clearly obvious. Seismically fast geological units correspond to electrically resistive regions. Seismically anisotropic regions also exhibit strong electric anisotropy. The degree and mechanisms behind the correlation is subject of ongoing research.

SUMMARY AND DISCUSSION

We have successfully developed and implemented an inversion algorithm that inverts data from DC–electrical experiments into anisotropic conductivity distributions. The presented algorithm is computationally effective and can deal with datasets of approximately 10000 data points on FE–meshes with more than 100000 node points. The success of the method is

attributed to the flexible implementation of model covariance information.

We have tested the algorithm on a dataset from a geologically well-studied test-site. At the same site a seismic crosswell experiment and geological and hydrological logs could be used to benchmark the anisotropic electric inversion images. The images derived using the new electrical inversion algorithm correlate well with both, the seismic images and the geological logs. We have thus demonstrated the feasibility and necessity of anisotropic resistivity inversion.

We believe that anisotropy effects will become important in non-geophysical fields of research that are also governed by Laplace equation. We can foresee applications in medical electrical resistance tomography (ERT), where muscular tissue is known to be electrically anisotropic and inverse heat flow problems, where aligned crystals account for anisotropic thermal conductivity.

ACKNOWLEDGMENTS

We would like to thank Prof. Gerhard Pratt of Queens University, Canada, for making the seismic inversion software available. We also thank Dr. Andrew Binley of Lancaster University for invaluable assistance in the acquisition of the electrical data.

REFERENCES

1. F. S. Grant and G.F. West, *Interpretation Theory in Applied Geophysics*, McGraw Hill, New York, 1965.

2. A. Binley, Shaw, B. Shaw and H. Siobhan, *Flow Pathways in Porous Media: Electrical Resistance Tomography and Dye Staining Image Verification*, Meas. Sci. Technol., **7**, 384-390 (1996).

3. W. Daily, A. Ramirez, D. LaBreque and J. Nitao, *Electrical Resistivity Tomography of Vadose Water Movement*, Water Resources Research, **28**, 1429-1442 (1992).

4. A. Weller, M. Gruhne, M. Seichter and F. D. Börner, *Monitoring Hydraulic Experiments by Complex Conductivity Tomography*, Europ. J. Env. Eng. Geoph., **1**, 209-228 (1996).

5. L. D. Slater, *Electrical Imaging of Fractures using Groundwater Salinity Changes*, Ground Water, **35**, 436-442, (1997).

6. R. Maillet, *The Fundamental Equations of Electrical Prospecting*, Geophysics, **12**, 529-556 (1947).

7. J. V. Herwanger, C. C. Pain, A. Binley and M. H. Worthington, *Diagnosing Anisotropy in Electrical Tomography*, submitted to Geopys. Prosp.

8. N. L. Jefferies, S. Clabburn, C. Tabb, V. M. B. Watkins, and A. V. Bromley, *GroundwaterFlow at Reskajeage Quarry, Cornwall: Acquisition of Borehole Data for the NAPSAC Fracture Network Program*, AEA Technology Report AEAT/ERRA-0087, 2000.

9. G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, Academic Press, 4th edition, 1995.

10. S. C. Constable, R. L. Parker and C. G. Constable, *Occam's Inversion: A Practical Algorithm for Generating Smooth Models from Electromagnetic Data*, Geophysics, **51**, 289-300, 1987.

11. H. Maurer, K. Holliger and D.E. Boerner, *Stochastic Regularisation: Smoothness or Similarity?*, Geophysical Research Letters, **25**, 2889-2892, 1998.

12. H.F.C. Velho and F.M. Ramos, *Numerical Inversion of Two-dimensional Geoelectric Conductivity Distributions from Magnetotelluric Data*, Braz. J. Geophys., **15(2)**, 133-143, 1997

13. J. P. Kaipio, V. Kolehmainen, M. Vauhkonen, and E. Somersalo, *Inverse Problems with Structural Prior Information*, Inverse Problems, **15**, 713-729, 1999

14. C. C. Pain, J. V. Herwanger, M. H. Worthington and C. R. E. de Oliveira, *Effective Multi-Dimensional Resistivity Inversion using Finite Element Techniques*, submitted to Geophys. J. Int.

15. C. C. Pain, J. V. Herwanger and J. Saunders, *Finite Element Anisotropic Resistivity Inversion*, manuscript in preparation.

16. C. H. Chapman and R. G. Pratt, *Traveltime Tomography in Anisotropic Media – I. Theory*, Geophys. J. Int., **109**, 1-19, (1992)

17. R. G. Pratt and C. H. Chapman, *Traveltime Tomography in Anisotropic Media – II. Application*, Geophys. J. Int., **109**, 20-37 (1992)

18. R. G. Pratt and M. S. Sams, *Reconciliation of Crosshole Seismic Velocities with Well Information in a Layered Sedimentary Environment*, Geophysics, **61**, 549-560 (1996).

19. J. V. Herwanger, M. H. Worthington, R. Lubbe, A. Binley and J. Khazanehdari, *A Comparison of Crosshole Electrical and Seismic Data in Fractured Rock*, submitted to Geophys. Prosp.

20. L. Thomsen, *Weak Elastic Anisotropy*, Geophysics, **51**, 1954-1966 (1986).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

POINTWISE ESTIMATION OF THE MATERIAL PROPERTIES OF A BEAM BY ELECTRONIC HOLOGRAPHY

Dan Borza[1], Eduardo Souza de Cursi[2]
Laboratoire de Mécanique de Rouen, UMR 6138 CNRS
Institut National des Sciences Appliquées de Rouen, INSA
Avenue de l'Université, BP 08
76801 Saint-Etienne du Rouvray CEDEX, France
[1] borza@insa-rouen.fr, [2] souza@insa-rouen.fr

ABSTRACT

This work presents a method for the identification of a *field* of material properties of a beam; the aim is to characterize the value of the material properties at *each point* of the structure.

The material property is given by a *function* defined everywhere on the structure, and may have different values at different points.

The identification procedure needs a large amount of information in order to proceed: the field of displacement of the beam under a given load is experimentally obtained by electronic holography, which furnishes dense enough data.

From the numerical standpoint, both the distribution of the moments and the measured field are finite dimensionnally approximated by using Galerkin's basis and the equations of the equilibrium are used in order to calculate the values of the unknown at each node.

We shall present numerical and experimental results. The results have been confirmed by electronic shearography.

NOMENCLATURE

a	point of application of the force
C(x,y)	local contrast in the image plane
EI	material parameter to be identified
h	length of a subinterval
I(x,y)	intensity distribution of the object image with interference fringes
$I_{OBJ}(x,y)$	intensity distribution of the object image without interference fringes
M	bending momentum
n	number of nodes
npx, npy	numbers of pixels along x and y
P	force applied on the beam
x,y	coordinates in the image plane
s	particle of the beam
T	tension
u',u''	derivatives with respect to s: $u' = du/ds$; $u'' = d^2u/ds^2$
w	vertical displacement of the beam
we	Experimental displacements
ℓ	length of the beam
α	arbitrary phase shift
φ	optical phase of object wave with respect to reference wave in the detector plane
λ	laser light wavelength

INTRODUCTION

The characterization of the material properties of structures is an important field in inverse problems. Many works may be found in the literature. The particular case of beams has been considered, with a special interest in fault detection. Dynamic and static measurements have been used in this field.

We introduce in this paper an original method, based on the use of a large amount of data related to the static displacements and an accurate estimation of the third order derivative of the measured field of displacements.

We adopt a point of view analogous to those of some works on fault detection: the material parameter is treated as a field defined everywhere on the structure. The field is discretized by using its value on a set of nodes. Usually, the number of nodes is connected to the number of points of measure. In our experiments, we consider data furnished by electronic holography, which leads to a large number of nodes.

LINEAR BENDING OF A BEAM

As previously observed, a beam is a mechanical structure which can be approximated by an unidimensional continuous medium. We do not develop here the different models which can be constructed (see, for instance, [1]) and we focus on the experimental situation concerned by the identification results, presented in Figure 1. In this situation, the beam is clamped at both the extremities and modifies its geometry under the action of an external force $P > 0$, concentrated at a point a of the beam. The force is chosen so that the equilibrium of the structure may be described by a linear bending model.

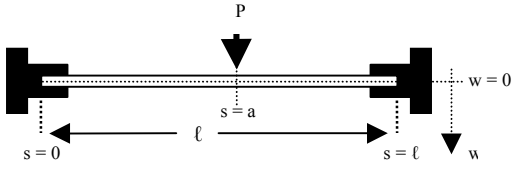


Figure 1 – The experimental situation

The linear bending model assumes that the sections of the beam may be approximatively considered as not deformed and the displacement field may be described by the single vertical displacement w. Such a model is classical and we do not present a complete description of the mechanical assumptions and approximations involved. More detailed information may be found in the litterature (see, for instance, [1]). We give below some elements concerning the model which will be used in the identification procedure.

The linear bending model

Geometrical description. The structure is described by a coordinate $s \in (0, \ell)$, where $\ell > 0$ is the length of the beam (see Figure 1). Thus each particle of the beam is associated to a point on this interval.

The vertical displacements along the beam are denoted by w: the particle s has the vertical displacement w(s).

External load. The external force P is applied at the particle s = a and causes the bending of the beam. In practical situations, the value of a is only approximately known.

Internal efforts and equilibrium. The internal efforts of the medium are given by a bending momentum M and a tension T. These unknowns are fields: $M, T : (0, \ell) \to R$.

Let us denote by the symbol ' the derivation with respect to s. The equilibrium is characterized by the equations

$$M' + T = 0 \ ; \ T' + P\delta_a = 0 \quad \text{on } (0, \ell) \qquad (1)$$

where δ_a is the Dirac's distribution concentrated at the point a. Thus, we have :

$$M(s) = -T_0(s-a) + M_0 \ ; \ T(s) = T_0 \ \text{on } (0, a) \qquad (2)$$

$$M(s) = -T_1(s-a) + M_1 \ ; \ T(s) = T_1 \ \text{on } (a, \ell) \qquad (3)$$

$$T_1 = T_0 - P \ ; \ M_1 = M_0 \qquad (4)$$

Constitutive Relation and material parameter. The bending moments are connected to the vertical displacements by

$$M = EI\, w'' \ \text{ on } (0, \ell) \qquad (5)$$

where EI is the material parameter. We consider EI as a field, $EI : (0, \ell) \to R$ and we introduce a method for its determination from measurements of w. Equations (2)-(3) yield that w" is affine on each interval where EI is constant. In addition, if we assume the EI is continuous on the neighbourhood of a, then (4) yields that

$$w''(a+) = w''(a-); \ w'''(a+) - w'''(a-) = P/EI(a) \qquad (6)$$

These properties are exploited in the sequel.

Boundary conditions. The extremities of the beam are clamped:

$$w(0) = w(\ell) = w'(0) = w'(\ell) = 0 \qquad (7)$$

IDENTIFICATION OF THE FIELD OF MATERIAL PROPERTIES

Our purpose is the determination of the field of material properties from measurements of w.

In [2], the values of M are determined by a Finite Element Method and w" is determined from the experimental data. Then (5) gives EI. Heree, we shall present an alternative approach, which has shown to be more stable for experimental data.

Discretization

The interval $(0, \ell)$ is discretized as follows: let $n > 0$ be a given integer and

$$h = \ell/n \ ; \ s_i = (i-1)h \ , \ i = 1, \ldots, n+1 \qquad (8)$$

We shall note

$$EI_i = EI(s_i) \ ; \ (w")_i = w"(s_i) \ ; \ M_i = M(s_i) \qquad (9)$$

Thus, we have

$$M_i = EI_i (w")_i \ , \ i = 1, \ldots, n+1 \qquad (10)$$

Let us introduce $I_i = (s_i - h/2, s_i + h/2) \cap (0, \ell)$. The field EI is approximated by a piecewise constant function on each I_i.

$$EI(s) \approx EI_i \quad \text{on} \quad I_i \qquad (11)$$

Identification

As previously remarked, equations (2)-(3) show that M is an affine function (for instance, we have $M' = -T_0$ on $(0, a)$). Thus, the approximation (11) implies that w''' is an affine function on each subinterval I_i. Thus,

$$w'''(s) \approx (w''')_i \quad \text{on} \ I_i \qquad (12)$$

Let ip verify $s_{ip} \le a < s_{ip+1}$. We have

$$(w''')_i = T_0/EI_i \ , \ i \le ip-1 \ ; \qquad (13)$$

$$(w''')_i = T_1/EI_i \ , \ i \ge ip+1 \ . \qquad (14)$$

Once the values of T_0 and T_1 have been determined, these equations are used for the determination of EI. These values can be determined by assuming that EI is constant in the neighborhood of a. Thus,

$$EI(a) = \frac{P}{w'''(a+) - w'''(a-)} \approx \frac{P}{w'''_{ip+1} - w'''_{ip}} \qquad (15)$$

$$T_0 = -EI(a) w'''_{ip} \quad ; \quad T_1 = -EI(a) w'''_{ip+1}$$

The use of this method implies the numerical evaluation of w'''. In practice, we have the values of the measured displacements $we_i = we(s_i)$ on a region $s_{min} \le s \le s_{max}$, corresponding to $i_1 \le i \le i_2$. In the sequel, we consider such a measured data and we introduce a filtering method for the determination of w'''.

Numerical approximation of w'''

As previously observed, the measurements will be affected by experimental noise. Thus, we must introduce an adapted procedure for the evaluation of the derivatives w'''. The numerical filtering below is based on dynamical programming and has been introduced in [3]. A more detailed presentation can be found in this reference. Let us note

$$z^1 = w \ ; z^2 = w' \ ; z^2 = w" \ ; z^3 = w"'; \qquad (16)$$

$$u = w"" \qquad (17)$$

We set $z = (z^1, z^2, z^3, z^4)$;

$$z_i = z(s_i) \ ; \ u_i = u(s_i) \qquad (18)$$

Then

$$z_{i+1} = Az_i + Bu_i \ ; \ z_{i_1} = \theta \qquad (19)$$

where

$$A = \begin{pmatrix} 1 & h & h^2/2 & h^3/6 \\ 0 & 1 & h & h^2/2 \\ 0 & 0 & 1 & h \\ 0 & 0 & 0 & 1 \end{pmatrix} \ ; \quad B = \begin{pmatrix} h^4/24 \\ h^3/6 \\ h^2/2 \\ h \end{pmatrix} \qquad (20)$$

Let us introduce

$$u = (u_{i_{ini}}, \ldots, u_{i_{fin}}) \qquad (21)$$

and denote by $\{z_i(u, \theta)\}_i$ the sequence defined by (19). We consider

$$J(u, \theta) = \sum_{i=i_{ini}}^{i_{end}} \left(we_i - z_i^1(u, \theta) \right)^2 + b_{reg} \sum_{i=i_{ini}}^{i_{end}} u_i^2 \qquad (22)$$

where $b_{reg} > 0$ is a given parameter and we denote by (u^*, θ^*) the solution of

$$(u^*, \theta^*) = \text{Arg Min } J \qquad (23)$$

We shall approximate

$$(w''')_i \approx (z^3(u^*, \theta^*))_i \qquad (24)$$

EXPERIMENTAL SETUP

We have realized a structure corresponding to the model by using a steel plate clamped on its two opposite sides with an out-of-plane force applied at the center of plate, as shown in Figure 2. The dimensions of the plate are lx × ly × lz, where lx = 0.15 m, ly = 0.03 m, lz = 0.001 m. The force was applied through a piezoelectrical actuator having a spherical cap. Between the PZT ceramic stack and the spherical cap a calibrated force measuring transducer was mounted. The whole system is positionned so that the applied force lies in a horizontal plane

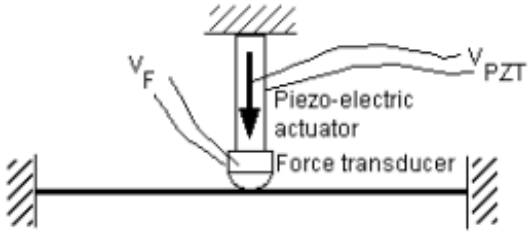


Figure 2. Loading mechanism

Besides the force transducer calibration, a preliminary series of tests was completed in order to check the linearity between the voltage applied to the piezoelectric actuator and the maximum value of the corresponding displacement map produced. Nine different values of the applied voltage had as results different holographic fringe patterns. Figure 3 presents the graph relating the applied voltage and the maximum number of fringes on the corresponding interferogram.

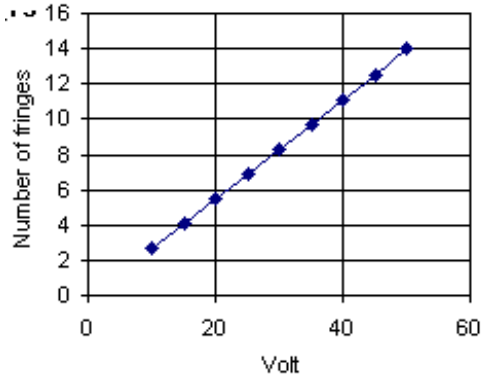


Figure 3. Voltage – displacement relationship

Three of these interferograms, for increasing values of the applied voltage, are presented in Figure 4.

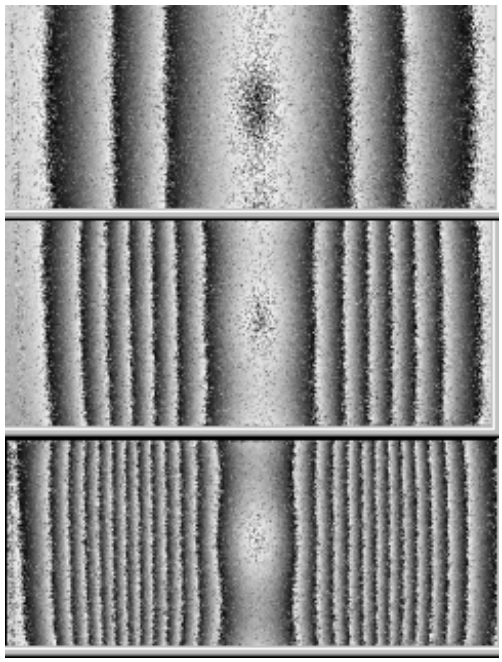


Figure 4. Three interferogram samples

The interferograms and the full-field displacement map at the surface of the tested plate are obtained by electronic holography (Figure 5).

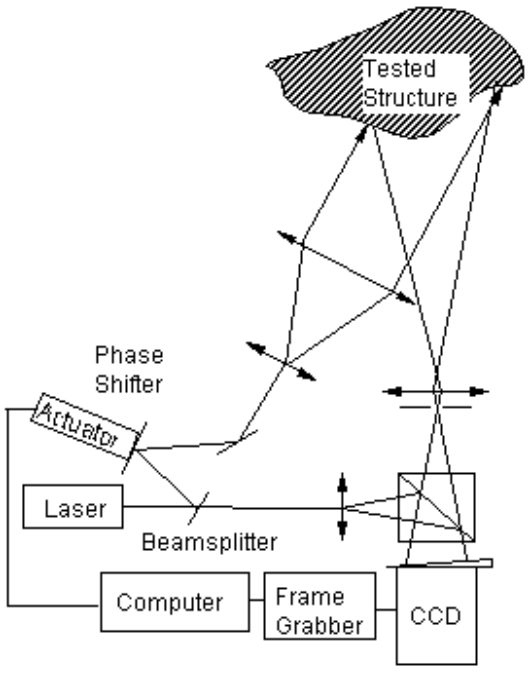


Figure 5. Holographic setup

The holographic system includes a frequency-doubled YAG laser in a typical 4-frames phase-stepped electronic holography configuration.

In order to obtain the full-field displacement data the plate clamping and loading mechanism is set in place and four reference frames are acquired, by applying a staircase voltage to the piezoelectric actuator.

The voltage step is chosen so as to produce a phase shift close to $\pi/2$ between the reference wave and the object wave. The intensity distribution corresponding to these images are given by relation (25):

$$I_i(x,y) = I_{OBJ}(x,y)\{1 + C(x,y)\cos[\varphi + \alpha]\}$$
$$\alpha = (i-1)\frac{\pi}{2}, i = 1,2,3,4 \quad (25)$$

The reference state is then recorded in the computer memory by calculating the two differences C_0 and S_0, given by eqs. (26) and (27):

$$C_0 = I_1 - I_3 = 2C(x,y)I_{OBJ}(x,y)\cos\varphi \quad (26)$$

$$S_0 = I_4 - I_2 = 2C(x,y)I_{OBJ}(x,y)\sin\varphi \quad (27)$$

After applying the desired force to the plate, the plate bends and the mutual phase between the object wave and the reference wave becomes $\varphi+\Delta\varphi$. Another four-frames bucket J_i is then acquired while the staircase voltage is being applied to the piezoelectric actuator. The expressions of these images are given by eq. (28):

$$J_i(x,y) = I_{OBJ}(x,y)\{1 + C(x,y)\cos[\varphi + \Delta\varphi + \alpha]\}$$
$$\alpha = (i-1)\frac{\pi}{2}, i = 1,2,3,4 \quad (28)$$

The two new differences C_d and S_d corresponding to the deformed plate are then computed; they are described by eqs. (29) and (30).

$$C_d = J_1 - J_3 = 2C(x,y)I_{OBJ}(x,y)\cos(\varphi + \Delta\varphi) \quad (29)$$

$$S_d = J_4 - J_2 = 2C(x,y)I_{OBJ}(x,y)\sin(\varphi + \Delta\varphi) \quad (30)$$

Phase Map

The modulo 2π phase difference corresponding to the plate deformation between the reference state and the actual state is calculated as:

$$\Delta\varphi = \operatorname{atan}\frac{S_0C_d - C_0S_d}{C_0C_d + S_0S_d} \quad (31)$$

It is illustrated in Figure 6. The force value is 1 mN.

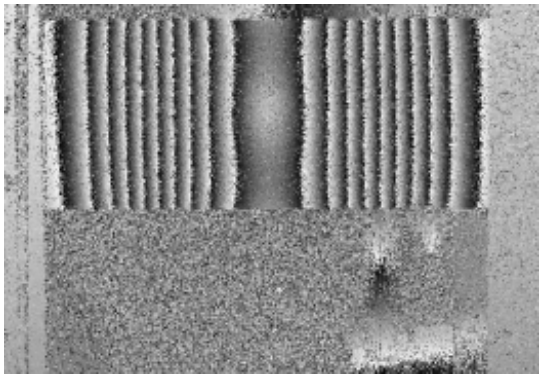


Figure 6. Phase map at the plate surface

The full-field displacement-related phase map is obtained, after applying an edge-preserving smoothing filter and appropriate masking, by phase unwrapping, as shown in Figure 7.

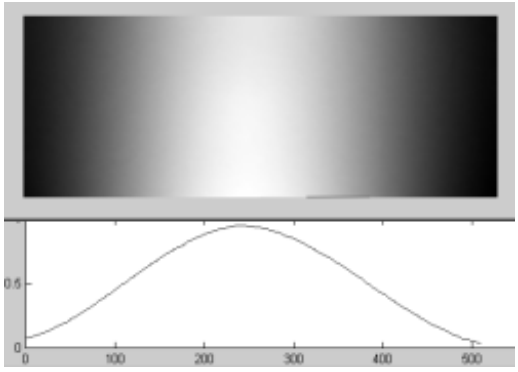


Figure 7. Unwrapped phase and profile

Displacement map

The normal displacement d(x,y) at a location (x,y) on the plate is related to the phase map $\Delta\varphi(x,y)$ by the approximate relation:

$$d(x,y) = \frac{4\pi}{\lambda}s(x,y)\Delta\varphi(x,y) \quad (32)$$

In equation (32) s(x,y) represents the sensibility vector at the current plate point (x, y). To provide

the necessary accuracy, the sensibility vector value variation across the plate surface has been calculated and taken into account, although it only exhibits small variations (between 0.985 and 0.992) across the object field.

After the phase unwrapping, the resulting phase map is spatially corrected for the perspective and lens distorsions; taking into account the extrinsic and intrinsic camera parameters. About 1.7 % of the total number of pixels situated near the clamped sides were excluded from this procedure, most of them on the right side of the image, because of the shade of the clamping device. Thus, the final map concern a part of plate such that $x_{min} \leq x \leq x_{max}$. The values of x_{min} and x_{max} are evaluated from the number of pixels excluded.

The final displacement map is scaled to a size of npx × npy pixels and provided as a binary matrix $D = (d_{ij})$. We have

$$d_{ij} = d(x_i,y_j),\ \ x_i = x_{min}+(i-1)hx,\ y_i = (j-1)hy\ , \quad (33)$$

$$hx\ = (x_{max} - x_{min})/npx\ ;\ hy = ly/npy\ . \quad (34)$$

hx and hy are respectively the horizontal and vertical length corresponding to one pixel. The matrix D is transmitted to the material properties identification computing system. The experiment uses npx = 646, npy = 248.

Displacement Derivatives. To check the consistency of the holographically obtained displacement field over the plate surface the horizontal derivative of the displacement map was numerically computed. The results have been compared with the experimentally obtained derivative field, obtained by electronic shearography.

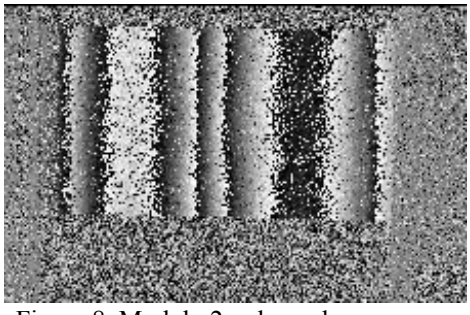


Figure 8. Modulo 2π phase shearogram

Figure 8 shows the initial modulo 2π phase shearogram, and Figure 9 shows the corresponding unwrapped phase and the profile of a horizontal line.

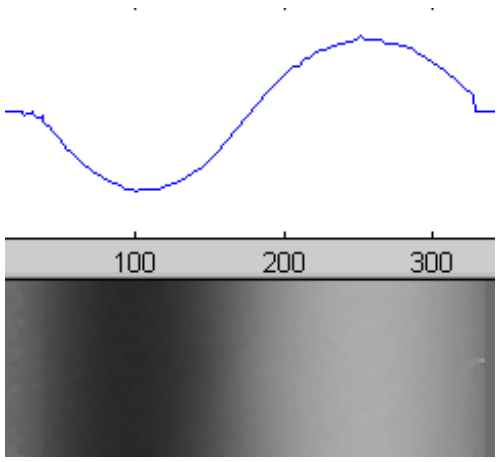


Figure 9.Unwrapped phase from shearogram

APPLICATION OF THE METHOD TO THE EXPERIMENTAL DATA.

We take s = x. The values of we_i are obtained from D by considering the mean value of each column:

$$we_i = \frac{1}{npy}\sum_{j=1}^{npy} d_{ij} \quad (35)$$

The values of x_{min} and x_{max} define the indices i_1 and i_2 :

$$s_{i_1} \leq x_{min} < s_{i_1+1}\ ;\ \ s_{i_2-1} < x_{max} \leq s_{i_2}\ . \quad (36)$$

The minimum of we_i gives the index ip.

$$we_{ip} = \min\ \{we_i\ ,\ i_1 \leq\ i \leq i_2\} \quad (37)$$

and the value of a is approximated by $a \approx s_{ip}$. w''' is evaluated by using (24). We present below the results for P = 0.2 N. Figure 10 shows the values of z^1, which estimates the displacement w. The result is coherent with the experimental data in Figure 7.

The values of z^2, which corresponds to the derivative w' are shown in Figure 11. The result is in coherence with the independent shearographic data in Figure 9.

The values of z^3 and z^4 which estimate w" and w"', respectively, are shown in Figures 11 and 12.

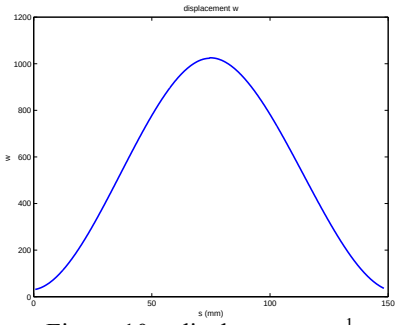


Figure 10 – displacements z^1

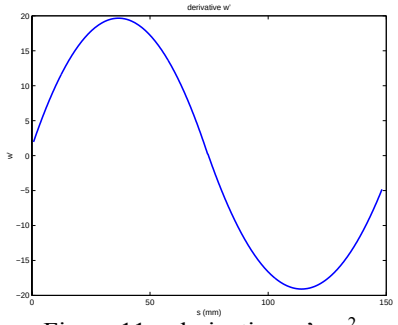


Figure 11 – derivative w'= z^2

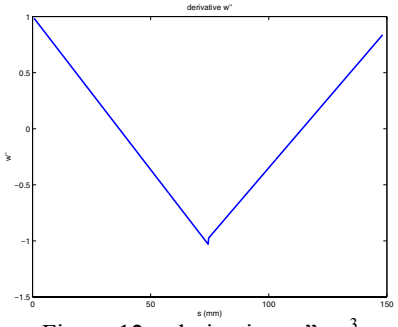


Figure 12 – derivative w"= z^3

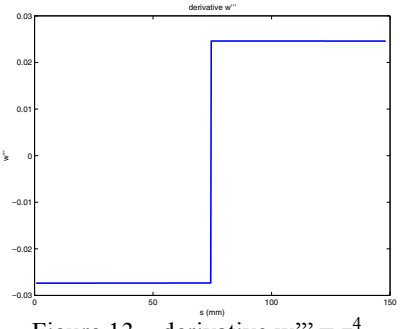


Figure 13 – derivative w"' = z^4.

The values of EI_i obtained are shown in the Figure 14.

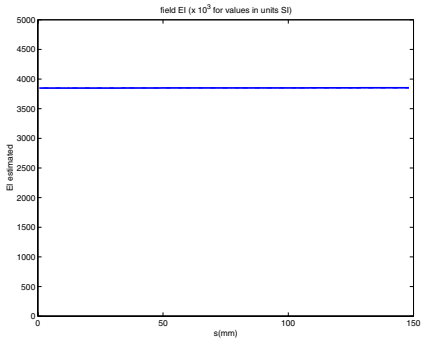


Figure 14 – Field EI.

The values of EI obtained are in accordance with the known values for a steel beam of the given rectangular section. We have EI_{mean} = 3850 (10^3 units SI).

The mean value has been compared to the value obtained by the identification of a single parameter EI for the whole beam. This corresponds to the special case where the field is constant. In this case, the displacements are given by two polynomials of the third degree:

$$w(s) = \begin{cases} a_0(s-a)^3 + b_0(s-a)^2 + c_0(s-a) + d_0 & , s < a \\ a_1(s-a)^3 + b_1(s-a)^2 + c_1(s-a) + d_1 & , s > a \end{cases} \quad (38)$$

The coefficients of each polynomial may be determined by using the data and the value of EI will be given by EI = P/(6(a_1-a_0)). This method gives EI = 3868 (10^3 units SI), what is in accordance with the preceding calculations.

The value of b_{reg} used has been $b_{reg} = 10^{10}$.

LOCALIZATION OF A BEAM DEFECT.

In this experiment we shall consider a beam having a defect represented by a local modification of the value of EI.

We consider a beam of length d = 0.15 m, with a force P = 0.2 N applied on its middle point a = 0.075. We assume that EI = 3800 ($\times 10^3$ units SI) along the beam, except in the region $0.025 \le s \le 0.0275$, where EI = 3000 ($\times 10^3$ units SI).

The field of moments M(s) is generated by using $M_1 = M_0 = -0.0375$; $T_0 = 0.1$; $T_1 = -0.1$. The exact value w corresponds to these parameters. We generate the experimental field by adding a random noise ε corresponding to 10 % of w: we = w + ε. We generate npx = 600 points.

The values of we are shown in Figure 15. Figure 16 shows the values of EI obtained with b_{reg}=0.01. The horizontal line gives the value of EI: the values obtained correspond to the exact value, except on the defective region. The defect is localized. The value of EI on the defective region may be estimated from the values of z^3 = w". The mean value of the calculated values of EI on (a, ℓ) is ei_{mean} = 3800 ($\times 10^3$ units SI). In Figure 17, we present the region where the values of EI differ from ei_{mean} by more than 10 %: it corresponds exactly to defective region.

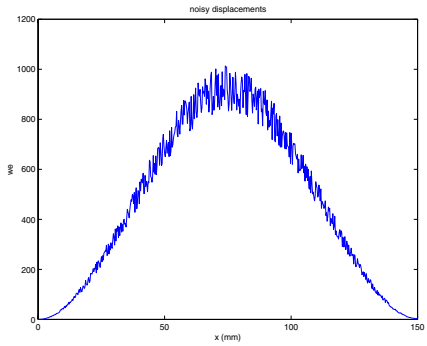


Figure 15. Noisy data for the displacements

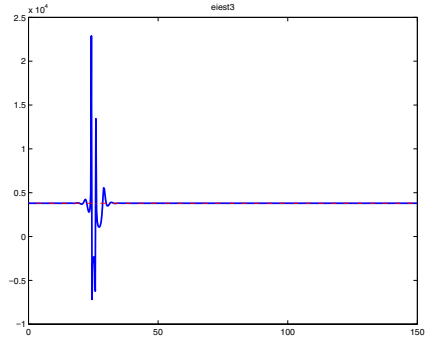


Figure 16 – Values of EI

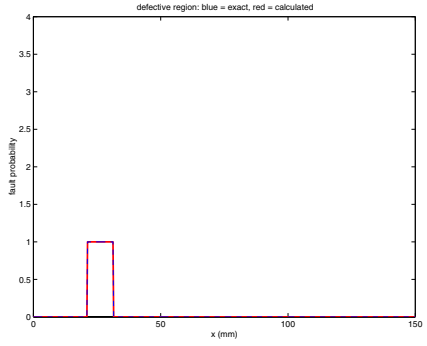


Figure 17 – Estimation of the defective region

Local information about the values of EI in the default zone may be obtained from z^3 = w", as shown in Figure 18.

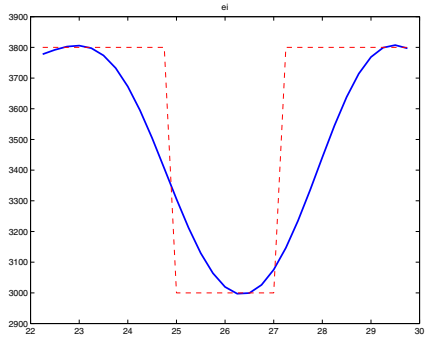


Figure 18 – EI obtained by using z^3 = w"

CONCLUDING REMARKS.

We have presented a method for the identification of a field of material properties of a beam. The material property is given by a function defined everywhere on the structure, and may have different values at different points. From the numerical standpoint, the field is discretized at a set of nodes (typically 600 ones). The method is based on the use of the third order derivatives of the displacements, which are calculated by an adapted filtering procedure.

The identification method has been tested on data experimentally obtained by electronic holography. It has also been tested on numerically generated data representing a defective beam. In both situations, it has shown to be effective for calculations. In addition, the results of the experimental case have been confirmed by comparison with the results furnished by electronic shearography.

Improvements and development will be matter of future work: extension to models of plates, simultaneous use of the second and third order derivatives.

REFERENCES.

1. J. Mandel, *Mécanique des Milieux Continus*, Gauthier-Villars, Paris, 1966

2. L. Ketata, *Etude et Identification du comportement macroscopique de structures en matériaux anisotropes à l'aide d'essais d'optique cohérente*, Ph. D. Thesis, University of Rouen, 1999.

3. D. M. Trujillo and H. R. Busby, *Practical Inverse Analysis in Engineering*, CRC Press, New York, 1997

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

INVESTIGATIONS ON DEFECT IDENTIFICATION IN METALLIC WALLS USING ARTIFICIAL NEURAL NETWORK AND THE FINITE ELEMENT METHOD

Naasson P. de Alcântara Jr. Alexandre M. de Carvalho Alfredo J. C. Ulson
Department of Electrical Engineering DEE/FEB
São Paulo State University / Unesp
Bauru, SP, Brazil
naasson@feb.unesp.br

ABSTRACT

This work presents an investigation on the use of the finite element method (FEM) and artificial neural networks (ANN) for the identification of defects on metallic walls (pipelines, metallic vessels, large metallic structures, etc.), due to the aggressive actions of the fluids contained by them, and/or atmospheric agents. The methodology used in this study consists in the simulation of a large number of defects in a metallic wall, considering its geometry and magnetic characteristics, by the finite element method. Both variations in the width and height of the defects are considered. Then, the obtained results are used to generate a set of vectors for the training of a perceptron multilayer artificial neural network. Finally, the obtained neural network is used to classify a group of new defects, simulated by the finite element method, but not belonging to the original dataset. The results on the simulated defects seem to support the proposed method, and encourage future works on this subject

INTRODUCTION

Metallic walls constitute important components of many kinds of industrial plants, such as gas pipelines, chemical pipelines, fuel vessels, sugar and alcohol plants etc. Generally, these walls are subject to the aggressive (corrosive) actions by the fluids contained by them, or even by atmospheric agents. So, these equipments must be periodically evaluated, in order to avoid operational interruptions and/or dangerous accidents. Usually, these evaluations are done using non-destructive techniques. Such techniques may involve the use of electromagnetic fields, which are induced in the metallic walls of the equipment under inspection.

More common techniques used in the inspection of metallic walls are based on eddy current systems. In this kind of analysis, electromagnetic devices are excited by an alternating current of a given frequency that induces a flow of eddy currents in the material under test. As the probe passes over the defect, the same causes a change in the flow of eddy currents. These changes are then detected by electronic sensors. The changes are generally proportional to the depth of the defect. So, we can estimate the depth of the defect by proper electronic calibration. Relative motion between the test probe and the material being inspected is a requirement of this type of inspection. Although the probe can be hand held as the piece under test is examined, this method is usually too slow, and unreliable. A very interesting alternative, introduced by Low, [1], is the use of the Finite Element Method (FEM) in conjunction with Artificial Neural Networks (ANN) for solving this kind of inverse problem.

In this paper we present an investigation on the use of FEM and ANN in the identifications of defects in metallic walls. The methodology consists of the following steps:

1. A large number of defects in a metallic wall is simulated using the finite element method.

2. The obtained results are then used to generate the training vectors for a multilayer perceptron artificial neural network.

3. The trained network is used to classify new defects in the wall, which not belong to the original dataset.

4. The network weights can be embeded in an electronic device, and used to identify defects in real pieces, with characteristics similar to those of the simulated ones.

For the methodology presented here, the measured values are independent of the relative motion between the probe and the piece under test. In other words, the movement is necessary only to change the position of the probes, to acquire the fields values, which are necessary to the construction of the defect pattern. Furthermore, the use of neural networks in conjunction with the finite element method permits a very good determination of both, width and height of the defect.

The kind of defect we have investigated in this work is corrosion on the inner surface of metallic tubes. For the purpose of the paper, the defects were classified in large, medium and small. The dataset was generated considering variations on width and height, resulting in approximately 550 finite element simulations.

THE FINITE ELEMENT METHOD IN THE ELECTROMAGNETIC FIELD ANALYSIS

In this section we present a brief resume of the application of the finite element method in magnetostatic field problems.

Two-dimensional magnetostatic field problems are described by the quasi-Poisson equation :

$$\frac{\partial}{\partial x}\left(\nu\frac{\partial A}{\partial x}\right)+\frac{\partial}{\partial y}\left(\nu\frac{\partial A}{\partial y}\right)=-J \qquad (1)$$

where :

A = is the magnetic potential vector, here presented as a scalar quantity, in A/m
J = density of current, in A/m^2.
ν = is the inverse of the magnetic permeability, μ.

The magnetic potential vector $\vec{A}$ does not have any physical meaningful, but it is a mathematical function used to obtain the magnetic flux density B.

Equation (1) has no analytical solution. So, its solution must be numerical, and the most popular technique for this kind of solution is the finite element method (FEM).

In terms of calculus of variations, the magnetostatic field problem can be formulated in terms of a functional of energy :

$$F=\iint\left(\int \nu B dB - J.A\right)dxdy \qquad (2)$$

where $B=\nabla\times\vec{A}$.

Minimization of (2) is done by proposing an approximating function for the magnetic potential vector, that is :

$$A(x,y)=\sum_{i=1}^{n}\phi_i.A_i \qquad (3)$$

where A_i is the value of the magnetic potential at the nodes of the finite element, and ϕ_i are the shape functions. For the first order triangular element (the element used in this work, and showed in figure 1), ϕ_i is :

$$\phi_i=\frac{a_i x+b_i y+c}{2\Delta} \qquad (4)$$

where the coefficients a_i, b_i and c_i are dependent of the node positions, and Δ is the area of the triangle.

The minimization is done substituting (3) in (2), and taking its derivatives in relation to the magnetic potential in the nodes.

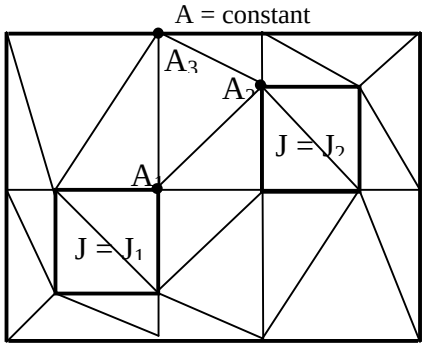


Figure 1 – Free representation of finite elements

For the magnetostatic case, after minimization, we have for each element the following 3x3 algebraic system of equations :

$$\frac{\nu}{4\Delta}\begin{pmatrix} b_1b_1+c_1c_1 & b_1b_2+c_1c_2 & b_1b_3+c_1c_3 \\ b_2b_1+c_2c_1 & b_2b_2+c_2c_2 & b_2b_3+c_2c_3 \\ b_3b_1+c_3c_1 & b_3b_2+c_3c_2 & b_3b_3+c_3c_3 \end{pmatrix}\begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}=\frac{\Delta}{3}\begin{pmatrix} J \\ J \\ J \end{pmatrix} \qquad (5)$$

Combining all the elementary matrices, we have the global system of equations :

$$(S)(A)=(R) \tag{6}$$

More details about the finite element theory can be found in [2] and [3].

THE METHODOLOGY FOR DEFECT IDENTIFICATION

First of all, an electromagnetic device was idealized to be used as an electromagnetic field exciter (Figure 2). In this paper, we have considered direct current in the coils. So, the material of the metallic wall must be ferromagnetic. Very low frequency currents in the coil must be used for non-ferromagnetic materials, and these will be studied in future works.

Surface swapping with the above described electromagnetic device are done to take deviations of the magnetic induction at equally stepped points on the external surface of the wall.

Each position of the swapping, with each defect dimensions, correspond to one simulation with the finite element program. In order to generate the training vectors for the neural network, a large number of defect shapes must be simulated. In this work, 40 defects have been simulated, with 13 positions of the sensor. So, more than 500 finite element simulations were done.

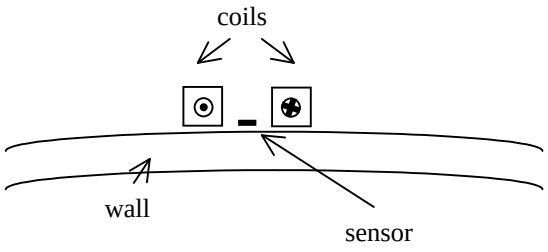


Figure 2 – Arrangement for the measurements

Figures 3a and 3b show the steps of the methodology used in this work.

Steps 1-4 correspond to the finite element analysis of the defects. In this work we used a 2D finite element program to simulate the defects in a metallic wall. Extensions of this work include the use of 3D finite element programs.

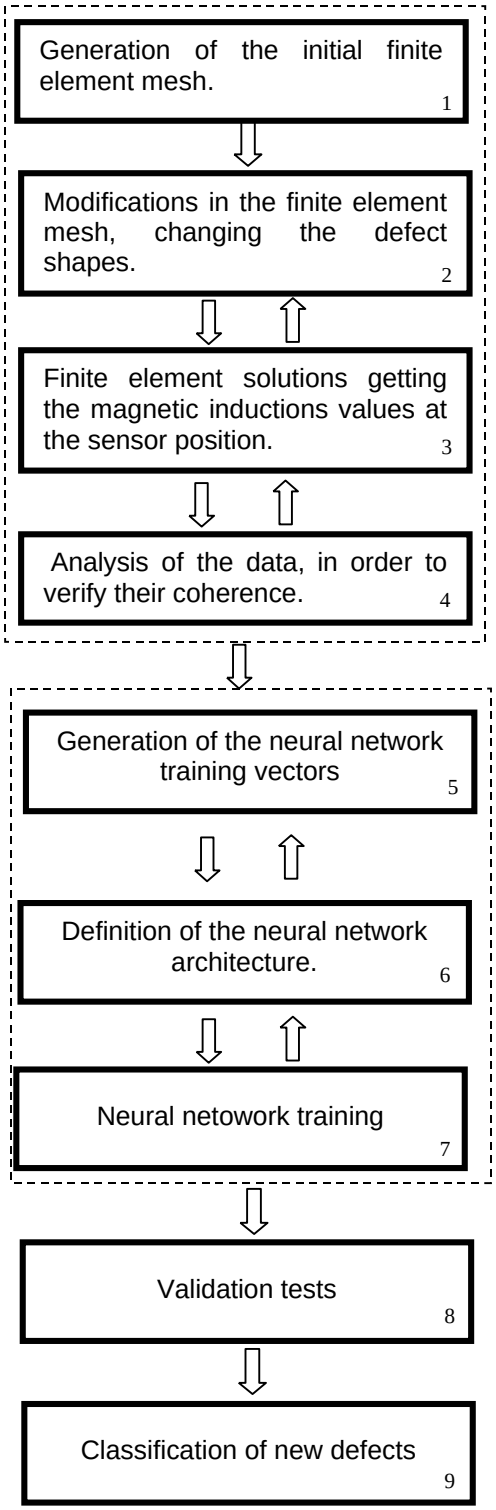


Figure 3 – Flowchart of the Used Methodology

The simulations were done for a hypothetic high pressure vessel, with 1500 mm of diameter and 10 mm thick. The material of the vessel is 1006 Steel (a magnetic material), and the permeability of the defects was set to the permeability of the air. Finite element meshes with 36000 elements and 18000 nodes, approximately, were used in the simulations. Figure 4 shows a field distribution for one of these simulations.

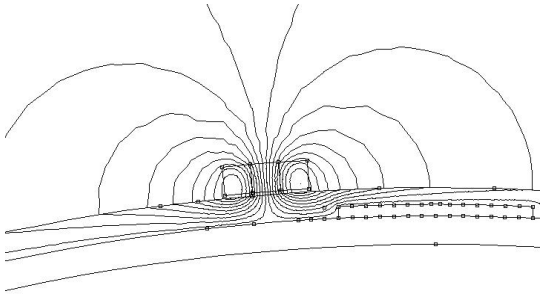


Figure 4 – Field mapping for a finite element simulation

During the phase of finite elements simulations, errors can appears, due to its massively nature. So, the results of the simulations must be carefully analyzed. This can be done, for instance, plotting in the same graphic the magnetic induction deviations for a set of defects. Figure 5 shows the deviation on the magnetic induction in the space between the coils for a set of defects, having the same height (4 mm), and width ranging from 9.87 mm to 98.24 mm, and with the sensor fixed at the middle of the defect.

In the step 5, we generate the training vectors for the neural network. In this work, we generated 40 vectors with 25 elements each one, with mirror symmetry in relation to the 13th element. Figure 6 and 7 shows the the graphics for two of these vectors, for the heights of 4 mm and 1 mm respectively. In this graphic, and all subsequent ones, magnetic inductions values are at vertical axes, and length are at horizontal axes.

For the purpose of training and classification, the defects were identified by : initial (number 1), serious (number 2) and critical (number 3). From the original 40 vectors, 36 vectors were used in the network training, and 4 vectors were used in their validation.

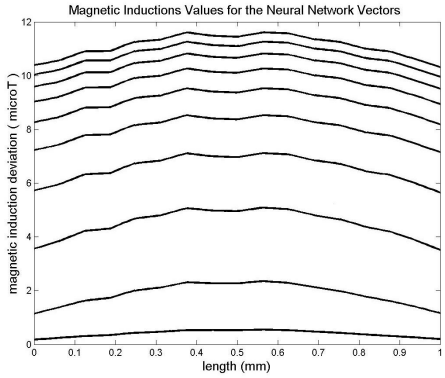


Figure 5 – Magnetic induction in the region between the coils, for a group of defects

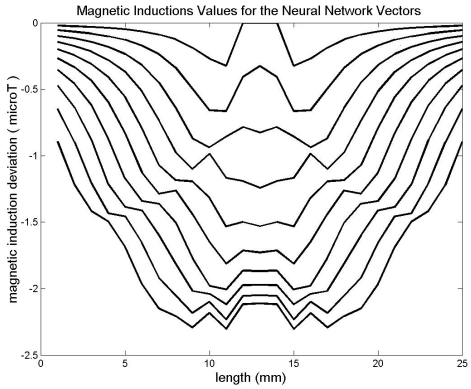


Figure 6 – Magnetic induction for the 25 elements of the vectors which correspond to the height of 4 mm

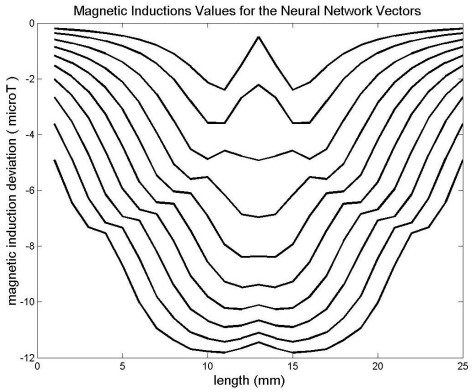


Figure 7 – Magnetic induction for the 25 elements of the vectors which correspond to the height of 1 mm

The neural network architecture chosen was a multilayer perceptron, trained with the Levemberg-Marquadt, [4]. Several network configurations were tried, and better results have been obtained by a network constituted by three layers with 15, 10 and 5 neurons, as we can see in Table 1

Table 1 – Errors for several neural network training sessions

1st hidden layer	2nd hidden layer	3rd hidden layer	Mean Squared Error	Standard Deviation
15	-	-	13,86	18,11
15	-	-	12,46	23,27
15	-	-	12,30	24,12
20	-	-	12,02	18,75
20	-	-	11,71	17,56
20	-	-	9,83	14,89
15	5	-	5,19	10,21
15	5	-	0,57	0,55
15	5	-	0,35	0,55
15	10	-	2,61	3,31
15	10	-	0,67	0,79
15	10	-	0,53	0,75
15	10	5	0,17	0,20
15	10	5	0,09	0,13
15	10	5	0,005	0,002

NEW CLASSIFICATIONS

After the neural network training and respective validations, new defects were simulated by the finite element method, for posteriori classification by the network. Table 2 shows the defects dimensions (height and width), expected defect number and obtained defect number, by the neural network.

Table 2 – Simulation results

Defect height (mm)	**Defect width (mm)**	**Expected number**	**Obtainded number**
3	35,26	2,0	2,0001
1	84,23	1,0	1,000
4	81,55	3,0	2,9999

As we can see, the results obtained by the neural network for these examples, although very simple, were very good.

CONCLUSIONS

In this paper we presented an investigation on the use of the finite element method and artificial neural networks for identification of defects in metallic walls, present in industrial plants. For a given metallic wall characteristics, defects can be simulated by the finite element method, and the fields results used in the preparation of training vectors for artificial neural networks. The main contribution of the proposed methodology is the possibility of better identification of shape of the defects. The network can be embedded in an electronic device in order to identify defects in real metallic walls. Again, we can So the association of FEM and ANN techniques seems to be an useful alternative for non destructive evaluations. Future works are intended to be done in this field, such as the use of more realistic finite element problems, computer parallel programming, in order to get quickly solutions.

ACKNOWLEDGMENTS

The authors thank to Dr. David C. Meekers for sharing the 2D FEM modeler FEMM free of charge.

REFERENCES

1. T. S. Low & B. Chao, The Use of Finite Elements and Neural Networks for the solution of Inverse Electromagnetic Problems, *IEEE Transactions on Magnetics*, 28(5), (1992).

[2] P. P. Silvester, R. L. Ferrari, 1996, *Finite Elements for Electrical Engineers* U.K., Cambridge,1996.

[3]S. S. Rao, *The Finite Elements Methods in Engineering* U. K, Oxford , 1989.

[4] S. Haykin, *Neural Networks-A Comprehensive Course* ,USA, Prentice Hall, 1999.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

NONLINEAR IDENTIFICATION OF MECHANICAL SYSTEMS USING ORTHOGONAL FUNCTIONS

Ricardo P. Pacheco
School of Mechanical Engineering, FEMEC
Federal University of Uberlândia, UFU
Uberlândia, MG, Brazil
rpacheco@mecanica.ufu.br

Valder Steffen, Jr
School of Mechanical Engineering, FEMEC
Federal University of Uberlândia, UFU
Uberlândia, MG, Brazil
vsteffen@mecanica.ufu.br

ABSTRACT

Real world mechanical systems present nonlinear behavior and in many cases simple linearization in modeling the system would not lead to satisfactory results. Coulomb damping and cubic stiffness are typical examples of system parameters currently used in nonlinear models of mechanical systems. This paper uses orthogonal functions to represent input and output signals. These functions are easily integrated by using a so-called operational matrix of integration. Consequently, it is possible to transform the nonlinear differential equations of motion into algebraic equations. After mathematical manipulation the unknown linear and nonlinear parameters are determined. Numerical simulations confirm the above methodology.

INTRODUCTION

Real systems, in general, have a non-linear dynamical behavior. However, most of these systems can be studied through an approach based on the linear system theory, according to which the superposition principle can be applied. The error that arises from this type of approach depends on the non-linearity degree of the system under analysis. When systems with high non-linearity are concerned, the application of the theory for linear systems is not acceptable. So, in these situations, specific methods for non-linear system analysis must be employed.

In reference [1] the authors present a parametric identification method in which time series are used to extract the dynamical characteristics of the system and predict the time response. Systems presenting Coulomb friction and non-linear stiffness, occurring separately or simultaneously, are studied. This parametric identification procedure uses the AVD model (acceleration, velocity and displacement) and models the dry friction force using the velocity. The AVD model of the friction mechanism is independent of the excitation level and can predict accurately the time response due to random excitations since the condition of continuous motion is satisfied, i. e., stick-slip motion does not occur. Displacement and velocity signals should be used besides the acceleration in order to obtain accurately the non-linear terms. The precision in the friction force identification is better when using higher excitation force levels.

In reference [2] a wavelet-based procedure is developed to identify mechanical parameters of discrete non-linear structural systems. The methodology allows the parameter estimation of a prior known dynamical models as well as the identification of classes of suitable non-linear models based on input-output data. The method relies on a wavelet-based discretization of the non-linear governing differential equation of motion. The inertia terms of the system have to be known, a priori, in order to identify the other parameters, what may limit the use of this technique in certain applications.

In this paper a methodology to identify physical parameters of non-linear systems, through orthogonal functions, is presented. Different types of non-linearity can be addressed for both free or forced systems, since the mathematical model is known. Numerical simulations testify the efficiency of the technique and show its applicability for single and multi-degree of freedom systems.

ORTHOGONAL FUNCTION REVIEW

A set of functions $\phi_i(t)$, i = 1, 2, 3, ... is said to be orthogonal in the interval [a,b] if:

$$\int_a^b \phi_m(t)\,\phi_n(t)\,dt = K_{mn} \qquad (1)$$

where: $\begin{cases} K_{mn} = 0 \text{ if } m \neq n \\ K_{mn} \neq 0 \text{ if } m = n \end{cases}$

If K_{mn} is the Kronecker's delta, the set of functions $\phi_i(t)$ is said orthonormal. The following property, related to the successive integration of the vectorial basis, holds for a set of *r* orthonormal functions:

$$\underbrace{\int_0^t \cdots \int_0^t}_{\text{n times}} \{\phi(\tau)\} (d\tau)^n \cong [P]^n \{\phi(t)\} \qquad (2)$$

where $[P] \in \Re^{r,r}$ is a square matrix with constant elements, called operational matrix and $\{\phi(t)\} = \{\phi_0(t) \quad \phi_1(t) \quad ... \quad \phi_{r-1}(t)\}^T$ is the vectorial basis of the orthonormal series

References [3] and [4] give details about the vectorial basis and operational matrix related to the various types of orthogonal functions considered in the present paper.

PARAMETER ESTIMATION TECHNIQUE THOUGH ORTHOGONAL FUNCTIONS

The equation of motion of a N-D.O.F. non-linear system, submitted to any external force $\{f(t)\}$, can be described by:

$$[M]\{\ddot{x}(t)\} + [C]\{\dot{x}(t)\} + [K]\{x(t)\} + \\ + \{g(x(t), \dot{x}(t))\} = \{f(t)\} \qquad (3)$$

where $[M]$, $[C]$ and $[K]$ are, respectively, the N order mass, damping and stiffness matrices, $\{x(t)\}$ is the displacement vector, $\{g(x(t), \dot{x}(t))\}$ is the non-linear restoring force vector, which is a function of the displacement and velocity, and $\{f(t)\}$ is the excitation force vector. Depending on the nature and magnitude of the non-linear forces in $\{g(x(t), \dot{x}(t))\}$ and the vibration level of the system, the non-linear term can be ignored and a linear modal analysis can be performed with negligible errors. However, in this study, cases are addressed in which the non-linear effects can not be neglected.

The non-linear force vector can assume different forms. Without loss of generality, a formulation will be developed for a one-D.O.F. system with cubic stiffness and mixed damping (viscous and dry friction damping). In this case, Eq. (3) is written in the following way:

$$M\,\ddot{x}(t) + C\,\dot{x}(t) + K\,x(t) + K_3\,x(t)^3 + \\ + f_d\,\text{sign}(\dot{x}(t)) = f(t) \qquad (4)$$

where K_3 is the cubic stiffness coefficient, f_d is the dry friction force and $\text{sign}(\dot{x}(t))$ is defined as:

$$\text{sign}(\dot{x}(t)) = \begin{cases} +1 & \text{for } \dot{x}(t) > 0 \\ 0 & \text{for } \dot{x}(t) = 0 \\ -1 & \text{for } \dot{x}(t) < 0 \end{cases} \qquad (5)$$

Let $y(t) = x(t)^3$ and $z(t) = \text{sign}(\dot{x}(t))$, substituting in Eq. (4), results:

$$M\,\ddot{x}(t) + C\,\dot{x}(t) + K\,x(t) + \\ + K_3\,y(t) + f_d\,z(t) = f(t) \qquad (6)$$

Integrating Eq. (6) twice in the interval $[0; t]$, one can obtain:

$$M(x(t) - x(0) - \dot{x}(0)\,t) + C\left(\int_0^t x(\tau)\,d\tau - x(0)\,t\right) + \\ + K\int_0^t\int_0^t x(\tau)\,d\tau^2 + K_3\int_0^t\int_0^t y(\tau)\,d\tau^2 + f_d\int_0^t\int_0^t z(\tau)\,d\tau^2 = \\ = \int_0^t\int_0^t \{f(\tau)\}d\tau^2 \qquad (7)$$

where $x(0)$ and $\dot{x}(0)$ are, respectively, the displacement and velocity initial conditions.

The signals $x(t)$, $y(t)$, $z(t)$ and $f(t)$ are expanded in truncated orthogonal function series with *r* terms, i.e.:

$$\begin{aligned} x(t) &\cong \{X\}_{(1,r)}\,\{\phi(t)\}_{(r,1)} \\ y(t) &\cong \{Y\}_{(1,r)}\,\{\phi(t)\}_{(r,1)} \\ z(t) &\cong \{Z\}_{(1,r)}\,\{\phi(t)\}_{(r,1)} \\ f(t) &\cong \{F\}_{(1,r)}\,\{\phi(t)\}_{(r,1)} \end{aligned} \qquad (8)$$

where $\{X\}$, $\{Y\}$, $\{Z\}$ and $\{F\}$ are the vectors containing the coefficients of the series expansion.

Substituting Eqs. (8) in Eq. (7), one obtains:

$$\begin{aligned} &M\left(\{X\}\{\phi(t)\} - x(0) - \dot{x}(0)\,t\right) + \\ &+ C\left(\int_0^t \{X\}\{\phi(\tau)\}\,d\tau - x(0)\,t\right) + \\ &+ K\int_0^t\int_0^t \{X\}\{\phi(\tau)\}\,d\tau^2 + K_3\int_0^t\int_0^t \{Y\}\{\phi(\tau)\}\,d\tau^2 + \\ &+ f_d\int_0^t\int_0^t \{Z\}\{\phi(\tau)\}\,d\tau^2 = \int_0^t\int_0^t \{F\}\{\phi(\tau)\}\,d\tau^2 \end{aligned} \tag{9}$$

Considering that $\{e\}^T\{\phi(t)\} = 1$ and $t = \{e\}^T[P]\{\phi(t)\}$ [5], where $\{e\}_{(r,1)}$ is a vector with constant elements whose form depends on the orthogonal function used. For instance, for Fourier, Chebyshev, Legendre, Jacobi and Walsh series $\{e\} = \{1 \;\; 0 \;\; \cdots \;\; 0\}^T$ and for the Block-Pulse function $\{e\} = \{1 \;\; 1 \;\; \cdots \;\; 1\}^T$. Substituting these expressions in Eq. (9), results:

$$\begin{aligned} &M\left(\{X\}\{\phi(t)\} - x(0)\{e\}^T\{\phi(t)\} - \dot{x}(0)\,\{e\}^T[P]\{\phi(t)\}\right) + \\ &+ C\left(\int_0^t \{X\}\{\phi(\tau)\}\,d\tau - x(0)\,\{e\}^T[P]\{\phi(t)\}\right) + \\ &+ K\int_0^t\int_0^t \{X\}\{\phi(\tau)\}\,d\tau^2 + K_3\int_0^t\int_0^t \{Y\}\{\phi(\tau)\}\,d\tau^2 + \\ &+ f_d\int_0^t\int_0^t \{Z\}\{\phi(\tau)\}\,d\tau^2 = \int_0^t\int_0^t \{F\}\{\phi(\tau)\}\,d\tau^2 \end{aligned} \tag{10}$$

Applying the property for the integration of orthogonal functions (Eq. 2)and equating the coefficients of $\{\phi(t)\}$ in Eq. (10), one can obtain the following algebraic equation:

$$\begin{bmatrix} M \\ -Mx(0) \\ -M\dot{x}(0) - Cx(0) \\ C \\ K \\ K_3 \\ f_d \end{bmatrix}^T \begin{bmatrix} \{X\} \\ \{e\}^T \\ \{e\}^T[P] \\ \{X\}[P] \\ \{X\}[P]^2 \\ \{Y\}[P]^2 \\ \{Z\}[P]^2 \end{bmatrix} = \left[\{F\}[P]^2\right] \tag{11}$$

Let:

$$[H] = \begin{bmatrix} M \\ -Mx(0) \\ -M\dot{x}(0) - Cx(0) \\ C \\ K \\ K_3 \\ f_d \end{bmatrix}^T, \quad [J] = \begin{bmatrix} \{X\} \\ \{e\}^T \\ \{e\}^T[P] \\ \{X\}[P] \\ \{X\}[P]^2 \\ \{Y\}[P]^2 \\ \{Z\}[P]^2 \end{bmatrix} \quad \text{and} \quad [E] = \left[\{F\}[P]^2\right].$$

Equation (11) can be written in the following compact form:

$$[H]_{(1,7)}\,[J]_{(7,r)} = [E]_{(1,r)} \tag{12}$$

The expression below gives an estimate of matrix $[H]$ by using the least square method:

$$[H] = [E][J]^T\left([J][J]^T\right)^{-1} \tag{13}$$

Equation (13) is solved by using the Singular Value Decomposition Method and it is possible to determine the unknown parameters, as well as the displacement and velocity initial conditions.

As presented above, Eq. (12) is valid for a one-D.O.F. system with cubic stiffness and mixed damping (viscous and dry friction damping). If only one type of non-linearity holds, the term referred to the other type of non-linearity is neglected and Eq. (12) remains valid. When multi-degree of freedom systems are concerned, the development of the formulation follows the same procedure adopted above, except for the case in which cubic stiffness is considered. In that case, the formulation is slightly changed, since various non-linear terms appear in the equations of motion (the relative motion of the different masses have to be computed). However the general technique applied is the same.

In the case of free systems, the procedure is similar and the following algebraic equation is obtained:

$$\begin{bmatrix} x(0) \\ \dot{x}(0) + M^{-1}Cx(0) \\ -M^{-1}C \\ -M^{-1}K \\ -M^{-1}K_3 \\ -M^{-1}f_d \end{bmatrix}^T \begin{bmatrix} \{e\}^T \\ \{e\}^T[P] \\ \{X\}[P] \\ \{X\}[P]^2 \\ \{Y\}[P]^2 \\ \{Z\}[P]^2 \end{bmatrix} = \{X\} \qquad (14)$$

or, in a compact form:

$$[H]_{(1,6)}\,[J]_{(6,r)} = [E]_{(1,r)} \qquad (15)$$

where $[H] = \begin{bmatrix} x(0) \\ \dot{x}(0) + M^{-1}Cx(0) \\ -M^{-1}C \\ -M^{-1}K \\ -M^{-1}K_3 \\ -M^{-1}f_d \end{bmatrix}^T$, $[J] = \begin{bmatrix} \{e\}^T \\ \{e\}^T[P] \\ \{X\}[P] \\ \{X\}[P]^2 \\ \{Y\}[P]^2 \\ \{Z\}[P]^2 \end{bmatrix}$

and $[E] = \{X\}$.

One can observe, from the equations above, that the unknown parameters can not be identified separately but as a combination of them. However, if the mass or another parameter is previously known, all other parameters can be determined separately.

CASE STUDIES

One-D.O.F. Mechanical System With Mixed Damping (Viscous And Dry Friction Damping)

This case corresponds to the system represented by Eq. (4), for which the term associated with the non-linear stiffness is not taken into account, with the following parameters: $M = 1\,\text{kg}$, $C = 20\,\text{Ns/m}$, $K = 10000\,\text{N/m}$, $f_d = 1 \text{ and } 3\,\text{N}$.

At first, a swept-sine excitation with an RMS value of 10 N ($F_{rms} = 10.0\,\text{N}$) from 10 to 20 Hz was used. The response, considering $f_d = 1\,\text{N}$, was sampled at a frequency of 1700 Hz using the fourth-order Runge-Kutta method. The parameters and dry friction force identified are shown in Tables 1.1 and 1.2.

Table 1.1. Swept-sine excitation – One-D.O.F system (f_d = 1 N)

Orthogonal Function	M [kg]	C [Ns/m]
Fourier (r=51)	0.996 (0.5%)(1)	20.18 (0.9%)
Chebyshev (r=30)	0.995 (0.5%)	20.10 (0.5%)
Legendre (r=35)	0.993 (0.7%)	20.40 (2.0%)
Jacobi (r=30)	0.994 (0.6%)	20.33 (1.7%)
Block-Pulse(2) (r=512)	0.993 (0.7%)	20.18 (0.9%)

Table 1.2. Swept-sine excitation – One-D.O.F system (f_d = 1 N)

Orthogonal Function	K [N/m]	f_d [N]
Fourier (r=51)	10029 (0.3%)	1.005 (0.5%)
Chebyshev (r=30)	10019 (0.2%)	1.020 (2.0%)
Legendre (r=35)	10003 (0.0%)	0.993 (0.7%)
Jacobi (r=30)	9997 (0.0%)	1.079 (7.9%)
Block-Pulse(2) (r=512)	10045 (0.5%)	1.004 (0.4%)

Obs.: 1) (•) relative error
2) Walsh functions presented the same results as the Block-Pulse functions

When the value of the dry friction force is increased keeping the same excitation force, i. e., increasing the ratio (f_d/F_{rms}) the errors in the identified parameters, in general, are greater. This is shown in Tables 2.1 and 2.2 for $f_d = 3\,\text{N}$ and the same excitation force ($F_{rms} = 10.0\,\text{N}$).

Analyzing Tables 1.1, 1.2, 2.1 and 2.2, one can say that, in general, Fourier series had the best performance. On the other hand, Jacobi polynomials presented difficulties with respect to dry friction force identification.

It was also applied a random force in the range 10 to 25 Hz using the same intensity of excitation level ($F_{rms} = 10.0\,\text{N}$) and two different values for the dry friction force. The results, only for Fourier series, are shown in Tables 3.1 and 3.2. The response predictions are presented in Fig. 1 and 2, for $f_d = 1\,\text{N}$ and $f_d = 3\,\text{N}$, respectively.

Table 2.1. Swept-sine excitation – One-D.O.F system (f_d = 3 N)

Orthogonal Function	M [kg]	C [Ns/m]
Fourier (r=51)	0.990 (1.0%)	20.42 (2.1%)
Chebyshev (r=30)	0.987 (1.3%)	20.82 (4.1%)
Legendre (r=35)	0.987 (1.3%)	20.13 (0.6%)
Jacobi (r=30)	0.987 (1.3%)	19.27 (3.6%)
Block-Pulse (r=512)	0.987 (1.3%)	20.47 (2.3%)

Table 2.2. Swept-sine excitation – One-D.O.F system (f_d = 3 N)

Orthogonal Function	K [N/m]	f_d [N]
Fourier (r=51)	9995 (0.1%)	2.990 (0.3%)
Chebyshev (r=30)	9974 (0.3%)	2.914 (2.9%)
Legendre (r=35)	9957 (0.4%)	3.103 (3.4%)
Jacobi (r=30)	9946 (0.5%)	3.378 (12.6%)
Block-Pulse (r=512)	10012 (0.1%)	2.980 (0.7%)

Table 3.1. Random excitation – One-D.O.F system

Orthogonal Function	M [kg]	C [Ns/m]
Fourier (f_d = 1 N, r=91)	0.978 (2.2%)	19.41 (3.0%)
Fourier (f_d = 3 N, r=101)	0.957 (4.3%)	19.54 (2.3%)

Table 3.2. Random excitation – One-D.O.F system

Orthogonal Function	K [N/m]	f_d [N]
Fourier (f_d = 1 N, r=91)	9967 (0.3%)	1.059 (5.9%)
Fourier (f_d = 3 N, r=101)	9859 (1.4%)	3.045 (2.9%)

The results shown in the Tables above indicate that the accuracy in the identification was higher when swept-sine excitation was used. However, even for random excitation (Tables 3.1 and 3.2), the results obtained for the response prediction (Fig. 1 and 2) can be considered acceptable.

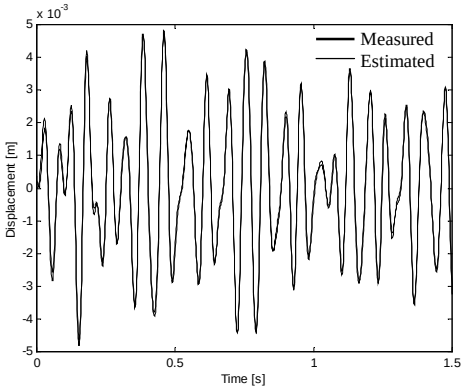


Figure 1. Response prediction – One-D.O.F system (f_d = 1 N)

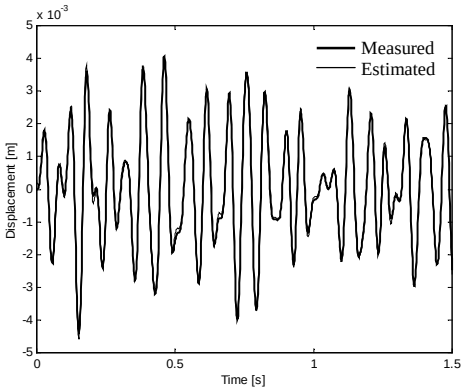


Figure 2. Response prediction – One-D.O.F system (f_d = 3 N)

One-D.O.F. Duffing Oscillator With Mixed Damping

In this application, all terms presented in Eq. (4) are considered. The parameters of the system are given by $M = 1\,\text{kg}$, $C = 20\,\text{Ns/m}$, $K = 10000\,\text{N/m}$, $K_3 = 5\times10^9\,\text{N/m}^3$ and $f_d = 1\,\text{N}$.

A harmonic excitation force such as $f(t) = F_0 \sin(2\pi f_0 t)$ with $f_0 = 21\,\text{Hz}$, $F_{rms} = 20\,\text{N}$ and a sampling frequency of $5115\,\text{Hz}$ was used. The results with the identified parameters and the respective relative errors for the Duffing oscillator are shown in Table 4.

Table 4. Identified parameters – One-D.O.F. Duffing oscillator with mixed damping

	Fourier (r=75)	Block-Pulse (r=512)
M [kg]	1.000 (0.0%)	0.997 (0.4%)
C [Ns/m]	18.71 (6.4%)	18.70 (6.5%)
K [N/m]	10041 (0.4%)	10030 (0.3%)
K_3 [N/m^3]	5.004×10^9 (0.1%)	5.001×10^9 (0.0%)
f_d [N]	1.037 (3.7%)	1.035 (3.5%)

Two-D.O.F. Mechanical System With Mixed Damping

A two-D.O.F. mechanical system was tested in order to illustrate the application of the methodology to multi-degree of freedom systems. The corresponding physical model is presented in Fig. 3.

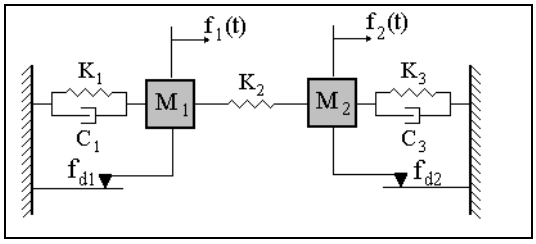


Figure 3. Two-D.O.F. mechanical system with mixed damping

The physical parameters are given by:

$$M_1 = M_2 = 1.0\,\text{kg}, \qquad C_1 = C_2 = 10.0\,\text{Ns/m},$$
$$K_1 = K_2 = K_3 = 10000.0\,\text{N/m}, \; f_{d1} = f_{d2} = 1.0\,\text{N}$$

A swept-sine excitation in the band 10 to 23 Hz is applied to mass 1 and another force in the band 17 to 30 Hz is applied to mass 2. The response was sampled at a frequency of 1023 Hz. For both excitation forces $F_{rms} = 20.0\,\text{N}$. The identified values for the parameters and the dry friction force are shown in Tables 5.1 and 5.2.

A band limited (10 to 30 Hz) random excitation force ($F_{rms} = 10.0\,\text{N}$) was also applied to the above two-D.O.F. Mechanical system. Some identified parameters presented greater errors when compared with the previous case, as shown in Table 6. However, comparing the predicted and measured responses, a reasonable agreement was found, as shown in Figures 4 and 5.

Table 5.1. Swept-sine excitation – Mixed damping two-D.O.F system

	Fourier (r=75)	Legendre (r=51)
M_1	1.002 (0.2%)	1.001 (0.1%)
M_2	0.996 (0.4%)	0.995 (0.5%)
C_1	9.49 (5.1%)	9.55 (4.5%)
C_3	9.99 (0.1%)	10.00 (0.0%)
K_1	10155 (1.6%)	10168 (1.7%)
K_2	9965 (0.4%)	9926 (0.7%)
K_3	9896 (1.0%)	9904 (1.0%)
f_{d1}	1.047 (4.7%)	0.973 (2.7%)
f_{d2}	1.026 (2.6%)	1.043 (4.3%)

Table 5.2. Swept-sine excitation – Mixed damping two-D.O.F system

	Jacobi (r=46)	Block-Pulse (r=512)
M_1	0.998 (0.2%)	1.000 (0.0%)
M_2	0.996 (0.4%)	0.994 (0.6%)
C_1	9.67 (3.3%)	9.48 (5.2%)
C_3	10.04 (0.4%)	9.99 (0.1%)
K_1	10147 (1.5%)	10163 (1.6%)
K_2	9917 (0.8%)	9972 (0.3%)
K_3	9916 (0.8%)	9900 (1.0%)
f_{d1}	0.973 (2.7%)	1.054 (5.4%)
f_{d2}	1.048 (4.8%)	1.027 (2.7%)

Table 6. Random excitation – Mixed damping two-D.O.F system

	Legendre (r=105)
M_1	1.098 (9.8%)
M_2	1.009 (0.9%)
C_1	9.13 (8.7%)
C_3	10.11 (1.1%)
K_1	10082 (0.8%)
K_2	10002 (0.0%)
K_3	9242 (7.6%)
f_{d1}	1.098 (9.8%)
f_{d2}	1.009 (0.9%)

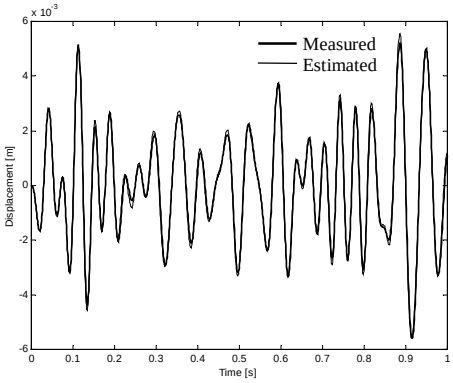


Figure 4. Response prediction $x_1(t)$ – Two-D.O.F system (random excitation)

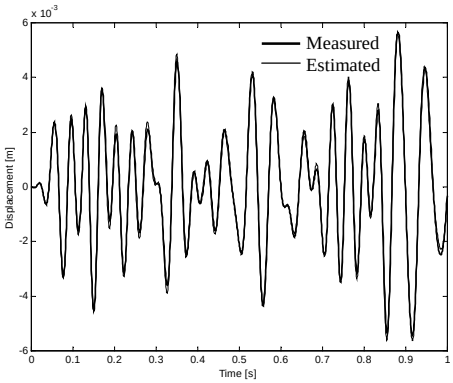


Figure 5. Response prediction $x_2(t)$ – Two-D.O.F system (random excitation)

CONCLUSIONS

A technique for parameter identification of non-linear systems, based on orthogonal functions, has been developed. It can be applied to systems with different types of non-linearity, since the mathematical model of the system is known.

The studied cases show that good results can be obtained either for free and forced systems.

Among the orthogonal functions tested, Fourier, Legendre, Block-Pulse and Walsh series presented a superior performance when compared with Chebyshev and Jacobi polynomials.

The efficiency of the methodology encourages further studies related to inverse problem identification, particularly the case of force identification.

REFERENCES

1. Q. Chen and G. R. Tomlinson, Parametric Identification of Systems with Dry Friction and Nonlinear Stiffness Using a Time Series Model, *Journal of Vibration and Acoustics*, ASME **118**, 1996, pp. 252-263.

2. R. Ghanem and F. Romeo 2001, A Wavelet-Based Approach for Model and Parameter Identification of Non-Linear Systems, *International Journal of Non-Linear Mechanics*, vol. **36**, 2001, pp. 835-859.

3. R. P. Pacheco and V. Steffen Jr., On Time Domain Identification Using Orthogonal Functions, *DETC'99 Proc. (in CD) of the 1999 ASME Design Engineering Technical Conferences – Design for the Next Millenium*, Las Vegas-NV, 1999.

4. R. P. Pacheco, V. Steffen Jr. and D. A. Rade, Time Domain-Based Identification of Modal Parameters and Excitation Forces Using Orthogonal Functions, *COBEM/99 Proc. (in CD) of the 15th Brazilian Congress of Mechanical Engineering*, Águas de Lindóia, Brazil, 1999.

5. R. P. Pacheco, *Mechanical Systems Identification Through Time Domain Methods*, Ph.D. Thesis, Federal University of Uberlândia, 2001 (in Portuguese).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ESTIMATION OF INTERNAL SOURCES IN NATURAL WATERS USING REMOTE SENSING DATA

E. S. Chalhoub, H. F. de Campos Velho
Instituto Nacional de Pesquisas Espaciais
Laboratório Associado de Computação e Matemática Aplicada
P. O. Box 515, 12201-970 São José dos Campos, SP, Brazil
[ezzat, haroldo]@lac.inpe.br

ABSTRACT

An inverse analysis for the estimation of internal sources in natural waters, using remote sensing data, is presented. The analysis involves a forward model that utilizes an analytical discrete-ordinates method for solving the radiative transfer equation and an inverse model which contains an algorithm for least-squares estimation that is iteratively solved for retrieving the internal source profile by using the Levenberg-Marquardt optimizer. The experimental data are simulated with synthetic data (exit radiances) that are calculated at the surface of the water and corrupted with noise. The results show that the internal sources can be recovered with good accuracy, even for high noisy data.

Keywords: Inverse hydrologic optics, internal sources, remote sensing, analytical discrete-ordinates methods.

NOMENCLATURE

Following is a list of the most important symbols used in this work:

I	intensity (radiance) of the radiation field
I_0	incident beam strength
L	order of the anisotropy
M	number of optimization functions
N	quadrature order
p	phase function
P	Legendre polynomial
P^m	associated Legendre function
R	number of spatial regions
S	inhomogeneous source term
S_0	internal source of radiation
z	geometrical thickness of the medium
β	expansion coefficient
ζ	optical thickness of the medium
Θ	direction of propagation of the radiation in the medium
λ	photon wavelength
μ	cosine of the polar angle (measured from the *positive* τ axis)
μ_0	cosine of polar angle of incidence
ϖ	albedo for single scattering
τ	optical variable
φ	azimuthal angle
φ_0	azimuthal angle of incidence

Subscripts/Superscripts

g	wavelength interval index
l	expansion order of the phase function
m	Fourier component index
r	spatial region index

INTRODUCTION

The inverse analysis of radiation in a participating medium has a broad range of applications, including, among others, remote sensing of the atmosphere and the determination of radiative properties in natural waters. McCormick in his articles [1–3] presents reviews of methods for solving the inverse radiation problems, such as the estimation of optical properties, the thickness of a medium and the presence of a spatially distributed source. In hydrologic optics, the estimation of the radiative properties can be performed by using either data from *in situ* or remote sensing data.

One of the most important challenges in inverse hydrologic optics is to retrieve the properties of the physical system from remote sensed data. There are only few works in the literature on this subject, most of them are based on the Gershun's equation and the estimation of the apparent optical property named irradiance attenuation coefficient [4] or diffuse attenuation coefficient [5]. Another simplification for remote sensing estimates of bio-optical properties is to consider an homogeneous ocean [6–8].

Differently from our early estimations with *in situ* radiometric measurements [9–15], the source term estimation is performed using remote sensing data, that are represented by the exit radiances. The method presented in this paper is completely different from the one presented in Ref. [16], which is based on the reciprocity principle. However, there is no tests in this reference, not even a numerical one, for the validation of that methodology.

The inverse analysis involves the following two basic steps: (a) a forward problem solution and (b) an inverse problem solution. In the first step, the radiative transfer equation is solved by an analytical discrete-ordinates method [17–19] to determine the exact (synthetic) exit radiances, and, in the second, an algorithm [20,21] for least-squares estimation is iteratively utilized to retrieve the spatially distributed sources. In the analysis, the experimental data are simulated with synthetic data corrupted with noise from 1 to 5%.

PRELIMINARY ANALYSIS

For a multispectral problem, we consider the equation of transfer

$$\mu\frac{\partial}{\partial\tau}I(\tau,\mu,\varphi,\lambda)+I(\tau,\mu,\varphi,\lambda)=\varpi(\lambda)\times\int_{-1}^{1}\int_{0}^{2\pi}\int_{\lambda}p(\cos\Theta,\lambda)I(\tau,\mu',\varphi',\lambda')\mathrm{d}\lambda'\mathrm{d}\varphi'\mathrm{d}\mu'+S_0(\tau,\lambda),\quad(1)$$

subject to the boundary conditions

$$I(0,\mu,\varphi,\lambda)=L(\mu,\varphi,\lambda)=I_0(\lambda)\delta(\mu-\mu_0)\delta(\varphi-\varphi_0)\quad(2a)$$

and

$$I(\zeta,-\mu,\varphi,\lambda)=R(\mu,\varphi,\lambda)=0,\quad(2b)$$

where $I(\tau,\mu,\varphi,\lambda)$ denotes the intensity (radiance) of the radiation field, $\tau\in(0,\zeta)$ the optical variable, with ζ representing the optical thickness of the medium, $\mu\in[-1,1]$ and $\varphi\in[0,2\pi]$, respectively, the cosine of the polar angle (measured from the *positive* τ axis) and the azimuthal angle, that specify the direction of propagation Θ of the radiation in the medium, and λ the photon wavelength. In addition, $\varpi(\lambda)\in[0,1]$ is the albedo for single scattering, where $\varpi=b/(a+b)$ with a and b representing the absorption and scattering coefficients, $p(\cos\Theta,\lambda)$ is the phase function for scattering from $\{\mu',\varphi',\lambda'\}$ to $\{\mu,\varphi,\lambda\}$, $S_0(\tau,\lambda)$ an internal source of radiation, and $L(\mu,\varphi,\lambda)$ and $R(\mu,\varphi,\lambda)$ are the distributed incident radiances at the boundaries. The incident beam is characterized by a beam strength $I_0(\lambda)$ and a beam direction (μ_0,φ_0). An outline of the physical process is depicted in Fig. 1.

In this work, we discretize Eqs. (1) and (2) in the wavelength variable and then consider all wavelength-dependent values as being averages over a wavelength interval (band) $\Delta\lambda_g$. Thus, for a generic variable $F(\lambda)$, we have

$$F_g=F(\lambda_g)=\frac{1}{\Delta\lambda_g}\int_{\Delta\lambda_g}F(\lambda)\mathrm{d}\lambda,\quad(3)$$

where λ_g is an average wavelength in the interval g. To further simplify the calculations, we consider that a particle can only be scattered to within the same interval, and so we write our original equation of transfer, for a specific wavelength λ_g, as,

$$\mu\frac{\partial}{\partial\tau}I_g(\tau,\mu,\varphi)+I_g(\tau,\mu,\varphi)=\varpi_g\times\int_{-1}^{1}\int_{0}^{2\pi}p(\cos\Theta)I_g(\tau,\mu',\varphi')\mathrm{d}\varphi'\mathrm{d}\mu'+S_{0,g}(\tau),\quad(4)$$

subject to the boundary conditions

$$I_g(0,\mu,\varphi)=I_{0,g}\delta(\mu-\mu_0)\delta(\varphi-\varphi_0)\quad(5a)$$

and

$$I_g(\zeta,-\mu,\varphi)=0.\quad(5b)$$

Here, the phase function $p(\cos\Theta)$, for scattering from $\{\mu',\varphi'\}$ to $\{\mu,\varphi\}$, is represented by a finite Legendre polynomial expansion given in terms of the cosine of the scattering angle Θ,

$$p(\cos\Theta)=\frac{1}{4\pi}\sum_{l=0}^{L}\beta_l P_l(\cos\Theta),\quad\beta_0=1\quad\text{and}\quad|\beta_l|<2l+1\ \text{ for }\ 0<l\le L,\quad(6)$$

where β_l and P_l are, respectively, the coefficient and the Legendre polynomial in the L^{th}-order expansion of the phase function.

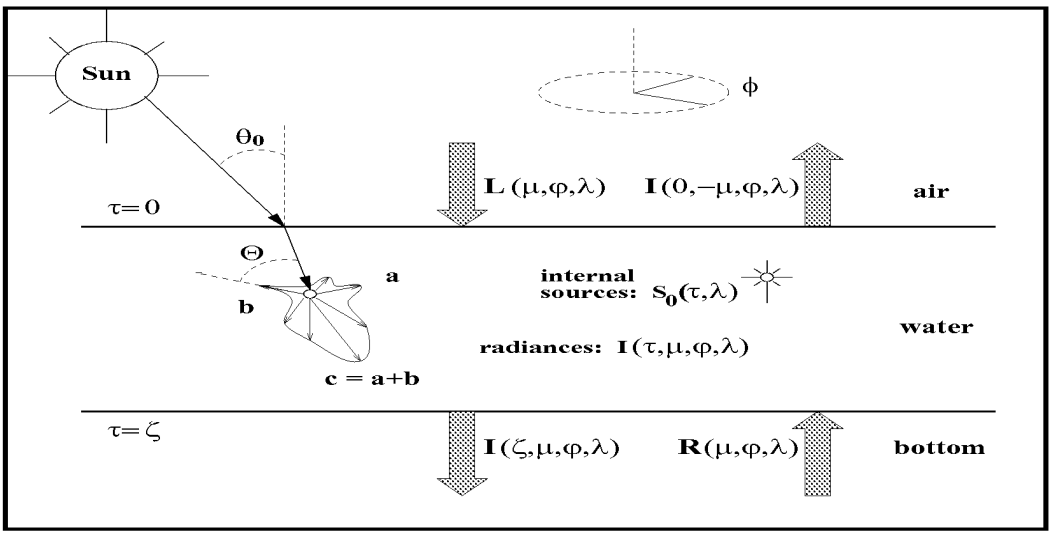


Figure 1: Pictorial representation of the radiative process in natural waters.

FORWARD PROBLEM SOLUTION

Following Chandrasekhar [22], we then write the intensity of the radiation field as

$$I_g(\tau,\mu,\varphi) = I_{u,g}(\tau,\mu,\varphi) + I_{s,g}(\tau,\mu,\varphi), \quad (7)$$

where $I_{u,g}(\tau,\mu,\varphi)$ is the unscattered component which satisfies a version of Eq. (4) with zero right-hand side and boundary conditions similar to Eqs. (5), and $I_{s,g}(\tau,\mu,\varphi)$ is the scattered component that must satisfy

$$\mu\frac{\partial}{\partial\tau}I_{s,g}(\tau,\mu,\varphi) + I_{s,g}(\tau,\mu,\varphi) = \varpi_g \times \int_{-1}^{1}\int_{0}^{2\pi} p(\cos\Theta)I_{s,g}(\tau,\mu',\varphi')\mathrm{d}\varphi'\mathrm{d}\mu' + S_g(\tau,\mu,\varphi), \quad (8)$$

for $\tau \in [0,\zeta]$, $\mu \in [-1,1]$ and $\varphi \in [0,2\pi]$, and the boundary conditions

$$I_{s,g}(0,\mu,\varphi) = I_{s,g}(\zeta,-\mu,\varphi) = 0, \quad (9)$$

for $\mu \in (0,1]$ and $\varphi \in [0,2\pi]$. The inhomogeneous source term $S_g(\tau,\mu,\varphi)$ is given by

$$S_g(\tau,\mu,\varphi) = S_{0,g}(\tau) + \varpi_g \times \int_{-1}^{1}\int_{0}^{2\pi} p(\cos\Theta)I_{u,g}(\tau,\mu',\varphi')\mathrm{d}\varphi'\mathrm{d}\mu'. \quad (10)$$

Continuing, we make use of the Fourier *cosine* decomposition [22]

$$I_{s,g}(\tau,\mu,\varphi) = \frac{1}{2}\sum_{m=0}^{L}(2-\delta_{0,m})I_g^m(\tau,\mu) \times \cos[m(\varphi-\varphi_0)] \quad (11)$$

along with the addition theorem [23] for the Legendre polynomials to deduce that the original problem can be reduced to the problem of solving, for $m = 0, 1, \ldots, L$,

$$\mu\frac{\partial}{\partial\tau}I_g^m(\tau,\mu) + I_g^m(\tau,\mu) = \frac{\varpi_g}{2}\sum_{l=m}^{L}\beta_l P_l^m(\mu) \times \int_{-1}^{1} P_l^m(\mu')I_g^m(\tau,\mu')\mathrm{d}\mu' + S_g^m(\tau,\mu), \quad (12)$$

where

$$P_l^m(\mu) = \left[\frac{(l-m)!}{(l+m)!}\right]^{1/2}(1-\mu^2)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}\mu^m}P_l(\mu) \quad (13)$$

denotes an associated Legendre function, and the inhomogeneous source term is given by

$$S_g^m(\tau,\mu) = 2S_{0,g}(\tau)\delta_{0,m} + \frac{\varpi_g}{2}\frac{I_{0,g}}{\pi}e^{-\tau/\mu_0}\sum_{l=m}^{L}\beta_l P_l^m(\mu_0)P_l^m(\mu), \quad (14)$$

subject to the boundary conditions,

$$I_g^m(0,\mu) = I_g^m(\zeta,-\mu) = 0, \quad (15)$$

for $\mu \in (0,1]$. It is clear that once we solve the problems formulated by Eqs. (12) and (15), for $m = 0, 1, \ldots, L$, we can compute the scattered component of the intensity with Eq. (11).

Thus far we have considered a problem formulated as a single region. The extension of this method to multi-region geometry is based on the work performed by Dias and Garcia [24,25]. So we use, as mentioned in Refs. [24] and [25], "an iterative approach that is based on solving the problem one region at a time and using spatial sweeps to connect these solutions and guide them to convergence." For the treatment of the space variable, we consider a system of R regions, as shown in Fig. 2.

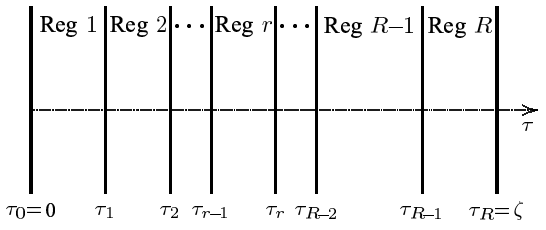


Figure 2. A system of R regions

The multi-region problem is then expressed, for $\mu \in (0,1]$ and $r = 1, 2, \ldots, R$, as

$$\mu \frac{\partial}{\partial \tau} I^m_{r,g}(\tau,\mu) + I^m_{r,g}(\tau,\mu) = \frac{\varpi_{r,g}}{2} \times \sum_{l=m}^{L} \beta_{l,r} P^m_l(\mu) \int_{-1}^{1} P^m_l(\mu') I^m_{r,g}(\tau,\mu') \mathrm{d}\mu' + S^m_{r,g}(\tau,\mu), \quad (16)$$

subject, for $\mu \in (0,1]$, to the boundary conditions,

$$I^m_{1,g}(\tau_0,\mu) = I^m_{R,g}(\tau_R,-\mu) = 0, \quad (17)$$

and to the interface conditions, for $r = 1, 2, \ldots, R-1$,

$$I^m_{r,g}(\tau_r,\pm\mu) = I^m_{r+1,g}(\tau_r,\pm\mu), \quad (18)$$

where $S^m_{r,g}(\tau,\mu)$ is given by

$$S^m_{r,g}(\tau,\mu) = 2S_{0,g}(\tau)\delta_{0,m} + \frac{\varpi_{r,g}}{2}\frac{I_{0,g}}{\pi} e^{-\tau/\mu_0} \times \sum_{l=m}^{L} \beta_{l,r} P^m_l(\mu_0) P^m_l(\mu). \quad (19)$$

To define our discrete-ordinates version of the problem posed by Eqs. (16) to (18), we utilize a quadrature of order N with nodes $\{\mu_j\}$ and weights $\{\eta_j\}$ to approximate the integral in Eq. (16). The selected quadrature scheme is the double quadrature of order $N = 2n$ obtained by applying a standard Gauss-Legendre scheme of order n to each of the half-intervals $[0,1]$ and $[-1,0]$. By using the elementary solutions of the discrete-ordinates equations and their orthogonality property developed in Ref. 26, we can write the general discrete-ordinates solution of order N as

$$I^m_{r,g}(\tau,\mu_j) = \sum_{k=1}^{n} \Big[A_{k,r,g}\Phi_{r,g}(\nu_k,\mu_j) e^{-(\tau-\tau_{r-1})/\nu_k} + B_{k,r,g}\Phi_{r,g}(-\nu_k,\mu_j) e^{-(\tau_r-\tau)/\nu_k} \Big] + \sum_{k=1}^{n} \left[\mathfrak{A}_{k,r,g}(\tau)\Phi_{r,g}(\nu_k,\mu_j) + \mathfrak{B}_{k,r,g}(\tau)\Phi_{r,g}(-\nu_k,\mu_j) \right]. \quad (20)$$

In Eq. (20), ν_k and $-\nu_k$, $k = 1, 2, \ldots, n$, denote, respectively, the inverses of the *positive* and the *negative* eigenvalues of the $N \times N$ matrix $\mathbf{\Xi}^{-1}(\mathbf{I} - \mathbf{W_{r,g}})$, where $\mathbf{\Xi} = \mathrm{diag}\{\mu_1, \mu_2, \ldots, \mu_N\}$, $\mathbf{I}$ is the identity matrix of order N, and $\mathbf{W_{r,g}}$ is an $N \times N$ matrix with elements

$$W_{i,j,r,g} = \frac{\varpi_{r,g}}{2}\eta_j \sum_{l=m}^{L} \beta_{l,r} P^m_l(\mu_i) P^m_l(\mu_j). \quad (21)$$

The elementary solutions $\Phi_{r,g}(\nu_k,\mu_j)$ and $\Phi_{r,g}(-\nu_k,\mu_j)$ present in Eq. (20) are, respectively, the jth components of the eigenvectors $\mathbf{\Phi_{r,g}}(\nu_k)$ and $\mathbf{\Phi_{r,g}}(-\nu_k)$, associated, respectively, with the eigenvalues $1/\nu_k$ and $-1/\nu_k$. Also the coefficients $\{\mathfrak{A}_{k,r,g}(\tau)\}$ and $\{\mathfrak{B}_{k,r,g}(\tau)\}$ of the particular solution can be expressed as [26]

$$\mathfrak{A}_{k,r,g}(\tau) = \frac{1}{N_{r,g}(\nu_k)} \sum_{i=1}^{N} \eta_i \Phi_{r,g}(\nu_k,\mu_i) \times \int_{\tau_{r-1}}^{\tau} S_{r,g}(x,\mu_i) e^{-(\tau-x)/\nu_k} \mathrm{d}x \quad (22a)$$

and

$$\mathfrak{B}_{k,r,g}(\tau) = -\frac{1}{N_{r,g}(-\nu_k)} \sum_{i=1}^{N} \eta_i \Phi_{r,g}(-\nu_k,\mu_i) \times \int_{\tau}^{\tau_r} S_{r,g}(x,\mu_i) e^{-(x-\tau)/\nu_k} \mathrm{d}x, \quad (22b)$$

with

$$N_{r,g}(\pm\nu_k) = \sum_{i=1}^{N} \eta_i \mu_i [\Phi_{r,g}(\pm\nu_k,\mu_i)]^2. \quad (23)$$

Note that the coefficients of the homogeneous solution $\{A_{k,r,g}\}$ and $\{B_{k,r,g}\}$ are the solutions to the linear system of N algebraic equations obtained by imposing that the general solution expressed by Eq. (20) satisfies the boundary and interface conditions of the problem. These calculations are reported in detail in Refs. 18 and 19.

INVERSE PROBLEM SOLUTION

The inverse problem is formulated as an optimization problem for the minimization of the norm of the differences between the measured data and the data obtained through the forward model presented in the previous section.

To simplify our solution, we split the internal source of radiation $S_{0,g}(\tau)$ in two functions

$$S_{0,g}(\tau) = S_0(\tau)\ Q(\lambda_g) = S_0(\tau)\ Q_g, \qquad (24)$$

and considered the function with spectral dependency Q_g a known function.

While the internal source of radiation $S_{0,g}(\tau)$ is unknown, the expansion coefficients $\{\beta_{l,r}\}$ of the phase function for anisotropic scattering, the single scattering albedo $\{\varpi_{r,g}\}$ and the optical thickness ζ, as well as the measured (exact) values of the exit radiances $I_g(0,-\mu,\varphi)$ are considered available. The inverse problem is iteratively solved by an algorithm for least-squares estimation where, in each iteration, values of the exit radiances $I_g(0,-\mu,\varphi)$ are computed from estimated values of the internal source profile $S_0(\tau)$. Here the calculated radiances are obtained from a modified and adapted package of subroutines taken from the forward model code, which we here refer to as subroutine *Peesna*, and the estimated parameters are obtained with the IMSL subroutine *Dbclsf* [27]. Basically, these calculations constitute our inverse model.

The IMSL subroutine *Dbclsf* uses a modified Levenberg-Marquardt method [20,21] of minimization and an active set strategy [28] to solve nonlinear least squares problems subject to simple bounds on the variables. The problem is stated as follows [27,29]:

$$\min_{x\in\Re^K} \frac{1}{2} Y^T(x)Y(x) = \min_{x\in\Re^K} \frac{1}{2}\sum_{i=1}^{M} [y_i(x)]^2,$$
$$d_j \le x_j \le u_j, \qquad (25)$$

where $M \ge K$, $Y : \Re^K \to \Re^M$, $y_i(x)$ is the *i*-th component function of $Y(x)$, and d_j and $u_j, j = 1,\ldots,K$, are the lower and upper bounds, respectively. The functions $y_i, i = 1,\ldots,M$, represent the differences between the experimental radiances and radiances that are calculated through each call to subroutine *Peesna*, and $x_j, j = 1,\ldots,K$, the unknown variables to be estimated by the IMSL subroutine *Dbclsf*. We implemented two techniques for retrieving the internal source profile. In the first one, referred to as the *average-value* technique, we consider within each region r, an average source value $\bar{S}_0$ represented by the variable x_j, and thus we estimate $K = R$ variables. And in the second one, referred to as the *three-coefficient* technique, the source $S_0(\tau)$ is determined by the quadratic expression

$$S_0(\tau) = x_1 + x_2(\tau/\zeta) + x_3(\tau/\zeta)^2, \qquad (26)$$

where the coefficients x_1, x_2 and x_3 are the only variables to be estimated. Note that, using the *average-value* technique for retrieving the source becomes very expensive, computationally speaking, as a large number of regions R is needed to generate a smooth profile.

NUMERICAL SOLUTION

Two problems were chosen to test the inverse model. The parameters that define the chosen problems were based on the simulation presented in Mobley's book [30], *Light and Water - Radiative Transfer in Natural Waters*, in Section 11.8 - "A Simulation of Case 1 Water". In addition, we considered an internal source generated by bioluminescent organisms.

The single scattering albedo $\varpi_{r,g}$, is calculated by the expression

$$\varpi_{r,g} = \frac{b_{r,g}}{c_{r,g}} = \frac{b_{r,g}}{b_{r,g} + a_{r,g}}, \qquad (27)$$

where $c_{r,g}, a_{r,g}$ and $b_{r,g}$ are the attenuation, absorption and scattering coefficients, respectively. The absorption and scattering coefficients, which are expressed in m^{-1}, are approximated by the equations [30]

$$a_{r,g} = \left[a_g^w + 0.06\ a_g^c\ C^{0.65}(z)\right] \times \left[1 + 0.2\ e^{-0.014(\lambda_g - 440)}\right] \qquad (28)$$

and

$$b_{r,g} = \left(\frac{550}{\lambda_g}\right) 0.30\ C^{0.62}(z), \qquad (29)$$

where a_g^w is the absorption coefficient of pure water, a_g^c a nondimensional, statistically derived

chlorophyll-specific absorption coefficient, and $C(z)$ the chlorophyll concentration, in $mg\ m^{-1}$, which is approximated as a background value plus a Gaussian [30]

$$C(z) = C_0 + \frac{h}{s\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{z-z_{max}}{s}\right)^2}, \qquad (30)$$

where the geometric depth z is expressed in meters. The values of the parameters a_g^w, a_g^c, for three chosen values of λ_g (500, 550 and 600), and those of C_0, h, s and z_{max}, obtained from Ref. 30, are shown in Table 1.

Table 1. Parameters obtained from Ref. [30]

Parameter	Value
$a^w(\lambda_g{=}500; 550; 600)$	$0.026; 0.064; 0.245\ \mathrm{m}^{-1}$
$a^c(\lambda_g{=}500; 550; 600)$	$0.668; 0.357; 0.236$
C_0	$0.2\ \mathrm{mg\ m}^{-3}$
h	$144\ \mathrm{mg\ m}^{-2}$
s	$9\ \mathrm{m}$
z_{max}	$17\ \mathrm{m}$

Note that the inherent optical properties, $c_{r,g}, a_{r,g}$ and $b_{r,g}$, are regarded as being averages within each region r, and that the correspondence between the optical depth and geometric depth is given by the expression $\tau(z) = \tau_{r-1} + (z - z_{r-1})c_{r,g}$, where

$$\tau_{r-1} = \sum_{i=1}^{r-1}(z_i - z_{i-1})c_{i,g}. \qquad (31)$$

For both problems, we consider $\varphi_0 = 0$, $\mu_0 = 1$, the Henyey-Greenstein parameter $f = 0.924$, the quadrature order $N = 130$, a water layer thickness of 40 meters, equidistantly divided into $R = 5$ regions. With the f factor, applied to all regions, the β_l values in Eq. (6) are obtained in two steps. First the code calculates the Henyey-Greenstein phase function [31]

$$\tilde{\beta}_{HG}(f;\Theta) \equiv \frac{1}{4\pi}\frac{1-f^2}{(1+f^2-2f\cos\Theta)^{3/2}}\ , \qquad (32)$$

then iteratively searches for a scattering order L that generates, through Eq. (6), a phase function whose graphic representation compares well with the graphic representation of the Henyey-Greenstein phase function given by Eq. (32), *i.e.*, the iteration process is stopped when corresponding points on the two graphs do not differ by more than $\pm 1\%$. The β_l values of Eq. (6) are determined through the expression [32]

$$\beta_l = (2l+1)\ f^l. \qquad (33)$$

Note that we consider $\beta_{l,1} = \beta_{l,2} = \beta_{l,\ldots} = \beta_{l,R}$, for the chosen multi-region problems.

With these input parameters, the simulated measured exit radiances $I(\tau = 0, -\mu, \varphi = 0)$, were determined by the forward model code at five values of μ $(-.96, -.97, -.98, -.99 \text{ and } -1)$, for each one of the three chosen values of λ_g, totalizing, in Eq. (25), $M = 15$ optimization functions. We used a constant source profile ($S_0(\tau) = 0.5$) and a sine source profile ($S_0(\tau) = \sin(\pi\tau/\zeta)$), in the first and second problems, respectively, considering in both problems $Q_g = 1.0$, in Eq. (24).

In order to simulate measured radiances Z_m containing measurement errors, the calculated data Z_e were corrupted with noise by using the IMSL subroutine *Drnnor* [33], which generates pseudo-random numbers from a standard normal distribution. Thus we have

$$Z_m = Z_e(1 + \kappa\ \xi), \qquad (34)$$

where κ is the percent noise and ξ is a random variable calculated by subroutine *Drnnor*.

We used both techniques, the *average-value* technique and the *three-coefficient* technique, to retrieve the internal source profiles. Besides the excessive computational time involved to solve the problems, when using the *average-value* technique, we were not able to obtain results with a good degree of accuracy. In Figs. 3 and 4 we show the estimated source profiles, using the *three-coefficient* technique for four selected experimental errors, $\kappa = 0, 1, 2$ and 5%. The values in parentheses indicate the percentage deviations of the calculated areas under the estimated curves from those under the exact curves.

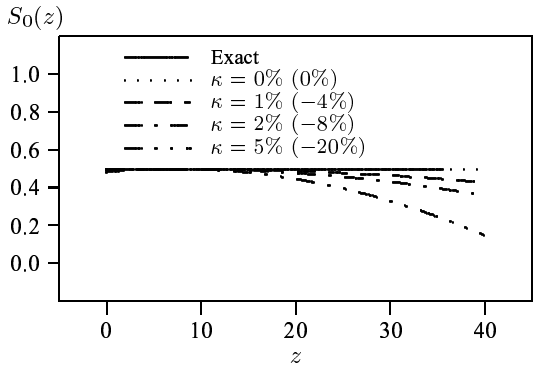


Figure 3: Estimation of the constant source profile.

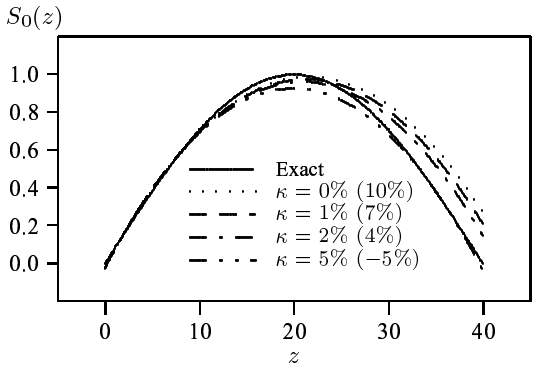


Figure 4: Estimation of the sine source profile.

FINAL COMMENTS

The inverse problem of estimating the internal sources in natural waters, using remote sensing data, is solved by an analytical discrete-ordinates method [17–19] and a modified Levenberg-Marquardt method [20,21,27] for the adopted forward model and inverse technique, respectively.

In the solution of the two problems chosen to test the implemented inversion technique, the analyses were performed by using simulated measurements containing random errors varying from 0 to 5%. We note that, as shown in Fig 3 and 4, the differences between the exact and estimated source profiles increase with depth. In the opinion of the authors, this bias is due to the physical nature of the problem, where the radiances exponentially decay within water, so the signal to noise relation greatly degrades for greater depths.

In order to solve our inverse problem, various modeling aspects had to be formulated, including the consideration of splitting the internal source in two functions, one carrying the spatial dependency and the other the spectral dependency, where the latter was considered as a known function. This simplification was applied in order to reduce the computational time involved in the solution.

Two techniques, the *average-value* technique and the *three-coefficient* technique, were implemented for retrieving the internal source profiles. The *average-value* technique was considered inadequate, due to its ineffectiveness and inaccuracy. With the *three-coefficient* technique, we were able to recover the desired profiles with an acceptable degree of accuracy, even with input data containing significant measurement errors and with a small number of measurement points.

We note that other techniques can be used to solve this type of inverse problems, such as the Tikonov regularization techniques or the principle of maximum entropy (in its various forms), Kalman filtering technique and the variational methods, and that we consider applying them in future works.

The estimation of the optical properties by using remote sensing data is still a challenge to the scientific community. This work represents the first effective step forward for the multispectral inversion. It is important to point out that there is no need in our analysis to impose an homogeneous ocean or any other constrained assumptions. Although we only present preliminary results of our research, we are able to conclude that it is possible to develop techniques for estimating the desired properties, using remote sensing data.

To close this work, we finally note that we expect soon to be able to extend our analysis for the estimation of other optical properties, such as the absorption and scattering coefficients.

ACKNOWLEDGEMENTS

This work was supported in part by FAPESP, São Paulo State Foundation for Research Support, Brazil, through a Thematic Project grant process 96/07200-8. One of the authors (ESC) wishes also to thank CNPq, National Counsel for Scientific and Technological Development, for the financial aid granted, through Research grant process 380465/00-0.

REFERENCES

1. N. J. McCormick, Recent developments in inverse scattering transport methods, *Transp. Theory Stat. Phys.*, **13**, 15 (1984).

2. N. J. McCormick, Methods for solving inverse problems for radiation transport—an update, *Transp. Theory Stat. Phys.*, **15**, 759 (1986).

3. N. J. McCormick, Inverse radiative transfer problems: a review, *Nuclear Sci. Eng.*, **112**, 185 (1992).

4. H. R. Gordon, Can the Lambert-Beer law be applied to the diffuse attenuation coefficient of ocean water?, *Limnol. Oceanogr.*, **34**, 1389 (1989).

5. D. L. Wooddruff, R. P. Stumpf, J. A. Scope and H. W. Pearl, Remote estimation of water clarity in optically complex estuarine waters, *Remote Sens. Environ.*, **61**, 290 (1997).

6. H. R. Gordon, Radiative transfer in the ocean: a method for determination of absorption and scattering properties, *Appl. Optics*, **15**, 2611 (1976).

7. R. W. Gould Jr. and R. A. Arnone, Remote sensing estimates of inherent optical properties in a coastal environment, *Remote Sens. Environ.*, **61**, 290 (1997).

8. M. K. Pinkerton, C. C. Trees, J. Aiken, A. J. Bale, G. F. Moore, R. G. Barlow and D. G. Cummings, Retrieval of near-surface bio-optical properties of the Arabian sea from remotely sensed ocean colour data, *Deep-Sea Res. Pt. II*, **46**, 549 (1999).

9. S. Stephany, Reconstruction of Optical Properties and Bioluminescent Sources in Natural Waters, Sc. D. thesis (in Portuguese), Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil (1997).

10. S. Stephany, F. M. Ramos, H. F. Campos Velho and C. D. Mobley, Reconstruction of bioluminescence sources in natural waters, International Conference on Computational Engineering Science, In: Atluri SN, Yagawa G. editors, Proceedings of ICES, 1997, p. 447.

11. S. Stephany, F. M. Ramos, H. F. Campos Velho and C. D. Mobley, A methodology for internal light sources estimation, *Comput. Model Simulat. Eng.*, **3**, 161 (1998).

12. S. Stephany, H. F. Campos Velho, F. M. Ramos and C. D. Mobley, Identification of Inherent Optical Properties and Bioluminescence Source Term in a Hydrologic Optics Problem, *J. Quant. Spectrosc. Radiat. Transfer*, **67**, 113 (2000).

13. E. S. Chalhoub, H. F. Campos Velho, F. M. Ramos and J. C. R. Claeyssen, Phase function estimation in natural waters using discretes ordinate method and maximum entropy principle, *Hybrid Methods in Engineering*, **2** (2000).

14. E. S. Chalhoub and H. F. Campos Velho, Simultaneous estimation of radiation phase function and albedo in natural waters, *J. Quant. Spectrosc. Radiat. Transfer*, **62**, 137 (2001).

15. E. S. Chalhoub and H. F. Campos Velho, Estimation of the optical properties of seawater from measurements of exit radiance, *J. Quant. Spectrosc. Radiat. Transfer*, **72**, 551 (2002).

16. H. R. Gordon, Remote sensing marine bioluminescence: the role of the in-water scaler irradiance, *Appl. Optics*, **23**, 1694 (1984).

17. E. S. Chalhoub, The Discrete-Ordinates Method for Solving Azimuthally-Dependent Transport Problems, Sc. D. thesis (in Portuguese), Universidade de São Paulo, Instituto de Pesquisas Energéticas e Nucleares, São Paulo, Brazil (1997).

18. E. S. Chalhoub and R. D. M. Garcia, The Equivalence between Two Techniques of Angular Interpolation for the Discrete-ordinates Method, *J. Quant. Spectrosc. Radiat. Transfer*, **64**, 517 (2000).

19. E. S. Chalhoub, Discrete-ordinates solution for radiative transfer problems, *J. Quant. Spectrosc. Radiat. Transfer*, accepted (2002).

20. K. Levenberg, A Method for the Solution of Certain Problems in Least Squares, *Quart. Appl. Math.*, **2**, 164 (1944).

21. D. Marquardt, An Algorithm for Least-squares Estimation of Nonlinear Parameters, *J. Soc. Indust. Appl. Math.*, **11**, 431 (1963).

22. S. Chandrasekhar, *Radiative Transfer*, Oxford University Press, London, 1950.

23. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

24. A. F. Dias, The P_N method for shielding calculations in multislab geometry, Sc. D. thesis (in Portuguese), Universidade de São Paulo, Instituto de Pesquisas Energéticas e Nucleares, São Paulo, Brazil, 1999.

25. A. F. Dias and R. D. M. Garcia, Coupled scalar and vector P_N methods for solving multigroup transport problems in multislab geometry, *Ann. Nucl. Energy*, **27**, 1607 (2000).

26. L. B. Barichello, R. D. M. Garcia and C. E. Siewert, Particular solutions for the discrete-ordinates method, *J. Quant. Spectrosc. Radiat. Transfer*, **64**, 219 (2000).

27. DBCLSF, *IMSL MATH/LIBRARY Users Manual*, Version 2.0, IMSL, Houston, 1991.

28. P. E. Gill and W. Murray, Minimization subject to bounds on the variables, NPL Report NAC 72, National Physical Laboratory, England, 1976.

29. M. N. Özişik and M. R. B. Orlande, *Inverse heat transfer: fundamentals and applications*, Taylor and Francis, London, 2000.

30. C. D. Mobley, *Light and Water - Radiative Transfer in Natural Waters*, Academic Press, San Diego, 1994.

31. L. C. Henyey and J. L. Greenstein, Diffuse radiation in the galaxy, *Astrophys J.*, **93**, 70 (1941).

32. G. W. Kattawar, A three-parameter analytic phase function for multiple scattering calculations, *J. Quant. Spectrosc. Radiat. Transfer*, **15**, 839 (1975).

33. IMSL, *IMSL STAT/LIBRARY Users Manual*, Version 2.0, IMSL, Houston, 1991.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

PASSIVE ELECTRIC POTENTIAL CT METHOD USING PIEZOELECTRIC MATERIAL FOR CRACK IDENTIFICATION

Daiki Shiozawa
Department of Mechanical Engineering and Systems
Graduate School of Engineering, Osaka University
2-1, Yamadaoka, Suita, Osaka 565-0871 Japan
shiozawa@saos.mech.eng.osaka-u.ac.jp

Shiro Kubo
Dept. of Mechanical Engineering and Systems
Graduate School of Engineering, Osaka Univ.
2-1, Yamadaoka, Suita, Osaka 565-0871 Japan
kubo@mech.eng.osaka-u.ac.jp

Takahide Sakagami
Department of Mechanical Engineering and Systems
Graduate School of Engineering, Osaka University
2-1, Yamadaoka, Suita, Osaka 565-0871 Japan
sakagami@mech.eng.osaka-u.ac.jp

ABSTRACT

When the piezoelectric film is glued on the surface of a cracked material subjected to mechanical load, change in electric potential distribution is observed on the surface of film. Based on this phenomenon, the passive electric potential CT (computed tomography) method can be developed, which does not require electric current application for identifying cracks. This method may be applied to develop an intelligent structure with a function of self-monitoring of flaws and defects. For the crack identification from electric potential distribution, an inverse method based on the least residual method was applied, in which square sum of residuals are evaluated between the measured electric potential distributions and those computed by using the finite element method. Numerical simulations were carried out on identification of a through-thickness transverse crack. It was found that the location and size of the crack can be quantitatively identified by the proposed passive electric potential CT method.

NOMENCLATURE

a half length of crack
$[C]$ stiffness matrix
$\{D\}$ electric displacement vector
$[e]$ piezoelectric coefficient matrix
$\{E\}$ electric field vector
E_{elas} Young's modulus of substrate material
$\{F\}$ mechanical load vector
$[g]$ dielectric constant matrix
G_{elas} shear modulus of substrate material
h crack depth
H distance between locations taking peaks of electric potential
$[K_{uu}]$ mass matrix
$[K_{u\phi}]$ displacement electric stiffness matrix
$[K_{\phi\phi}]$ electric stiffness matrix
M number of measuring point
$\{Q\}$ electric load vector
R_s square sum of residual between measured and computed potential
t_{piezo} thickness of piezoelectric film
x_c crack location

Greeks

$\{\varepsilon\}$ strain vector
ϕ electric potential
ϕ_0 remote value of electric potential
$\phi_i^{(c)}$ electric potential values at i-th measuring point computed by the FEM
$\phi_i^{(m)}$ measured electric potential value at i-th measuring point
$\phi_{\max}$ peak value of electric potential
ν_{elas} Poisson's ratio of substrate material
ρ density
$\{\sigma\}$ stress vector

INTRODUCTION

Non-destructive and real-time damage monitoring technique is important for maintenance of in-service structures such as, aircrafts, space structures or nuclear power plants. Non-destructive crack identification is recognized as a domain/boundary inverse problem [1] which

deals with the estimation of an unknown boundary. Conventional NDT (non-destructive testing) methods such as ultrasonic method, radiation method, electric-potential method may not be applied for the purpose, since they have some limitations in their applications for automatic inspection, non-contact inspection or remote inspection in severe environment. Development of an 'intelligent structure' [2], which has a function of self-damage detection and monitoring, is required for solving the problems. The intelligent structure with self-damage monitoring will provide us continuous and real-time assessment of structural integrity and also gives us warning signals of damage propagation before catastrophic failure of the structure.

Piezoelectric film is a sensing device that generates an electrical charge proportional to a change in mechanical strain. Several investigations have been conducted on the development of intelligent structures using piezoelectric materials. Galea et al. [3] showed the possibility of the use of piezoelectric PVDF (poly vinylidene fluoride) film as a sensing device for detecting and monitoring damages in composite materials. Yin et al. [4] carried out numerical analyses to demonstrate the feasibility of applying PVDF film for damage detection in composites. Li et al. [5] made theoretical and numerical investigation on the development of crack identification technique for the structures on which piezoelectric material was installed.

The present authors proposed the active electric potential CT (computed tomography) method [6-8] for quantitative identification of two- and three-dimensional cracks, by using electric potential distributions observed under electric current aplications. The purpose of our study is the development of passive electric potential CT method for quantitative crack identification based on a change in distribution of electric potential observed on the surface of PVDF film, when a cracked material is subjected to mechanical load. In this paper, the effects of crack location and size on the electric potential distribution are investigated by the FEM (finite element method) analyses. Numerical simulations are carried out on the estimation of location and size of through thickness transverse crack, based on the FEM inverse analyses.

FINITE ELEMENT ANALYSIS

When cracked material is subjected to mechanical load and PVDF film is glued on the surface of the material, a change in electric potential distribution is observed on the surface of PVDF film. The FEM computer analysis scheme was developed [9] for coupled elastic and electric potential problem to investigate the relationship between crack parameters and electric potential distribution on PVDF film. The governing equations of the piezoelectric material can be written as [10];

$$\{\sigma\} = [C]\{\varepsilon\} - [e]^T \{E\} \tag{1}$$

$$\{D\} = [e]\{\varepsilon\} + [g]\{E\} \tag{2}$$

where $\{\sigma\}$ and $\{\varepsilon\}$ are stress and strain vector, $[C]$, $[e]$ and $[g]$ are stiffness matrix, piezoelectric coefficient matrix and dielectric constant matrix, respectively. $\{E\}$ is electric field vector. $\{D\}$ is electric displacement vector. The static FEM equation, based on Eqns. (1) and (2), is obtained as,

$$\begin{aligned} [K_{uu}]\{d\} + [K_{u\phi}]\{\phi\} &= \{F\} \\ [K_{u\phi}]\{d\} + [K_{\phi\phi}]\{\phi\} &= \{Q\} \end{aligned} \tag{3}$$

where $[K_{uu}]$, $[K_{u\phi}]$ and $[K_{\phi\phi}]$ are the mass matrix, displacement electric stiffness matrix and electric stiffness matrix, respectively. $\{F\}$ and $\{Q\}$ are the mechanical load vector and the electric load vector, respectively.

Table 1. Properties of the piezoelectric film

Elastic properties ($\times 10^{10}$N/m^2)	c_{11}	23.82×10^{-10}
	c_{12}	3.98×10^{-9}
	c_{13}	2.19×10^{-9}
	c_{33}	10.64×10^{-9}
	c_{44}	2.15×10^{-9}
Piezoelectric properties (C/m^2)	d_{31}	25×10^{-12}
	d_{33}	2×10^{-12}
	d_{15}	35×10^{-12}
Dielectric properties ($\times 10^{-9}$C/Vm)	g_{11}	1.15×10^{-10}
	g_{33}	1.15×10^{-10}
Density (10^3kg/m^3)	ρ	1.75
Thickness (mm)	t_{piezo}	0.04

Table 2. Properties of the elastic substrate material

E_{elas}	G_{elas}	ν_{elas}	ρ
70.56 (GPa)	26.46 (GPa)	0.33	7.6 (10^3 kg/m^3)

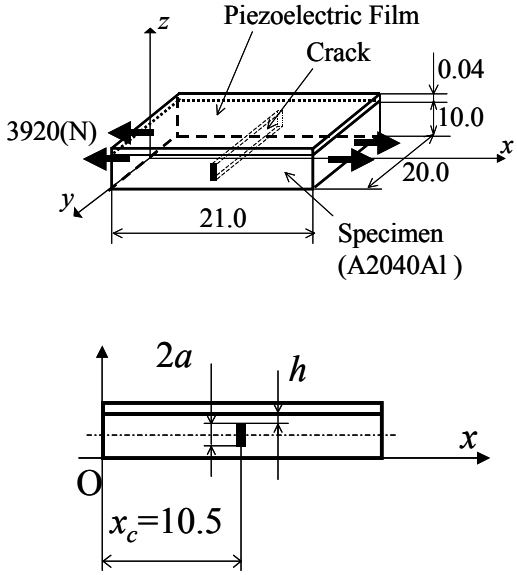


Fig. 1. Model used in analyses

From Eqn. (3), coupled property between an elastic field and an electric potential field is given in $[K_{u\phi}]$.

A model shown in Fig. 1 is employed for the FEM analyses. This model consists of an elastic substrate material and a PVDF film.

Crack parameters are chosen as follows: a is half length of the crack, h is crack depth from surface of the plate and x_c is crack location in the x-direction. The properties of the piezoelectric material [11] and the elastic material are shown in Tables 1 and 2, respectively. It was assumed that the potential on the interface between the elastic material and PVDF film is 0. Mechanical load was applied in the x-direction.

For evaluating the effect of crack length a on electric potential distribution, electric potential distributions for three combinations of crack parameters, *i.e.* (a, h) = (1, 2), (2, 2) and (3, 2) with h keeping constant, are compared. Furthermore, electric potential distribution on PVDF film was obtained in the case of no crack for investigating the effect of the existence of crack on the electric potential distribution.

The results of FEM calculations are shown in Fig. 2. It is found in Fig. 2 that the electric potential value is higher than the remote value ϕ_0, which is the same as the value for the case of no crack. The electric potential values show a symmetrical change with respect to the crack location $x = x_c = 10.5$. The electric potential distribution has two peaks taking a peak value $\phi_{\max}$. The location of local minimum between the two peaks of potential coincides with location of the crack. It is also found in Fig. 2 that the peak value of electric potential $\phi_{\max}$ increases with increase in crack length a.

For examining the effect of h on electric potential distribution, the electric potential distributions for five combinations of crack parameters, *i.e.* (a, h) = (2, 1), (2, 2), (2, 3), (2, 4) and (2,5) with keeping a constant, are compared. The results of FEM calculations are shown in Fig. 3. It is found that the peak value of electric potential $\phi_{\max}$ decreases with increase in crack depth h. It is also found that the distance between two peaks H increases with increase in crack depth h.

From Figs. 2 and 3, the following features were found in the relationship between electric potential distribution and crack parameters.

(a) When the plate has a crack, the electric potential values on PVDF film are higher than that of the plate without crack, and show a characteristic distribution.
(b) Electric potential distribution shows symmetrical shape with two peaks, and the location of transverse crack coinciding with the point of local minimum between the two peaks.
(c) The value of electric potential at the peak $\phi_{\max}$ changes with the crack length a: $\phi_{\max}$ is larger for the longer crack.
(d) The value of electric potential at the peak $\phi_{\max}$ changes with the crack depth h: $\phi_{\max}$ is larger for the smaller crack depth. The distance between two peaks H is larger for the larger crack depth.

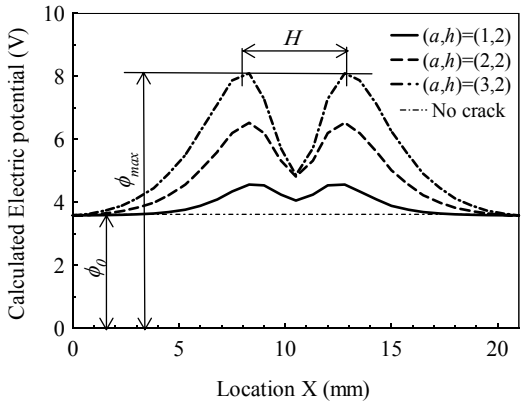


Fig. 2. Effect of crack length on electric potential distributions

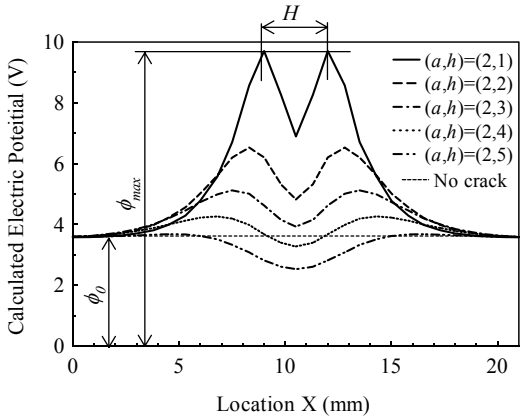


Fig. 3. Effect of crack depth on electric potential distributions

CRACK IDENTIFICATION

As the inverse analysis method for identification of cracks, the least residual method was applied. In this method, computed values $\phi^{(c)}$ are compared with the measured values $\phi^{(m)}$ to determine the most plausible crack location and size. As a criterion for crack identification the following square sum R_s of residual is calculated.

$$R_S(a,h,x_C) = \sum_{i}^{M} \left(\phi_i^{(c)}(a,h,x_C) - \phi_i^{(m)}\right)^2 \quad (4)$$

Here $\phi_i^{(m)}$ denotes measured electric potential value at the i-th measuring point, and $\phi_i^{(c)}(a,h,x_c)$ denotes the electric potential values at the i-th measuring point computed by the FEM, in which crack parameters are assumed to be a, h and x_c. M is the total number of measuring points. The combination of crack location and size, which minimized R_s, was determined as the most plausible one among all the assumed combinations of the crack location and size. In the numerical simulation of crack identification, crack parameters (a, h, x_c) were set to be (3.2, 2.1, 11.3).

In the actual applications, $\phi_i^{(m)}$ are obtained experimentally. In the present computer simulation, the measured values are obtained by the FEM analysis. Artificial noise was added to the computed values. Several noise levels, *i.e.* $\pm 0.5\%$, $\pm 1.0\%$ and $\pm 5.0\%$, were selected. On the surface of PVDF film, electric potential was measured at 49 points placed with an interval of 0.5mm as shown in Fig. 4.

For effective inverse analysis, the following hierarchical calculation steps were introduced.

(a) In the first step, crack parameters are roughly estimated. The crack location in the x-direction x_c is determined to be 11.5 from the location of local minimum between two peaks in the electric potential distribution. In the estimation of a and c, R_s is calculated for the combinations of three crack lengths and three crack depths as shown in Fig. 5. It is assumed that R_s is approximated by the following quadratic function of a and h.

$$R_S(a,h,x_c) = A + Ba + Cah + Da^2 + Eh + Fh^2 \quad (5)$$

Coefficients A, B, C, D, E and F are determined by the least-squares method from the values of R_s for the combinations of three crack lengths and three crack depths. The combination of a and h, which minimized this approximate function for R_s, is employed as the plausible combination in the rough estimation of crack parameters.

(b) In the second step, the combination of crack parameters, which gives the minimum R_s, is searched by using the modified Powell optimization method [12]. The crack parameters obtained in the above rough

estimation are used as the initial values of the crack parameters for the modified Powell method.

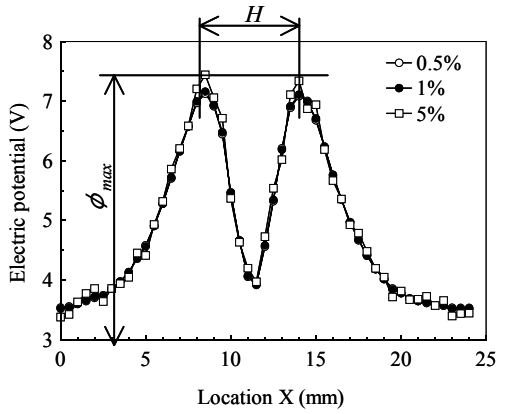


Fig. 4. Measured electric distributions

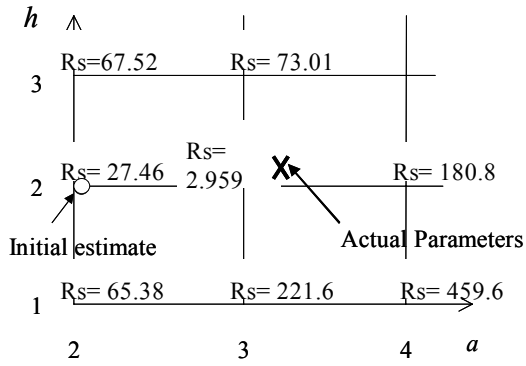


Fig. 5. Values of residual (Noise level of 5%)

The estimated crack parameters using the least residual method with the modified Powell method are shown in Table 3.

The combination of initial estimate of parameters used in the modified Powell method is shown by a circle in Fig. 5. R_s values are shown at the grid points of crack parameters. It is found from the table and the figure that the crack parameters can be estimated in a good accuracy and crack parameters can be identified within the error of 1.0%, when the noise level of observed electric potential distribution is lower than 1.0%.

The error in the estimated value of x_c is found to be smaller when compared with those in the other crack parameters, a and h. This is due to the relationship between crack parameters and electric potential distribution on PVDF film discussed in the foregoing: crack location x_c can be determined from the point of local minimum between two peaks of the electric potential distribution, so the accuracy in the estimation of x_c is not affected by absolute values of the electric potential value $\phi_{\max}$ at the peak. On the other hand, crack length a and depth h were determined from the distance of peaks H and magnitude of $\phi_{\max}$. Therefore, the accuracy in the estimation of crack parameters a and h is prone to the noise in measurements.

Table 3. Estimated parameters of cracks

Noise level		Crack parameters (mm)			Residual R_s
		a	h	x_c	
	Actual	3.2	2.1	11.3	
±0.5	Estimated	3.203	2.095	11.30	8.286
(%)	Error (%)	0.100	0.222	0.008	$\times 10^{-03}$
±1.0	Estimated	3.206	2.091	11.30	3.314
(%)	Error (%)	0.202	0.441	0.016	$\times 10^{-02}$
±5.0	Estimated	3.231	2.054	11.29	8.286
(%)	Error (%)	0.970	2.226	0.008	$\times 10^{-1}$

Error is defined as the ratio of the difference $\triangle$ between actual and estimated values to actual value.

CONCLUSIONS

The passive electric potential CT method was proposed, in which two- and three-dimensional cracks were identified. This method is based on a change in distribution of electric potential distribution observed on PVDF film glued on cracked material subjected to mechanical load. The electric potential distribution on PVDF film was investigated by the FEM. It was found that the electric potential distribution showed a characteristic change with crack location and size. This fact can be used for the identification of the crack. Numerical simulations were carried out for the estimation of location and size of through-

thickness transverse crack. It was found that crack parameters can be identified within the error level of 1.0%, when the noise level of observed electric potential distribution was lower than 1.0%.

Experimental examination on the applicability of the proposed method to the identification of cracks in a homogeneous body and bonded dissimilar bodies is now underway.

ACKNOWLEDGMENT

This work was partly supported by the Ministry of Education, Science, Sports and Culture, Japan under the Grant-in-Aid for Scientific Research,

REFERENCES

1. S. Kubo, “Inverse Problems Related to the Mechanics and Fracture of Solids and Structures”, JSME *International Journal, Series I* , **31**-2, pp.157-166(1988).

2. E. F. Crawley, and J. Luist, “Use of Piezoelctric Actuator as Elements of Intelligent Structures”, *AIAA Journal*, **25**, pp.1375-1385(1987).

3. S. C. Gelea, W. K. Chiu, and J. J. Paul, “Use of Piezoelectric Films in Detecting and Monitoring Damage in Composites”, *International Journal of Intelligent Material Systems and Structures*, **4**, pp. 330-336(1993).

4. L. Yin, X.-M. Wang, and Y.-P. Shen, “Damage-Monitoring in Composite Laminates by Piezoelectric Films”, *Computers and Structures*, **59**-4, pp.623-630(1995).

5. S-Q. Li, S. Kubo, and T. Sakagami, “Theoretical and Numerical Investigation on Crack Identification Using Piezoelectric Material-Embedded Structures”, *Material Science Research Int.*, **6**-1, pp.41-48(2000).

6. S. Kubo, T. Sakagami, and K. Ohji, "Electric Potential CT Method Based on BEM Inverse Analyses for Measurement of Three-Dimensional Cracks", *Computational Mechanics '86, Proc. of Int. Conf. on Computational Mechanics,* Vol.1, Springer, 1986, pp.V-339-V-344.

7. T. Sakagami, S. Kubo, T. Hashimoto, H. Yamawaki, and K. Ohji, "Quantitative Measurement of Two-Dimensional Inclined Cracks by the Electric-Potential CT Method with Multiple Current Applications", *JSME Int. J.,* Ser.I, Vol.31, No.1, Jan. 1988, pp.76-86.

8. S. Kubo, T. Sakagami, and K. Ohji, “The Electric Potential CT Method for Measuring Two- and Three-Dimensional Cracks”, *Current Japanese Materials Research- Vol.8 Fracture Mechanics,* Okamura, H. and Ogura, K. eds., Elsevier Applied Science , Society Material Science, Japan, pp.235-254(1991).

9. W. Cai, S. Q. Lo, Y. W. Yang, and Z. X. Liu, “The Finite Element Method for the Vibration Control of Piezoelectric Laminated Plate”, *Journal of Solid Mechanics*, (in Chinese), **18**-4, 76(1997).

10. IEEE, IEEE Standard on Piezoelectricity, *IEEE/ANSI Std.*, (1978).

11. T. Tashiro, H. Tadokoro, and M. Kobayashi, “Structure and Piezoelectricity of Poly(Vinylidene Fluoride)”, *Ferroelectrics*, **32**, pp.167-175(1981).

12. W. I. Zangwill, “Minimizing a Function without Calculating Derivatives”, *Computer J.*, **7**, pp.149-154(1964).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ESTIMATION OF THE RELEASE HISTORY OF A CONTAMINANT SOURCE IN 2-D GROUNDWATER SYSTEMS

Nerbe J. Ruperti Jr.
Waste Management Department, COREJ
Brazilian Nuclear Energy Commission, CNEN
Rio de Janeiro, RJ, Brazil
nruperti@cnen.gov.br

ABSTRACT

The purpose of this paper is to present an efficient inverse solution based on the use of an inverse sequential technique to estimate the unknown time-dependent mass flux release of a contaminant source into 2-D aquifers. One example was chosen to demonstrate the feasibility of such estimates using exact, noisy and noisy filtered input data to numerically simulate experimental measurements.

INTRODUCTION

Usually most simulations in the application of groundwater models are calibrated using trial and error instead of through the use of automated inverse modeling. Inverse modeling [1] has been used to estimate parameters, such as transverse and longitudinal dispersivities, from field-scale tracer experiments [2-4].

Other type of inverse problem involves the reliable assessment of water contamination due to unknown pollution releases, requiring the use of methods to estimate the location or the release history of the contaminant source [5,6]. Major sources of contamination are landfills, radioactive and hazardous waste disposals, industrial facilities, and runoff of fertilizer on agricultural land.

The solution of an inverse problem is commonly associated to the numerical solution of a forward problem. Throughout the last decades a great effort has been done in order to develop purely numerical and hybrid numerical-analytical techniques for the prediction of soil contamination. The Generalized Integral Transform Technique (GITT), fully described in reference [7], has been used as a reliable tool for both benchmark solutions and direct engineering simulations for different linear and nonlinear heat and mass transfer problems. Recently, the GIT technique has been applied to the solution of groundwater problems, such as the one-dimensional solute transport in unsaturated porous media [8] and simulations of two-dimensional contaminant transport in groundwater pathway [9].

The aim of the present work is to make use of an inverse sequential technique associated with the GIT technique to estimate the unknown release history of a contaminant source into 2-D aquifers. A test case was considered to simulate concentration measurements at given sampling locations by the solution of a direct problem. Then, the inversions are performed from these data, and the estimated mass flux and concentration profiles are compared with the imposed quantities.

INVERSE PROBLEM FORMULATION

Consider a finite section of an isotropic, homogeneous aquifer under saturated conditions with thickness H and length L, where the flow is horizontal and steady with an average pore velocity V_{aq}, as shown in Fig. 1. The contaminant source is located at the top and has a uniform width L_2-L_1. The following assumptions are made:

1. A cartesian coordinate system is used, as shown below in Fig. 1;
2. The mass flux being released from the source is an unknown function of time;
3. Transport is limited to a single specie that may decay or degrade as a function of time;
4. The properties of the medium are constant and known.

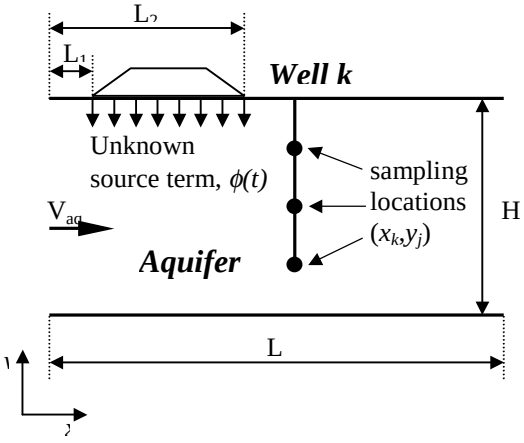


Figure 1. Schematic representation of the proposed inverse problem.

The dimensionless governing equations for the mass transport problem are:

$$R_d \frac{\partial C(x,y,t)}{\partial t} = D_x \frac{\partial^2 C}{\partial x^2} + Ar^2 D_y \frac{\partial^2 C}{\partial y^2} - V_{aq} \frac{\partial C}{\partial x} - R_d \lambda C, \quad 0 < x < 1;\ 0 < y < 1 \text{ and } t > 0 \quad (1.a)$$

with the known initial and boundary conditions, respectively, as:

$$C(x,y,0) = 0, \quad 0 < x < 1;\ 0 < y < 1 \quad (1.b)$$

$$\left.\frac{\partial C}{\partial x}\right|_{x=0} = \left.\frac{\partial C}{\partial x}\right|_{x=1} = 0, \quad 0 < y < 1;\ t > 0 \quad (1.c,d)$$

$$\left.\frac{\partial C}{\partial y}\right|_{y=0} = 0, \quad 0 < x < 1;\ t > 0 \quad (1.e)$$

and with the unknown boundary condition at *y=1*:

$$\left.\frac{\partial C}{\partial y}\right|_{y=1} = \frac{\phi(t)}{Ar\, D_y} \left[u_s(x - L_1) - u_s(x - L_2)\right], \quad 0 < x < 1;\ t > 0 \quad (1.f)$$

The following non-dimensional groups have been used:

$$t = \frac{D_0\, t^*}{L^{*2}};\ x = \frac{x^*}{L^*};\ y = \frac{y^*}{H^*};\ Ar = \frac{L^*}{H^*};\ C = \frac{C^*}{C_0}$$

$$\phi = \frac{L^*\, \phi^*}{D_0\, C_0};\ D_x = \frac{D_x^*}{D_0};\ D_y = \frac{D_y^*}{D_0};\ V_{aq} = \frac{L^*\, q_{aq}^*}{D_0\, \varepsilon};$$

$$V_{inj} = \frac{L^*\, q_{inj}^*}{D_0\, \varepsilon};\ \alpha = \frac{\alpha^*}{L^*};\ \lambda = \frac{\gamma\, L^{*2}}{D_0} \quad (2.a\text{-}l)$$

where the superscripts "*" identifies the dimensional variables, the subscript "0" denotes reference values of the parameters, *Ar* is the aspect ratio, λ is the dimensionless decay rate constant, $\phi(t)$ is the dimensionless time-dependent injection function, u_s is the step function and the Darcy velocity is represented by *q**. The dimensionless retardation factor R_d and the dispersion coefficients in the *x* and *y* directions, D_x and D_y [m^2/yr], are defined, respectively, as:

$$D_x = \alpha_L\, V_{aq} \quad (3.a)$$

$$D_y = \alpha_t\, V_{aq} \quad (3.b)$$

$$R_d = 1 + \frac{\rho_s K_d}{\varepsilon} \quad (3.c)$$

The longitudinal and transversal dispersivities are represented by α_L and α_t; ρ_s is the soil bulk density; K_d is the distribution coefficient and the dimensionless effective porosity is given by ε.

The inverse mass transfer problem herein considered consists in estimating the unknown contaminant release function, $\phi(t)$, based on concentration measurements, $Cw(x_k,y_j,t)$, taken at given sampling locations. The input data for this inverse problem, instead of being measured concentrations, are predicted from the solution of a direct problem for a known set of boundary conditions.

SOLUTION OF THE DIRECT PROBLEM

The numerical solution of Eqs. (1) for a given function, $\phi(t)$, can be obtained through the use of the Generalized Integral Transform Technique (GITT). The detailed numerical solution and its validation is fully described in [9].

The GIT technique, is a hybrid numerical-analytical approach based on the eigenfunction expansion of the original potential [7]. The hybrid nature of this approach allows for the automatic global error control along the solution process, towards an user prescribed accuracy target. The basic steps in applying the generalized approach

in the present work are: (a) choose the related auxiliary problem in *y* direction (a particular case of the classical Sturm-Liouville problem with second type boundary conditions); (b) develop the appropriate integral transform pair in *y* direction; (c) transform by a single integration the original *PDE* equation and its initial and boundary conditions into a system of *PDE*s; (d) numerically solve the *PDE* system by using the *DMOLCH* routine [10]; (e) invoke the inversion formula to construct the original potential.

The appropriate auxiliary problem in the *y* direction is chosen as:

$$\frac{d^2 Y_\ell(y)}{dy^2} + \eta_\ell^2\, Y_\ell(y) = 0 \qquad (4.a)$$

with boundary conditions:

$$\left.\frac{dY_\ell(y)}{dy}\right|_{y=0} = \left.\frac{dY_\ell(y)}{dy}\right|_{y=1} = 0 \qquad (4.b,c)$$

where the analytical solution yields eigenfunctions and eigenvalues, respectively,

$$Y_\ell(y) = \cos\eta_\ell\, y \qquad (5)$$

and,

$$\eta_\ell = (\ell - 1)\pi \quad \text{for} \quad \ell = 1, 2, 3, ... \qquad (6)$$

as well as norms:

$$M_\ell = \begin{cases} 1 & \text{for} \quad \eta_\ell = 0 \\ 1/2 & \text{for} \quad \eta_\ell \neq 0 \end{cases} \qquad (7)$$

Problem (4) allows the definition of the integral transform pair:

$$\overline{C}_\ell(x,t) = \int_0^1 \tilde{Y}_\ell(y)\, C(x,y,t)\, dy \text{ ; transform} \qquad (8)$$

$$C(x,y,t) = \sum_{\ell=1}^{\infty} \tilde{Y}_\ell(y)\ \overline{C}_\ell(x,t) \text{ ; inverse} \qquad (9)$$

The normalized eigenfunction $\tilde{Y}_\ell(y)$ is defined by $\tilde{Y}_\ell(y) = Y_\ell(y) / M_\ell^{1/2}$.

Following the formalism of the classical integral transform approach, a single transformation is operated in the differential equation (1) in *y* direction, i.e., $\int_0^1 \tilde{Y}_\ell(y)\, ...\, dy$; and making the appropriate use of the inversion formula (9) and the orthogonality property, one obtains:

$$R_d \frac{\partial \overline{C}_\ell(x,t)}{\partial t} = D_x \frac{\partial^2 \overline{C}_\ell}{\partial x^2} - Ar^2 D_y\, \eta_\ell^2\, \overline{C}_\ell - V_{aq} \frac{\partial \overline{C}_\ell}{\partial x} - R_d \lambda\, \overline{C}_\ell + Ar\ \phi(t)\ \left[u_s(x - L_1) - u_s(x - L_2)\right] E_\ell \qquad (10.a)$$

The same transformation procedure operated on the initial and boundary conditions provides:

$$\overline{C}_\ell(x,0) = 0 \qquad (10.b)$$

$$\left.\frac{\partial \overline{C}_\ell(x,t)}{\partial x}\right|_{x=0} = \left.\frac{\partial \overline{C}_\ell(x,t)}{\partial x}\right|_{x=1} = 0 \qquad (10.c)$$

where, E_ℓ is an analytical coefficient, obtained as follows:

$$E_\ell = \int_0^1 \tilde{Y}_\ell \delta(y-1)\, dy = \frac{1}{M_\ell^{1/2}} \cos\eta_\ell \qquad (11)$$

The expansion in Eq. (9) has to be truncated to a sufficiently large finite order N_y, so as to reach the requested accuracy target. The subroutine DMOLCH from the IMSL [10] was employed to numerically solve the resulting system of N_y partial differential equations, Eqs. (10). A first attempt was made to solve simultaneously the whole *PDE* system using DMOLCH. Due to workspace limitations and a high computational cost, another strategy was adopted. The *PDE* system was separated into small subsets of five equations and solved sequentially. Then, the $N_y/5$ subsets were joined to furnish the complete transformed potential field.

Once the transformed potentials $\overline{C}_\ell(x,t)$ are obtained for $\ell = 1, ..., N_y$, the original dimensionless concentration field can be evaluated through the use of the inverse formula, Eq. (9). Note that the transformed potential $\overline{C}_0(x,t)$ is the average concentration in *y* direction:

$$\overline{C}_0(x,t) = \int_0^1 C(x,y,t)\,dy \qquad (12)$$

For $\eta_0 = 0$ Eq. (10.a) simplifies to the partial differential equation:

$$R_d \frac{\partial \overline{C}_0(x,t)}{\partial t} = D_x \frac{\partial^2 \overline{C}_0}{\partial x^2} - V_{aq} \frac{\partial \overline{C}_0}{\partial x} - R_d \lambda \overline{C}_0 + Ar\ \phi(t)\ [u_s(x-L_1) - u_s(x-L_2)] \qquad (13)$$

that does not depend on the other transformed potentials. This characteristic allows the estimation of $\phi(t)$ based on the inverse solution of the above equation.

SOLUTION OF THE INVERSE PROBLEM

The sensitivity coefficient for the proposed inverse problem is defined by:

$$Z(x_k,t) = \frac{\partial \overline{C}_0(x_k,t;\phi)}{\partial \phi}, \quad k=1,2,\ldots,M \qquad (14)$$

where M is the total number of wells.

Function $Z(x,t)$ is calculated from its own partial differential equation, which can be derived from Eq. (13) to yield:

$$R_d \frac{\partial Z(x,t)}{\partial t} = D_x \frac{\partial^2 Z}{\partial x^2} - V_{aq} \frac{\partial Z}{\partial x} - R_d \lambda Z + Ar\ [u_s(x-L_1) - u_s(x-L_2)] \qquad (15.a)$$

with the initial and boundary conditions:

$$Z(x,t_{m-1}) = 0 \qquad (15.b)$$

$$\left.\frac{\partial Z(x,t)}{\partial x}\right|_{x=0} = \left.\frac{\partial Z(x,t)}{\partial x}\right|_{x=1} = 0 \qquad (15.c)$$

Beck's sequential method [11] was chosen to estimate the unknown function $\phi(t)$, using the observations of solute concentrations at given sampling wells. The data were averaged over the cross section perpendicular to the flow direction. The average concentration for well k is calculated from the measured concentrations at N_w equally weighted observation locations:

$$\overline{C}w_{k,m} = \sum_{j=1}^{Nw} \frac{1}{N_w} Cw(x_k, y_j, t_m), \quad k=1,2,\ldots,M \qquad (16)$$

where M is the total number of wells, and the sampling locations in y direction, y_j, are given by:

$$y_j = \frac{(2j-1)}{2N_w}, \qquad j=1,2,\ldots,N_w \qquad (17)$$

The least squares error function, for M wells and r future times, and with the temporary assumption that a constant mass flux ϕ_m is applied over the time interval $t_{m-1} \le t \le t_{m+r-1}$, can be defined by:

$$\varphi = \sum_{k=1}^{M} \sum_{j=1}^{r} \left[\overline{C}w_{k,m+j-1} - \overline{C}_0(x_k, t_{m+j-1}; \phi_m)\right]^2, \qquad (18)$$

Expanding the concentration field in a Taylor series about an assumed mass flux ϕ_t, the value of ϕ_m that minimizes φ is given by:

$$\hat{\phi}_m = \phi_t + \frac{\sum_{k=1}^{M} \sum_{j=1}^{r} \left[\overline{C}w_{k,m+j-1} - \overline{C}_0(x_k, t_{m+j-1}; \phi_t)\right] Z(x_k, j\Delta t)}{\sum_{k=1}^{M} \sum_{j=1}^{r} Z^2(x_k, j\Delta t)} \qquad (19)$$

The mass flux estimations can be compared with the true mass flux in order to evaluate the behavior of the proposed solution. The standard deviation between exact and estimated mass fluxes is given by the following expression:

$$S = \left[\frac{1}{N_t} \sum_{i=1}^{Nt} \left(\phi(t_i) - \hat{\phi}_i\right)^2\right]^{1/2} \qquad (20)$$

TEST CASE

A hypothetical test case was selected in order to verify the feasibility of the estimates taking into account the sampling locations listed in Table 1. The following input data were selected: the domain length, 1000 m; aquifer thickness, 50 m; source length, 150 m, located at 150 m $\le x^* \le$ 300 m on the top boundary (y^*= 50 m); injection Darcy velocity, q_{inj}^*= 1 m/year; aquifer Darcy velocity, q_{aq}^*= 20 m/year; porosity, ε =0.3; longitudinal and transversal dispersivities, respectively, α_L= 10 m and α_t= 0.5 m; retardation factor, R_d= 1 (K_d= 0). The dimensionless concentration being injected C_{inj} is equal to 1 and no degradation effect was considered (λ= 0).

Table 1. Position of the wells considered in the simulations.

Well	***Position,*** x_k^* ***(m)***
1	300
2	400
3	600
4	800
5	1000

The prescribed time-dependent function corresponding to the mass flux release, chosen to simulate the measurements at the sampling locations, is given by:

$$\phi(t) = V_{inj}\, C_{inj}\, \mathrm{g(t)} \tag{21}$$

where $\mathrm{g(t)}$ is given by,

$$g(t) = \begin{cases} \dfrac{(\mathrm{t - t_i})}{\mathrm{t_0}} e^{-\left((\mathrm{t-t_i})^2 V_{inj}\right)/(2h\,\mathrm{t_i})}, & \text{for } \mathrm{t < t_0 + t_i} \\ e^{-(2(\mathrm{t-t_i})-\mathrm{t_0})V_{inj}/2h}, & \text{for } \mathrm{t > t_0 + t_i} \end{cases} \tag{22}$$

Equations (22) are the exact solution of a leaching model corresponding to a radioactive waste repository where the contaminant is completely dissolved into liquid phase at $t = t_i$, and where a linear fault of the engineer barriers occurs between $t = t_i$ and $t = t_0$. The repository height is given by 'h'. The values here considered were: $h^* = 6$ m, $t_i^* = 10$ years and $t_0^* = 20$ years.

The simulated concentration measurements were obtained from the numerical solution of Eqs. (10) with $N_y = 800$, and by using subroutine DMOLCH with a user prescribed relative error criteria equal to 10^{-5}, an equally spaced mesh of 401 grid points, and with a time step $\Delta t = 0.5$ yr. Three situations were considered for the estimation of the mass flux histories: exact data (predicted values), simulated experimental data, and noisy filtered data. Error-free average concentrations are directly obtained from the solution of Eq. (13):

$$\overline{C}w_{k,m} = \overline{C}_0(x_k, t_m), \quad k=1,2,\ldots,M \tag{23}$$

Random errors are added to the exact data in order to simulate experimental measurements at each sampling location [4]:

$$Cw_n(x_k, y_j, t_m) = C(x_k, y_j, t_m) \pm \nu\sigma\,, \; k=1,2,\ldots,M, \quad j=1,2,\ldots,N_w \tag{24}$$

where ν is the standard Gaussian random variable with zero mean and unit standard deviation, and σ is the standard deviation of the errors added to the exact values, given by:

$$\sigma = \varepsilon_n\, C(x_k, y_j, t_m) \tag{25}$$

The relative error $\varepsilon_n = 0.05$ was used in this study.

It has been shown [12] that better results can be obtained if the data are smoothed prior to the inversion. The measured concentrations are replaced by linear combinations of past and future measurements. The filtered concentrations are evaluated by the following expression:

$$Cw_f(x_k, y_j, t_m) = \sum_{s=1}^{w} Z_s Cw_n(x_k, y_j, t_{m+s-(w+1)/2}), \quad k=1,2,\ldots,M, \;\; j=1,2,\ldots,N_w \tag{26}$$

where w, the width of the filter, is an odd integer and Z_s are the weighting coefficients. The width $w = 11$ was used in this work. The coefficients for the Gaussian filter used to smooth the contaminated data are:

$$Z_s = \frac{e^{-2\pi\left((w+1-2s)/(w-1)\right)^2}}{\sum_{i=1}^{w} Z_i}, \quad s=1,2,\ldots,w \tag{27}$$

This filtering technique is equivalent to the mollification procedure that has been used to stabilize some inverse problems.

Once the error contaminated concentrations, Cw_n and Cw_f, are obtained, the average concentrations for each well are calculated by Eq. (16).

NUMERICAL RESULTS

The performance of the proposed inverse method for different numbers of future times, r, considering the measurements taken at the first well ($x^* = 300$ m), was investigated first. Table 2 summarizes the values of the standard deviations of the estimated mass fluxes, S_e, S_n and S_f, given by Eq. (20), for exact, noisy, and noisy filtered data, respectively. The values of S_e correspond to

the deterministic bias of the inverse solution. For this inverse method, the introduced bias increases with r, and, in this case, it is much more important than the error introduced by the contaminated data. The smoothing filter does not contributed to the results obtained with the noisy data. In this test case, the smoothing has an opposite trend, increasing the bias of the inverse solution due to the linear combination of past and future observations. One can also observe that a proper choice of the number of observation locations at each sampling well, N_w, is much more important for the estimates. For lower values of r, the results obtained with N_w= 5 are much more accurate than those obtained with N_w= 3.

Table 2. Standard deviation of the estimates considering measurements at x^*= 300m.

r	S_e	S_n		S_f	
		N_w=3	N_w=5	N_w=3	N_w=5
3	1.786	18.71	7.127	21.45	9.780
5	5.404	18.53	7.511	21.73	10.90
7	12.89	21.04	13.40	23.88	15.49
9	20.80	25.73	20.74	27.90	21.91
11	28.66	31.54	28.28	33.08	28.91
13	36.32	37.82	35.75	38.88	36.02

Figure 2 shows a comparison between true and estimated fluxes at x^*= 300 m for N_w= 5. It is interesting to note that smoothed estimates can be obtained through two different strategies: by filtering of the contaminated data (w= 11) or by increasing the number of future times (r= 13). Higher values of r lead to an advance in time of the estimates.

Table 3 shows the values of the standard deviations of the estimated mass fluxes, S_e, S_n and S_f, for r= 7 considering all the possible combinations of the 5 wells. The best estimates are obtained with the measurements taken from the wells close to the source. Results are not available for the cases 4, 5 and 15, because the solution is not sufficiently stable to perform estimations from the data provided by wells 4 and 5 due to the low sensitivity coefficients at these locations. The information provided by wells 4 and 5 do not affect significantly the results obtained with the other wells. The best estimates were obtained in cases 6 and 16, corresponding to the simultaneous use of the sampling data provided by wells 1 and 2, and 1, 2, and 3, respectively.

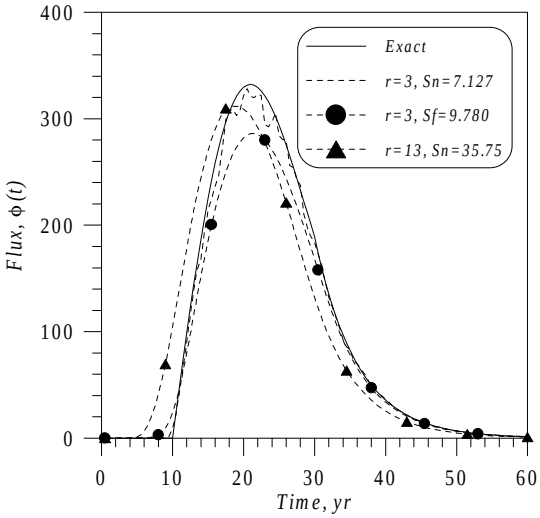


Fig. 2. Comparison between true and estimated mass fluxes at x*= 300 m.

The chosen filtering strategy have not improved the results shown in Table 3. The only enhancement was observed when using the measurements from well 3. The estimated fluxes obtained in case 3 are shown in Fig. 3, for exact and noisy filtered data. One can observe that even when the true values of the average concentrations are used, the solution is unstable. This is due to the existence of small inaccuracies in the solution of the forward problem, which are amplified by the inverse solution. One can notice that the filter largely attenuates these instabilities, indicating that it can be much more efficient when considering wells far from the source. Figure 3 also shows that the estimated flux is advanced in time for distant wells.

Table 3. Standard deviation of the estimates from different observation locations.

Case	*Wells*	S_e	S_n	S_f
1	***1***	12.89	13.40	15.49
2	***2***	27.58	27.04	27.01
3	***3***	74.25	481.9	67.13
4	***4***	-	-	-
5	***5***	-	-	-
6	***1,2***	9.925	10.23	12.64
7	***1,3***	13.25	13.66	15.64
8	***1,4***	12.89	13.40	15.49
9	***1,5***	12.89	13.40	15.49
10	***2,3***	28.13	27.54	27.41
11	***2,4***	27.59	27.05	27.01
12	***2,5***	27.58	27.04	27.01
13	***3,4***	74.30	482.7	67.25
14	***3,5***	74.26	481.9	67.13
15	***4,5***	-	-	-
16	***1,2,3***	9.832	10.14	12.58
17	***1,2,4***	9.926	10.23	12.64
18	***1,2,5***	9.925	10.23	12.64
19	***1,3,4***	13.25	13.66	15.64
20	***1,3,5***	13.25	13.66	15.64
21	***1,4,5***	12.89	27.54	15.49
22	***2,3,4***	28.13	27.54	27.41
23	***2,3,5***	28.13	27.54	27.41
24	***2,4,5***	27.59	27.05	27.01
25	***3,4,5***	74.27	482.7	67.25
26	***1,2,3,4***	9.832	10.14	12.58
27	***1,2,3,5***	9.832	10.14	12.58
28	***1,2,4,5***	9.926	10.23	12.64
29	***1,3,4,5***	13.25	13.66	15.64
30	***2,3,4,5***	28.13	27.54	27.41
31	***1,2,3,4,5***	9.832	10.14	12.58

The fluxes estimated in case 6 for noisy and noisy filtered data are presented in Fig. 4. Figure 5 presents a comparison between exact and estimated contour maps of the 2-D solution at t= 20 yr, t= 30 yr and t= 40 yr. The 2-D profiles were obtained through the GIT solution, Eqs. (9) and (10)), using the mass flux estimated in case 6, with noisy filtered data. A vertical exaggeration in graphical scales was used for the contour maps presented in order to facilitate their interpretation.

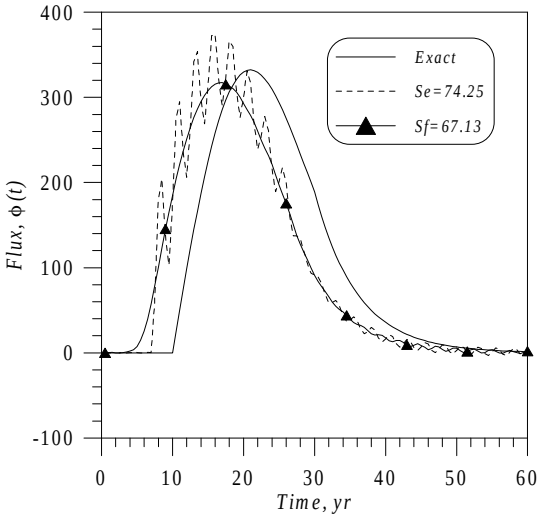


Fig. 3. Comparison between true and estimated mass fluxes for case 3.

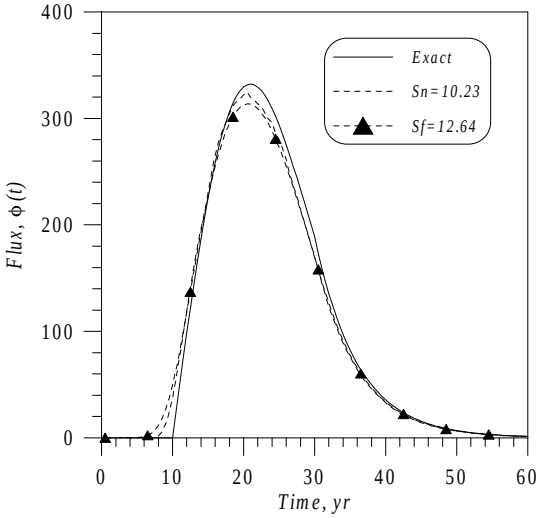


Fig. 4. Comparison between true and estimated mass fluxes for case 6, using noisy and noisy filtered data.

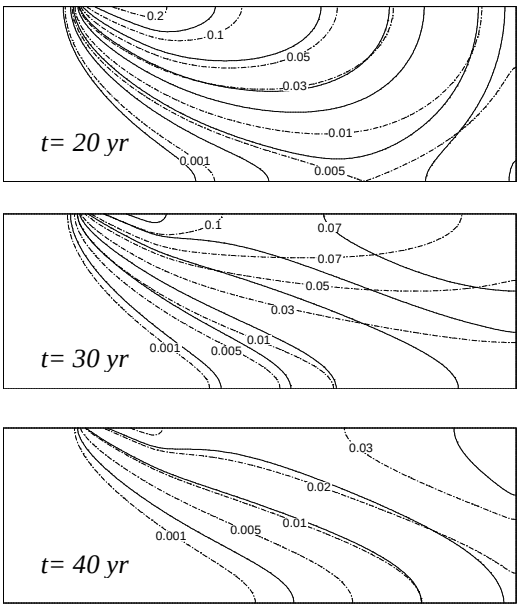


Fig. 5. Comparison between true (solid lines) and estimated (dotted lines) concentration contours at t= 20 yr, t= 30 yr and t= 40 yr, for r= 7, using noisy filtered data provided by wells 1 and 2 (case 6).

CONCLUSIONS

An efficient inverse solution based on the use of a sequential method associated with the numerical solution of a one-dimensional partial differential equation was proposed to estimate the unknown time-dependent mass flux release of a contaminant source. An example was chosen to investigate the performance of the proposed inverse solution as a function of the number of future times and well locations, using exact, noisy and noisy filtered data to numerically simulate experimental measurements. The feasibility of such estimates was studied, and showed that the use of a filtering strategy can be useful when considering data collected far from the contaminant source. It was also showed that the estimated fluxes are advanced in time when increasing the number of future times or the distance between the sampling location and the source.

Finally a comparison between 2-D exact and estimated concentration profiles at given times was made, showing that multidimensional estimates can be obtained by such efficient method.

ACKNOWLEDGEMENTS

The author would like to express special thanks to Dr. Marco A. Leal for his valuable technical cooperation throughout the last years in the field of numerical simulations in groundwater modeling.

REFERENCES

1. N.-Z. Sun, *Inverse Problems in Groundwater Modeling*, Kluwer Academic Publishers, Dordrecht, 1994, p. 337.

2. G. L. Moltyaner and R. W. D. Killey, Twin lake tracer tests: longitudinal dispersion, *Water Resour. Res.*, **24** (10), 1613-1627 (1988).

3. G. L. Moltyaner and R. W. D. Killey, Twin lake tracer tests: transversal dispersion, *Water Resour. Res.*, **24** (10), 1628-1637 (1988).

4. N. Sun, N.-Z. Sun, M. Elimelech and J. N. Ryan, Sensitivity analysis and parameter identifiability for colloid transport in geochemically heterogeneous porous media, *Water Resour. Res.*, **37** (2), 209-222 (2001).

5. J. Atmadja and A. C. Bagtzoglou, Pollution source identification in heterogeneous porous media, *Water Resour. Res.*, **37** (8), 2113-2125 (2001).

6. R. M. Neupauer, B. Borchers and J. L. Wilson, Comparison of inverse methods for reconstructing the release history of a groundwater contamination source, *Water Resour. Res.*, **36** (9), 2469-2475 (2000).

7. R. M. Cotta, Ed., *The Integral Transform Method in Thermal-Fluid Sciences and Engineering*, Begell House, New York, 1998.

8. C. Liu, J. E. Szecsody, J. M. Zachara and W. P. Ball, Use of the generalized integral transform method for solving equations of solute transport in porous media, *Adv. Water Resour.*, **23**, 483-492 (2000).

9. M. A. Leal and N. J. Ruperti Jr., A numerical study for the two-dimensional solute transport in groundwater via integral transform method, *Hybrid Meth. Engng.*, **2** (1), 111-129 (2000).

10. IMSL Library, MATH/LIB., Houston (1989).

11. J. V. Beck, B. Litkouhi and C. R. St. Clair, Efficient sequential solution of the nonlinear inverse heat conduction problem, *Num. Heat Transfer*, **5**, 275-286 (1982).

12. M. Raynaud, Combination of methods for the inverse heat conduction problem with smoothing filters, AIAA Paper No. 86-1243 (1986).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

MODELING INVERSE SUBSIDENCE DIFFUSION-CONVECTION IN GEOSTRUCTURES

Emmanouil G. Vairaktaris
Ioannis P. Vardoulakis
Department of Applied Mathematics and Physics,
Section of Mechanics
National Technical University of Athens, NTUA
Athens, Greece
mvairak@central.ntua.gr

Euripides Papamichos
SINTEF Petroleum Research
Trondheim, Norway and
Aristotle Univ. of Thessaloniki,
Dept. of Civil Engineering,
Thessaloniki, Greece
epapamic@civil.auth.gr

Vasilios A. Dougalis
Department of Mathematics
University of Athens, UOA
Athens, Greece
doug@math.uoa.gr

ABSTRACT

Compaction of a collapsible substratum due to effective stress increase may give rise to the formation of the well-known trap-door mechanism (Terzaghi 1936, Vardoulakis et al. 1981). According to early works, large-scale subsidence over a yielding underground geostructure is seen as a stochastic (Markov) process (Litwinszyn 1974, Dimova 1990). This process leads to the Einstein-Kolmogorov (E-K) integral equation. Under certain physical conditions and transformations of the coordinate system the E-K integral equation satisfies some partial differential equation of parabolic type, where the vertical coordinate replaces time. Initial condition of the direct problem is the base subsidence and solution yields the surface subsidence. The solution depends on a diffusivity coefficient, which determines the formation of the subsidence trough inside the body as well as on the surface. This is the Direct Subsidence-Diffusion - Convection (DSDC) problem. Considering the results of the DSDC problem in this paper we present the inverse SDC problem using two kinds of regularization (Lattés and Lions, 1969). In particular Lion's u_{xxxx}-method is compared to the presently proposed u_{xzz}-method. Stability, in the sense of the von Neumann condition is ensured where the amplification factor depends on the regularization parameter ε. A first approach to convergence is done in the sense of the norm of the amplification factor. Another convergence study is given in terms of the truncation error (Richtmyer and Morton, 1967).

NOMENCLATURE

B(m): Half width of the trap door
B*(m): Transformed depth, B*=f(H/2B)
b : Dimensionless height, maximum value of variable z
c : Diffusion coefficient depending on soil properties
$f_{\Delta x}$: Stability coefficient-function for the u_{xxxx}-regularization
H(m): Height of the depression trough
Im : Imaginary part of a complex number
L_n : Coefficient of the numerical algorithm depending on time (depth)
S : Sum of roots of the stability equation for the u_{xzz}-regularization
$\tilde{u}_{\varepsilon_j}^{n}$: Discretized form of the exact solution
w_0 : Normalized trap-door displacement
z_1 : Complex number depending on the numerical parameters of the u_{xzz}-regularization
β(deg): Angle of the depression trough
ε : Regularization parameter
κ : Coefficient of the von Neumann method of stability
ρ : Factor of amplification, considering von Neumann condition
Φ : Truncation error

INTRODUCTION

Compaction of a collapsible substratum due to oil-production (effective stress increase) and/or water-injection (capillary action) may give rise to the formation of the well-known trap-door mechanism [1] as shown in Figure 1.

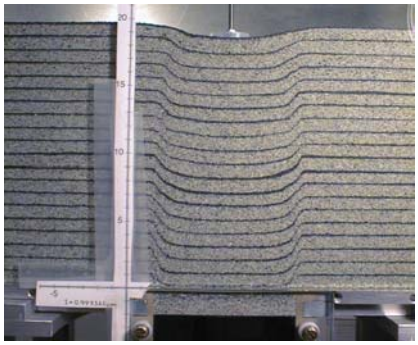


Fig1: Trap-door experiment

1g-experiments with sand ([2], [3]) have indicated that the trap-door displacement is convected practically upwards through a mechanism that localizes the soil displacements above the trap-door (Fig. 2). The boundaries of the subsidence trough are inclined inwards, reducing the extend of the depression in the vertical direction. The angle β of the trough boundaries is evolving as function of trap-door displacement. In particular it is found that the angle β of the trough boundaries is decreasing with trap-door displacement [2]; i.e. with increasing trap-door displacement the boundaries of the trough tend to become vertical. The trap-door displacement w_0 is causing the formation of a trough, which is a function of the position in vertical direction.

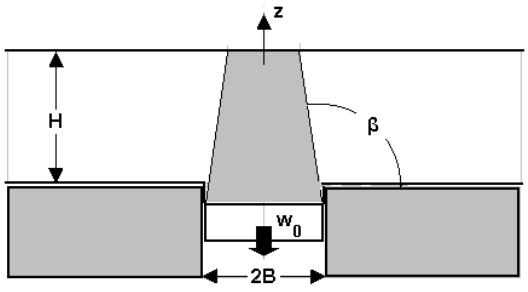


Fig.2: Trap-door mechanism

According to an early work of J. Litwinszyn ([4], [5]) we assume that large-scale subsidence over a yielding underground geo-structure is a stochastic (Marcov) process. In simple terms we assume that the displacement of a particle in vertical direction (i.e. in the direction of gravity) causes movement of particles mainly lying above it in a manner that particle vertical displacement is spread also horizontally. This assumption results in a mechanism of subsidence convection-diffusion, which leads to the Einstein-Kolmogorov integral equation [6]. It can be shown, that under some mathematical and physical conditions, the solution of the (E-K) equation satisfies a partial differential equation of parabolic type for the displacement, which corresponds to subsidence function w(x,z) in the space of the trough:

$$\frac{\partial w}{\partial z} = C\frac{\partial^2 w}{\partial x^2} \tag{1}$$

This is a diffusion equation, with the vertical coordinate z playing the role of time. Introducing a set of dimensionless variables considering several transformations of the mathematical quantities we end-up with the following initial, boundary value problem [7]:

$$\frac{\partial u}{\partial z} = \frac{c}{(1-z)^2}\frac{\partial^2 u}{\partial x^2} - \frac{x}{1-z}\frac{\partial u}{\partial x} \tag{2}$$

for

$$0 \le x \le 1 \quad 0 < z \le b, \quad b = -\left(\frac{H}{B}\right)\cot\beta \tag{2a}$$

and

$$u(x,0) = 1 \quad \text{(i.c.)} \tag{2b}$$

$$u(\pm 1, z) = 0 \quad \text{(b.c.)} \tag{2c}$$

$$u_x(0,z) = 0 \quad \text{(b.c.)} \tag{2d}$$

This is a diffusion-convection partial differential equation with variable coefficients. The model contains a free parameter, the diffusivity coefficient c, which is fitted to the experimental results [7]. Below we show a typical computational example, concerning the numerical solution of the i.b.v. problem, inside a prescribed trough (H, 2B, and β). The data that have been

used for this example refer to small-scale model tests performed by the authors [3]. In particular we use data corresponding to H/2B=2, w_0=0.062 and β=105^o (Fig.3).

This initial – boundary value problem will be used for posing the corresponding inverse SDC problem.

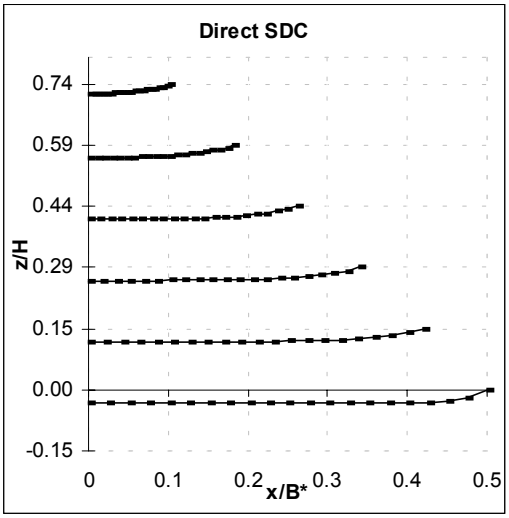


Fig.3: Results of the direct problem

STABILITY OF THE DSDC PROBLEM

In the DSDC problem the formation of the subsidence trough depends on angle β. For large values of this angle the trough boundaries converge below the surface. In this case the differential equation cannot be solved near the point of intersection, because of unbounded coefficients of the derivatives of x, equation (2). For bounded coefficients, the numerical solution of second order linear parabolic problems, using finite differences, admits a sufficient condition for stability [8]:

$$2\lambda\, a(x,z) < 1 \qquad (3)$$

$\lambda = \Delta z / \Delta x^2$ is depending on the discretization and a(x,z) is the coefficient of the second order derivative in x. In the considered problem this condition has the form:

$$2\lambda \frac{c}{(1-z)^2} < 1 \qquad (4)$$

For any given value of the factor $1/(1-z)^2$ we may choose λ so that the above stability condition is met. Here the numerical solution is obtanied for c=0.1, b ≈ 0.75 and λ=0.2.

Compatibility of initial and boundary conditions is ensured also by interpolating values of trap-door subsidence for z=0 in x=1.

INVERSE SDC PROBLEM

Oil-production or water injection in situ, results in surface subsidence that can be large enough to cause severe damages in the surface constructions. The difficulty in large-scale problems lies in the fact that for given surface subsidence the corresponding base displacement is not known. The problem of computing the base displacement using as "initial" conditions the surface subsidence corresponds to the solution of inverse in "time" (depth) SDC problem.

Inverse problems are in general mathematically ill-posed, which means that existence, uniqueness and stability of the solution cannot be ensured. Several regularization methods have been developed for this kind of problems. Due to the dual nature (diffusion-convection) of the considered problem we introduce here mainly two methods of regularization (u_{xxxx}, u_{xzz}) of the Inverse Subsidence Diffussion-Convection problem (ISDC). Notice that the first regularization scheme is essentially motivated by Lions' Method of Quasireversibility [9] .

FIRST REGULARIZATION METHOD

As initial condition for the ISDC problem we use the results of the corresponding DSDC. For inverse parabolic problems, Lions' Method of Quasireversibility suggests the use of the 4th order derivative in x, as regularization term. The initial - boundary value problem is described in treaties (5)-(5e) where u is the solution to the direct subsidence diffusion – convection problem and u_ε is the solution of the inverse problem:

$$\frac{\partial u_\varepsilon}{\partial z} = -\frac{c}{(1-b+z)^2}\frac{\partial^2 u_\varepsilon}{\partial x^2} + \frac{x}{1-b+z}\frac{\partial u_\varepsilon}{\partial x} - \varepsilon\frac{\partial^4 u_\varepsilon}{\partial x^4} \qquad (5)$$

for

$$0 \le x \le 1 \quad ; \quad 0 < z \le b \qquad (5a)$$

and

$$u_\varepsilon(x,0) = u(x,b) \quad \text{(i.c.)} \tag{5b}$$
$$u_\varepsilon(\pm 1,z) = 0 \quad \text{(b.c.)} \tag{5c}$$
$$\Delta u_\varepsilon(\pm 1,z) = 0 \quad \text{(b.c.)} \tag{5d}$$
$$(u_\varepsilon)_x(0,z) = 0 \quad \text{(b.c.)} \tag{5e}$$

Due to the increased order of the governing equation for the inverse problem, extra boundary conditions are needed. Condition (5d) is proposed by Lions, and reflects a zero curvature requirement [9].

NUMERICAL CONSIDERATIONS AND RESULTS

The numerical solution of the above-mentioned inverse initial, boundary value problem is obtained by the finite differences method. The following explicit algorithm has been used:

$$u_j^{n+1} = u_j^n + \frac{\Delta z}{\Delta x(1-z)} x(j)\left(u_{j+1}^n - u_{j-1}^n\right) - \\ - \frac{c\Delta z}{\Delta x^2(1-z)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) - \\ - \frac{\varepsilon\Delta z}{\Delta x^4}\left(u_{j+2}^n - 4u_{j+1}^n + 6u_j^n - 4u_{j-1}^n + u_{j+2}^n\right) \tag{6}$$

The results of the ISDC problem are obtained for $0.005 < \varepsilon < 0.01$. For values of regularization parameter outside this interval the solution of the inverse problem diverges significantly from the corresponding solution of the direct problem. We remark that the central subsidence (x=0) approaches the solution of the direct problem for small values of ε. However, independent of ε and from the start of the inverse solution, the curvature of displacement line close to the right boundary diverges from the results of the direct problem. This observation will be elaborated below.

As is mentioned by Lions [9] the explicit algorithm (6) is deficient: For large values of the time-like variable (depth), the solution ISDC problem diverges significantly from the data of the corresponding DSDC. Let b be the depth for which the direct problem has been solved. Due to the aforementioned divergence of results, the initial condition of the inverse problem could not be placed "earlier" as the value $z \approx 0.25$ b. The limited time solution of the ISDC problem, using Lions' regularization method, results from the strong diffusive character of the 4th order regularization term, u_{xxxx}. This pathology is depicted in Figures 4 and 5, where we show the comparison between the solutions of direct (solid line) and the regularized inverse problem (dotted line). Notice that the x-coordinate re-scaled by the factor $B^* = (H/2B)B = 2B$.

In future works an implicit algorithm for the above-mentioned i. -b. value problem will be considered.

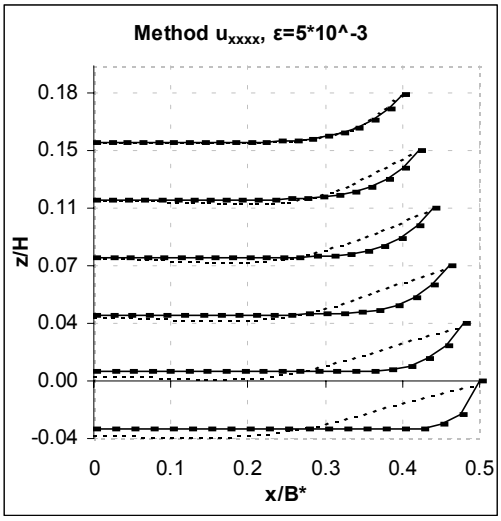


Fig. 4: Comparison of direct and inverse subsidence solution using Lion's regularization

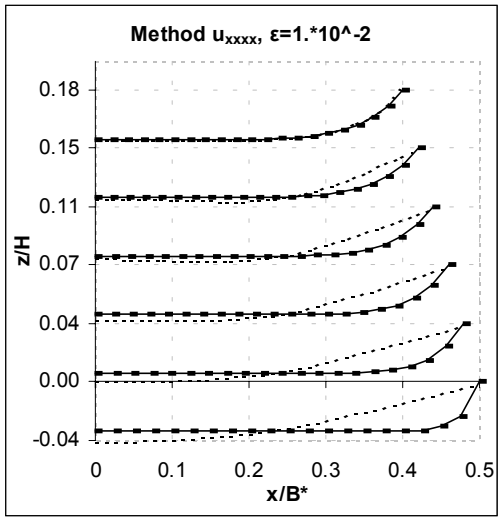


Fig. 5: Comparison of direct and inverse subsidence solution using Lion's regularization

SECOND REGULARIZATION METHOD

Above observations have prompted the use of a mixed regularization term that has derivatives due to both x and z. The choice that has been made is a u_{xzz}-regularization, considering appropriate initial and boundary conditions:

$$\frac{\partial u_\varepsilon}{\partial z}+\varepsilon\frac{\partial^3 u_\varepsilon}{\partial z^2\partial x}= -\frac{c}{(1-b+z)^2}\frac{\partial^2 u_\varepsilon}{\partial x^2}+\frac{x}{1-b+z}\frac{\partial u_\varepsilon}{\partial x} \quad (7)$$

for

$$0\le x\le 1 \quad ; \quad 0<z\le b \quad (7a)$$

and

$$u_\varepsilon(x,0)=u(x,b) \quad \text{(i.c.)} \quad (7b)$$
$$(u_\varepsilon)_z(x,0)=0 \quad \text{(i.c.)} \quad (7c)$$
$$u_\varepsilon(1,z)=0 \quad \text{(b.c.)} \quad (7d)$$
$$(u_\varepsilon)_x(0,z)=0 \quad \text{(b.c.)} \quad (7e)$$

It can be observed that the difference in the conditions between the direct and the inverse SDC problem, is concerning (7c). This choice is preferred because, in numerical terms, this results in a pure downward subsidence for the first level of the trough.

NUMERICAL ASPECTS - RESULTS

The numerical solution of the above-mentioned initial, boundary-value problem, equations (7)-(7e), was obtained using the method of finite differences. The algorithm that has been used is the following:

$$\left(\frac{1}{\Delta z}-\frac{\varepsilon}{\Delta x\Delta z^2}\right)u_j^{n+1}+\left(\frac{\varepsilon}{\Delta x\Delta z^2}\right)u_{j+1}^{n+1}= \left(-\frac{c}{\Delta x^2(1-z)^2}\right)u_{j-1}^{n}+ \left(\frac{1}{\Delta z}-\frac{2\varepsilon}{\Delta x\Delta z^2}+\frac{x}{\Delta x(1-z)}+\frac{2c}{\Delta x^2(1-z)^2}\right)u_j^{n}+ \left(\frac{2\varepsilon}{\Delta x\Delta z^2}+\frac{x}{\Delta x(1-z)}-\frac{c}{\Delta x^2(1-z)^2}\right)u_{j+1}^{n}+ \left(\frac{\varepsilon}{\Delta x\Delta z^2}\right)u_j^{n-1}+\left(-\frac{\varepsilon}{\Delta x\Delta z^2}\right)u_{j+1}^{n-1} \quad (8)$$

As already mentioned, as initial condition are used the results of the solution of the direct problem. Unlike the first regularization method, in the present one the "depth" of the solution used as initial data equals to the whole "depth" of the solution of the direct problem. i.e. z=b.

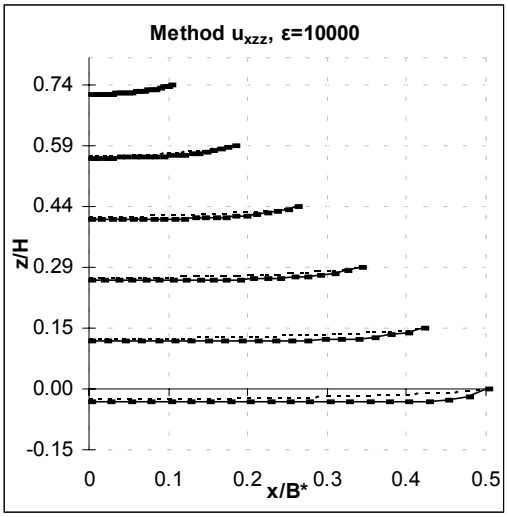


Fig.6: Comparison of direct and inverse subsidence solution using u_{xzz} regularization

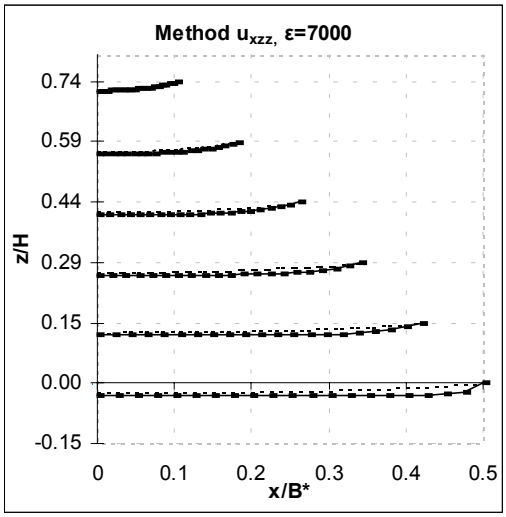


Fig7: Comparison of direct and inverse subsidence solution using u_{xzz} regularization

The results of the ISDC problem using u_{xzz}-regularization are shown in Figures 6 and 7. These results are obtained for values of the regularization parameter ε between 6000 and

10000. A choice of the regularization parameter greater than 10000 gives approximately the same results in accuracy of 10^{-5} with the one of ε =10000. For values of ε lower than 6000, the numerical results diverge significantly from the direct solution.

VON NEUMANN CONDITION

The von Neumann condition claims that the stability of the numerical solution, can be insured, if, under certain conditions which depend on the numerical parameters of the problem, holds [10]:

$$|\rho| \le 1 + \kappa\, \Delta z \tag{9}$$

with

$$0 \le \kappa \le M \tag{10}$$

for Δx, Δz sufficiently small and with $|\rho|$ being the absolute value of the amplification factor. The right-hand inequality corresponds to the stability condition and the left-hand one corresponds to first study of convergence due to the expanding character of the initial–boundary value problem.

APPLICATIONS TO THE ISDC

1. The u_{xxxx}- Regularization

The von Neumann stability condition imposes as variables the increments Δx and Δz. Within a stability analysis, the time-like variable (1-b+z) and the space-like variable x of the original problem are treated as parameters. Thus for given discretization and at the "time" step n and for the "space" point j we denote with Ln the variable (1-b+z) and with x_j the variable x.

In equation (6) we set

$$u_j^n = A\rho^n e^{imj\Delta x} \tag{11}$$

Following the formulation given above by equations (5) - (5e) we get similarly that:

$$|\rho| = |1 + \Delta z\, f_{\Delta x}| \le 1 + \Delta z |f_{\Delta x}| \tag{12}$$

$$f_{\Delta x} = I\frac{x_j}{L_n \Delta x}\sin\Delta x - \frac{\sin^2\frac{\Delta x}{2}}{\Delta x^2}\left(\frac{4c}{L_n^2} - \frac{16\varepsilon}{\Delta x^2}\sin^2\frac{\Delta x}{2}\right)$$

where

$$\begin{aligned} &L_n \in [0.18,\ 1.0] \\ &c = 0.1 \\ &\varepsilon \in [-\infty, +\infty] \\ &x_j \in [0\ ,\ 1.0] \end{aligned} \tag{13}$$

For the left-hand side of the inequality (10) to be true we must insure that:

$$\varepsilon < \frac{c}{L_n^2} - \frac{1}{c_{\Delta x}^2} - \frac{\sqrt{1 - \mathrm{Im}^2(f_{\Delta x})}}{c_{\Delta x}^2}$$
$$c_{\Delta x} = 4\frac{\sin^2(\Delta x/2)}{\Delta x^2} \tag{14}$$

The measure of $f_{\Delta x}$ is a decreasing function of Δx in the interval (0, 1). Thus the lower upper bound of the measure of $f_{\Delta x}$ is:

$$\lim_{\Delta x \to 0} |f_{\Delta x}| = \sqrt{\frac{x_j^{\,2} L_n^2 + c^2 - c\varepsilon L_n^2 + \varepsilon^2 L_n^4}{L_n^4}} \tag{15}$$

The previous quantity is an indicative amplification coefficient for the time-depth change of the $u_\varepsilon(x,z)$. The bigger previous coefficient is the more divergence exists between the numerical solution and the real solution.

Since we have insured the stability of the problem, we are going to study the rate of convergence of the algorithm. It is obvious that the factor with which the rate of convergence becomes measurable is the measure of $f_{\Delta x}$, which depends on $\sin(\Delta x)$. So it can be proved that

$$|f_{\Delta x}| = O(\Delta x) \tag{16}$$

2. The u_{xzz} - Regularization

Considering equation (11) in equation (8) we result in:

$$\begin{aligned} &\left[\frac{1}{\Delta z} + \frac{\varepsilon}{\Delta x \Delta z^2}\left(e^{i\Delta x} - 1\right)\right]\rho^2 + \\ &+\left[\left(-\frac{1}{\Delta z} - \left(\frac{2\varepsilon}{\Delta x \Delta z^2} + \frac{x_j}{L_n \Delta x}\right)\left(e^{i\Delta x} - 1\right)\right)\right]\rho + \\ &+\left[\frac{4c}{L_n^2 \Delta x^2}\sin^2\left(\frac{\Delta x}{2}\right)\right]\rho + \\ &+\left[\frac{\varepsilon}{\Delta x \Delta z^2}\left(e^{i\Delta x} - 1\right)\right] = 0 \end{aligned} \tag{18}$$

Let $\rho_{1,2}$ be the complex roots of equation (18). According to the stability condition we are interested in the behavior of these roots as Δx and Δz are tending to 0. For this limit, the roots of equation (18) tend to 1, and therefore $|\rho_{1,2}|$ does the same. Thus the condition (9) reduces to the right hand side of the inequality (10), as soon as the coefficient κ is positive.

First we observe that this coefficient is bounded as Δx and Δz are tending to 0:

$$|\kappa| \to \left| \frac{0.5I}{\varepsilon} \pm \sqrt{\frac{4x_j}{L_n \varepsilon} - \frac{1}{\varepsilon^2} + I\frac{4c}{L^2{}_n \varepsilon^2}} \right| \tag{19}$$

since the measure of the above quantity is bounded for big enough ε without restrictions rising from the limit of Δx and Δz to 0. This ensures stability.

Secondly we have to examine if the coefficient κ is positive, i.e. if

$$\frac{|\rho| - 1}{\Delta z} \geq 0 \tag{20}$$

The prove of this consideration is concerning the sum of roots of the equation (18) and simply refers that

$$\|S\| = \|\rho_1 + \rho_2\| \geq 2 \Rightarrow \max\|\rho_i\| \geq 1 \quad i = 1,2 \tag{21}$$

which is proved based on the derivation

$$\|S\| = \|2 + z_1\|, \qquad \mathrm{Re}(z_1) > 0 \tag{22}$$

TRUNCATION ERROR OF INVERSE PROBLEM - CONVERGENCE

We are interested for the rate of convergence between numerical and exact solution as well as for the rate of convergence of the solutions individually. In previous paragraphs the rate of convergence was examined in the terms of the values of the absolute value of the amplification factor ρ.

In this paragraph the rate of convergence will be examined in the terms of the truncation error [10]. The truncation error for:

a) Lion's u_{xxxx} -regularization method is:

$$\Phi_1(\tilde{u}_\varepsilon) = \frac{\Delta z}{2}\left(\frac{\partial^2 \tilde{u}_\varepsilon}{\partial z^2}\right)^n_j - \frac{\Delta x^2}{6L_n}\left(x_j\left(\frac{\partial^3 \tilde{u}_\varepsilon}{\partial x^3}\right)^n_j - \frac{c}{2L_n}\left(\frac{\partial^4 \tilde{u}_\varepsilon}{\partial x^4}\right)^n_j\right) = O(\Delta z) + O(\Delta x^2) \tag{23}$$

b) the u_{xzz} -regularization method is:

$$\Phi_2(\tilde{u}_\varepsilon) = \frac{\Delta z}{2}\left(\frac{\partial^2 \tilde{u}_\varepsilon}{\partial z^2}\right)^n_j + \frac{x_j \Delta x}{L_n}\left(\frac{\partial^2 \tilde{u}_\varepsilon}{\partial x^2}\right)^n_j = O(\Delta z) + O(\Delta x) \tag{24}$$

where $\tilde{u}_{\varepsilon}{}^n_j$ corresponds to the discretized form of the exact solution.

According to the stability condition for the direct problem $\Delta z << \Delta x$. Considering the equations (23) and (24) we can conclude that the main factor, which influences the truncation error, is the one that concerns Δx. The point in space, which is influenced significantly by the numerical truncation error, is that close to the right boundary. At this point the 2nd derivative in x for u_{xzz}-regularization and the 4th derivative in x for u_{xxxx}-regularization are important, since for small "times" the time parameter L_n has small enough values. This can be verified in the graphs of the solution close to the right boundary, where there is significant difference between the solution of the inverse and direct problem. Notice that central subsidence is not affected of the truncation error.

CONCLUSIONS

In summary we can mention the following:

- The numerical solution of the ISDC concerning u_{xzz} – regularization can be derived for larger depth comparing with u_{xxxx}-regularization due to the strong diffusive character of the 4th order derivative in x. In future works an implicit scheme for the numerical solution of the u_{xxxx}-regularization must be concerned.
- Stability in the sense of the von Neumann condition for both regularizations is ensured due to right choices of ε and "time" b of the problem that affects the parameter L_n.

- The measure of the coefficients κ, $f_{\Delta x}$ is not concerning only the stability of the algorithm but also the rate of convergence between numerical and exact solution. It is affected in turn from the above-referred parameters.
- Convergence in the terms of the Truncation error, is satisfactory except of the boundary x=1, where both regularization methods have significant differences concerning the solution of the direct problem.

ACKNOWLEDGEMENTS

The authors wish to thank Norsk Agip and Statoil for supporting this research.

REFERENCES

1. K.v.Terzaghi (1936). Stress distribution in dry and saturated sand above a yielding trap door. *Proc. Int. Conf. Soil Mech., Cambridge Mass.,*Vol. I, 307-311.
2. Vardoulakis, I., Graf, B and Gudehus, G. (1981). *Trap-door problem with dry sand: A statical approach based upon model test kinematics*, International Journal for Numerical and Analytical Methods in Geomechanics, Vol. 5, 57-78
3. Papamichos, E., Vardoulakis, I., Heil, L.K., (2000). *Overburden Modeling Above a Compacting Reservoir Using a Trap Door Apparatus*. Phys. Chem. Earth (A), Vol. 26
4. Dimova, V.L., *Some Direct and Inverse Problems in Applied Geomechanics*, University of Mining & Geology, Sofia, 1990.
5. Litwinszyn, J., *Stochastic Methods in the Mechanics of Granular Bodies*. Springer-Verlag, Wien, 1974.
6. Tikhonov, A.N. and Samarskii, A.A., *Equations of Mathematical Physics*, Dover, 1963.
7. Vardoulakis, I., Vairaktaris, E. (2002), *Modeling Subsidence Diffusion-Convection in Geostructures,* International Journal for Numerical and Analytical Methods in Geomechanics, (to appear), J. Wiley and sons
8. John, F., *Partial Differential Equations.* Springer-Verlag, 1982
9. Lattés, R. and Lions, J.L. *The method of quasi-reversibility.* American Elsevier Pub. Co., New York, 1969.
10. Richtmyer, R.D. and Morton, K.W., *Difference methods for initial-value problems*. J. Wiley and sons, 1967

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

VIBRATION-BASED IDENTIFICATION OF ISOTROPIC MATERIAL PROPERTIES BY QUASI-BINARY ELECTRONIC HOLOGRAPHY AND FINITE ELEMENT MODELLING

Dan N. Borza
Department of Mechanical Engineering, LMR
Institut National des Sciences Appliquées de Rouen
76800, Saint-Etienne du Rouvray, France
borza@insa-rouen.fr

ABSTRACT

The first objective is presenting a novel interferometric method, the quasi-binary holography, developped to assist the hybrid, experimental-numerical identification of material properties. The quasi-binary electronic holography is a vibration measurement method extending the vibration amplitude measurement range by a factor of two with respect to the well-known time-average method. The fringe contrast is also highly improved, so the reliability is higher, and the full laser power is available for the measurement, without losses as in stroboscopic techniques. The second objective is to evaluate the importance of the number of vibration modes used in the identification. Finite element updating using the first 5, 10, 20 or 40 eigenfrequencies of a uniform rectangular plate allows finding an optimal number of nodes in the mesh and the material properties. The whole updating process is converging towards material properties values for which the average eigenfrequency error of the numerical model is less than 0.2 %.

NOMENCLATURE

C(x,y) local contrast in the image plane
I(x,y) intensity distribution of the object image with interference fringes
x,y coordinates in the image plane
α arbitrary phase shift
φ optical phase of object wave with respect to reference wave in the detector plane
λ laser light wavelength
$I_{OBJ}(x,y)$ intensity distribution of the object image without interference fringes
$J_0(z)$ first-order, zeroth kind Bessel function of argument z

INTRODUCTION

A great deal of work has already been reported on the subject of identification of material parameters by numerical/experimental methods applied to vibrating objects. Experimental data may be obtained by using microphones [1], piezo-electric transducers [2], coherent optical techniques [3], or other means.

Vibration measurement by full-field, nob-contact coherent optical techniques presents the interest of not disturbing the tested object, while simultaneously acquiring a large amount of data to be used in the identification process. They are taking into account the true properties of the vibrating structure and also allow damage detection. The most widely used coherent optical technique, electronic holography, achieves real-time measurement of the full-field out-of-plane vibration amplitude field at the surface of steady-state vibrating structures, and is simultaneously providing the values of the eigenfrequencies. The amplitude informations are presented as fringe patterns superimposed on the object image.

The fringe visibility and the processing of the fringe pattern in order to quantitatively calculate the vibration amplitudes are greately reduced by two factors. The first one is the fringe pattern being eventually locally undersampled, as in the case of nodal lines running close to each other. The second reason is the limited spatial resolution of the speckled pattern and its low fringe contrast, determined by the fringe function. The fringe function, whose argument is linearly related to the vibration amplitude at any object point, represents the function which is modulating the intensity of the object image. Its relative minima and maxima are the centers of the loci of iso-amplitude points at the object surface. The type of fringe function is given by the interferometric method being used.

ELECTRONIC HOLOGRAPHY

Phase stepped electronic holography (also known as Electronic Speckle Pattern Interferometry, or Digital Speckle Pattern Interferometry) is an interferometric whole-field displacement measuring technique [4] based on recording with the help of a CCD camera the primary interference fringes in an on-axis interferometric setup. The general principle is illustrated by the lay-out in Figure 1.

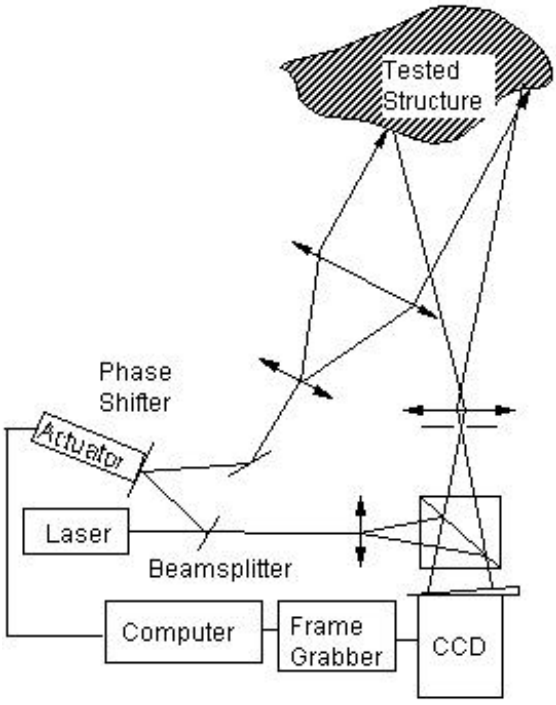


Figure 1. Electronic holography setup

The holographic system uses a continuous wave laser whose output is separated into the object illumination beam and the reference beam. On the path of either one or the other of these beams there is a mirror attached to a piezo-electric actuator. A staircase voltage applied to the actuator is producing equal successive pathlength variations of magnitude $\frac{\lambda}{4}$. Accordingly, a $\frac{\pi}{2}$ phase shift is produced between the reference wave and the object wave which interfere on the CCD camera during each frame.

Assuming that the object undergoes a steady-state harmonic vibration at a frequency either great enough with respect to the frame acquisition frequency (25 Hz) or an integral multiple of this, the current i-th image $I_i(x,y)$ is described by the relation:

$$I_i(x,y) = I_{OBJ}(x,y)\times \\ \times\{1 + C(x,y)\cos[\varphi(x,y) + \alpha]J_0[\varphi_d(x,y)]\}$$

$$\alpha = (i-1)\frac{\pi}{2},\ i = 1, 2, 3, 4 \qquad (1)$$

In electronic holography with directions of object illumination and observation close to the normal, the phase φ_d is direcly related to the out-of-plane vibrational amplitude d(x,y) at any visible point of the object surface by the approximate relation:

$$\varphi_d(x,y) = \frac{4\pi}{\lambda}d(x,y) \qquad (2)$$

The holographic processor calculates and stores the two differences C_1 and S_1, given by eq. (3) and (4):

$$C_1 = I_1 - I_3 = 2C(x,y)\cos\varphi I_{OBJ}(x,y)J_0(\varphi_d) \qquad (3)$$

$$S_1 = I_4 - I_2 = 2C(x,y)\sin\varphi I_{OBJ}(x,y)J_0(\varphi_d) \qquad (4)$$

Fringe function for time-average method

In time-averaged electronic holography [5], the real-time image displayed by the monitor is given by:

$$I_{TAV} = C_1^{\,2} + S_1^{\,2} \qquad (5)$$

Taking into account eq. (3) and (4), this expression becomes:

$$I_{TAV} = CI_{OBJ}J_0^2[\varphi_d(x,y)] \qquad (6)$$

The time-averaged interferogram displayed on the monitor and described by eq. (6) is refreshed after each four-frames cycle. It shows the image

of the object covered by an iso-amplitude fringe pattern corresponding to the Bessel-type fringe function:

$$F(d) = J_0^2\left[\frac{4\pi}{\lambda} d(x,y)\right] \tag{7}$$

Such a time-averaged hologram, representing the vibration amplitude map of a rectangular plate, is shown in Figure 2. The contrast of the fringe pattern is decreasing with increasing fringe order, which makes difficult the fringe processing by tracking the dark and bright fringe centers. The other current difficulty in fringe processing, namely the existence of undersampled, high fringe density regions, has deliberately been avoided by choosing one of the lowest modes and low vibration amplitudes, so as to produce a hologram allowing the visual counting of fringes. For higher modes and amplitudes, that is often impossible.

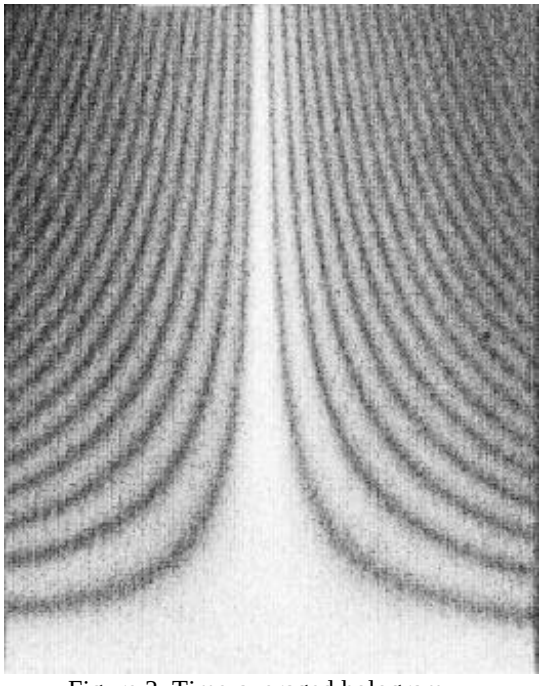


Figure 2. Time-averaged hologram

Other fringe functions

The contrast of fringe patterns obtained in vibration measurement may be improved by using the stroboscopic principles [6], but this involves using a much higher power laser to compensate the important loss produced by strobing. The fringe function in this case is given by:

$$F(d) = \cos^2\left[\frac{4\pi}{\lambda} d(x,y)\right] \tag{8}$$

Phase imaging [7] is another useful method providing directly the φ_d modulo 2π distribution, but the procedure is rather tedious and does not work in real time, like the other methods.

Quasi-binary holography

This new interferometric method [8] produces in real-time a fringe pattern given by:

$$I_{QUB}(x,y) = A + B \times \mathrm{sgn}\{J_0[\varphi_d(x,y)]\} \tag{9}$$

A and B are constants. Figure 3 shows the image obtained from a quasi-binary hologram for the same vibration state (same amplitudes and same frequency) already shown in Figure 2.

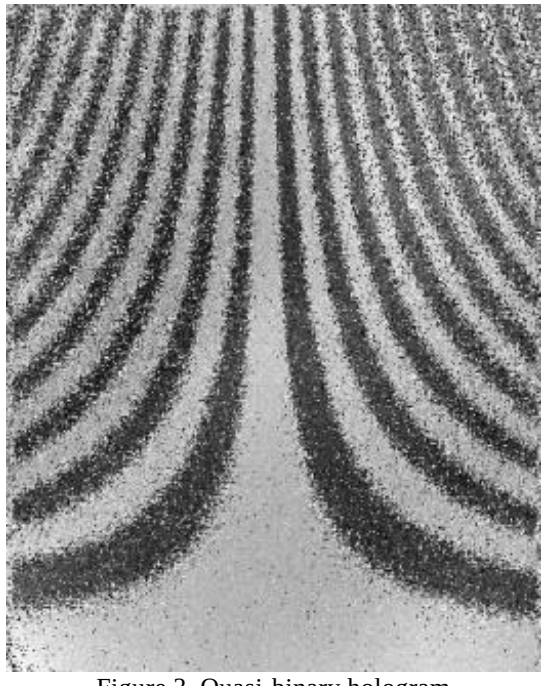


Figure 3. Quasi-binary hologram

As shown by Eq. (9), the fringe pattern in the image presented in Figure 3 is quasi-binary; the limits of each fringe are defined only by the zero crossings of the Bessel function $J_0(\varphi_d)$.

The number of fringes is halved with respect to that of the image in Figure 2,, so fringe counting becomes possible in those regions of the object where the time-averaged speckled fringe pattern would otherwise be spatially undersampled. The contrast is constant for any fringe order, which makes possible an efficient quantitative processing of the fringe pattern, while in the image in Figure 2 the contrast is close to zero for the higher order fringes, in the two upper corners, and the signal-to-noise ratio is about unity.

Comparison of the quasi-binary method with the time-averaged method

The two most important metrological characteristics, fringe contrast and spatial sampling frequency of the fringe pattern, are shown in Figure 4, as obtained through numerical simulation.

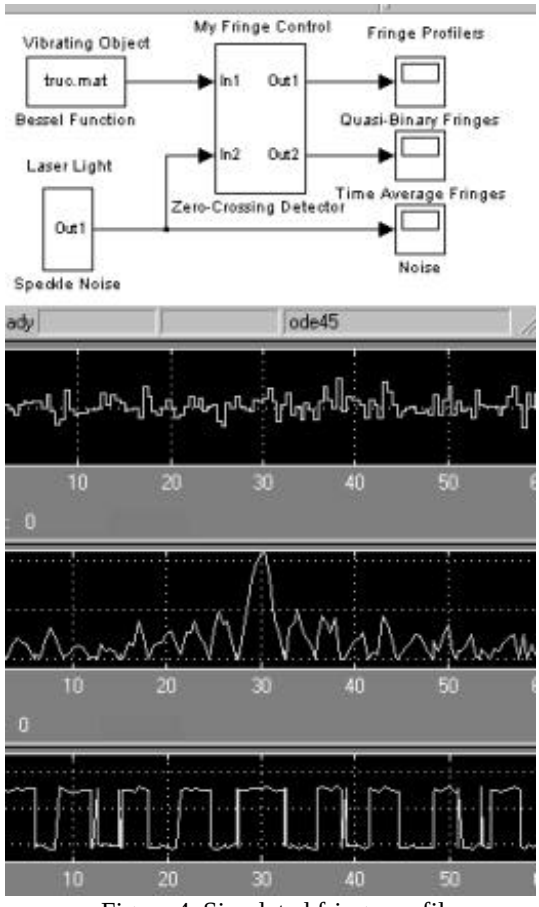


Figure 4. Simulated fringe profile

The upper drawing in Fig. 4 shows the model.

The three output screens in the lower part of Fig. 4 represent: the upper one – the speckle noise, the middle one – the profile of Bessel – type (time average holography) fringes, and the lower one – the profile of quasi – binary holography fringes.

These simulated characteristics are entirely confirmed by the experimental results presented in Figure 5

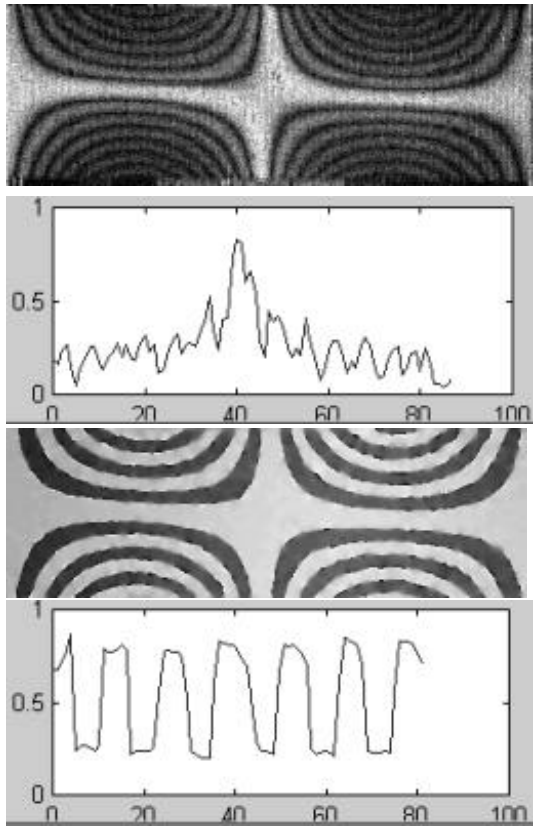


Figure 5. Experimental fringe patterns

The two images in the upper part of Figure 5 present a time-averaged hologram with a medium fringe density (a total of 14 noisy cyles per 90 pixels), and its intensity profile along a vertical line. The other two images, in the lower part of Figure 5, represent the equivalent quasi-binary hologram and the profile of the same vertical line.

By using simple linear filtering and thresholding procedures, the image obtained by the quasi-binary method may be transformed to a binary image, allowing further morphological processing in order to obtain the full-field vibration amplitude field.

The new interferometric method was used during the material properties identification work for measuring experimentally the vibration amplitude distributions. The high contrast of the resulting interferograms allowed a precise recognition of each mode shape.

MATERIAL PROPERTIES IDENTIFICATION

The general principle used in identifying the material properties is updating a finite element model so as to make its results converge toward the experimental results obtained by quasi-binary electronic holography. By using a sufficient number of vibration modes of a uniform rectangular plate, not only the material properties, but also the optimal number of nodes and the shape factor of the finite elements used in discretization may be found.

The experimental data used in the identification were the resonant frequencies, provided that the order of the numerically found modes and their shapes are the same as in the holographic results.

The numerical normal mode shapes and frequencies are found as solution of the undamped eigenvalue problem.

Experimental data

The first 40 vibration modes and their frequencies were computed numerically, using a general software program [9]. The experimentally measured amplitude maps were obtained by quasi-binary electronic holography. The excitation was provided by a small loudspeaker placed behind the tested plate. The position of the loudspeaker was adjustable, so as to be able to avoid coupling of modes whose frequencies are very closed

The fine frequency adjustment of the numerical signal generator was used to measure the resonant frequencies, and make sure a single mode is excited.

The first three modes of the plate are shown in the images presented in Fig. 6. Some of the higher modes are presented in Fig. 7 and Fig. 8.

The images on the left column are the quasi-binary holograms corresponding to the numerically predicted modes illustrated on the right column. The experimentally measured frequencies are indicated under each experimental amplitude distribution. For the higher modes, the number of elements may become a limiting factor for the correct representation of the numerically calculated modes.

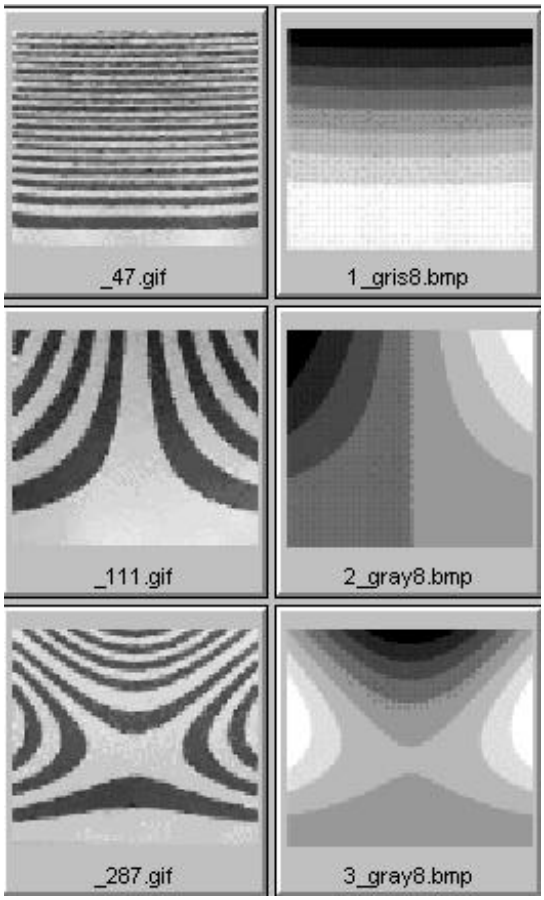


Figure 6. First three modes

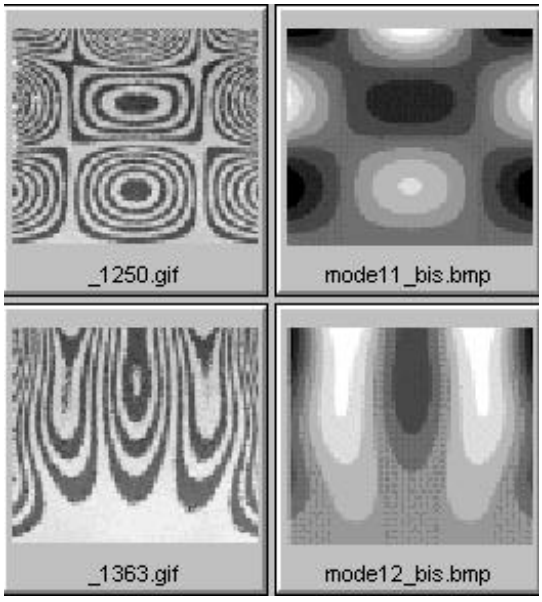


Figure 7. Modes 11 - 12

At medium frequencies (modes 11 - 12) the finite element representation is still correct, as shown in Figure 8. At higher frequencies, as shown in Figure 9, only the nodal lines were used for recognizing in the numerical model the mode shapes and theit order, and the time-averaged holography was used instead of quasi-binary holography, because of its higher sensitivity.

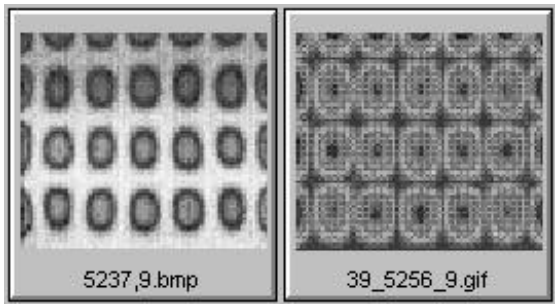


Figure 8. Mode 39

Along with the amplitude distributions, the first 40 eigenfrequencies were calculated ($f_{EF}^{(i)}$) by the finite element program and measured ($f_{exp}^{(i)}$) by holographic measurement.

The material parameters to be found for the rectangular steel plate are the Young modulus E and the Poisson coefficient ν. For the finite element mesh using tria3 elements, the parameters are the number of nodes along each dimension of the plate.

The discrete cost function being minimized during the parameters optimization procedure is the mean value of the relative error, as in eq. (10).

$$e_{mean_N} = \frac{\sum_{i=1}^{N} \left| \frac{f_{exp}^{(i)} - f_{EF}^{(i)}}{f_{exp}^{(i)}} \right|}{N} \qquad (10)$$

The mean value is a global function integrating the overall quality of the model. The maximum value of the relative error given by eq. (11) was also tested with good results:

$$e_{max_N} = \max\left(\left| \frac{f_{exp}^{(i)} - f_{EF}^{(i)}}{f_{exp}^{(i)}} \right| \right)_i \qquad (11)$$

Parameters identification

An initial identification procedure was carried in order to find the values of E and ν assuring values of $e_{mean_{40}}$ and $e_{max_{40}}$ of about 1 %. These values were then used in a full optimization procedure in an attempt to find the best shape of the tria3 elements used in the discretization.

Optimum Mesh. The graph in Figure 7 shows the values of $e_{mean_{40}}$ as a function of the number of nodes along each side of the plate, NX and NY. As expected, the smallest values of $e_{mean_{40}}$ and $e_{max_{40}}$ occur when the values of NX, NY are so that the rectangular discretizing elements become isosceles. Those values are situated near the right line superimposed on the graph.

The absolute minimum value of $e_{mean_{40}}$ corresponds to NX=11 and NY=12 nodes. Such a mesh is dense enough to allow a fair representation of the highest frequency modes.

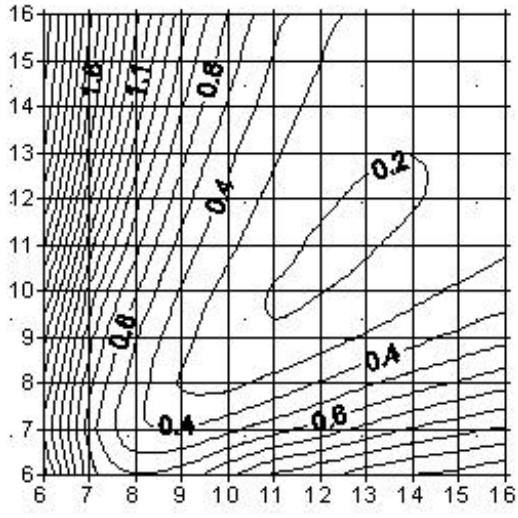


Figure 7. Optimum mesh

Material properties. Several minimization procedures for e_{mean_N} and e_{max_N} were carried out, for N = 5; N = 10; N = 20 and N = 40 vibration modes, in order to examine the influence of the medium and high frequency modes on the material properties values and on the two error functions. The results obtained for e_{mean_N} are shown in Table 1.

Table 1. Influence of modes number N on identification

N	5	10	20	40
E (GPa)	203.8	203.8	205.2	205.5
ν	0.256	0.258	0.256	0.257
e_{mean_N} (%)	0.13	0.16	0.18	0.19

The four values of N included in Table 1 correspond, respectively, to maximum eigenfrequencies of 407 Hz, 1010 Hz, 2570 Hz and 5350 Hz.

The material properties values converged, in all cases, towards optimum values very close to each other (less than 1 % difference for both E and ν).

As an example, Figure 8 presents the isovalue contours of $e_{mean_{20}}$, whose minimum value, 0.18, corresponds to an abscissa of value E = 205.2 GPa and to a value of Poisson coefficient of 0.256. When using e_{max_N} instead of e_{mean_N} as function to be minimized, the curves converge towards a point (E, ν) slightly different from the previous one. The isovalues of $e_{max_{20}}$ are shown in Figure 9.

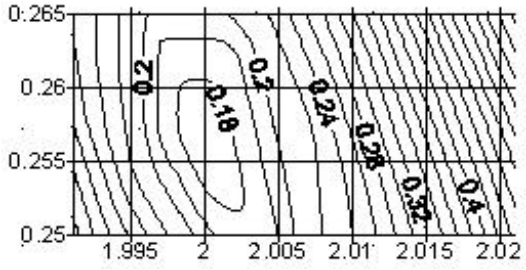


Figure 8. Isovalues of $e_{mean_{20}}$

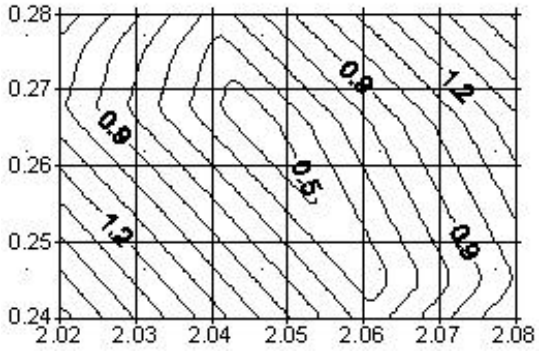


Figure 9. Isovalues of $e_{max_{20}}$

The values of E and ν are, in this case, 204.5 GPa (less than 0.5 % difference with respect to the value found on the basis of $e_{mean_{20}}$), respectively 0.265 (about 3 % difference).

Verification. The values of E and ν found by using the hybrid procedure described before have been checked using an experimental procedure inspired by [3], based on measuring the eigenfrequencies of the "O" and "X" modes of a free square uniform plate (Figure 10).

The method may be applied for isotropic and for orthotropic materials. The procedure also involves the use of the values of two particular eigenfrequencies, f_{20} and f_{02}, having two nodal lines parallel to one side or another.

The measurements of amplitude distributions for the free plate have been done by using time-averaged electronic holography.

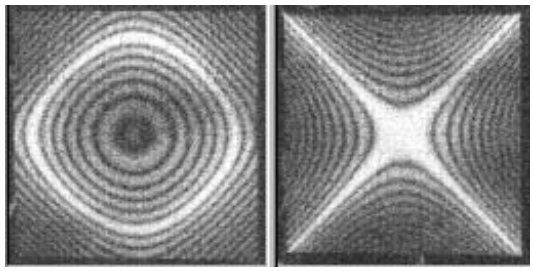


Figure 10. The "O" and "X" modes of a free square plate

Since the "O" mode whose frequency is needed to calculate f_{20} and f_{02} has a very wide resonant curve, the accuracy of these eigenfrequencies determination has been increased by taking into account, within a least squares procedure, the frequencies of several higher modes whose resonant curves are less wide; the amplitude maps corresponding to some of those higher modes are illustrated in Figure 11.

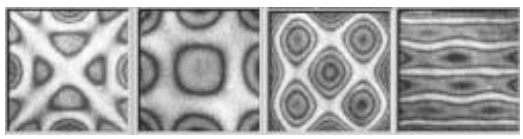


Figure 11. Higher modes of the free plate

The value of Young modulus thus found, E=206 GPa, is very close (less than 1 % difference) to the values already presented in

Table 1, and the Poisson coefficient value, ν=0.29, is about 12 % higher.

The scales shown in the plots in Figures 8 and 9 show that the accuracy characterizing the value of Poisson coefficient is lower than that concerning the Young modulus.

From the experimental point of view, measuring by holography the vibration amplitudes of a "freely" suspended plate may prove difficult, imposing some very strict precautions to be taken to prevent from air currents and other laboratory disturbances, like thermal gradients or vibro-acoustic perturbations.

CONCLUSIONS

A novel interferometric method, the quasi-binary holography, has been described. The experimental tests prove the theoretical predictions [8] concerning its higher fringe contrast and wider measurement range.

When necessary, fringe processing of the quasi-binary patterns may be done easily through the fringe tracking method followed by the interpolation of the values corresponding to the limits of the bright and dark fringes.

The use of the quasi-binary holographic method along with the time-average method and the finite element modelling had lead to very accurate experimental data. These data were used for material properties identification as well as for the finite element mesh optimization. The complete understanding of the optimal mesh, illustrated in Figure 7, needs more investigation.

Overall eigenfrequencies errors below 0.2 % have been obtained for the frequency range of 47 Hz – 5350 Hz, corresponding to the 40 first vibration modes, while the use of only five modes further reduces that error to 0.13 %, at the expense of higher errors for the unused modes.

Both discrete cost functions tested, e_{mean_N} and e_{max_N}, lead to similar results; however, the first one provides more sensitivity, as suggested by Figure 8, and integrates better the over-all behaviour.

The quantitative results for the materials properties E and ν have been confirmed by a different, direct identification procedure, using time-average electronic holography and appropriate processing of formulas for the eigenfrequencies of a free square plate. They had also been confirmed by another direct identification procedure using beams.

REFERENCES

1. A. L. Araujo, C. M. Mota Soares, M. J. Moreira de Freitas, Characterization of material parameters of composite plate specimens using optimization and experimental vibrational data, *Composites part B*, **27B**, p. 85 – 91, 1996

2. P. Pedersen, P. S. Frederiksen, Identification of orthotropic material moduli by a combined experimental/numerical method, *Measurement*, **10**, p. 113 – 118, 1992

3 S. Hurlebaus, Nondestrictuve Evaluation of Composite Laminates, *NDT Net*, **4**, 3, 1999

4. K. Creath, *Phase-Shifting Holographic Interferometry*, Holographic Interferometry, ed. P. R. Rastogi, Springer Series in Optical Sciences, **68**, p. 142-145, 1994

5. R. Jones and C. Wykes, *Holographic and Speckle Interferometry*, Cambridge University Press, Cambridge , 1989

6. P. Hariharan and B. F. Oreb, Stroboscopic holographic Interferometry, *Opt. Comm.*, **26**, p. 83 – 86, 1986

7. K. A. Stetson and J. Wahid, Real-Time Phase Imaging for Nondestructive Testing, *Experimental Techniques*, **22** (3), 1998, p. 15-18

8. D. N. Borza, Stepped-Amplitude Modulation Interferometry - A New Real-Time Mechanical Vibrations Measurement Technique, *Proc. of Intl. Conference "Interferometry in Speckle Light", Lausanne*, Springer Verlag, 2000, p. 205 - 210

9. E. Balmès, *Structural Dynamics Toolbox*, Scientific Software Group, Sèvres, 1997

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

IDENTIFING COUNTER-GRADIENT TERM IN CONVECTIVE PLANETARY BOUNDARY LAYER

Débora R. Roberti
Department of Physics
Federal University of Santa Maria, UFSM
Santa Maria, RS, Brazil
droberti@bol.com.br

Haroldo F. de Campos Velho
Lab. for Computing and Appl. Math. (LAC)
National Institute for Space Research (INPE)
São José dos Campos (SP) Brazil
haroldo@lac.inpe.br

Gervásio A. Degrazia
Department of Physics
Federal University of Santa Maria, UFSM
Santa Maria, RS, Brazil
degrazia@ccne.ufsm.br

ABSTRACT

Since Deardorff (1966) derived the counter-gradient term contribution, a good representation has been searched for this term. An inverse problem methodology is used to identify this property. The inverse analysis is performed minimizing an objective function: the square difference between experimental and model data added to a regularization function. Two regularization functions are used: second-order maximum entropy, and Tikhonov regularization of second-order. The scheme is tested with synthetic noise data. A new formulation for the counter-gradient is also presented.

Keywords: Counter-gradient term estimation, Inverse problems, Meteorological models.

NOMENCLATURE

b	Constant in Eq. (2)
$\overline{c}(z,t)$	Mean concentration ($\mathrm{g\,m^{-3}}$)
$C_{i,n}$	Approximated solution for concentration
h	Boundary layer height (m)
$J(\gamma)$	Objective function
K_{zz}	Vertical eddy diffusivity ($\mathrm{m^2\,s^{-1}}$)
$R(\gamma)$	Square difference term
$S(\gamma)$	Entropy regularization function
u_*	Friction velocity ($\mathrm{m\,s^{-1}}$)
w	Wind vertical component ($\mathrm{m\,s^{-1}}$)
w_*	Convective scale for velocity ($\mathrm{m\,s^{-1}}$)
$\overline{w'c'}$	Vertical turbulent flux of quantity c
γ	Counter-gradient term ($\mathrm{g\,m^{-2}\,s^{-1}}$)
λ	Regularization parameter
σ_w^2	Variance of the vertical velocity ($\mathrm{m^2\,s^{-2}}$)
χ_*	Scale for mean quantity ($\mathrm{g\,m^{-3}}$)
Δt	Time discretization (s)
Δz	Vertical space discretization (m)
Ψ	Nondimensional dissipation function
$\Omega(\gamma)$	Regularization function

INTRODUCTION

Terms denoting the turbulence can be represented by Reynolds fluxes. The process to parameterize the Reynolds tensors is called the closure problem. In the first order closure, turbulent fluxes are given by the product between the gradient of the mean quantity and an eddy diffusivity. A new formulation for first order closure is constituted considering the contribution of the counter-gradient term. Since Deardorff [5] has derived an expression for counter-gradient term and he also performed some experimental procedure to identify this term intense research has been done. Turbulence is always present in the Planetary Boundary Layer (PBL), where several stability conditions are found: convective, neutral, and stable. For convective conditions, the largest transporting eddies may have a similar size as the boundary layer height itself and, in particular, the flux can be in the opposite direction of the local gradient of the mean quantity. For neutral and stable conditions, the flux of a quantity is proportional to

the local gradient of that quantity. In fact, the counter-gradient represents nonlocal influences on the mixing by turbulence, and as this term is small in stable conditions, it is, therefore, neglected in these conditions [12].

Holtslag and Moeng [11] (HM91) derived a counter gradient term from turbulent heat flux, and their parameterization for the transport term is based on the results from large eddy simulation (LES). Cuijpers and Holtslag [4] (CH98) generalized the expression of the HM91 by writing the nonlocal term as a function of vertical velocity variance and an integral form for the flux.

The purpose of this paper is two fold. Firstly, a new approach to counter-gradient term is introduced. The new formulation is based on the parameterization presented in [4], however the convective scale for a scalar is a space average of the mean quantity inside of the PBL; additionally, the mixing length is computed by Taylor's theory [7]. Secondly, the counter-gradient term is identified by numerical procedures, from experimental data, based on an inverse problem methodology.

An implicit inversion technique is used for determining the counter-gradient term by a numerical scheme. The inverse problem is formulated as an optimization problem, where the objective function is defined as the least-squares fit between model results and experimental data. A stabilizer (or regularization) operator is added to the objective function with help of a Lagrange multiplier (also called regularization parameter). Iteration proceeds until objective function converges to a specified limit value. Synthetic data with Gaussian white noise corruption are used to simulate experimental data.

NEW APPROACH FOR THE COUNTER-GRADIENT TERM

The first order closure approach considering the counter-gradient term can be expressed as

$$\overline{w'\chi'} = -K_{zz}\left(\frac{\partial \chi}{\partial z} - \gamma_\chi\right) \qquad (1)$$

where $\overline{w'\chi'}$ is the vertical turbulent Reynolds flux, k_{zz} is vertical eddy diffusivity, γ_χ is the counter-gradient term, w is the vertical component of the velocity of wind, and χ is mean quantity (mass or temperature).

Cuijpers and Holtslag [4] have presented an expression for the counter-gradient, given by

$$\gamma_\chi = b\,\frac{w_*^2\,\chi_*}{\sigma_w^2\,h} \qquad (2)$$

being h the PBL height, σ_w^2 is the velocity variance, w_* is the convective scale for velocity, b is a constant and χ_* is a scale for mean quantity:

$$\chi_* \equiv \frac{1}{hw_*}\int_0^h \overline{w'\chi'}\,dz \qquad (3)$$

Combining Eqs. (1), (2) and (3) an integral equation is resulted for the turbulent flux $\overline{w'\chi'}$. One procedure to avoid this difficulty is to consider the counter-gradient term as zero in the first step of the time integration, and after that, having an estimative for the turbulent flux, the Reynolds fluxes can be computed from Eq. (1) and (2) using the flux $\overline{w'\chi'}$ previously computed [13]. In order to circumvent this constraint in the first step of time integration, a new formulation for convective scale χ_* is proposed. In the new formulation χ_* is given by

$$\chi_* = \overline{\chi} = \frac{1}{h}\int_0^h \chi(z)\,dz \qquad (4)$$

The constant b in Eq. (2) in Ref. [4] is 1.5, whereas in the new formulation this constant is $b = 0.07$.

There are many expressions for the velocity variance σ_w^2, see [7, 11, 18] and Appendix. Different approaches for the counter-gradient term are displayed in Figure 1, the Holtslag and Boville's [12], (HB-93) parameterization is also shown.

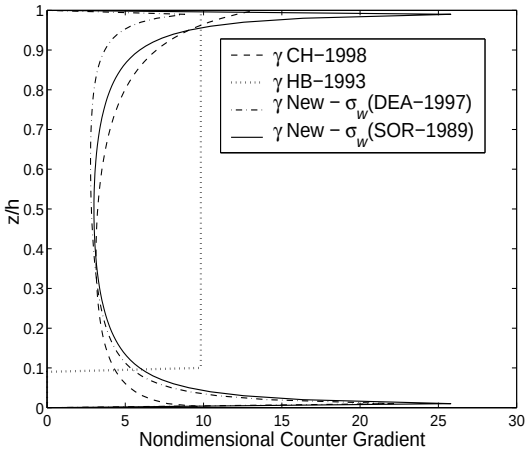


Figure 1: Different parameterizations for the non dimensional counter-gradient.

ESTIMATION OF COUNTER-GRADIENT TERM BY NUMERICAL PROCEDURES

As pointed out in [3], fully computational methods can be used to estimate some properties in turbulent flux. In particular, it is suggested that an inverse analysis could also be a good strategy to identify the counter-gradient term.

To exemplify the inverse methodology, the scheme is illustrated considering a one-dimensional pollutant dispersion model. The diffusion of passive scalars is given by mass conservation principle. A special case occurs in the dispersion of the instantaneous area sources. In this case, the diffusion equation can be simplified to

$$\frac{\partial \overline{c}}{\partial t} = \frac{\partial}{\partial z}\left[K_{zz}\left(\frac{\partial \overline{c}}{\partial z} - \gamma_c\right)\right] \tag{5}$$

with following initial and boundary conditions

$$\overline{c}(z,0) = \begin{cases} 0.1z & \text{if } 0 < z \leq h/2 \\ 100 - 0.1z & \text{if } h/2 < z < h \end{cases} \tag{6}$$

$$K_{zz}\frac{\partial \overline{c}}{\partial z} = 0 \text{ at } z = 0 \text{ and } z = h$$

where $\overline{c}$ is the mean concentration to be measured, k_{zz} is the vertical eddy, γ_c is the counter-gradient term to be estimated, h is the height of the atmospheric boundary layer. Our numerical model consists of Eq. (5) numerically solved by central finite difference method of second order $O(\Delta z^2)$ in space and the explicit Euler method in time. The approximate solution is denoted by $C_{i,n} \approx \overline{c}(z_i, t_n)$.

Inverse Model

In order to establish the inverse analysis, it is assumed that measurements $C_{i,n}^{Exp}$ are available at $i = 0, 1, ..., N_z$ vertical points, and at $n = 1, 2, ..., N_t$ time steps. The vector $\gamma = \{\gamma(z_i) | i = 0, 1, ..., N_z - 1)$ denotes the discrete counter-gradient term to be obtained by a nonlinear constrained minimization problem:

$$\min \quad J(\kappa, \gamma); \qquad l_i \leq \gamma_i \leq u_i \tag{7}$$

where the objective function is given by

$$J(\lambda, \gamma) = R(\gamma) + \lambda\Omega[\gamma] \tag{8}$$

with Ω being a regularization function and λ a positive parameter, called Lagrange multiplier. The bounds l_i and u_i are chosen to allow the inversion to lie within some physical limits. The least square difference between experimental data and the calculated values is represented by

$$R(\gamma) = \sum_{n=1}^{N_t}\sum_{i=0}^{N_Z}\left[C_{i,n}^{exp} - C_{i,n}(\gamma)\right]^2 \tag{9}$$

The regularization operator can be expressed by Tikhonov scheme [20] as

$$\Omega(\gamma) = \sum_{m,j=0}^{p} \kappa_{m,j}\|\gamma^{(m)}\|_2^2 \tag{10}$$

here $\gamma^{(m)}$ denotes the m-th difference. In general the parameter $\kappa_{m,j}$ is chosen as $\kappa_{m,j} = \delta_{m,j}$ (Kronecker delta) and the regularization is named Tikhonov-j regularization operador, where j denotes the order of the regularization.

Another regularization technique is given by an entropic scheme [3, 16, 17]

$$\Omega(\gamma) = \sum_{m=0}^{p} \kappa_m S^m(\gamma); \tag{11a}$$

$$S^m(\gamma) \equiv -\sum_{q=1}^{N_q} s_q \log(s_q) \tag{11b}$$

with

$$s_q = r_q^{(m)} / \sum_{l=1}^{N_q} r_l^{(m)} \tag{12}$$

and $r_q^{(m)}$ represents the m-th difference among the parameter vector [2, 17]. The function S^m attain its global maximum when all s_q are the same, i.e., a uniform distribution with $S_{max}^m = \log(N_q)$, in contrast, the lowest entropy value $S_{min}^m = 0$ is reached when all elements s_q but one are set to zero. This scheme is based on Jaynes' criterium of inference [14], called *maximum entropy principle* (MaxEnt).

Optimization Algorithm

The optimization problem is iteratively solved by the quasi-newtonian optimizer routine from the NAG Fortran Library [9], with variable metrics. This algorithm is designed to minimize an arbitrary smooth function subject to constraints (simple bound, linear or nonlinear constraints), using a sequential programming method.

This routine has been successfully used in several previous works: in geophysics, hydrologic optics, heat transfer, and meteorology.

RESULTS FOR COUNTER-GRADIENT TERM ESTIMATION

In order to verify the method presented in previous Section, some numerical experiments were performed. The use of synthetic data is a standard procedure to test a methodology in inverse problems, emulating experimental data. Therefore, the synthetic data are obtained from a concentration computed in the forward problem added to a random fluctuation.

The expression for the eddy diffusivities in the diffusion equation (5) is given by [7]

$$\frac{K_{zz}}{w_* h} = 0.15\Psi^{1/3}\left[1 - \exp\left(-4\frac{z}{h}\right) - 0.0003\left(8\frac{z}{h}\right)\right]^{4/3} \quad (13)$$

where w_* is the convective scale for velocity, h is height of the atmospheric boundary layer, and $\Psi = 0.913$ is the non-dimensional dissipation function, as computed by Degrazia [8].

In all simulations the following discretizations were $\Delta z = 10$ m with $N_z = 100$, and $\Delta t = 0.4$ s with $N_t = 1000$; and $\gamma(z) = 0$ is assumed as initial guess. For meteorological parameters the following values were used: $h = 1000$ m; $w_* = 0.6$ ms^{-1}; $u_* = 1.8$ ms^{-1} (friction velocity).

For the estimation of the counter-gradient term three levels of noise were used: 1%, 2.5%, and 5%. The best results were obtained with second order regularization operator, for Tikhonov and entropic regularization. The reconstruction using Tikhonov-2 and MaxEnt-2 were similar. In this paper only inversions with MaxEnt-2 are shown. Figure 2 shows the inversion for different levels of noise. As expected, by increasing the level of noise the inversion become poorer. The numerical values for regularization parameters are: 2.5×10^{10} for 1% of noise, 4.5×10^{10} for 2.5% of noise, and 4 x 10^{10} for 5% of noise. As expected, the reconstruction is degraded as the level of noise enhances.

In order to become clear the role of regularization parameter Fig. 3 shows the estimation for three different λ. Clearly, as $\lambda \to 0$ some spurious solutions (oscillations) appear in the inversion (Fig. 3a), even for small values of λ (Fig. 3b), for $\lambda \to \infty$ the optimization is only focused on the regularization term (see Fig. 3d). Good inversions are obtained with appropriate values for λ, as shown in Fig. 3c. An important feature is a good choice of the regularization parameter.

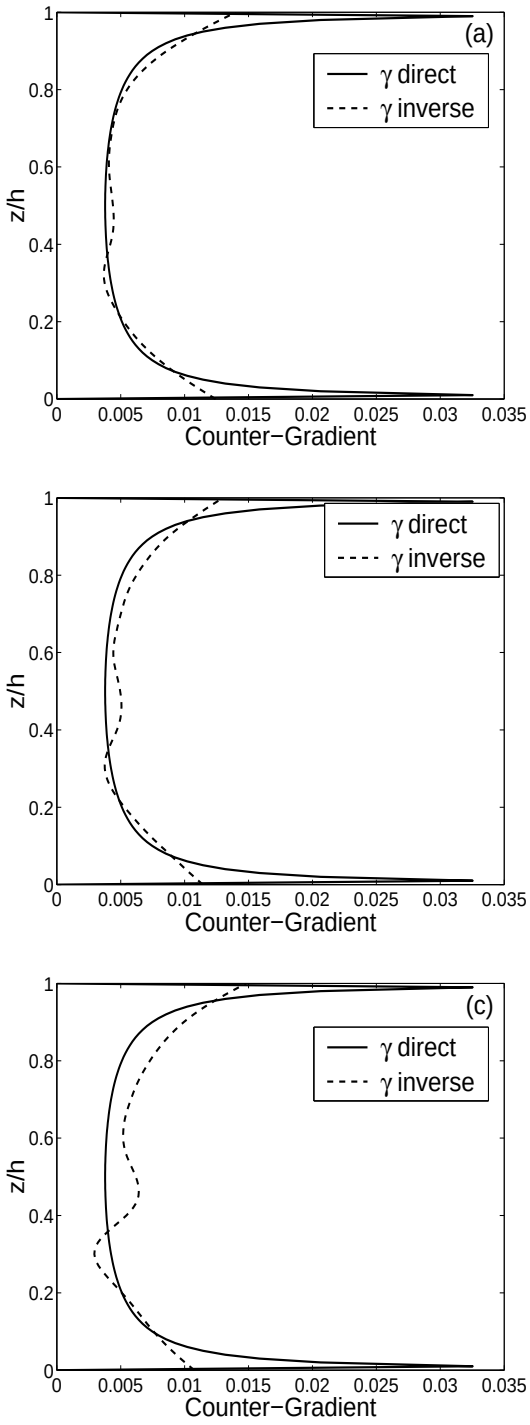


Figure 2: New Counter-gradient estimation by second-order maximum entropy with σ_*^2 given by Sorbjan (1989): (a) noise 1%, with $\lambda = 2.5$ x 10^{10}, (b) noise 2.5%, with $\lambda = 4.5$ x 10^{10}, and (c) noise 5%, with $\lambda = 4.0$ x 10^{10}.

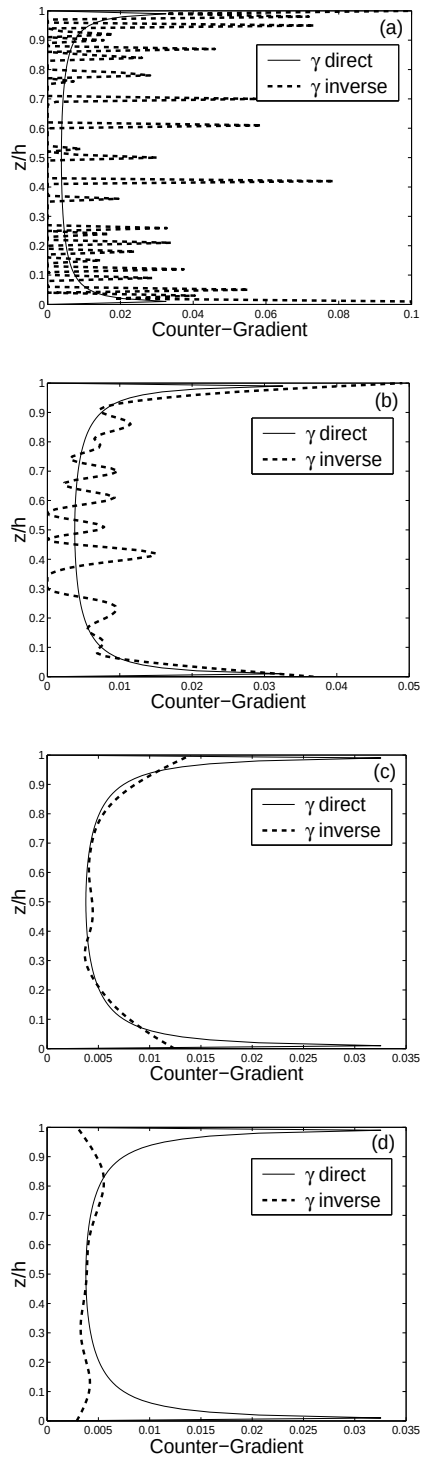


Figure 3: Influence of the regularization parameter to 1% of noise: (a) without regularization; (b) $\lambda = 1.0 \times 10^7$; (c) $\lambda = 2.5 \times 10^{10}$; (d) $\lambda = 1.29 \times 10^{11}$.

Many schemes have been proposed to find the value of the regularization parameter which gives a fine balance between square difference and regularization term. Some of these techniques are: Morozov's discrepancy principle [1, 15], the L-curve and the generalized cross validation [1]. Here, the Hansen's procedure [10], essentially the maximum curvature of the L-curve, was used producing good result. Figure 4 displays the L-curve for different Lagrange multipliers, for estimation with 1% of noise. From the plot it can be seen that $\lambda \approx 2.5 \times 10^{10}$ is a good choice.

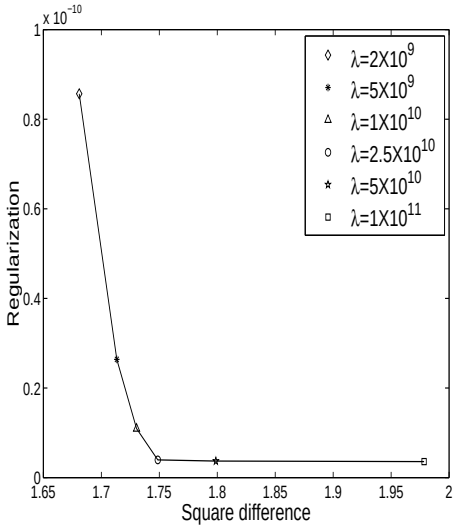


Figure 4: L-curve for reconstruction of counter-gradient with 1% of noise.

PERFORMANCE OF THE NEW COUNTER-GRADIENT FORMULATION

The forward problem described by Eqs. (5) and (6) was also used to verify the behaviour of the new counter-gradient approach given by expressions (2) and (4).

A relevant remark is to note that there is a small difference when the concentration is computed from counter-gradient in which χ_* is given by Eq. (3) or when the counter-gradient is calculated by Eq. (4). In our simulation, the variance of the vertical velocity appearing in formula (2) is that calculated according to Degrazia et al. [7], from analytical integration of the spectrum of kinetic energy. Other formulas for the variance are

shown in the Appendix.

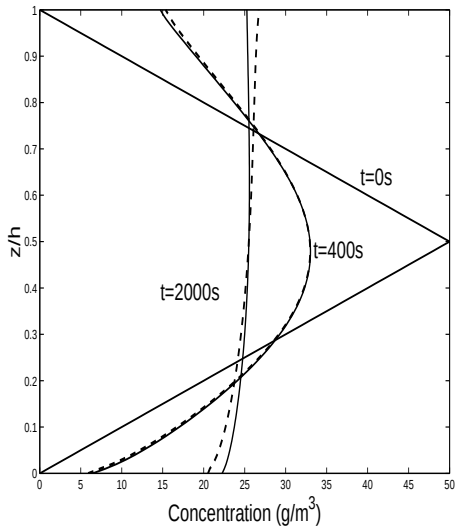


Figure 5: Concentration profiles at three times t=0s, t=400s and t=2000s: the solid line represents the diffusion without counter-gradient term, the dashed line represents the use of counter-gradient term given by Eqs. (2), (4) and (A.3).

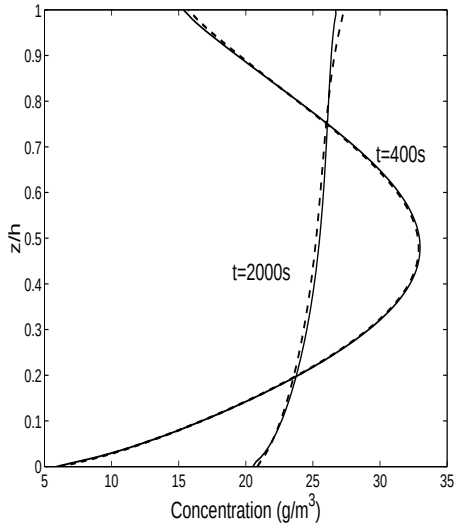


Figure 6: Concentration profiles for three time periods: the solid line with counter-gradient term given by Eqs. (2), (4) and (A.3); the dashed line the counter-gradient is given by Ref. [4] (Eqs. (2), (3) and (A.2)).

Figure 5 shows the simulation of the diffusion with and without the presence of the counter-gradient correction. If there is no transport in the counter-gradient direction, the PBL will be homogeneous in less period of time than when the counter-gradient contribution is present in the process. Figure 6 displays the concentration profile at two times, using the Cuijpers-Holtslag's approach Eqs. (2) and (3), with σ_w given by Eq. (A.2) and the counter-gradient term proposed here — Eqs. (2) and (4), with σ_w given by Eq. (A.3). A small difference is seen from Figure 6.

CONCLUSION

The methodology proposed for estimating the counter-gradient term was effective to procude nice reconstructions of this function. Good results were obtained even for a high level of noise with second order Tikhonov regularization and second-order maximum entropy principle, showing that the inverse model is robust related to the noise in experimental data. The determination of the regularization parameter by L-curve, following the procedure presented by Hansen [10], permitted to find out an appropriated value for both regularization operators.

The determination of turbulent properties by inverse procedure from available experimental data is useful under situations out of the scope of the theorical assumptions, e.g., sometimes flat terrain hypothesis is assumed to derive some property; therefore, such properties can be estimated on complex terrain by using inverse analysis. However, as it can be verified in Tables 1 and 2, the differences between exact and estimated values (Table 2) — $O(10^{-4})$ — are of same order of magnitude of the counter-gradient models (Table 1) — $O(10^{-4})$. Therefore, the inverse technique can not be used to select the best parameterization for the counter-gradient term.

Table 1: Difference between several formulations for the counter-gradient term.

$\rho_{Model} = \|\gamma_{Model} - \gamma^{New}_{DEA97}\|_2^2$	
γ_{Model}	ρ_{Model}
γ_{CH98}	$\rho_{CH98} = 3.5 \times 10^{-4}$
γ_{HB93}	$\rho_{HB93} = 6.4 \times 10^{-3}$
γ^{New}_{S89}	$\rho^{New}_{S89} = 1.5 \times 10^{-4}$

The proposed modification for the counter-gradient in the Cuijpers-Holtslag's formulation [4] worked well, producing similar results of those obtained with CH-98 expression [4]. The CH-98's approach has been compared with experimental data [4], thus this is an indirect validation of our expression for the counter-gradient term. Therefore, the following formula for the counter-gradient term

$$\gamma(z) = 0.085 \left(\frac{q_i}{\Psi}\right) \left(\frac{h}{z}\right)^{2/3} \left(\frac{\overline{\chi}}{h}\right) \qquad (14)$$

can be use in operational air-pollutant dispersion models, as well as in meteorological simulators under convective condition for the planetary boundary layer.

Table 2: Square error between exact and estimated counter-gradient terms for several levels of noise.

$\rho = \|\gamma_{exact} - \gamma_{estimated}\|_2^2$		
1% noise	2.5% noise	5% noise
3.1×10^{-5}	7.5×10^{-5}	2.6×10^{-4}

REFERENCES

1. M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics, 1999.

2. H.F. de Campos Velho, A.A.M. Holtslag, G.A. Degrazia and R.A. Pielke, *New parameterizations in RAMS for vertical turbulent fluxes*, Technical Report, Department of Atmospheric Sciences, Colorado State University, Fort Collins (CO), USA, 1998.

3. H.F. de Campos Velho, M.R. de Moraes, F.M. Ramos, G.A. Degrazia and D. Anfossi, An automatic methodology for estimating eddy diffusivity from experimental data, *Nuovo Cimento*, **23-C**, 65 (2000).

4. J.W.M. Cuijpers and A.A.M. Holtslag, Impact of skewness and nonlocal effects on scalar and boundancy fluxes in convective boundary layers, *J. Atmos. Sci.*, **55**, 151 (1998).

5. J.W. Deardorff, The countergradient heat flux in the lower atmosphere and in the laboratory, *J. Atmos. Sci.*, **23**, 503 (1966).

6. G.A. Degrazia and O.L.L. Moraes, New model for eddy diffusivity in a stable boundary layer, *Boundary Layer Meteorol.*, **58**, 205 (1992).

7. G.A. Degrazia, H.F. de Campos Velho and J.C. Carvalho, Nonlocal exchange coefficients for the convective boundary layer derived from spectral properties, *Beitr. Phys. Atmos.*, **70**, 57 (1997).

8. G.A. Degrazia, Modelling dispersion from elevated sources in a planetary boundary layer dominated by moderate convection, *Nuovo Ciment.*, **21C**, 345 (1998).

9. E04UCF routine, *NAG Fortran Library*, Mark 17, Oxford, UK (1995).

10. P.C. Hansen, Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Rev.*, **34**, 561 (1992).

11. A.A.M. Holtslag and C.H. Moeng (1991): Eddy diffusivity and countergradient transport in the convective boundary layer, *J. Atmos. Sci.*, **48**, 1690 (1991).

12. A.A.M. Holtslag and B.A. Boville, Local versus nonlocal boundary layer diffusion in a global climate model, *J. Climate*, **6**, 1825 (1993).

13. A.A.M. Holtslag, Personal communication, (1998).

14. E.T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.*, **106**, 620 (1957).

15. V.A. Morozov, On the solution of functional equations by the method of regularization, *Soviet Math. Dokl.*, **7**, 414 (1966).

16. W.B. Muniz, F.M. Ramos and H.F. de Campos Velho, Entropy and Tikhonov-based regularization techniques applied to the backwards heat equation, *Comp. Math. Appl.*, **40**, 1071 (2000).

17. F.M. Ramos, H.F. de Campos Velho, J.C. Carvalho and N.J. Ferreira, Novel approaches on entropic regularization, *Inverse Probl.*, **15**, 1139 (1999).

18. Z. Sorbjan, *Structure of the Atmospheric Boundary Layer*, Prentice Hall, 1989.

19. G.I. Taylor, Diffusion by Continuous Movements, *Proc. London Math. Soc.*, **20**, 196 (1921).

20. A.N. Tikhonov and V.I. Arsenin, *Solutions of Ill-posed Problems*, John Wiley & Sons, 1977.

APPENDIX: VERTICAL VELOCITY VARIANCE PARAMETERIZATIONS

Expressions for velocity variance in convective boundary layer (CBL) are provided in this section. These values are used to compute the counter-gradient term given by Eq. (2). Table 1 shows the difference between the counter-gradient term evaluated by several variance expressions.

Sorbjan [18] (page 113), (SOR89): Using $D = 0$, $c_{wt} = 0.5$, $c_{wb} = 1$ and $R = -0.2$

$$\sigma_w^2 = 1.8 \left(\frac{z}{h}\right)^{2/3} \left(1 - \frac{z}{h}\right)^{2/3} w_*^2 \ . \qquad (A.1)$$

Holtslag and Moeng [11] for CBL (HM 91):

$$(\sigma_w^2)^{2/3} = \left[1.6u_*^2 \left(1 - \frac{z}{h}\right)\right]^{3/2} + 1.2w_*^3 \left(\frac{z}{h}\right) \left(1 - 0.9\frac{z}{h}\right)^{3/2} \ .(A.2)$$

Degrazia et al. for CBL [7] (DEA 97)

$$\sigma_i^2 = \frac{0.98c_i}{(f_m)_{n,i}^{2/3}} \left(\frac{\Psi}{q_i}\right)^{2/3} \left(\frac{z}{h}\right)^{2/3} w_*^2 \qquad (A.3)$$

where $c_i = 0.3$ for ***u***-component and 0.4 for *v, w* components, $(f_m)_{n,w} = 0.33$, spectral peak frequency in neutral stratification, $(f_m)_i$ is the model spectral peak. The stability function q_i given by:

$$q_i = (f_m)_i (f_m)_{n,i}^{-1} = \frac{z/(\lambda_m)_i}{(f_m)_{n,i}}$$

$$(\lambda_m)_i = \begin{cases} 1.5h; & \text{at } i = u, v \\ 1.8h\,\phi; & \text{at } i = w \end{cases}$$

$$\phi = \{1 - \exp[-4(z/h) - 0.0003[8(z/h)]\} \ . \qquad (A.4)$$

4^{th} International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil 2002

IDENTIFICATION OF AQUIFER TRANSMISSIVITY FROM INTERIOR POINT OBSERVATION

Kazuei Onishi
Department of Mathematical Sciences
Ibaraki University, Mito 310–8512, Japan
onishi@mito.ipc.ibaraki.ac.jp

Kazuya Yasuhara
Department of Urban and Civil Engineering
Ibaraki University, Hitachi 316–8511, Japan
yasuhara@civil.ibaraki.ac.jp

Satoshi Murakami
Department of Urban and Civil Engineering
Ibaraki University, Hitachi 316–8511, Japan
murakami@civil.ibaraki.ac.jp

Yoko Ohura
Faculty of Information Management
Kyushu Institute of Information Science
Dazaifu 818–0117, Japan
ohura@kiis.ac.jp

Kentaro Iijima
Graduate School of Science and Engineering
Ibaraki University, Mito 310-8512, Japan
nm0101a@mcs.ibaraki.ac.jp

ABSTRACT

We are given the volumetric discharge $f(x)$ from the aquifer and hydraulic heads $h^{(j)}$ of wells for $j = 1, 2, ..., m$ at m interior observation points $\boldsymbol{x}^{(j)}$ in a vast area of the Kanto plain north of Tokyo megalopolis, that area now suffers from subsidence. Our Problem consists of identifying the transmissivity $\theta(\boldsymbol{x})$ of the confined aquifer in the area in order to conduct groundwater flow simulation for the purpose of planning the optimal rate of groundwater abstraction in the area to make the subsidence damages minimum.

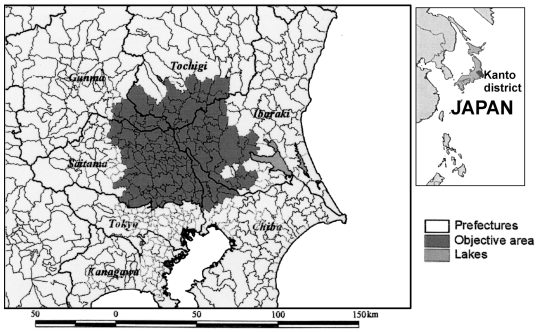
We consider the identification problem in the form of a variational problem for unknown transmissivity, which in turn recasts the problem into a system of primary and the adjoint problems in the form of conventional boundary value problems of the steady seepage equation. Some regularization techniques are introduced in the identification. With an initial guess of the transmissivity distribution, an iterative process of identification starts in order to attain the convergent transmissivity by numerically solving the boundary value problems using the finite element method.

INTRODUCTION

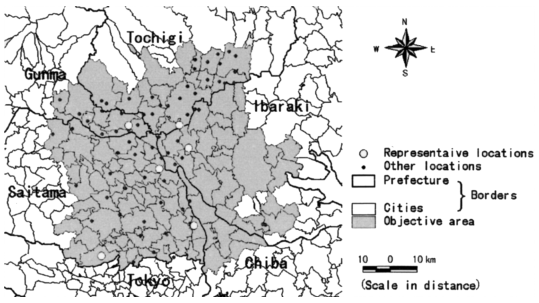
In the northern Kanto area of Japan (see Figure 1(a)), we are given water discharge strength $f(\boldsymbol{x})$ and hydraulic heads $\bar{h}^{(j)}$ for $j = 1, 2, \cdots, m$ at m observation points $\boldsymbol{x}^{(j)}$ as shown in Figure 1(b). Our problem consists of identifying the transmissivity $\theta(\boldsymbol{x})$ in the area, based on the data. Among the m hydraulic heads, we take some of them to determine Dirichlet data along the boundary of the domain of analysis, and we retain some of them for varidation of our calculated results, leaving n hydraulic heads as live data of the interior measumement.

MATHEMATICAL MODEL

We consider an inverse problem of coefficient identification: Given a domain $\Omega \subset \boldsymbol{R}^2$ [the unit of length in meters], in which groundwater flows, a water discharge strength $f(\boldsymbol{x})$ [$\mathrm{m}^3/\mathrm{m}^2 \cdot$ day], a hydraulic head $\bar{h}(\boldsymbol{x})$ [m] on the whole boundary $\Gamma = \partial\Omega$, and internal hydraulic heads $\bar{h}^{(j)}$ [m] for $j = 1, 2, \cdots, n$ at n internal points $\boldsymbol{x}^{(j)}$ of the domain Ω, we will find the transmissivity $\theta(\boldsymbol{x})$ [m^2/day] in Ω, that



(a) Objective area



(b) Groundwater level monitoring locations

Figure 1: Target problem

satisfies [1]

$$\nabla \cdot \theta(\boldsymbol{x}) \nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega$$

subject to

$$h|_\Gamma = \bar{h}(\boldsymbol{x}) \,,$$
$$h(\boldsymbol{x}^{(j)}) = h^{(j)} \quad \text{for} \quad j = 1, 2, \cdots, n \,.$$

Let $\hat{\theta}$ [m^2/day] be some reference transmissivity, independent on unknown θ. As a proper transmissivity, we seek such $\theta(\boldsymbol{x})$ that minimizes the functional

$$J(\theta) = \hat{\theta}^2 \sum_{j=1}^{n} |h(\boldsymbol{x}^{(j)}; \theta) - \bar{h}^{(j)}|^2$$

under the constraints

$$\nabla \cdot \theta(\boldsymbol{x}) \nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega \,,$$
$$h|_\Gamma = \bar{h}(\boldsymbol{x}) \,.$$

In order to identify the proper $\theta(\boldsymbol{x})$, we consider the following minimizing process [2] with a suitably chosen dimensionless numbers α_k $(k = 1, 2, 3, \cdots)$:

1 *Given* $\theta_0|_\Omega$.

2 *For* $k = 0, 1, 2, \cdots$, *until satisfied, do:*

$$\theta_{k+1} = \theta_k - \alpha_k J'(\theta_k) \,.$$

We expect that $\theta(\boldsymbol{x}) = \lim_{k \to \infty} \theta_k(\boldsymbol{x})$. The functional derivative $J'(\theta)$ [m^2/day] is defined from the first variation via

$$J(\theta + \delta\theta) - J(\theta) = \int_\Omega J'(\theta) \delta\theta \, d\Omega + o(\| \, \delta\theta \, \|_{L^2(\Omega)})$$

for any $\delta\theta$ as $\| \, \delta\theta \, \|_{L^2(\Omega)} \to 0$.

In fact, put $\delta h(\boldsymbol{x}; \theta) = h(\boldsymbol{x}; \theta + \delta\theta) - h(\boldsymbol{x}; \theta)$. We see

$$\begin{aligned}
&J(\theta + \delta\theta) - J(\theta) \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \Big\{ |h(\boldsymbol{x}^{(j)}; \theta + \delta\theta) - \bar{h}^{(j)}|^2 \\
&\qquad\qquad - |h(\boldsymbol{x}^{(j)}; \theta) - \bar{h}^{(j)}|^2 \Big\} \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \Big\{ h(\boldsymbol{x}^{(j)}; \theta + \delta\theta) \\
&\qquad\qquad + h(\boldsymbol{x}^{(j)}; \theta) - 2\bar{h}^{(j)} \Big\} \\
&\qquad \Big\{ h(\boldsymbol{x}^{(j)}; \theta + \delta\theta) - h(\boldsymbol{x}^{(j)}; \theta) \Big\} \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \Big\{ \delta h(\boldsymbol{x}^{(j)}; \theta) \\
&\qquad + 2[h(\boldsymbol{x}^{(j)}; \theta) - \bar{h}^{(j)}] \Big\} \delta h(\boldsymbol{x}^{(j)}; \theta) \\
&= \int_\Omega \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x}; \theta) - \bar{h}^{(j)}] \delta(\boldsymbol{x} - \boldsymbol{x}^{(j)}) \\
&\qquad \delta h(\boldsymbol{x}; \theta) d\Omega + o(\| \, \delta\theta \, \|_{L^2(\Omega)})
\end{aligned}$$

with the Dirac measure $\delta(\cdot)$ [$1/\mathrm{m}^2$].

We introduce $v(\boldsymbol{x})$ [m^3/day] as a solution of the problem;

$$\nabla \cdot \theta(\boldsymbol{x}) \nabla v = \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x}; \theta) - \bar{h}^{(j)}] \\ \delta(\boldsymbol{x} - \boldsymbol{x}^{(j)}) \quad \text{in} \quad \Omega, \quad v|_\Gamma = 0 \,.$$

Let $\boldsymbol{n}$ denote outward directed unit normal to

the boundary Γ. Since $\delta h|_\Gamma = 0$, we have

$$
\begin{aligned}
&J(\theta+\delta\theta) - J(\theta) \\
&= \int_\Omega (\nabla\cdot\theta(\boldsymbol{x})\nabla v)\delta h(\boldsymbol{x};\theta)d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) \\
&= \int_\Gamma \theta(\boldsymbol{x})\frac{\partial v}{\partial n}\delta h(\boldsymbol{x};\theta)d\Gamma \\
&\quad - \int_\Omega \theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) \\
&= -\int_\Omega \theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) .
\end{aligned}
$$

By subtracting side by side of the equations;

$$
\begin{aligned}
&\nabla\cdot(\theta+\delta\theta)\nabla h(\boldsymbol{x};\theta+\delta\theta) + f(\boldsymbol{x}) = 0 , \\
&\nabla\cdot\theta\nabla h(\boldsymbol{x};\theta) + f(\boldsymbol{x}) = 0 ,
\end{aligned}
$$

we know that

$$
\nabla\cdot\theta\nabla\delta h(\boldsymbol{x};\theta) + \nabla\cdot\delta\theta\nabla h(\boldsymbol{x};\theta+\delta\theta) = 0 .
$$

Therefore we have

$$
\begin{aligned}
&\int_\Omega v\big\{\nabla\cdot\theta\nabla\delta h(\boldsymbol{x};\theta) \\
&+ \nabla\cdot\delta\theta\nabla h(\boldsymbol{x};\theta+\delta\theta)\big\}d\Omega = 0 .
\end{aligned}
$$

From the integration by parts we have

$$
\begin{aligned}
&\int_\Gamma v\theta\frac{\partial\delta h}{\partial n}d\Gamma - \int_\Omega \theta\nabla v\cdot\nabla\delta h d\Omega \\
&\quad + \int_\Gamma v\delta\theta\frac{\partial h}{\partial n}(\boldsymbol{x};\theta+\delta\theta)d\Gamma \\
&\qquad - \int_\Omega \delta\theta\nabla v\cdot\nabla h(\boldsymbol{x};\theta+\delta\theta)d\Omega \\
&= -\int_\Omega \theta\nabla v\cdot\nabla\delta h d\Omega \\
&\qquad - \int_\Omega \delta\theta\nabla v\cdot\nabla h(\boldsymbol{x};\theta+\delta\theta)d\Omega = 0 ,
\end{aligned}
$$

from which we can see

$$
\begin{aligned}
&J(\theta+\delta\theta) - J(\theta) \\
&\quad = \int_\Omega \nabla v\cdot\nabla h(\boldsymbol{x};\theta+\delta\theta)\delta\theta d\Omega \\
&\qquad + o(\| \delta\theta \|_{L^2(\Omega)}) \\
&\quad = \int_\Omega \nabla v\cdot\nabla h(\boldsymbol{x};\theta)\delta\theta d\Omega \\
&\quad + \int_\Omega \nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)\delta\theta d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) \\
&\quad = \int_\Omega \nabla v\cdot\nabla h(\boldsymbol{x};\theta)\delta\theta d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) .
\end{aligned}
$$

Accordingly we can obtain

$$
J'(\theta) = \nabla v\cdot\nabla h(\boldsymbol{x};\theta) \quad \text{in} \quad \Omega .
$$

Algorithm

1 *Given* $\theta_0|_\Omega$.

2 *For* $k = 0, 1, 2, \cdots$, *until satisfied, do:*

2.1 *Solve* $\nabla\cdot\theta_k(\boldsymbol{x})\nabla h(\boldsymbol{x};\theta_k) + f(\boldsymbol{x}) = 0$ *in* Ω *with* $h|_\Gamma = \bar{h}$ *to find* $h(\boldsymbol{x}^{(j)};\theta_k)$ *for* $j = 1, 2, \cdots, n$.

2.2 *Solve* $\nabla\cdot\theta_k(\boldsymbol{x})\nabla v_k(\boldsymbol{x}) = \hat{\theta}^2\sum_{j=1}^{n} 2[h(\boldsymbol{x};\theta_k) - \bar{h}^{(j)}]\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})$ *in* Ω *with* $v_k|_\Gamma = 0$.

2.3 *Calculate* $J'(\theta_k) = \nabla v_k\cdot\nabla h(\boldsymbol{x};\theta_k)$ *in* Ω.

2.4 *Update* $\theta_{k+1} = \theta_k - \alpha_k J'(\theta_k)$.

REGULARIZATION

Dirichlet Integral Regularizer

We consider the case in which the given internal hydraulic heads $\bar{h}^{(j)}$ for $j = 1, 2, \cdots, n$ at n internal points $\boldsymbol{x}^{(j)}$ are contaminated with measurement errors. In this case, as a proper transmissivity, we seek such $\theta(\boldsymbol{x})$ that minimizes the functional

$$
\begin{aligned}
J(\theta) = \hat{\theta}^2\sum_{j=1}^{n}&|h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}|^2 \\
&+ \eta|\Omega|\int_\Omega |\nabla\theta(\boldsymbol{x})|^2 d\Omega
\end{aligned}
$$

under the constraints

$$
\begin{aligned}
&\nabla\cdot\theta(\boldsymbol{x})\nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega , \\
&h|_\Gamma = \bar{h}(\boldsymbol{x})
\end{aligned}
$$

with the dimensionless regularization parameter η. Here, $|\Omega|$ denotes the area of the domain Ω.

In order to find $J'(\theta)$, we put $\delta h(\boldsymbol{x};\theta) = h(\boldsymbol{x};\theta+\delta\theta) - h(\boldsymbol{x};\theta)$. We assume that $\dfrac{\partial\theta}{\partial n} = 0$.

We can see

$$
\begin{aligned}
&J(\theta+\delta\theta)-J(\theta)\\
&\quad=\hat{\theta}^2\sum_{j=1}^{n}\{|h(\boldsymbol{x}^{(j)};\theta+\delta\theta)-\bar{h}^{(j)}|^2\\
&\qquad\qquad-|h(\boldsymbol{x}^{(j)};\theta)-\bar{h}^{(j)}|^2\}\\
&\qquad+\eta|\Omega|\int_\Omega\left\{|\nabla(\theta+\delta\theta)|^2-|\nabla\theta|^2\right\}d\Omega\\
&\quad=\hat{\theta}^2\sum_{j=1}^{n}\Big\{h(\boldsymbol{x}^{(j)};\theta+\delta\theta)+h(\boldsymbol{x}^{(j)};\theta)-2\bar{h}^{(j)}\Big\}\\
&\qquad\qquad\Big\{h(\boldsymbol{x}^{(j)};\theta+\delta\theta)-h(\boldsymbol{x}^{(j)};\theta)\Big\}\\
&\qquad+\eta|\Omega|\int_\Omega\left(2\nabla\theta\cdot\nabla\delta\theta+|\nabla\delta\theta|^2\right)d\Omega\\
&\quad=\hat{\theta}^2\sum_{j=1}^{n}\Big\{\delta h(\boldsymbol{x}^{(j)};\theta)\\
&\qquad\qquad+2[h(\boldsymbol{x}^{(j)};\theta)-\bar{h}^{(j)}]\Big\}\delta h(\boldsymbol{x}^{(j)};\theta)\\
&\qquad+\eta|\Omega|\Big\{\int_\Gamma 2\frac{\partial\theta}{\partial n}\delta\theta d\Gamma-\int_\Omega 2(\Delta\theta)\delta\theta d\Omega\\
&\qquad\qquad+\int_\Omega|\nabla\delta\theta|^2d\Omega\Big\}\\
&\quad=\int_\Omega\hat{\theta}^2\sum_{j=1}^{n}2[h(\boldsymbol{x};\theta)-\bar{h}^{(j)}]\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})\\
&\qquad\qquad\delta h(\boldsymbol{x};\theta)d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)}).
\end{aligned}
$$

We now introduce $v(\boldsymbol{x})$ as a solution of the problem;

$$
\nabla\cdot\theta(\boldsymbol{x})\nabla v=\hat{\theta}^2\sum_{j=1}^{n}2[h(\boldsymbol{x};\theta)-\bar{h}^{(j)}]\\
\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})\quad\text{in}\quad\Omega\,,\quad v|_\Gamma=0
$$

as before. Since $\delta h|_\Gamma=0$, we have

$$
\begin{aligned}
&J(\theta+\delta\theta)-J(\theta)\\
&\quad=\int_\Omega(\nabla\cdot\theta(\boldsymbol{x})\nabla v)\delta h(\boldsymbol{x};\theta)d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&\quad=\int_\Gamma\theta(\boldsymbol{x})\frac{\partial v}{\partial n}\delta h(\boldsymbol{x};\theta)d\Gamma\\
&\qquad-\int_\Omega\theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&\quad=-\int_\Omega\theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&\quad=\int_\Omega\nabla v\cdot\nabla h(\boldsymbol{x};\theta+\delta\theta)\delta\theta d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&\quad=\int_\Omega\nabla v\cdot\nabla h(\boldsymbol{x};\theta)\delta\theta d\Omega\\
&\qquad+\int_\Omega\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)\delta\theta d\Omega\\
&\qquad-\int_\Omega 2\eta|\Omega|(\Delta\theta)\delta\theta d\Omega+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&\quad=\int_\Omega\{\nabla v\cdot\nabla h(\boldsymbol{x};\theta)-2\eta|\Omega|\Delta\theta\}\,\delta\theta d\Omega\\
&\qquad\qquad+o(\|\,\delta\theta\,\|_{L^2(\Omega)})\,.
\end{aligned}
$$

Accordingly we can obtain

$$
J'(\theta)=\nabla v\cdot\nabla h(\boldsymbol{x};\theta)-2\eta|\Omega|\Delta\theta\quad\text{in}\quad\Omega\,.
$$

Algorithm

1 *Given* $\theta_0|_\Omega$.

2 *For* $k=0,1,2,\cdots$, *until satisfied, do:*

- **2.1** *Solve* $\nabla\cdot\theta_k(\boldsymbol{x})\nabla h(\boldsymbol{x};\theta_k)+f(\boldsymbol{x})=0$ *in* Ω *with* $h|_\Gamma=\bar{h}$ *to find* $h(\boldsymbol{x}^{(j)};\theta_k)$ *for* $j=1,2,\cdots,n$.
- **2.2** *Solve* $\nabla\cdot\theta_k(\boldsymbol{x})\nabla v_k(\boldsymbol{x})=\hat{\theta}^2\sum_{j=1}^{n}2[h(\boldsymbol{x};\theta_k)-\bar{h}^{(j)}]\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})$ *in* Ω *with* $v_k|_\Gamma=0$.
- **2.3** *Calculate* $J'(\theta_k)=\nabla v_k\cdot\nabla h(\boldsymbol{x};\theta_k)-2\eta|\Omega|\Delta\theta_k$ *in* Ω.
- **2.4** *Update* $\theta_{k+1}=\theta_k-\alpha_kJ'(\theta_k)$.

Variance Regularizer

The statistical mean $E[\theta]$ [m^2/day] and the variance $V[\theta]$ [m^4/day^2] of the transmissivity $\theta(\boldsymbol{x})$ over the domain Ω are defined respectively

by

$$E[\theta] = \frac{1}{|\Omega|}\int_\Omega \theta(\boldsymbol{x})d\Omega$$
$$\text{and} \quad V[\theta] = \frac{1}{|\Omega|}\int_\Omega |\theta(\boldsymbol{x}) - E[\theta]|^2 d\Omega$$

with the area $|\Omega|$ of the domain Ω. We notice that $V[\theta] = E[\theta^2] - E[\theta]^2$. As an alternative proper transmissivity, we seek such $\theta(\boldsymbol{x})$ that minimizes the functional

$$J(\theta) = \hat{\theta}^2 \sum_{j=1}^n |h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}|^2 + \eta|\Omega|V[\theta]$$

under the constraints

$$\nabla\cdot\theta(\boldsymbol{x})\nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega\,,$$
$$h|_\Gamma = \bar{h}(\boldsymbol{x})$$

with the regularization parameter η.

In order to find $J'(\theta)$, we put $\delta h(\boldsymbol{x};\theta) = h(\boldsymbol{x};\theta+\delta\theta) - h(\boldsymbol{x};\theta)$. We can see

$$\begin{aligned}
&J(\theta+\delta\theta) - J(\theta)\\
&= \hat{\theta}^2 \sum_{j=1}^n \{|h(\boldsymbol{x}^{(j)};\theta+\delta\theta) - \bar{h}^{(j)}|^2\\
&\qquad - |h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}|^2\}\\
&\quad + \eta|\Omega|\{E[(\theta+\delta\theta)^2]\\
&\qquad - E[\theta+\delta\theta]^2 - (E[\theta^2] - E[\theta]^2)\}\\
&= \hat{\theta}^2 \sum_{j=1}^n \big\{h(\boldsymbol{x}^{(j)};\theta+\delta\theta) + h(\boldsymbol{x}^{(j)};\theta) - 2\bar{h}^{(j)}\big\}\\
&\qquad \Big\{h(\boldsymbol{x}^{(j)};\theta+\delta\theta) - h(\boldsymbol{x}^{(j)};\theta)\Big\}\\
&\quad + \eta|\Omega|\{E[2\theta\delta\theta] - 2E[\theta]E[\delta\theta]\\
&\qquad + E[\delta\theta^2] - E[\delta\theta]^2\}\\
&= \hat{\theta}^2 \sum_{j=1}^n \Big\{\delta h(\boldsymbol{x}^{(j)};\theta)\\
&\qquad + 2[h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}]\Big\}\delta h(\boldsymbol{x}^{(j)};\theta)\\
&\quad + \eta|\Omega|\,\{E[2(\theta - E[\theta])\delta\theta] + V[\delta\theta]\}\\
&= \int_\Omega \hat{\theta}^2 \sum_{j=1}^n 2[h(\boldsymbol{x};\theta) - \bar{h}^{(j)}]\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})\\
&\qquad \delta h(\boldsymbol{x};\theta)d\Omega\\
&\quad + \eta\int_\Omega 2(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)}).
\end{aligned}$$

We introduce $v(\boldsymbol{x})$ as a solution of the problem;

$$\nabla\cdot\theta(\boldsymbol{x})\nabla v = \hat{\theta}^2 \sum_{j=1}^n 2[h(\boldsymbol{x};\theta) - \bar{h}^{(j)}]\delta(\boldsymbol{x}-\boldsymbol{x}^{(j)})$$
$$\text{in} \quad \Omega\,, \qquad v|_\Gamma = 0\,.$$

Since $\delta h|_\Gamma = 0$, we have

$$\begin{aligned}
&J(\theta+\delta\theta) - J(\theta)\\
&= \int_\Omega (\nabla\cdot\theta(\boldsymbol{x})\nabla v)\delta h(\boldsymbol{x};\theta)d\Omega\\
&\quad + \eta\int_\Omega 2(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&= \int_\Gamma \theta(\boldsymbol{x})\frac{\partial v}{\partial n}\delta h(\boldsymbol{x};\theta)d\Gamma\\
&\quad - \int_\Omega \theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega\\
&\quad + \eta\int_\Omega 2(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&= -\int_\Omega \theta(\boldsymbol{x})\nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)d\Omega\\
&\quad + \eta\int_\Omega 2(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&= \int_\Omega \nabla v\cdot\nabla h(\boldsymbol{x};\theta+\delta\theta)\delta\theta d\Omega\\
&\quad + \int_\Omega 2\eta(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&= \int_\Omega \nabla v\cdot\nabla h(\boldsymbol{x};\theta)\delta\theta d\Omega\\
&\quad + \int_\Omega \nabla v\cdot\nabla\delta h(\boldsymbol{x};\theta)\delta\theta d\Omega\\
&\quad + \int_\Omega 2\eta(\theta - E[\theta])\delta\theta d\Omega + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\\
&= \int_\Omega \big\{\nabla v\cdot\nabla h(\boldsymbol{x};\theta) + 2\eta(\theta - E[\theta])\big\}\delta\theta d\Omega\\
&\qquad + o(\|\,\delta\theta\,\|_{L^2(\Omega)})\,.
\end{aligned}$$

Accordingly we can obtain

$$J'(\theta) = \nabla v\cdot\nabla h(\boldsymbol{x};\theta) + 2\eta(\theta - E[\theta]) \quad \text{in} \quad \Omega\,.$$

Algorithm

1 *Given* $\theta_0|_\Omega$.

2 *For* $k = 0, 1, 2, \cdots$, *until satisfied, do:*

2.1 *Solve* $\nabla\cdot\theta_k(\boldsymbol{x})\nabla h(\boldsymbol{x};\theta_k) + f(\boldsymbol{x}) = 0$ *in* Ω *with* $h|_\Gamma = \bar{h}$ *to find* $h(\boldsymbol{x}^{(j)};\theta_k)$ *for* $j = 1, 2, \cdots, n$.

2.2 *Solve* $\nabla \cdot \theta_k(\boldsymbol{x})\nabla v_k(\boldsymbol{x}) = \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x};\theta_k) - \bar{h}^{(j)}]\delta(\boldsymbol{x} - \boldsymbol{x}^{(j)})$ *in* Ω *with* $v_k|_\Gamma = 0$.

2.3 *Calculate* $J'(\theta_k) = \nabla v_k \cdot \nabla h(\boldsymbol{x};\theta_k) + 2\eta(\theta_k - E[\theta_k])$ *in* Ω.

2.4 *Update* $\theta_{k+1} = \theta_k - \alpha_k J'(\theta_k)$.

MIXED BOUNDARY DATA

Suppose that we are given a domain $\Omega \subset \boldsymbol{R}^2$ in which groundwater flows, water discharge strength $f(\boldsymbol{x})$, a hydraulic head $\bar{h}(\boldsymbol{x})$ on an arc Γ_h of the boundary $\Gamma = \partial\Omega$, a boundary discharge $\bar{q}(\boldsymbol{x})$ on the arc $\Gamma_q = \Gamma \setminus \Gamma_h$ in the direction of the exterior normal $\boldsymbol{n}$, and internal hydraulic heads $\bar{h}^{(j)}$ for $j = 1, 2, \cdots, n$ at n internal points $\boldsymbol{x}^{(j)}$ of the domain Ω. Our problem consists of identifying the transmissivity $\theta(\boldsymbol{x})$ in Ω, that satisfies

$$\nabla \cdot \theta(\boldsymbol{x})\nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega$$

subject to

$$h|_{\Gamma_h} = \bar{h}(\boldsymbol{x}), \quad -\theta(\boldsymbol{x})\frac{\partial h}{\partial n}|_{\Gamma_q} = \bar{q},$$
$$h(\boldsymbol{x}^{(j)}) = h^{(j)} \quad \text{for} \quad j = 1, 2, \cdots, n.$$

As a proper transmissivity, we seek such $\theta(\boldsymbol{x})$ that minimizes the functional

$$J(\theta) = \hat{\theta}^2 \sum_{j=1}^{n} |h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}|^2$$

under the constraints

$$\nabla \cdot \theta(\boldsymbol{x})\nabla h(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \quad \text{in} \quad \Omega,$$
$$h|_{\Gamma_h} = \bar{h}(\boldsymbol{x}), \quad \text{and} \quad -\theta(\boldsymbol{x})\frac{\partial h}{\partial n} = \bar{q}(\boldsymbol{x}).$$

We consider the following minimizing process with a suitably chosen numbers α_k ($k = 1, 2, 3, \cdots$):

1 *Given* $\theta_0|_\Omega$.

2 *For* $k = 0, 1, 2, \cdots$, *until satisfied, do:*

$$\theta_{k+1} = \theta_k - \alpha_k J'(\theta_k).$$

The derivative $J'(\theta)$ is defined from the first variation via

$$J(\theta+\delta\theta) - J(\theta) = \int_\Omega J'(\theta)\delta\theta d\Omega + o(\| \delta\theta \|_{L^2(\Omega)})$$

for any $\delta\theta$ as $\| \delta\theta \|_{L^2(\Omega)} \to 0$.

In fact, put $\delta h(\boldsymbol{x};\theta) = h(\boldsymbol{x};\theta+\delta\theta) - h(\boldsymbol{x};\theta)$. We see

$$\begin{aligned}
&J(\theta+\delta\theta) - J(\theta) \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \{ |h(\boldsymbol{x}^{(j)};\theta+\delta\theta) - \bar{h}^{(j)}|^2 \\
&\qquad - |h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}|^2 \} \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \Big\{ h(\boldsymbol{x}^{(j)};\theta+\delta\theta) + h(\boldsymbol{x}^{(j)};\theta) - 2\bar{h}^{(j)} \Big\} \\
&\qquad \Big\{ h(\boldsymbol{x}^{(j)};\theta+\delta\theta) - h(\boldsymbol{x}^{(j)};\theta) \Big\} \\
&= \hat{\theta}^2 \sum_{j=1}^{n} \Big\{ \delta h(\boldsymbol{x}^{(j)};\theta) + 2[h(\boldsymbol{x}^{(j)};\theta) - \bar{h}^{(j)}] \Big\} \\
&\qquad \delta h(\boldsymbol{x}^{(j)};\theta) \\
&= \int_\Omega \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x};\theta) - \bar{h}^{(j)}]\delta(\boldsymbol{x} - \boldsymbol{x}^{(j)}) \\
&\qquad \delta h(\boldsymbol{x};\theta) d\Omega + o(\| \delta\theta \|_{L^2(\Omega)})
\end{aligned}$$

with the Dirac measure $\delta(\cdot)$.

We introduce $v(\boldsymbol{x})$ as a solution of the problem;

$$\nabla \cdot \theta(\boldsymbol{x})\nabla v = \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x};\theta) - \bar{h}^{(j)}]$$
$$\delta(\boldsymbol{x} - \boldsymbol{x}^{(j)}) \quad \text{in} \quad \Omega,$$
$$v|_{\Gamma_h} = 0, \quad \text{and} \quad \frac{\partial v}{\partial n}|_{\Gamma_q} = 0.$$

Since $\delta h|_{\Gamma_h} = 0$, we have

$$\begin{aligned}
&J(\theta+\delta\theta) - J(\theta) \\
&= \int_\Omega (\nabla \cdot \theta(\boldsymbol{x})\nabla v)\delta h(\boldsymbol{x};\theta) d\Omega \\
&+ o(\| \delta\theta \|_{L^2(\Omega)}) \\
&= \int_\Gamma \theta(\boldsymbol{x})\frac{\partial v}{\partial n}\delta h(\boldsymbol{x};\theta) d\Gamma \\
&- \int_\Omega \theta(\boldsymbol{x})\nabla v \cdot \nabla \delta h(\boldsymbol{x};\theta) d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}) \\
&= - \int_\Omega \theta(\boldsymbol{x})\nabla v \cdot \nabla \delta h(\boldsymbol{x};\theta) d\Omega + o(\| \delta\theta \|_{L^2(\Omega)}).
\end{aligned}$$

By subtracting side by side of the equations;

$$\nabla \cdot (\theta + \delta\theta)\nabla h(\boldsymbol{x}; \theta + \delta\theta) + f(\boldsymbol{x}) = 0\,,$$
$$\nabla \cdot \theta \nabla h(\boldsymbol{x}; \theta) + f(\boldsymbol{x}) = 0\,,$$

we know that

$$\nabla \cdot \theta \nabla \delta h(\boldsymbol{x}; \theta) + \nabla \cdot \delta\theta \nabla h(\boldsymbol{x}; \theta + \delta\theta) = 0\,.$$

Therefore we have

$$\int_\Omega v\big\{\nabla \cdot \theta \nabla \delta h(\boldsymbol{x}; \theta) + \nabla \cdot \delta\theta \nabla h(\boldsymbol{x}; \theta + \delta\theta)\big\} d\Omega = 0\,.$$

Moreover, by subtracting side by side of the equations;

$$-(\theta + \delta\theta)\frac{\partial h}{\partial n}(\boldsymbol{x}; \theta + \delta\theta) = \bar{q}(\boldsymbol{x})\,,$$
$$-\theta \frac{\partial h}{\partial n}(\boldsymbol{x}; \theta) = \bar{q}(\boldsymbol{x})$$

on Γ_q, we know that

$$-\theta \frac{\partial \delta h}{\partial n}(\boldsymbol{x}; \theta) - \delta\theta \frac{\partial h}{\partial n}(\boldsymbol{x}; \theta + \delta\theta) = 0\,.$$

From the integration by parts we have

$$\int_\Gamma v \left(\theta \frac{\partial \delta h}{\partial n}(\boldsymbol{x}; \theta) + \delta\theta \frac{\partial h}{\partial n}(\boldsymbol{x}; \theta + \delta\theta)\right) d\Gamma$$
$$- \int_\Omega \theta \nabla v \cdot \nabla \delta h d\Omega$$
$$- \int_\Omega \delta\theta \nabla v \cdot \nabla h(\boldsymbol{x}; \theta + \delta\theta) d\Omega$$
$$= - \int_\Omega \theta \nabla v \cdot \nabla \delta h d\Omega$$
$$- \int_\Omega \delta\theta \nabla v \cdot \nabla h(\boldsymbol{x}; \theta + \delta\theta) d\Omega = 0\,,$$

from which we can see

$$J(\theta + \delta\theta) - J(\theta)$$
$$= \int_\Omega \nabla v \cdot \nabla h(\boldsymbol{x}; \theta + \delta\theta)\delta\theta d\Omega$$
$$+ o(\| \delta\theta \|_{L^2(\Omega)})$$
$$= \int_\Omega \nabla v \cdot \nabla h(\boldsymbol{x}; \theta)\delta\theta d\Omega$$
$$+ \int_\Omega \nabla v \cdot \nabla \delta h(\boldsymbol{x}; \theta)\delta\theta d\Omega$$
$$+ o(\| \delta\theta \|_{L^2(\Omega)})$$
$$= \int_\Omega \nabla v \cdot \nabla h(\boldsymbol{x}; \theta)\delta\theta d\Omega + o(\| \delta\theta \|_{L^2(\Omega)})\,.$$

Accordingly we can obtain

$$J'(\theta) = \nabla v \cdot \nabla h(\boldsymbol{x}; \theta) \quad \text{in} \quad \Omega\,.$$

Algorithm

1 *Given* $\theta_0|_\Omega$.

2 *For* $k = 0, 1, 2, \cdots$, *until satisfied, do:*

2.1 *Solve* $\nabla \cdot \theta_k(\boldsymbol{x})\nabla h(\boldsymbol{x}; \theta_k) + f(\boldsymbol{x}) = 0$ *in* Ω *with* $h|_{\Gamma_h} = \bar{h}$ *and* $-\theta_k \frac{\partial h}{\partial n}|_{\Gamma_q} = \bar{q}$ *to find* $h(\boldsymbol{x}^{(j)}; \theta_k)$ *for* $j = 1, 2, \cdots, n$.

2.2 *Solve* $\nabla \cdot \theta_k(\boldsymbol{x})\nabla v_k(\boldsymbol{x}) = \hat{\theta}^2 \sum_{j=1}^{n} 2[h(\boldsymbol{x}; \theta_k) - \bar{h}^{(j)}]\delta(\boldsymbol{x} - \boldsymbol{x}^{(j)})$ *in* Ω *with* $v_k|_{\Gamma_h} = 0$ *and* $\frac{\partial v}{\partial n}|_{\Gamma_q} = 0$.

2.3 *Calculate* $J'(\theta_k) = \nabla v_k \cdot \nabla h(\boldsymbol{x}; \theta_k)$ *in* Ω.

2.4 *Update* $\theta_{k+1} = \theta_k - \alpha_k J'(\theta_k)$.

EXAMPLES

Let Ω be the square $\Omega = (0, 3) \times (0, 3)$. Let $h(\boldsymbol{x}) = 1 + \frac{1}{18}(x_1^2 + x_2^2)$ and $\theta(\boldsymbol{x}) = 2 - \frac{1}{36}(x_1 + x_2)^2$. Corresponding to these synthetic $h(\boldsymbol{x})$ and $\theta(\boldsymbol{x})$, the abstraction $f(\boldsymbol{x}) = \frac{1}{9}\left\{\frac{1}{9}(x_1 + x_2)^2 - 4\right\}$ satisfies the seepage equation.

The domain Ω is divided into 3–noded linear triangular finite elements as shown in Figure 2, where internal observation points ($n = 15$) are indicated by black dots among nodes of the triangular elements. The exact value of $h(\boldsymbol{x})$ at all the nodes on the boundary Γ is prescribed as the Dirichlet data $\bar{h}$. The initial guess $\theta_0|_\Omega$ is set as $\theta_0(\boldsymbol{x}) = (1 + \epsilon(\boldsymbol{x}))\theta(\boldsymbol{x})$, where $\epsilon(\boldsymbol{x})$ is a randomly distributed error of the magnitude 5%.

Identified transmissivity θ_k by using the Dirichlet integral regularizer is shown in Figure 3(a). In comparison with the exact θ in Figure 3(b), the identification is fairly good.

The method is applied to our target problem in Figure 1. Source data on the water discharge strength from the aquifer and hydraulic heads are available in Hayano[3]. The initial

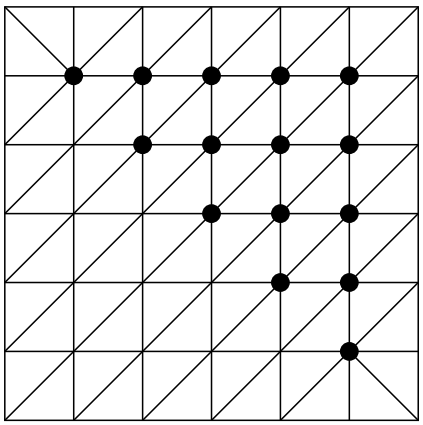
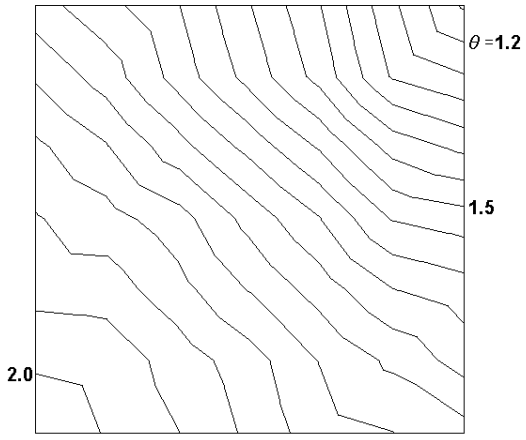
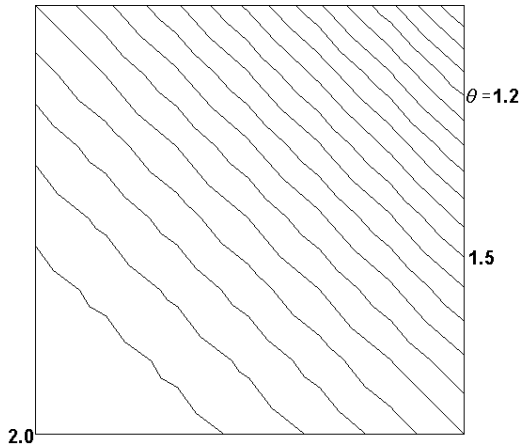


Figure 2: Finite element mesh (• observation points)



(a) Identified θ_k



(b) Exact θ

Figure 3: Hydraulic transmissivity θ

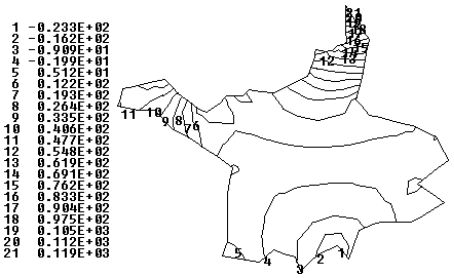


Figure 4: Calculated hydraulic head

guress θ_0 is obtained from [3] as well. Calculated hydraulic head is presented in Figure 4. When random errors of the magnitude 10% are added to θ_0, the calculated hydraulic head remains almost same.

CONCLUDING REMARKS

Variational methods are presented for identification of aquifer transmissivity from a set of measurement of hydraulic heads at internal points of the region. It is shown that idea of the method can be extended to the case of multiple interior observations and to identification of the transmissivity of transient state. The advantages of the present method are only the requirement for the standard finite element solver of the groundwater flow equation and the security of stability in the numerical identification process.

REFERENCES

[1] G. F. Pinder and W. G. Gray, *Finite Element Simulation in Surface and Subsurface Hydrology*, Academic Press, New York, 1977.

[2] K. Kobayashi, K. Onishi, and Y. Ohura, On identifying Dirichlet condition for 2D Laplace equation by BEM, *Engineering Analysis with Boundary Elements*, **17**(**3**), 223–230 (1996).

[3] T. Hayano, Under–graduate thesis, Ibaraki University (2001).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

EVOLUTIONARY IDENTIFICATION OF MATERIAL DEFECTS

Tadeusz Burczyński
Department for Strength of Materials and Computational Mechanics, Silesian University of Technology, Konarskiego 18a, 44-100 Gliwice, Poland
burczyns@polsl.gliwice.pl

Wacław Kuś
Department for Strength of Materials and Computational Mechanics, Silesian University of Technology, Konarskiego 18a, 44-100 Gliwice, Poland
wacok@rmt4.kmt.polsl.gliwice.pl

Piotr Orantek
Department for Strength of Materials and Computational Mechanics, Silesian University of Technology, Konarskiego 18a, 44-100 Gliwice, Poland
orantek@rmt4.kmt.polsl.gliwice.pl

Marek Nowakowski
Department for Strength of Materials and Computational Mechanics, Silesian University of Technology, Konarskiego 18a, 44-100 Gliwice, Poland
marek@rmt4.kmt.polsl.gliwice.pl

ABSTRACT

Evolutionary identification of multiple material defects (voids and cracks) in mechanical systems under dynamical loads is presented. The identification belongs to inverse problems and is treated here as an output (measurement) error minimization, which is solved using numerical optimization methods. The output error is defined in the form of a functional of boundary displacements. An evolutionary hybrid algorithm with the gradient mutation is employed to identification of internal defects. Numerical tests of identification internal defects for 2-D and 3-D problems are presented.

INTRODUCTION

Most of the catastrophic failure of mechanical structures were caused by appearance of material defects. There are several non-destructive methods, explored in condition monitoring to identification of such defects but only a few of them are able to find internal defects, which in some cases are very hardly detectable.

The goal of the proposed work is to develop and examine a solution technique for non-destructive crack and void identification. This technique is based on minimization approach performed by the evolutionary algorithm and using the boundary element method. Evolutionary algorithms were used in identification problems in [2], [3], [4], [7] and [8]. The paper deals with the identification of multiple internal defects in mechanical systems being under dynamical loads. In order to solve the defect identification problem the evolutionary hybrid approach is proposed [5]. This approach is based on a coupling of the evolutionary algorithm and the gradient algorithm. A special gradient mutation is employed, in which shape sensitivity information is used.

FORMULATION OF THE PROBLEMS

Consider a bounded body B with an external boundary S, containing an internal defect in the form of a void V of the boundary Γ (Fig. 1a) or a crack with the crack surface Γ (Fig. 1b). Let Ω denote the actual body (i.e. containing the defect): $\Omega = B \backslash V$ or $\Omega = B \backslash \Gamma$ and $\partial\Omega = S \cup \Gamma$. The displacement $\boldsymbol{u}$, strain $\boldsymbol{\varepsilon}$ and stress $\boldsymbol{\sigma}$ are related by well-known field equations of linear elastodynamics in the time domain:

$$\begin{aligned} \operatorname{div}\boldsymbol{\sigma} - \rho\ddot{\boldsymbol{u}} &= 0 \\ \sigma = \boldsymbol{C} : \boldsymbol{\varepsilon} & \\ \boldsymbol{\varepsilon} = \frac{1}{2}\left(\nabla \boldsymbol{u} + \nabla^T \boldsymbol{u}\right) & \end{aligned} \tag{1}$$

where ρ - a material density, $\mathbf{C}$ – a fourth-order elasticity tensor. Eqs (1) are completed with boundary and initial conditions. The given traction $\overline{\boldsymbol{p}}$ is imposed on a part of the external boundary S, while on the rest of S the displacement $\overline{\boldsymbol{u}}$ is known. The boundary Γ is traction-free and the initial rest is assumed. The traction vector $\boldsymbol{p} = \boldsymbol{\sigma}\mathbf{n}$ is defined in terms of the outward unit normal $\mathbf{n}$ to boundary S. In the crack case the displacement $\boldsymbol{u}$ is allowed to a jump across Γ; $[\![\boldsymbol{u}]\!] = \boldsymbol{u}^+ - \boldsymbol{u}^- \neq \mathbf{0}$.

If the body undergoes free vibration the governing equation is described as follows:

$$\mathrm{div}\boldsymbol{\sigma} + \omega^2 \rho \boldsymbol{u} = 0 \qquad (2)$$

where ω denotes a circular eigenfrequency of the body.

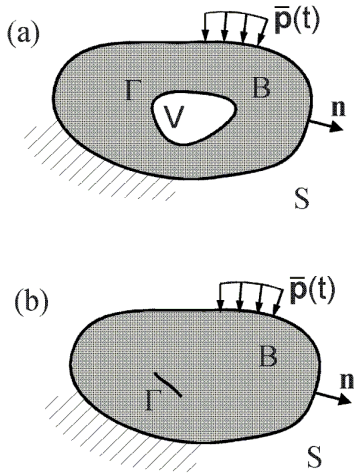


Fig. 1. A body with an internal defect:
a) void, b) crack

Consider the problem of finding the shape and position of an internal defect using elastodynamics experimental (or, for instance presented in this paper, simulated) data. The lack of information about V and Γ is compensated by some knowledge about $\boldsymbol{u}$ on S or ω (redundant boundary data). The usual approach for finding Γ is the minimization of some distance J between $\boldsymbol{u}$ or ω computed for an arbitrary internal defect and $\hat{\boldsymbol{u}}$ or $\hat{\omega}$ measured or simulated (computed for the actual defects), e.g.:

$$J = w_1 J_1 + w_2 J_2 \qquad (3)$$

where w_1 and w_2 are weight coefficients, J_1 and J_2 are defined as follows:

$$J_1 = \frac{1}{2}\sum_{i=1}^{N}\left(\hat{\omega}_i - \omega_i\right)^2 \qquad (4)$$

$$J_2 = \int_0^T \int_{S_m} \varphi\left[\boldsymbol{u}(\boldsymbol{x},t)\right] dSdt = \frac{1}{2}\int_0^T \int_S \left[\hat{\boldsymbol{u}}(\boldsymbol{x},t) - \boldsymbol{u}(\boldsymbol{x},t)\right]^2 dSdt \qquad (5)$$

where ω_i indicates i-th circular eigenfrequency of the body, $\boldsymbol{u}(\boldsymbol{x},t)$ is a displacement vector of the point $\boldsymbol{x}$ on the boundary S at time t.

The minimization of J with respect to Γ needs in turn, for efficiency, the evaluation of the value of J and its gradient with respect to perturbations of Γ.

EVOLUTIONARY IDENTIFICATION METHODS

A *hybrid evolutionary* algorithm is applied to the identification of an internal defect with a boundary Γ. The hybrid algorithm, which connects evolutionary and gradient algorithms together [5], is considerably more efficient than the classical genetic algorithm and its application makes the results more accurate. The objective function (3) is called a fitness function. The hybrid algorithm minimizes the fitness function with respect to defect shape parameters. A vector chromosome characterizes the solution:

$$\boldsymbol{z} = \{z_1, z_2 \ldots z_i \ldots z_n\} \qquad (6)$$

where z_i are genes which parameterize the defect. The genes are real numbers on which constraints are imposed in the form:

$$z_{iL} \le z_i \le z_{iR} \quad ; \text{i=1,2,...n} \qquad (7)$$

The evolutionary algorithm starts with an initial generation. This generation consists of N chromosomes generated in a random way. Every gene is taken from the feasible domain. Evolutionary operators: mutation and crossover modify the initial generation. The next stage is an evaluation of the fitness function for every chromosome and the selection is employed. The selection is performed in the form of the ranking selection or the tournament selection [6]. The next generation is created and operators work for this generation and the process is repeated. The algorithm is stopped if the chromosome, for which the value of the fitness function is zero, has been found. An effectiveness of the evolutionary

algorithm depends on its operators, which can be defined in a different way.

The crossover operator swaps some chromosome of the selected parents in order to create the offspring. Simple, arithmetical and heuristic crossover operators are used.

The *simple crossover* needs two parents and produces two descendants. The simple crossover may produce the offspring outside the design space. To avoid this, a parameter $\alpha \in [0,1]$ is applied. For a randomly generated crossing parameter *i* it works as follows (chromosomes $\mathbf{z}_1$, $\mathbf{z}_2$ are parents):

$$\begin{aligned} &\text{p1}: \ \mathbf{z}_1 = \{z_1, z_2, \ldots, z_i, \ldots, z_n\} \\ &\text{p2}: \ \mathbf{z}_2 = \{e_1, e_2, \ldots, e_i, \ldots, e_n\} \end{aligned} \tag{8}$$

d1:

$$\mathbf{z}_1^{'} = \{z_1, \ldots, z_i, +\alpha e_{i+1} + (1-\alpha) z_{i+1}, \ldots, \alpha e_n + (1-\alpha) z_n\} \tag{9}$$

d2:

$$\mathbf{z}_2^{'} = \{e_1, \ldots, e_i, +\alpha z_{i+1} + (1-\alpha) e_{i+1}, \ldots, \alpha z_n + (1-\alpha) e_n\} \tag{10}$$

The *arithmetical crossover* gives two descendants, which are a linear combination of two parents

$$\mathbf{z}_1^{'} = \alpha \mathbf{z}_1 + (1-\alpha)\mathbf{z}_2 \,; \quad \mathbf{z}_2^{'} = \alpha \mathbf{z}_2 + (1-\alpha)\mathbf{z}_1 \tag{11}$$

The *heuristic crossover* produces a single offspring:

$$\mathbf{z}_1^{'} = r(\mathbf{z}_2 - \mathbf{z}_1) + \mathbf{z}_2 \tag{12}$$

where *r* is a random value from the range [0,1] and $J(\mathbf{z}_2) \le J(\mathbf{z}_1)$.

Four kinds of mutation operators: uniform, boundary, non-uniform and gradient mutation, are used:

$$\begin{aligned} &\text{before mutation: } \mathbf{z}_1 = \{z_1, z_2 \ldots, z_i, \ldots, z_n\} \\ &\text{after mutation: } \mathbf{z}_1^{'} = \{z_1, z_2 \ldots, z_i^{'}, \ldots, z_n\} \end{aligned} \tag{13}$$

The *uniform mutation:* children are allowed to move freely within the feasible domain and the gene $z_i^{'}$ takes any arbitrary value from the range $[z_{iL}, z_{iR}]$.

The *boundary mutation:* the chromosome can take only boundary values of the design space, $z_i^{'} = z_{iL}$ or $z_i^{'} = z_{iR}$.

The *non-uniform mutation:* This operator depends on generation number t and is employed in order to tune of the system

$$\mathbf{z}_i^{'} = \begin{cases} z_i + \Delta(t, z_{iR} - z_i) & \text{if a random digit is 0} \\ z_i - \Delta(t, z_i - z_{iL}) & \text{if a random digit is 1} \end{cases} \tag{14}$$

where the function Δ takes value from the range [0, e].

A special type of mutation, so called the *gradient mutation*, is applied. This mutation is characterized by a full genetic interference, which means a modification of genes making use of information about the fitness function gradient. This single-argument operator changes any chromosome on the ground of the fitness function gradient:

$$\mathbf{z}^{'} = \mathbf{z} + \Delta\mathbf{z} \tag{15}$$

where $\Delta\mathbf{z} = \beta\mathbf{h}$, while β is a coefficient determining a step increment in a search direction $\mathbf{h}$. The search direction $\mathbf{h} = \mathbf{h}(\nabla_z J)$ depends on the fitness function gradient $\nabla_z J$. In the paper the steepest descent method is proposed for evaluation the direction $\mathbf{h} = -\nabla_z J$.

The sensitivity of *J* with respect to shape parameters of the defect is calculated using the adjoint variable method and the boundary element method. For the case of the void the first derivative can be obtained according the following formulas:

- for free vibration:

$$\frac{dJ_1}{dz} = -\sum_{i=1}^{N} (\hat{\omega}_i - \omega_i) \frac{1}{2\omega_i} \int_{\Gamma_d} \left[\boldsymbol{\sigma}(\boldsymbol{u}) \cdot \boldsymbol{\varepsilon}(\boldsymbol{u}) - \omega^2 \rho \boldsymbol{u} \cdot \boldsymbol{u} \right] n_k \theta_k^q \, dS_m \tag{16}$$

- for transient vibration:

$$\frac{dJ_2}{dz} = \int_0^T \int_\Gamma [\boldsymbol{\sigma}(\boldsymbol{u}) : \nabla(\boldsymbol{v}) - \rho \dot{\boldsymbol{u}} . \dot{\boldsymbol{v}}] \, \boldsymbol{\theta} . \mathbf{n} \, \mathrm{dSdt} \tag{17}$$

where: $\boldsymbol{\theta}$ is a shape transformation velocity defined at points on the boundary Γ for defect shape parameters $\mathbf{z}$, $\mathbf{v}$ is a solution of the adjoint

problem, described by Eq. (1) with the following boundary and terminal conditions:

$$\boldsymbol{p}(\mathbf{v}) = -\frac{\partial \varphi}{\partial \mathbf{u}} \quad \text{on S;} \quad \boldsymbol{p}(\mathbf{v}) = \mathbf{0} \quad \text{on } \Gamma; \qquad (18)$$

$$\mathbf{v} = \dot{\mathbf{v}} = \mathbf{0} \quad \text{in } \Omega, \text{ at } t = T$$

Eq. (17) can not be used for the defect in a form of a crack because of the singularities, which arise at crack tips. Assume that the dynamic stress intensity factors (DSIF) at tip $\mathbf{x}^i$ associated with the solutions of the primary and adjoint problems, respectively, are known: $K_I^u(t;\mathbf{x}^i)$, $K_{II}^u(t;\mathbf{x}^i)$, $K_I^v(t;\mathbf{x}^i)$, $K_{II}^v(t;\mathbf{x}^i)$. Now the sensitivity expression for the crack can be derived using DSIF [1], [7]:

(19)

$$\frac{dJ_2}{dz} = \int_0^T \int_{\mathbf{r}} [[\sigma(\boldsymbol{u}) : \nabla(\boldsymbol{v}) - \rho \dot{\boldsymbol{u}}.\dot{\boldsymbol{v}}]]\, \boldsymbol{\theta}.\mathbf{n}\, dS dt -$$

$$\frac{1}{} \sum_{i=1}^{2} \int_0^T \left[\begin{matrix} \left(K_I^u(t;\mathbf{x}^i)\, K_I^v(t;\mathbf{x}^i) + K_{II}^u(t;\mathbf{x}^i)\, K_{II}^v(t;\mathbf{x}^i)\right)\left(\begin{smallmatrix} i \\ \tau \end{smallmatrix}\right) \\ -\left(K_I^u(t;\mathbf{x}^i)\, K_{II}^v(t;\mathbf{x}^i) + K_{II}^u(t;\mathbf{x}^i)\, K_I^v(t;\mathbf{x}^i)\right)\left(\begin{smallmatrix} i \\ n \end{smallmatrix}\right) \end{matrix} \right] dt$$

where $\begin{smallmatrix} i \\ \tau \end{smallmatrix}$, $\begin{smallmatrix} i \\ n \end{smallmatrix}$ are the tangent and normal components of the crack tip transformation velocity.

DEFECT PARAMETRIZATION

The material defect is parametrized as an elliptical flaw (Fig. 2). In this case the chromosome, for the *i-th* flaw, consists of five genes

$$\mathbf{z}^i = \{z_1, z_2, z_3, z_4, z_5\} \qquad (20)$$

where z_1=x and z_2=y are co-ordinates of the center of the flaw, $z_3=r_1$ and $z_4=r_2$ are radii of the flaw and $z_5=\alpha$ is an angle.

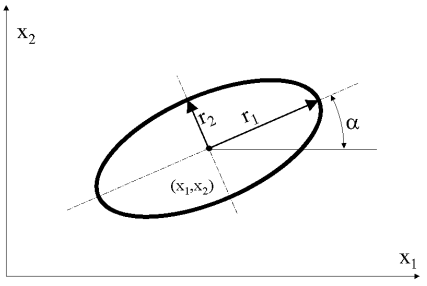


Fig.2 The elliptical flaw

From the elliptical flaw one can obtain special material defects as:

- circular void if $r_1=r_2=r$ and $\alpha=0$,
- crack if $r_2 \le r_{min}$, where r_{min} is a prescribed admisible small value.

In the case if $r_1 \le r_{min}$ and $r_2 \le r_{min}$ the defect does not exist.

For the case when the body contains *n* defects the chromosome takes the form

$$\mathbf{z} = \{\mathbf{z}^1, \mathbf{z}^2, ..\mathbf{z}^i. ,\mathbf{z}^n\} \qquad (21)$$

where $\mathbf{z}^i$ is the vector which contains genes for the *i-th* elliptical flaw.

NUMERICAL EXAMPLES

Numerical tests of identification have been carried out for 2-D and 3-D structures with internal defects. An identification procedure of the defect is based on the evolutionary programming and employs information on the gradient of the objective functional.

Example 1

A 2-D structure, showed in the Fig. 3 contains two internal defects. The actual parameters of an elliptic void are: $\mathbf{z}^2=\mathbf{z}(2)=\{50, 25, 5, 2.5, 2.5\}$, where the first two parameters are co-ordinates of the ellipse center, next - two radii of the ellipse and the last one – the angle between the x_1 axis and first radius. The actual crack parameters are: $\mathbf{z}^1=\mathbf{z}(1)=\{20, 30, 5, 0, 0.25\}$ and are defined as for the ellipse. The identification task is to find a number of defects and their shape having displacements $\hat{\mathbf{u}}(\mathbf{x},t)$ in 33 sensor points, showed in the Fig. 3.

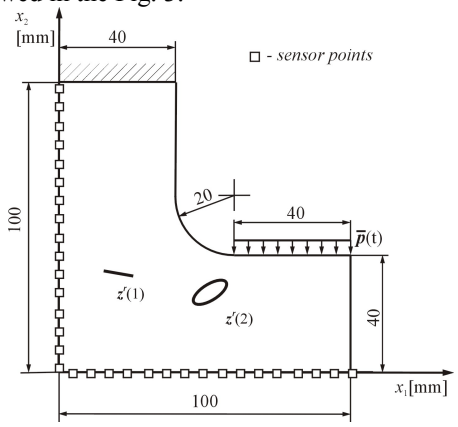


Fig.3 The 2D structure with an internal crack and void

The structure is loaded by $\mathbf{p}(t)=p_0 \sin\omega t$ (p_0=40 kN/m, ω=15708 rad/s) in time $t\in[0, 600\mu s]$ and has the following material properties: the Young modulus E = 0.2E12 Pa, the Poisson's ratio ν = 0.3 and the density ρ = 7800 kg/m^3. The multiple defect identification has been solved with the assumption, that the body contains: 2 defects, 1 defect or no defect. The proposed hybrid algorithm was used to solve this example. The chromosome consists of 10 genes, where first 5 parameterize first ellipse, and last 5 the second ellipse. In the case if one of genes, which is an ellipse radius is less than r_{min} = 2mm, the ellipse becomes a crack, when the both radii are less than r_{min} the ellipse disappears. The population contains 2000 chromosomes. The tournament method of selection was used. The solution was obtained for the case with no noise in a 100 generations and for noisy data in 120 generation. A value of the fitness function for the best chromosome found in each generation is showed in the Fig. 5, while Fig. 4 presents the best solution of the first and the last generation.

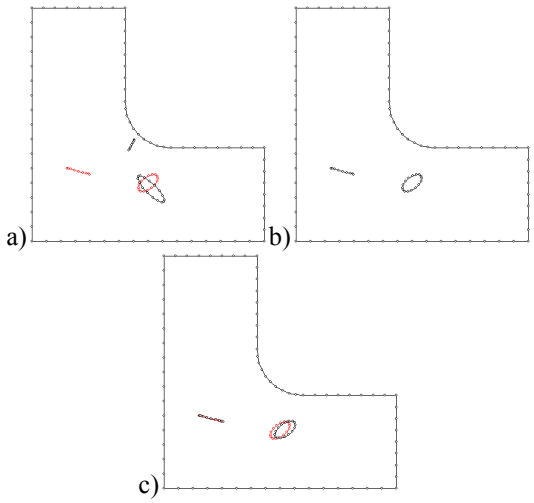


Fig. 4 Identification results: a) 1st generation, b) 100th generation, c) 120th generation (noisy data)

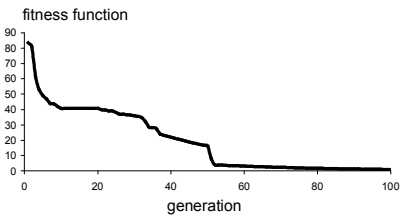


Fig. 5 The fitness function for the best chromosome of each generation

Example 2

The 2-D structure, showed in the Fig. 3 contains now two circular voids and one elliptical. Their actual shape parameters are the following: $\mathbf{z}^1$={70, 20, 3, 3, 0}; $\mathbf{z}^2$={20, 70, 2, 2, 0}; $\mathbf{z}^3$={20, 20, 6, 3, 1}. The identification task is to find a number of defects, their size and coordinates having: (i) eigenvalues ω_i, i=1,2,3; (ii) displacements $\mathbf{u}(x,t)$ in 21 boundary sensor points. The chromosome had 15 genes, because algorithm could find max 3 voids. If its value of radius is less than the critical value r_{min} then the void vanishes. The population consists of 3000 chromosomes. A value of the fitness function for the best chromosome of each generation is showed in the Fig. 6, but the best solutions in four various generations are shown in Fig. 7.

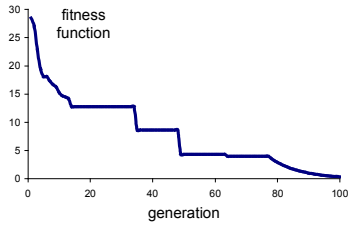


Fig. 6 The fitness function for the best chromosome of each generation

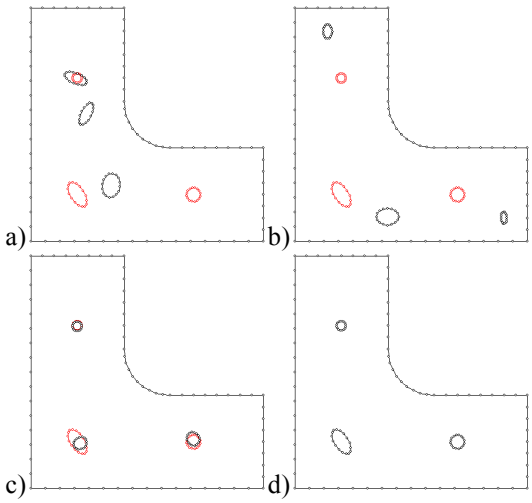


Fig. 7 Identification results for generation number: a) 1st, b) 10th , c) 50th and d) 100th .

Example 3

A 3-D structure – the cube with a 20 [cm] side, showed in the Fig. 8, has one wall supported, while the opposite one is subjected to the harmonic load $p=p_o \sin\omega t$. The load is uniformly distributed on the wall and has different direction in each quarter ($p_0=15000[N/m^2]$, $\omega=31[rad/s]$). The mass density of the structure is $\rho=100[kg/m^3]$, the shear modulus $G=1*10^6[Pa]$ and the Poisson's ratio $\nu=0.25$. The structure contains two internal defects in a form of spherical voids, which parameters – coordinates of centers and radii – are given in the Tab. 1 as actual parameters. The hybrid algorithm, using values of amplitudes in 64 sensor points, placed uniformly on four walls, carried out the identification of defects. The population contains 200 chromosomes. The result, obtained in the 200 generations, is presented in the Fig. 9.

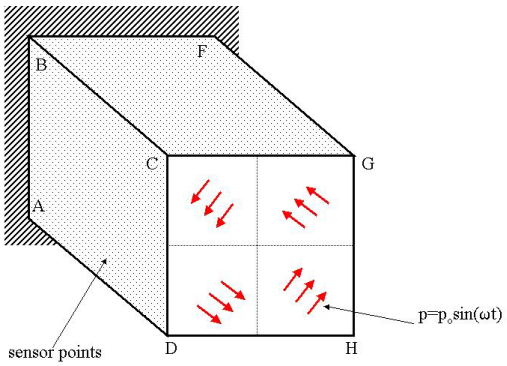


Fig. 8 The 3-D structure:
loads and location of the sensor points

Table 1. Parameters of the defects

defect parameter	actual	final
$x_1(1)$	5.00	4.99
$x_2(1)$	15.00	14.98
$x_3(1)$	15.00	14.98
r(1)	2.00	2.00
$x_1(2)$	15.00	15.83
$x_2(2)$	5.00	4.25
$x_3(2)$	15.00	14.22
r(2)	2.00	2.12

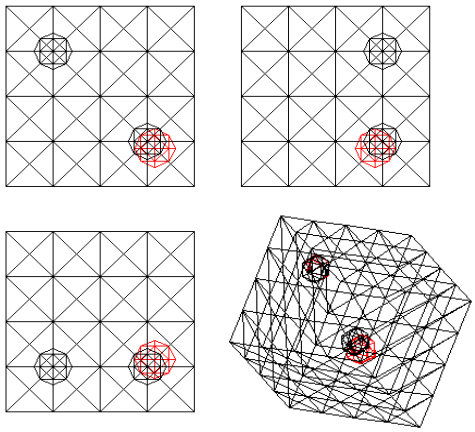


Fig. 9 Identification results
of two spherical defects

CONCLUSIONS

In the present work the problem of evolutionary identification of the internal defects was presented. The evolutionary hybrid algorithm based on the gradient mutation is employed in identification of voids and cracks. The gradient mutation operators were evaluated using sensitivity analysis of the fitness function. The presented results of defect identification are very accurate for displacement data simulated numerically for actual positions and shapes of defects. Even in the case of noisy data used for evolutionary calculation, the results remain reasonably accurate. This algorithm is considerably more efficient than the simple evolutionary algorithm and its application makes the results more accurate.

REFERENCES

1. M. Bonnet, T. Burczyński and M. Nowakowski, Sensitivity analysis for shape perturbation of cavity or internal crack using BIE and adjoint variable approach, *International Journal for Solids and Structures* (in press).

2. T. Burczyński, W. Beluch, A. Długosz, P. Orantek, and M. Nowakowski, Evolutionary methods in inverse problems of engineering mechanics, in: *Inverse Problems in Engineering Mechanics II* (eds. M. Tanaka and G.S. Dulikravich), Elsevier Science Ltd, London 2000, p.553-562.

3. T. Burczyński, W. Beluch, A. Długosz, W. Kuś, M. Nowakowski and P. Orantek, Evolutionary coputation in optimization and

identification, *Computer Assisted Mechanics and Engineering Sciences*, **9**, 3-20, (2002).

4. T. Burczyński, W. Kuś, M. Nowakowski and P. Orantek, Evolutionary algorithms in nondestructive identification of internal defects, *Proc. 5th Conference on Evolutionary Algorithms and Global Optimization,* Jastrzębia, Góra 2001, p.48-55.

5. T. Burczyński and P. Orantek, Evolutionary algorithms aided by sensitivity information, in: *Artificial Neural Nets and Genetic Algorithms* (ed.s V.Kurkova et al.), Springer 2001, 272-275.

6. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolutionary Programs*. Springer-Verlag, AI Series, New York 1992.

7. M. Nowakowski, Sensitivity analysis and identification of internal boundaries using boundary element method, *Ph.D. Thesis, Silesian University of Technology*, Gliwice 2000 (in Polish).

8. G.E. Stavroulakis and H. Antes, Flaw identification in elastomechanics: BEM simulation with local and genetic optimization, *Structural Optimization* **16**, 162-175 (1998).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

RECOVERY OF CRACKS USING A POINT-SOURCE RECIPROCITY GAP FUNCTION

Carlos J. S. Alves
Departmento de Matemática
Instituto Superior Técnico
1049-001 Lisboa, Portugal
carlos.alves@math.ist.utl.pt

Jalel Ben Abdallah
Institut Supérieur des Sciences Appliquées et de Technologie de Sousse and *Ecole Nationale d'Ingénieurs de Tunis*
BP37, 1002 Tunis Belvédére, Tunisia
jalel.benabdallah@enit.rnu.tn

Mohamed Jaoua
Ecole Nationale d'Ingénieurs de Tunis, BP37,
1002 Tunis Belvédére, Tunisia
mohamed.jaoua@enit.rnu.tn

ABSTRACT

In this paper we are interested in the recovery of cracks from boundary measurements. We will be making use of a function that we will call *point-source reciprocity gap function*, which comes as a particular case of the reciprocity gap functional, applied to point sources. This function can be calculated in each point of the outer domain, and we will show that the analytic continuation of this function to the inner domain may provide a tool to the identification of the cracks inside, especially using functions that we will call *cracklets*.

INTRODUCTION

Identifying the location and shape of cracks inside a material is an inverse problem with major applications in industry, related to other non destructive inverse problems. We will state the problem as an inverse heat conduction problem in the steady-state case (electric conduction inverse problems, for instance, would obviously be treated in the same way). The main idea is to consider the reciprocity gap functional, introduced in [1], in the special case of point-sources. This allows the introduction of a function instead of a functional, the *point-source reciprocity gap function*, which can be used to retrieve the crack, and some numerical methods are suggested.

The inverse problem here addressed has been treated in [2], and more recently in [3], where conditions to the uniqueness of identification, for insulating cracks, is proven. In the case of conductive cracks, also in [4] the result was proved for any connected crack, imposing positive boundary Dirichlet data and measuring a single flux. This was proved using a maximum principle argument.

It is well known that not all fluxes are *identifying*, as has been stated in [2]. A sufficient condition for a flux to be identifying is that the singular part of the solution does not vanish (cf. [5]). In that case, it has been proved in [6] that the subset of the crack where the jump vanishes can be neglected.

An identifying flux is the starting point of any recovery task. It has been proved that a flux producing a singularity is a necessary condition for stability (e.g.. [5], [7]). In fact, fluxes not producing singularities are those orthogonal to a dual singular function, which are thus highly unlikely to meet.

We will slightly address the problem of identification and suggest numerical methods to retrieve the shape and location of a single crack. Numerical experiments are presented and show that the method using *cracklets* allows an a priori location of the crack and its main orientation, even if a significant amount of noise is added.

CRACK RECOVERY PROBLEM

Let W be an open bounded set in $\mathbf{R}^d$ (d = 2 or 3), with boundary Γ, and let s be a *crack* inside the body. By *crack* we will understand any piecewise C^1 curve (orientable surface, in $\mathbf{R}^3$). We are interested in recovering this unknown crack by means of boundary measurements, that is by setting some flux f on the boundary and measuring the resulting temperature f. It has been noticed in [1] that the presence of cracks generates a so called reciprocity gap, a suitable handling of which may give rise to fast recovery algorithms. However this has mostly been worked out so far in the case of 3D planar or 2D line-segment cracks. In such cases, the reciprocity gap

provides us with explicit formulae that localize the host plane or line, and this constitutes the starting point for the numerical part of the algorithm (e.g. [5], [6]). In the present paper, we are investigating an alternative use of it, extending its generality to other crack shapes.

Let us first recall some definitions. The steady state heat problem we are dealing with is the following:

$$(P)\quad \begin{cases} \Delta u_s = 0 & \text{in } W \setminus s \\ \partial_n u_s = f & \text{on } \Gamma \\ \partial_n u_s = 0 & \text{on } s \end{cases}$$

where f is a known flux prescribed on the boundary Γ, and we hold some measurements on the boundary, that is, we assume that:

$$u_s = f \quad \text{on} \quad \Gamma, \tag{1}$$

is known. Our goal is to retrieve the crack s from the pair (f, f).

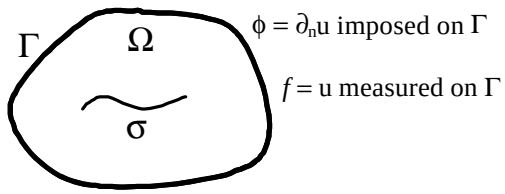


Figure 1. The unknown crack s inside W.

Definition 1. Let f be a flux and f_1, f_2 the two measurements produced by any two cracks, s_1 and s_2 (respectively). We say that the flux f is an *identifying flux*, if $f_1 = f_2 \Leftrightarrow s_1 = s_2$.

This definition means that no other crack (picked up in a suitable class of admissible cracks to be precised later on), than the actual one, may produce the same measurements on the boundary. Since we assume that a crack s is orientable, we can consider it as having two sides. On each one of these sides a trace of u is defined, and they will be called u^- and u^+. Using this notation, the jump of the solution on the crack s is $[u_s] = u^- - u^+$.

We can easily deduce the following result on the jump.

Lemma 1. Assume that f is an identifying flux, then $supp[u_s] = s$.

Proof. Suppose that $[u_s]$ vanishes in some open subset w of s, and let t be the crack s deprived of w. Since u_s is continuous across w, as well as its normal derivative, it is harmonic in $W \setminus t$ and hence solves:

$$\begin{cases} \Delta u_s = 0 & \text{in } W \setminus t \\ \partial_n u_s = f & \text{on } \Gamma \\ \partial_n u_s = 0 & \text{on } t \end{cases}$$

as well as $u_s = f$ on Γ. The flux f is therefore not an identifying one since the crack t is producing the same measurements on the boundary as the ones of s. ♦

Thanks to the above lemma, the set

$$s^o = s \setminus \partial s$$

(i.e. the crack without its boundary) can be characterized by $s^o = \{x \in s : |[u_s](x)| > 0\}$.

Now, given any function v, harmonic in W, the reciprocity gap states that the scalar

$$\int_G (f v - f \partial_n v)\, ds$$

is not vanishing as it would be if the domain was safe of cracks, and moreover its value can be related by a simple integration by parts to an integral on the crack itself, involving the jump of the solution u_s:

$$\int_G (f v - f \partial_n v)\, ds = \int_s [u_s] \partial_n v\, ds, \qquad \forall v \in \{w \in H^1(W): \Delta w = 0 \text{ in } W\}. \tag{2}$$

POINT-SOURCE RECIPROCITY GAP FUNCTION

Reciprocity gap algorithms are based on the choice of an appropriate set of harmonic functions v in order to derive relevant information on the crack from formula (2). We now remark that given any point $x \in W^C$, which is the open outer domain (without the boundary), the point source defined by the Green function G is harmonic in the inner domain W, and formula (2) applies.

Note that, in the 2D case,

$$G(x, y) = -\frac{1}{2\pi} \log|x - y|, \tag{3}$$

and in the 3D-case,

$$G(x, y) = \frac{1}{4\pi |x - y|}. \tag{4}$$

Definition 2. We introduce the *point-source reciprocity gap* as a function defined by

$$g_s(x) = \int_\Gamma \left(f(y)G(x,y) - f(y)\partial_{n_y}G(x,y) \right) ds_y \quad (5)$$

The above function is harmonic in $W \cup W^C$, since it is the sum of a single layer and a double layer potential, with densities on Γ. It should be pointed out that this part does not depend on the crack, but only on the given and measured data on the external boundary Γ. It is the restriction of the reciprocity gap functional, as introduced in [1], by considering only the Green functions placed at the points $x \in W^C$. Note that in this case the reciprocity gap is a function and not a functional. It is clear that this function g_s verifies

$$\Delta g_s = 0 \text{ in } \mathbf{R}^d \setminus \Gamma,$$

with the jumps $[g_s] = f$, $[\partial_n g_s] = f$ on Γ, and appropriate asymptotic conditions (depending on the dimension d) when $r = |x| \to \infty$.

Now, let v_s be the analytic function defined (for any $x \in \mathbf{R}^d \setminus s$) by

$$v_s(x) = \int_s [u_s](y)\partial_{n_y}G(x,y)\, ds_y . \quad (6)$$

This is the double layer representation for an exterior Laplace problem, generated by the crack s with the density of $[u_s]$.

Lemma 2. We have $g_s = v_s$ in W^C. Thus, the analytic extension of g_s from W^C to $W \setminus s$ must be v_s.

Proof. If $x \in W^C$ then $G(x, y)$ is harmonic for all $y \in W$, and from (5) it follows $g_s(x) = v_s(x)$. ♦

Theorem 1. *In* $\mathbf{R}^d$ *we have* $g_s + u_s c_W = v_s$.

Proof. From Lemma 2 the equality follows in W^C. Note also that

$$g_s(x) = \int_\Gamma \left(f(y)\, G(x,y) - f(y)\, \P_{n_y}G(x,y) \right) ds_y$$

$$= \int_{W-s} \left(\Delta u_s(y)\, G(x,y) - u_s(y)\, \Delta_y G(x,y) \right) dy +$$

$$+ \int_s [u_s](y) \P_{n_y}G(x,y)\, ds_y$$

$$= \int_{W-s} u_s(y)\delta(x,y)dy + \int_s [u_s](y)\P_{n_y}G(x,y)\, ds_y$$

Thus, if $x \in W \setminus s$, we have

$$g_s(x) = - u_s(x) + v_s(x).$$

Therefore, in Γ, $g_s^+ = v_s$ and $g_s^- = - u_s + v_s$, meaning that $[g_s]_\Gamma = g_s^+ - g_s^- = u_s = f$, as mentioned before. Also,

$$[\partial_n g_s]_\Gamma = (\partial_n g_s)^+ - (\partial_n g_s)^- = \partial_n u_s = f .$$

Finally, in s, we have $[g_s]_s = [v_s]_s - [u_s]_s = 0$. ♦

We used the notion of *analytic singular support* (e.g. [8]), $\text{sing}_A\text{supp}(g_s)$, as the complement of the largest open set in $\mathbf{R}^d$ where g_s is an analytic function. It is clear that

$$\text{sing}_A\text{supp}(v_s) \subseteq s \quad \textit{and that} \quad [v_s] = [u_s],$$

therefore $\text{sing}_A\text{supp}(v_s) = \text{supp}\,[u_s]$.

From Lemma 2, g_s^*, the unique analytic extension of g_s from W^C to s^C is v_s, and we conclude the following result.

Corollary 1. If we impose an identifying flux f then the crack s is perfectly determined by

$$s = \text{sing}_A\,\text{supp}(g_s^*) \quad (7)$$

Proof. Immediate, by Lemma 1, because

$$\text{sing}_A\text{supp}(g_s^*) = \text{sing}_A\text{supp}(v_s) = \text{supp}\,[u_s] = s . \blacklozenge$$

Remarks:

(i) Note that even if the flux is not identifying we can conclude that

$$\text{sing}_A\text{supp}(g_s^*) = \varsigma \subseteq s .$$

This means that on $s \setminus \varsigma$ we have null jump, $[u_s] = 0$, and therefore the function is analytic there. A way to overcome this problem is to impose that the jumps belong to the space

$$\mathbf{C}(s) = \{ q \in H_{00}^{1/2}(s) : q(x) \neq 0 \text{ a.e. on } s \} \quad (8)$$

Notice that is not a major restriction. Fluxes not producing singularities are those orthogonal to a dual singular function (cf. [9]), which are thus highly unlikely to meet. In addition, if missing, a component in the dual singular function will anyway arise from computational errors.

Therefore, if the chosen flux f only generates jumps $[u_s] \in \mathbf{C}(s)$, then the crack s is perfectly determined by $s = \text{sing}_A\text{supp}(g_s^*)$.

(ii) From Theorem 1, we also conclude that the solution u is the sum of a function $-g_s$, harmonic in W, with a function v_s. Thus, when doing the extension of g_s from W^C, the function obtained g_s^* coincides with u_s only if $u_s = v_s$. In the following, when we consider solutions u_s of the type v_s, then we are really recovering u_s, if the analytic extension of g_s is made. ♦

Example 1.
To show the effect of this procedure, we consider $W =]-1, 1[^2$ and

$$s = \{{}^1/_5\,(-4+7t,\ 4-7t):\ t \in [\,0,\ 1]\}$$

Given the pairs (f_1, f_1) and (f_2, f_2) generated by two different solutions,

$$u_1(x) = \int_s \partial_n G(x, y)\,ds_y,$$

and

$$u_2(x) = u_1(x) + (x_1^2 - x_2^2)/20$$

In Figure 2 we plotted the solution u_2 in $[-3,3]^2$. It becomes clear that the information that we will retrieve in the pair (f_2, f_2) will be added to the undesirable effect of the harmonic function

$$h(x_1, x_2) = (x_1^2 - x_2^2)/20\,.$$

This effect could be much worse, if we added a more significant harmonic function (in fact, the division by 20 was done just to keep the jump visible in the plot, on Figure 2). However, this effect will disappear if we calculate g_s. The reciprocity gap function will act like a *filter*, keeping only the relevant information, on the jump (that identifies the crack), and suppressing the harmonic part.

In fact one can see this as a consequence of Theorem 1, since we have a decomposition of the solution in terms of

$$u = v_s + h$$

where v_s is defined as in (6) and h is an harmonic function in W. We note that by Theorem 1, h is given by $-g_s$.. Thus, in numerical terms, the computation of the reciprocity gap function g_s appears to be important not only in the outer, but also in the inner domain, in order to suppress the harmonic part.

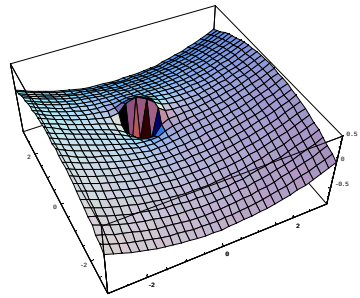


Figure 2. Plot of u_2, extended outside W.

In Figure 3 we plotted g_2 associated to u_2, and in Figure 4 we plotted g_1 associated to u_1. The only difference between the plots of g_1 and g_2 lies inside W, as predicted in Theorem 1. Since u_1 will be of the form v_s, as in (6) with $[u_s] = 1$, we will have a null g_1 in W. Again, we remark that this reduction to v_s situations will help the reconstruction, because a less clear jump on u_2 will became more evident on u_1.

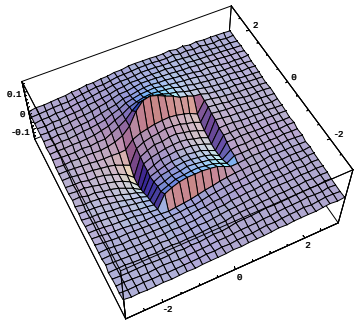


Figure 3. Plot of g_2, the reciprocity gap function associated to u_2.

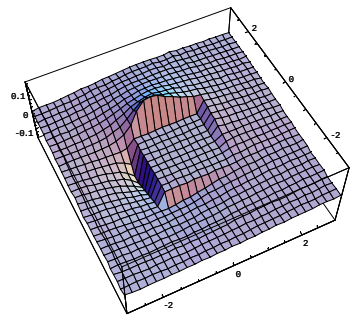


Figure 4. Plot of g_1, the reciprocity gap function associated to u_1.

FLAT CRACKS

In the case σ is a plane crack (or a line crack in 2D), i.e. $s \subset P$, where P is the plane of the crack, we can establish a criteria which is similar to a plane crack identification result derived in [10], in the context of wave scattering.

Theorem 2.

$$s \text{ is a plane crack in } P \Leftrightarrow g_{s\,(f)}(x) = 0, \quad \forall\, x \in P \setminus s\,, \forall\, f\,.$$

Proof.

($\Rightarrow$) For instance, in the 2D case, suppose ν is the normal to $P \supset s$, then

$$g_\sigma(x) = \int_s [u_s](y)\, \nu \cdot (x - y) / \left(2\pi |x - y|^2\right) ds_y$$

Therefore, if $x \in P \setminus s$, for all $y \in s \subset \Pi$, we get

$$\nu \cdot (x - y) = 0.$$

($\Leftarrow$) If s is not a plane crack, we can always take a flux f that produces a jump $[u_s]$ not orthogonal to $\partial_{n_y} G(x, \cdot)$ in $L^2(s)$. ♦

Remark: Likewise, if s is an almost plane crack, with $|\nu \cdot (x - y)/|x - y|| \le \varepsilon$, for all $y \in s$, $x \in P \setminus W$, we were able to derive the following estimate

$$\exists\, K_s > 0 : \quad |g_s(x)| \le K_s\, \varepsilon\,.$$

Thus, one expects $|g_s(x)|$ to be small when $x \in P \setminus s$. ♦

Example 2.

Consider a domain $W =]-1, 1[^2$ and a non flat crack defined by

$$s = \{ {}^1/_5\, (-4 + 6t + 3\cos(4t), 4 - 6t) : t \in [0, 1]\}.$$

We take measurements on $\Gamma = \partial W$ given by the traces and normal traces of

$$u(x) = \int_s \partial_{n_y} G(x, y)\, ds_y.$$

In Figure 5 we plotted the solution u and its extension to $[-3, 3]^2$. In that plot one clearly sees the jump of the solution field in the crack. Since s is an almost flat crack, and by the previous remark, one expects $|g_s|$ to be almost null along some line P that approaches the crack s. This fact can be seen in Figure 6, where the $|g_s|$ is plotted in $[-3, 3]^2$. Inside the domain $W =]-1, 1[^2$ the field g is null, as predicted in Theorem 1.

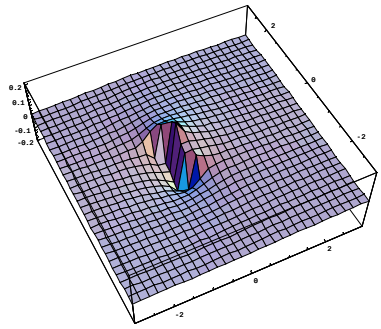


Figure 5. Plot of the solution u of Example 2 and the analytic extension outside W.

In Figure 7 we show the same results obtained in Figure 6, but now using a density plot, and also tracing a line that allows to put into evidence that the minimal values of $|g|$ are along a curve that crosses the (non planar) crack.

The inner white square corresponds to W and the crack is plotted inside this square. Although we will be taking squares for the domain W one should keep in mind that this not a restriction to any of the methods that we are presenting. They hold for any regular shape of W.

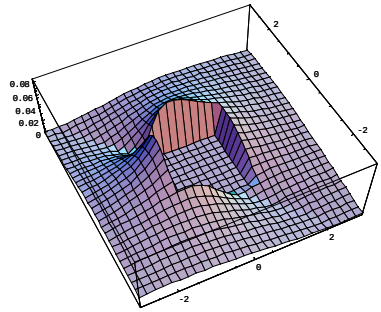


Figure 6. 3D-plot of $|g_s|$. Note that g is null inside W and almost null along a predicted line P .

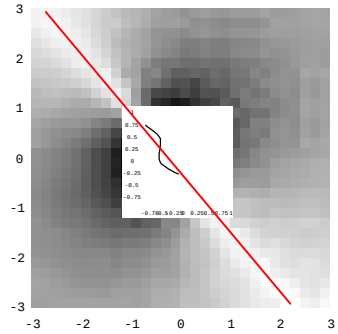


Figure 7. Density plot of $|g_s|$. The predicted P is plotted to point out that it crosses the crack s .

LOW SENSIBILITY TO NOISY DATA

The data given by the reciprocity gap function g_s smoothes the possible noise that arises in the measurement of f or even the noise in the input field f . In fact, suppose that the values with noise are $\boldsymbol{f} = f + \varepsilon_f$ and $\boldsymbol{f} = f + \varepsilon_f$, then

$$\boldsymbol{g_s}(x) = g_s(x) + \\ + \int_\Gamma \left(\varepsilon_f(y) G(x,y) - \varepsilon_f(y)\, \partial_{n_y} G(x,y)\right) ds_y.$$

Since we assume the noise to be random, we consider $\int_\Gamma \varepsilon_f$ and $\int_\Gamma \varepsilon_f$ to be almost null. Thus, if dist(x,Γ) is not too small, we can avoid the singularity of the integral, bounding $|G(x, y)|$ and $|\partial_{n_y} G(x, y)|$. Thus,

$$|\boldsymbol{g_s}(x) - g_s(x)| \le \max |G(x,y)| \left|\int_\Gamma \varepsilon_f(y)\, ds_y\right| \\ + \max |\partial_{n_y} G(x,y)| \left|\int_\Gamma \varepsilon_f(y)\, ds_y\right|$$

may be quite small. In the next example we present a case in which the result is not too perturbed even adding up 40% random noise.

Example 3.
Consider the same domain as before, and a planar crack defined by

$$\sigma = \{{}^1\!/_5\,(-4+7t,\ 4-7t): t \in [0, 1]\}$$

like in Example 1. We have added up to 40% noise in the measurements of f and also in the input data f, given by the solution

$$u(x) = \int_S \partial_{n_y} G(x,y)\, ds_y.$$

Since the solution is of the form v_s we notice that we are in a *favorable situation.* Any significant harmonic perturbation would lead to worst results, since the noise would be added to the proportions of the "non filtered data".

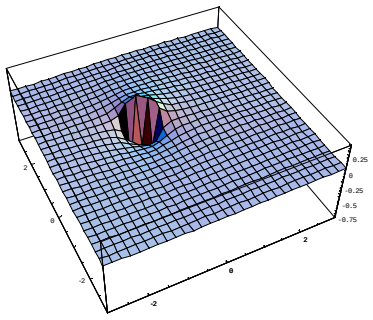


Figure 8. Plot of the solution u of Example 3.

In Figure 8 we plotted the solution in $[-3, 3]^2$, and in Figure 9 we plotted $|\boldsymbol{g_s}|$. One can see an almost null field on the direction of the line crack, which becomes more clear in the density plot shown in Figure 10.

We should also note that in Figure 9 the values of $|\boldsymbol{g_s}|$ are not too much perturbed by the high random noise. We only see small oscillations, and in the inner domain W the $|\boldsymbol{g_s}|$ is almost null (as it should be if there was no noise).

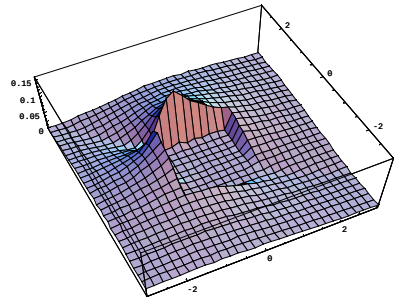


Figure 9. 3D-plot of $|\boldsymbol{g_s}|$ of Example 3.

In fact, if we were not in the *favorable situation*, of having picked up a solution of the form v_s we would get valuable information also. In that case, a 40% random noise could generate a huge perturbation on the data pair $(f,\ f)$, compromising the direct use of those values, but not the use of the reciprocity gap function $\boldsymbol{g_s}$. In fact, the values of $\boldsymbol{g_s}$ in the inner domain W would be a perturbation of the harmonic part, and this could be used to bring us again to the v_s case, by subtracting that harmonic contribution. Thus, we may keep ourselves in these *favorable situations.*

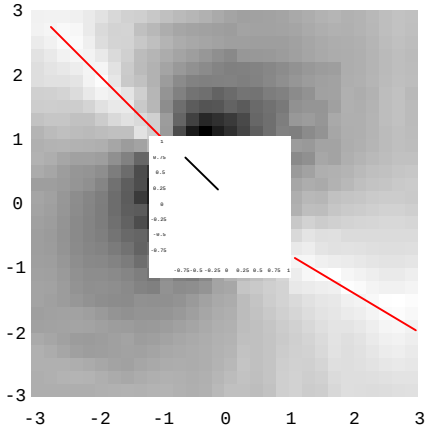


Figure 10. Density plot of $|\boldsymbol{g_s}|$ of Example 3.

In Figure 10 we plotted not one, but two lines, corresponding to the small values of $|\boldsymbol{g_s}|$, and the only significant difference appears on the line more distant to the crack (the extension of that line would not cross exactly the crack region). This is a natural effect of noise, because the crack is more distant to that region (the small values of the field will be more affected by the noise).

USING CRACKLETS

We now propose a different approach to this problem that consists in fitting the reciprocity gap function, in some points on the outer domain W^C, to a type of functions that we will call *cracklets.*

If one uses only piecewise constant densities, we get v_s approximated by

$$v_h(x) = \sum_{i,j=1}^{N} q_{ij} \int_{s_{ij}} \partial_{ny} G(x, y)\mathrm{d}y = \sum_{i,j=1}^{n} q_{ij}\xi_{ij}(x),$$

and those functions ξ_{ij} will be called *cracklets.*

For instance, in the 2D case, one can explicitly calculate the functions with an elementary crack $\tau = [(0, 0), (1, 0)]$.

In fact, given the *elementary cracklet* function

$$\Xi(x) = \int_0^1 \partial_{y2} G(x, y)\mathrm{d}y_1 = \arctan\left(\frac{1-x_1}{x_2}\right)+\arctan\left(\frac{x_1}{x_2}\right)$$

we can define any cracklet $\xi_{[a,b]}$ on the segment $[(a_1, a_2);(b_1, b_2)]$ by a rigid transform.

Minimization algorithm to find one single crack

Given an unknown crack s inside a domain W, we take a discrete set of data points given by the point-source reciprocity gap calculation,

$$D = \{(x_i, g_s(x_i)): x_i \in W^C\}.$$

If we assume that the crack is almost flat, then it makes sense to approach this data by nonlinear least squares using the cracklets $\xi_{[a,b]}$. The simplest approach in this context is to use functions of the form

$$\xi_c(x) = \arctan\left(\frac{c_0-c_1-c_2x_1-c_3x_2}{c_4+c_5x_1+c_6x_2}\right)+\arctan\left(\frac{c_1+c_2x_1+c_3x_2}{c_4+c_5x_1+c_6x_2}\right)$$

such that $c = (c_0, ..., c_6)$ minimizes

$$\boldsymbol{Q}s(c) = \sum_i |g_s(x_i) - \xi_c(x_i)|^2,$$

for an arbitrary number of points x_i in W^C.
The parameters $c_0, ..., c_6$ were given by a standard nonlinear minimization algorithm (we used the routine NonlinearFit of *Mathematica* from *Wolfram Research*).

Example 4.
We present our results for three different cracks placed inside $W =]-1, 1[^2$ (note again that there is no geometrical restriction on W, we only took a square for simplicity).

$$s_1 = \{(-3/10 + t/2 + 3t/8 \sin(t), -3/10 + t/4): t \in [0,1]\},$$

$$s_2 = \{(-1/2 + t/4 \sin(4t), -1/2 + t/2): t \in [0, 1]\},$$

$$s_3 = \{1/5(-4 + 6t + 3\cos(4t), 4 - 6t): t \in [0, 1]\}.$$

In Figure 11 we plotted the field ξ_c that minimizes $\boldsymbol{Q}s_1(c)$, using 75 random points plotted in the region $[-3, 3]^2 \backslash W$.

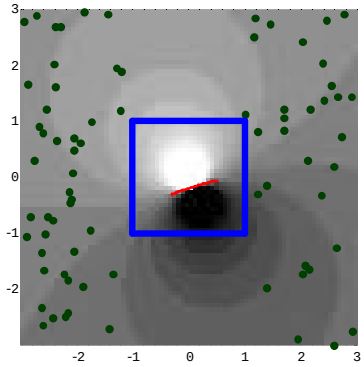


Figure 11. Density plot of the fitting for crack s_1.

The (bold) inner square defines the domain W and the line crack is plotted inside (gray line).
Although the reconstruction of v_s is very good, it is not completely accurate (we obtained a bigger jump instead of a larger crack). The crack is placed in the jump area (transition between black and white density, in the picture), and one may see that the orientation is the correct one.
In Figure 12, we plotted the same test for the crack s_2, noticing that this crack can hardly be considered as an "almost flat" one. Anyway, the density plots put the jump in a zone that it is basically the crack zone, and since we are approximating v_s with a single crack, these can be considered fairly good results.

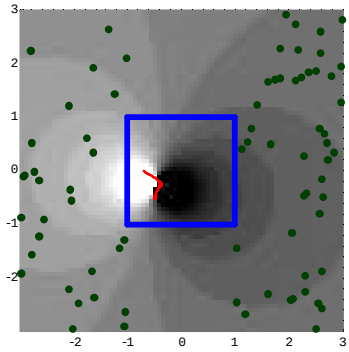


Figure 12. Density plot of the fitting for crack s_2.

Finally, in Figure 13, we plotted the same test for the crack s_3, used in Example 2, which is larger than the previous one, and it is clearly not flat. Moreover, we also took a 5% random noise. Again, we obtained good results, concerning the identification of the crack zone, and also of its main orientation. The predicted orientation is similar to the one that we obtained in Example 2 while taking the minimum of $|g_s|$ and a possible approximation by P. However, the reconstruction is not completely accurate since it predicts a bigger jump and a smaller crack.

This effect may be overcome in the future. One possibility is in the choice of a more adequate set of cracklets for the fitting. One other possibility is to impose more than one flux.

In fact, we notice that higher jumps will be placed in *heated* zones of the crack. Different fluxes will produce different zones of *heating* and one may combine the information in a reconstruction procedure.

One may see in these last three figures that the central zones of the cracks were preferred for the reconstruction. This may be related to the vanishing behavior of the jump in the extremities of the crack.

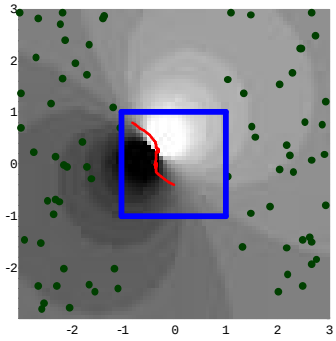


Figure 13. Density plot of the fitting for crack s_3.

CONCLUSIONS

i) The point-source reciprocity gap function allows to filter relevant information, not only by *vanishing noise* but because at the same time it filters the harmonic contribution that it is not relevant for crack detection.

ii) In the case of almost flat cracks, the line for which $|g_s|$ is minimal may allow the identification of the main direction of the crack.

iii) The use of a single cracklet allows the identification of the zone where the crack lies and its main direction. Extensions, using several cracklets may lead to the detection of several cracks. Moreover, the reconstruction was made using a single flux.

iv) These techniques are not limited to the geometry of the domain, and they can be extended to the 3D case. Applications to other partial differential equations are also possible, using their fundamental solution.

Acknowledgments

This work was partially supported by a project ICCTI(Portugal)/SERST(Tunisia), by SERST within the LAB-STI-02, and by project FCT-POCTI-34735/99 (Portugal).

REFERENCES

1. Andrieux S., Ben Abda A. *Inverse Problems* **12**, 553 (1996).
2. Friedman, A. and Vogelius, M. *Ind. Univ. Math. J.* **38**, 497 (1989).
3. Alessandrini G. and DiBenedetto *Ind. Univ. Math. J.* **46**, 1 (1997).
4. Alves C.J.S., Ha Duong T., Penzel F. *in Inverse Problems in Engineering Mechanics II* (Editors: M. Tanaka and G. Dulikravich), *Elsevier Publ.*, 213 (2000).
5. Bannour T., Ben Abda A., Jaoua M. *Inverse problems* **13**, 899 (1997).
6. Ben Abda A., Ben Ameur, H., Jaoua M. *Inverse problems* **15**, 67 (1999).
7. Andrieux S., Ben Abda A., Jaoua M. *Math. Meth. in the Appl. Sci.* **21**, 895 (1998).
8. Egorov Y. V. and Shubin M.A. *Foundations of the classical theory of PDEs*, Springer, Berlin. (1998)
9. Grisvard, P. *Elliptic problems in non smooth domains*, Pitman, Boston (1985).
10. Alves C.J.S., Ha Duong T. *Inverse Problems* **13**, 1161. (1997).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

3D INVERSE ANALYSIS MODEL USING SEMI-ANALYTICAL DIFFERENTIATION FOR MECHANICAL PARAMETER ESTIMATION

R. Forestier, Y. Chastel, E. Massoni
Centre de Mise en Forme des Matériaux, Ecole Nationale Supérieure des Mines de Paris, BP 207-F-06904 Sophia-Antipolis, e-mail : Romain.Forestier@cemef.cma.fr

ABSTRACT

An inverse method is developed in order to estimate constitutive parameters of a material from compression tests. The direct model used to simulate mechanical tests is FORGE3®. It solves a transient thermo-mechanical problem using a finite element method. From velocity, pression and temperature fields, any output of a mechanical test can be computed and compared with experimental data. A Gauss-Newton algorithm is implemented to solve the least-square problem associated with the inverse problem. The optimisation module is coupled with a semi-analytical sensitivity analysis method. This method is fast and stable when using a remeshing algorithm. A confidence interval estimator is proposed. The stability of the optimisation module and the confidence interval estimation are tested for numerical test cases. Finally, constitutive parameters of a steel grade are estimated for two elastic-viscoplastic constitutive laws.

KEYWORDS

Parameter estimation, semi–analytical derivatives, confidence intervals, mechanical tests.

NOMENCLATURE

D^c	computed data
D^e	measured data
E	elasticity tensor
$\boldsymbol{G}$	Gauss-Newton matrix
h	global heat transfer coefficient
k	conductivity
n	outward normal vector
NbMeas	number of measurements
NbPar	number of parameters
p	pressure field
R	discrete mechanical residual
T	temperature
T_{ext}	external temperature
T_i	interface temperature
T_p	sample temperature
v	velocity field
v_{die}	die velocity
X^n	coordinate vector of the mesh at time t_n
λ	thermo-mechanical parameter vector
ϕ	cost function
$\dot{\varepsilon}$	strain rate tensor
$\dot{\varepsilon}^{vp}$	viscoplastic strain rate tensor
$\dot{\varepsilon}^{e}$	elastic strain rate tensor
$\dot{\bar{\varepsilon}}$	generalised strain rate
$\bar{\varepsilon}$	generalised strain
σ	Cauchy stress tensor
σ_F	flow stress
τ	shear stress
ρc	heat capacity
$\Omega(t)$	domain defined by the sample at time t
$\partial\Omega_F$	free surface of the sample
$\partial\Omega_C$	contact surface between the sample and the dies

INTRODUCTION

Numerical simulation of metal forming processes requires thermo-mechanical data such as rheological, tribological or thermal parameters. Therefore, the identification of these parameters is crucial. This paper deals with a method able to determine the parameters of a thermo-mechanical model taking into account the evolution of the geometry, the forces and the temperature during a mechanical test involving large strain.

Classical thermo-mechanical parameter estimation techniques from laboratory tests are based on simple analytical models assuming that the material flow is homogeneous. When inhomogeneous material flow is involved in a mechanical test, these classical techniques cannot be used. The inverse analysis approach consists in coupling a direct model with an optimisation module allowing the simultaneous and automatic identification of the whole set of thermo-mechanical parameters. Optimisation module is generally based on zero order methods (genetic algorithms [1], simplex methods [2]) or gradient methods [3]. The computation of the cost function gradient can be done using an analytical method [4], a finite difference method or a semi-analytical method [5][6].

In this work an inverse analysis technique based on a finite element model is used for the identification of thermo-mechanical parameters from test measurements.

The 3D finite element model, FORGE3®, solves a strongly coupled thermo-mechanical equilibrium problem using an incremental approach. Since the discrete system is non–linear, it is solved using an

iterative procedure based on Newton-Raphson algorithm. State variables are updated using a Lagrangian formulation, and automatic remeshing algorithm is employed to avoid element degeneracy.
The inverse model is defined as the minimisation of an objective function representative of the difference between the experimental information (force, geometry, temperature) and the corresponding computed values, formulated in a least square sense. The optimisation procedure is based on a Gauss-Newton algorithm completed by an accurate computation of the sensitivity matrix. The differentiation is done using a semi-analytical method, which only requires linear problem resolutions. Confidence intervals are provided for each identified parameter. Validation of the proposed approach is first done using an artificial experimental data base. A comparison with the finite difference evaluation of the sensitivity matrix is also provided especially when remeshing is necessary, i.e. for large strain. In a second step, a set of actual experimental results is analysed in terms of constitutive laws varying from a simple one with low number of parameters to more sophisticated laws with large numbers of parameters.

DIRECT MODEL

The aim of the inverse method proposed in this paper is to estimate rheological parameters from a mechanical test. The direct model used to link measurements with the parameters values is Forge3®, a finite element based software devoted to the simulation of forming processes. A quasi-static thermo–mechanical problem has to be solved to simulate the mechanical tests.
Mechanical problem :

$$\begin{cases} div\ \sigma = 0 \quad on\ \Omega(t) \\ div\, v = 0 \quad on\ \Omega(t) \\ \left.\begin{cases} (v - v_{die}).n \le 0 \\ \sigma_n \le 0 \\ (v - v_{die}).n\sigma_n = 0 \end{cases}\right| \ on\ \partial\Omega_C(t) \\ \sigma n = 0 \quad on\ \partial\Omega_F(t) \end{cases} \tag{1}$$

with $\sigma_n = \sigma n.n,\ \partial\Omega_C(t) \cap \partial\Omega_F(t) = \phi$

Problem (1) has to be completed by a constitutive equation (2) and by a friction law (3).

$$\left.\begin{array}{l} (Prandtl\ Reuss\ relation) \\ \dot{\varepsilon} = \dot{\varepsilon}^e + \dot{\varepsilon}^{vp} \\ \dot{\sigma} = E\,\dot{\varepsilon} \\ \sigma_{FL} = \sigma_{FL}(\lambda, \bar{\varepsilon}, \dot{\bar{\varepsilon}}, T) \\ (plasticity\ criterium) \\ g(\lambda, \sigma, \bar{\varepsilon}, \dot{\bar{\varepsilon}}, T) = \bar{\sigma} - \sigma_{FL} \le 0 \end{array}\right\} \tag{2}$$

with $\tau = g(\lambda, v, v_{die}, T), \dot{\bar{\varepsilon}} = \left(\dot{\varepsilon}^{vp} : \dot{\varepsilon}^{vp}\right)^{1/2},\ \bar{\varepsilon} = \int_{(0,t)} \dot{\bar{\varepsilon}}\, dt$ (3)

The constitutive and the friction laws are written using a vector of mechanical parameters $\lambda = (\lambda_1, ..., \lambda_{NPar})^T$.
Thermal problem :

$$\begin{cases} \rho c \dfrac{dT}{dt} - \nabla.(k\nabla T) = \sigma : \dot{\varepsilon} \ \ on\ \Omega(t) \\ -k\nabla T.n = h(T_p - T_{ext}) \quad on\ \partial\Omega_F(t) \\ T = T_i \quad on\ \partial\Omega_F(t) \end{cases} \tag{4}$$

An explicit Euler scheme is used for time discretisation. Mechanical and thermal problem are coupled because the velocity and the temperature are unknown in problems (1) and (4). Then, global problem (1)+(4) is solved using a splitting method (i.e. (1) is solved and its solution is used to solve (4)). At each time increment, problem (1) solution is obtained from a P1+/P1 mixed finite element method in velocity and pressure [7] and problem (4) is solved using a P1 finite element method [8]. Then, the unknowns are the discrete velocity (V^n), pressure (P^n) and temperature (T^n) at time t_n. The thermo–mechanical discrete problem can be written as:

$$\left|\begin{array}{l} at\ time\ t_n \\ mechanical\ problem: \\ R(\lambda, V^n, P^n, \bar{\varepsilon}^n, T^{n-1}) = 0 \\ thermal\ problem: \\ C\dfrac{\Delta}{\Delta t} T^n + KT^n = Q(\lambda, V^n, \bar{\varepsilon}^n) \\ Explicit\ Euler\ scheme: \\ X^{n+1} = X^n + (t_{n+1} - t_n) V^n \\ \bar{\varepsilon}^{n+1} = \bar{\varepsilon}^n + (t_{n+1} - t_n)\dot{\bar{\varepsilon}}^n \end{array}\right. \tag{5}$$

The discrete mechanical problem is highly non-linear and solved with Newton-Raphson algorithm.

$$\left|\begin{array}{l} \dfrac{\partial R\left(\lambda, V^n, P^n, \bar{\varepsilon}^n\right)}{\partial\left(V^n, P^n\right)} \begin{pmatrix} \Delta V^n \\ \Delta P^n \end{pmatrix} = -R\left(\lambda, V^n, P^n, \bar{\varepsilon}^n\right) \\ \left(V^{n+1}, P^{n+1}\right) = \left(V^n, P^n\right) + \alpha_{ls}\left(\Delta V^n, \Delta P^n\right) \end{array}\right. \tag{6}$$

α_{ls} is a real obtained from a line search algorithm. Most forming processes simulated by Forge3® involve high strain and the elements can degenerate. To prevent the mesh from degenerating, an automatic remeshing algorithm is used [8]. Once $(V^n, P^n, T^n)_{n=1,Nincr}$ has been computed, the software provides the evolution of measurable data (load, torque, shape, etc…). The link between the computed data and the experimental data is done using the inverse model.

PARAMETER ESTIMATION TECHNIQUE

Parameter estimation methods aim at obtaining computed data which fit experimental data using a mathematical or a numerical model. The accuracy of the parameters depends strongly on the accuracy of the direct model results. The use of a realistic model is then very important. The parameter estimation problem is expressed as weighted least-square problem

$$\begin{cases} find\ \lambda^{opt} \in \Lambda\ such\ as \\ \phi(\lambda^{opt}) = \underset{\Lambda}{Min}\,\phi(\lambda) \\ with \\ \phi(\lambda) = \left(D^c(\lambda) - D^e\right)^T \boldsymbol{W} \left(D^c(\lambda) - D^e\right) \end{cases} \tag{7}$$

where D^c and D^e are the computed and the experimental data and $\boldsymbol{W}$ a weight matrix (symmetric positive-definite and generally chosen such as $\boldsymbol{W} = \frac{1}{\left\|D^e\right\|^2}\boldsymbol{I}$). This problem is solved using a Gauss–Newton algorithm

$$\begin{array}{l} Do\ while \left|\nabla\phi\right| > \varepsilon_{stop} \\ \left| \begin{array}{l} \boldsymbol{G}(\lambda_k)\delta\lambda_k = -\nabla\phi(\lambda_k) \\ \boldsymbol{G}(\lambda_k) = \dfrac{dD^c(\lambda_k)}{d\lambda}^T \boldsymbol{W} \dfrac{dD^c(\lambda_k)}{d\lambda} \\ \lambda_{k+1} = \lambda_k + \alpha\delta\lambda_k\ (line\ search) \end{array} \right. \end{array} \tag{8}$$

where $\boldsymbol{G}$ (Gauss-Newton matrix) is an approximation of the hessian matrix of ϕ. This method is of the first order and requires the computation of $\frac{dD^c(\lambda_k)}{d\lambda}$, .i.e. sensitivity matrix obtained from the sensitivity analysis.

SENSITIVITY ANALYSIS

Finite difference method

The finite difference method is the simplest approach to compute the sensitivity matrix. In this method, the direct model is simply considered as a 'black box' and the derivative is estimated from a first order approximation.

$$\frac{dD^c(\lambda_k)}{d\lambda} = \frac{D^c(\lambda_k + \Delta\lambda_k) - D^c(\lambda_k)}{\Delta\lambda_k} + 0\left(\left\|\Delta\lambda_k\right\|\right) \tag{9}$$

with $\Delta\lambda_k$ is a perturbation vector.

For *NbPar* parameters, it is necessary to perform *NbPar+1* simulations. Then, this method is slow because the direct model is highly non-linear (and then each computation of the direct model requires Newton-Raphson iterations). But this method is useful to validate more complex sensitivity analysis methods for simple test cases (without remeshing for example).

Semi-analytical method

Semi-analytical sensitivity analysis method is also based on a first order approximation of the sensitivity matrix

$$\begin{cases} at\ \ time\ t_n \\ \dfrac{dD^c}{d\lambda} = \dfrac{D^c(\lambda_k^p, V^{n,p}, P^{n,p}) - D^c(\lambda_k, V^n, P^n)}{\Delta\lambda_k} + O\left(\left\|\Delta\lambda_k\right\|\right) \\ \left(V^{n,p}, P^{n,p}\right) = \left(V^n, P^n\right) + \Delta\lambda_k \dfrac{d}{d\lambda}\left(V^n, P^n\right) \\ \lambda_k^p = \lambda_k + \Delta\lambda_k \end{cases} \tag{10}$$

$\frac{d}{d\lambda}\left(V^n, P^n\right)$ is obtained solving a tangent system (11) associated with (5).

$$\begin{array}{l} at\ time\ t_n \\ \left| \begin{array}{l} tangent\ mechanical\ problem: \\ \dfrac{\partial R(\lambda_k, V^n, P^n, \bar{\varepsilon}^n, T^{n-1})}{\partial(V^n, P^n)} \dfrac{d}{d\lambda}\begin{pmatrix} V^n \\ P^n \end{pmatrix} \\ \qquad = -\dfrac{\partial R(\lambda_k, V^n, P^n, \bar{\varepsilon}^n, T^n)}{\partial\boldsymbol{\lambda}} \\ tangent\ thermal\ problem: \\ C\dfrac{\Delta}{\Delta t}\dfrac{dT^n}{d\lambda} + K\dfrac{dT^n}{d\lambda} = \dfrac{dQ(\lambda_k, V^n, \bar{\varepsilon}^n)}{d\lambda} \\ tangent\ explicit\ Euler\ scheme: \\ \dfrac{dX^{n+1}}{d\lambda} = \dfrac{dX^n}{d\lambda} + \left(t_{n+1} - t_n\right)\dfrac{dV^n}{d\lambda} \\ \dfrac{d\bar{\varepsilon}^{n+1}}{d\lambda} = \dfrac{d\bar{\varepsilon}^n}{d\lambda} + \left(t_{n+1} - t_n\right)\dfrac{d\dot{\bar{\varepsilon}}^n}{d\lambda} \end{array} \right. \end{array} \tag{11}$$

Finally, $\frac{\partial R(\lambda_k, V^n, P^n, \bar{\varepsilon}^n, T^{n-1})}{\partial\lambda}$ and $\frac{dQ(\lambda_k, V^n, \bar{\varepsilon}^n)}{d\lambda}$ are computed using a first order scheme

$$\frac{\partial R(\lambda_k, V^n, P^n, \bar{\varepsilon}^n, T^{n-1})}{\partial\lambda} \approx \frac{R(\lambda_k^p, V^n, P^n, \bar{\varepsilon}^{n,p}, T^{n-1}) - R(\lambda_k, V^n, P^n, \bar{\varepsilon}^n, T^{n-1})}{\Delta\lambda} \tag{12}$$

and

$$\frac{dQ(\lambda_k, V^n, \bar{\varepsilon}^n)}{d\lambda} \approx \frac{Q(\lambda_k^p, V^{n,p}, \bar{\varepsilon}^{n,p}) - Q(\lambda_k, V^n, \bar{\varepsilon}^n)}{\Delta\lambda} \tag{13}$$

with

$$\bar{\varepsilon}^{n,p} = \bar{\varepsilon}^n + \Delta\lambda_k \frac{d\bar{\varepsilon}^n}{d\lambda_k} \tag{14}$$

The scheme gathering systems (10), (11), (12) (13) and (14) is called a semi–analytical method because it mixes finite difference schemes with analytical derivatives. Even though problem (5) is highly non–linear, tangent problem (11) is always linear. Moreover, all the terms used in the semi–analytical scheme are computed by the direct model : no formal derivative has to be calculated. Then the sensitivity analysis module upgrades with the direct model and is compatible with all the constitutive laws used by FORGE3®. An other advantage of this formulation is its stability when remeshing is needed. Some examples show that a standard finite difference scheme is unstable with remeshing [6]. Moreover, if the direct model is parallelized, no additional effort is required to obtain a parallel sensitivity analysis.

CONFIDENCE INTERVAL ESTIMATION

The use of a first order inverse method for parameter estimation is interesting because it provides more accurate results than an analytical model. But an other interesting feature of the inverse approach is the sensitivity matrix computation, which gives an idea of the quality and efficiency of the inverse model. For example, if we assume that the difference between the optimal computed data and the experimental data is due to additive un-correlated Gaussian perturbations with zero means and constant deviations [11], it is then possible to give a confidence interval of the parameters.

$$\begin{aligned} &Prob\left(\lambda_i \in I\right) \geq 0.95 \\ &I = \left(\lambda_i^{opt} - 1.98\sqrt{V_{ii}}\,;\lambda_i^{opt} + 1.98\sqrt{V_{ii}}\,;\right) \\ &V = \frac{\left[G(\lambda^{\boldsymbol{opt}})^{-1}\right]\ \phi(\lambda^{\boldsymbol{opt}})}{NbMeas - NbPar} \\ &\mathrm{cov}(\lambda_i, \lambda_j) = V_{ij} = V_{ji} \end{aligned} \qquad (15)$$

Then, if we norm $\boldsymbol{G}^{-1}$, we obtain an estimation of the respective correlations between the parameters. Correlation is the main difficulty of parameter estimation problems [10]. An important correlation between few parameters can make the result of the estimation meaningless. A solution to this problem can be to use an optimal experiment procedure [10] [11] or to change of constitutive law.
An other indicator of the quality of the estimation is given by Gauss-Newton Matrix $\boldsymbol{G}$:

$$G_{ij} = \left\langle \frac{dD}{d\lambda_i}, \frac{dD}{d\lambda_j} \right\rangle = \sum_{k=1}^{k=NbMeas} \frac{dD_k}{d\lambda_i} \frac{dD_k}{d\lambda_j} \qquad (16)$$

where the brackets represent the dot product of IR^{NbMeas}. Then the cosine between two derivatives is expressed as follows :

$$C_{ij} = \cos\left(\frac{dD}{d\lambda_i}, \frac{dD}{d\lambda_j}\right) = \frac{G_{ij}}{\sqrt{G_{ii} G_{jj}}} \qquad (17)$$

If the cosinus between two derivatives is close to 1 or –1, then $\boldsymbol{G}$ is ill-conditioned and confidence intervals become large. This means that the influences of both parameters on the data are similar. If $\cos\left(\frac{dD}{d\lambda_i}, \frac{dD}{d\lambda_j}\right) = \pm 1$, then :

$$\exists a \in IR \,/\, \frac{dD_k}{d\lambda_i} = a \frac{dD_k}{d\lambda_j} \ \forall k \in [1, NbMeas]_{IN}$$

and G is not definite (only symmetric positive).

VALIDATION OF THE OPTIMISATION MODULE

Stability of the optimisation module

The inverse module is used to obtain constitutive parameters from four uniaxial compression tests done on cylindrical samples (Figure 1).

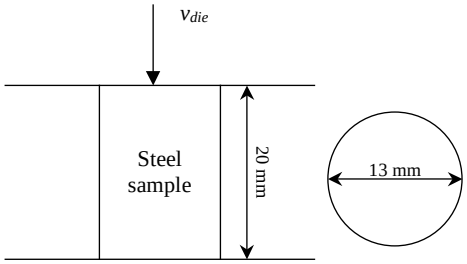


Figure 1. Geometry of the sample

An example of a simulation result for a compression test is given in Figure 2.

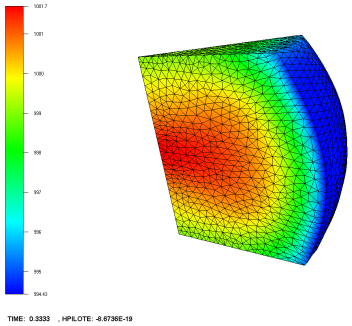


Figure 2. Simulation result for the uniaxial compression of a cylinder using FORGE3® (cylinder section)

The validation of the inverse model is done using an artificial experiment which is generated by the direct model.
The average strain rate , given by equation (18), is used for piloting the dies ($\dot{\bar{\varepsilon}}_{av}$ is constant during the tests)

$$\dot{\bar{\varepsilon}}_{av} = \frac{v_{die}}{h_{die}} \qquad (18)$$

The experimental plan is given in Table 1.

Table 1. Experimental plan

	0.02s^{-1}	**10.s^{-1}**
1000°C	+	+
1200°C	+	+

The first constitutive model used for the estimation is the Norton-Hoff law.

$$\sigma_{FL} = 2K \exp\left(\beta / T\right)(\bar{\varepsilon})^n \left(\dot{\bar{\varepsilon}}\right)^m \qquad \lambda = (K, m, n, \beta) \tag{19}$$

Then, the data used for the estimation is the load (denoted L). The values of the parameters used to generate the experiment is called nominal values. The optimisation module is initiated with a random set of parameters. The nominal values are K=860KPa.s^{-m}, m=0.2, n=0.2, β =6250K^{-1}. The aim of the test is to find back the nominal values using the inverse method. The values estimated by the optimisation module are compared with the nominal values in Table 2 for different initial parameter values.

Table 2. Results of the identification

	$\lambda_1^{ini} \Rightarrow \lambda_1^{opt}$	$\lambda_2^{ini} \Rightarrow \lambda_2^{opt}$	$\lambda_3^{ini} \Rightarrow \lambda_3^{opt}$
K	500=>860.12	600=>861.39	3000=>865.7
m	0.05=>0.2000	0.5=>0.2000	0.05=>0.2000
n	0.4=>0.1998	0.5=>0.1999	0.05=>0.1997
β	5000=>6249.3	8000=>6247.6	3000=>6241.3
ϕ_{final}	3.10^{-4}	3.10^{-4}	4.10^{-4}
Iterations	8	7	6

These results indicate that the optimisation module is stable and that the result of the estimation of Norton-Hoff coefficients does not depend on the initial values. The sensitivity analysis used to obtain these results is the semi-analytical method. No convergence was obtained with the finite difference method. Figure 3 represents the evolution of the derivative of the load with respect to K versus time increment obtained with the semi-analytical method and the finite difference method.

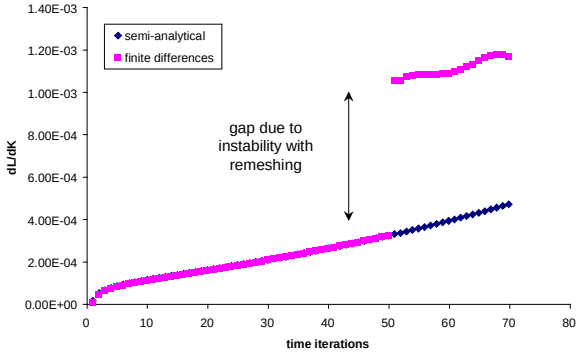


Figure 3. Evolution of the derivative of the load with respect to K

Before remeshing, the difference between the results obtained with both methods is negligible. But when remeshing is performed, the derivative obtained with the finite difference method shows a gap. This example indicates that the semi-analytical method is compatible with the use of a remeshing algorithm, contrary to the finite difference method. The semi-analytical method is then a fast (because it solves a linear problem) and stable method. Without remeshing, the accuracy of these methods are comparable.

Accuracy of the estimated values

The sensitivity analysis is necessary to use a gradient based method. But it gives also an important information about the accuracy of the result obtained at the end of the estimation. For example, if the data has a low sensitivity to a parameter, the estimation will not be accurate. Some problems can also be caused by an important correlation between parameters. At the end of the estimation, the cosine between the sensitivity vectors is computed :

$$\cos\left(\frac{dL}{dK}, \frac{dL}{dm}\right) = 0.42 \qquad \cos\left(\frac{dL}{dK}, \frac{dL}{dn}\right) = -0.84$$

$$\cos\left(\frac{dL}{dK}, \frac{dL}{d\beta}\right) = 0.998 \qquad \cos\left(\frac{dL}{dm}, \frac{dL}{dn}\right) = -0.39$$

$$\cos\left(\frac{dL}{dm}, \frac{dL}{d\beta}\right) = 0.39 \qquad \cos\left(\frac{dL}{dn}, \frac{dL}{d\beta}\right) = -0.83$$

where L is the load. $\frac{dL}{dK}$ and $\frac{dL}{d\beta}$ are almost colinear. This means that the optimisation module hardly decouples the effects of K and β on the data. The cost function with respect to K and β around the optimum shows a 'valley' (Figure 4).

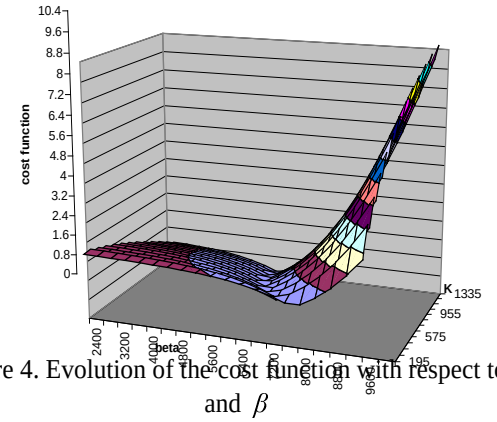


Figure 4. Evolution of the cost function with respect to K and β

The shape of the cost function makes the simultaneous estimation of K and β an ill–posed problem. It is linked to the colinearity of $\frac{dL}{dK}$ and $\frac{dL}{d\beta}$. Figure 5 represents the cost function versus K and m. There is no 'valley' and the optimum is easy to localise. The shape of the cost function near the optimum is linked to Gauss-Newton matrix (which estimates the hessian of the cost function).

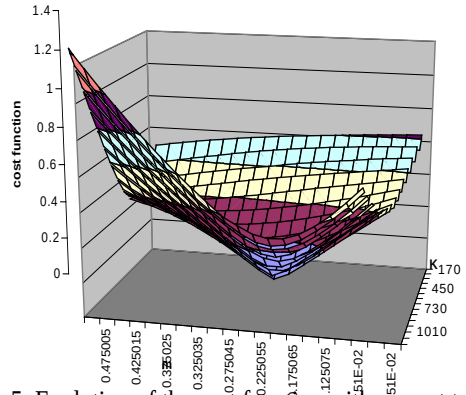


Figure 5. Evolution of the cost function with respect to K and m

Confidence interval validation

The artificial experiment is perturbed with independent gaussian noise with zero mean and a constant deviation (denoted $\{e_i\}$):

$$L_i^{ex} = L_i^{cal}(\lambda) + e_i, \quad i \in [1, NbMeas]_{IN} \qquad (20)$$

Hence, in this test, estimation (15) can be applied. We introduce a normalised deviation (21) for the noise :

$$\sigma_e^2 = \bar{\sigma}_e^2 \frac{\sum_{i=1}^{i=NbMeas} \left| D_i^{\exp} \right|^2}{N} \qquad (21)$$

where σ_e is the deviation and $\bar{\sigma}_e$ is the normalised deviation. The test is done for $\bar{\sigma}_e$=2.3%. The estimation of the constitutive parameters is given in Table 3.

Table 3. Result of the estimation

K	m	n	β
885.26	0.2018	0.2002	6214.47

Nominal values remain within confidence intervals Table 4. It can be noticed that confidence intervals for K and β are relatively wide, contrary the ones for m and n. This is due to the important correlation between K and β (c.f. matrix $\boldsymbol{C}$ (22)).

Table 4. Confidence intervals

Confidence Intervals
$K \in (851.1, 919.4)$
$m \in (0.199, 0.201)$
$n \in (0.198,\ 0.205)$
$\beta \in (6166.9, 6262.1)$

$$C = \begin{bmatrix} 1 & -0.55 & 0.022 & -0.995 \\ sym & 1 & 0.096 & 0.059 \\ sym & sym & 1 & 0.53 \\ sym & sym & sym & 1 \end{bmatrix} \qquad (22)$$

where $C_{ij} = cor(\lambda_i, \lambda_j) = \frac{V_{ij}}{\sqrt{V_{ii}}\sqrt{V_{jj}}}$

for $\lambda_1 = K, \lambda_2 = m, \lambda_3 = n, \lambda_4 = \beta$

So, we can conclude that a low angle between two sensitivity vectors (for example $\frac{dL}{dK}$ and $\frac{dL}{d\beta}$) can cause an important correlation between two parameters (for example K and β) and then wide confidence intervals. The simultaneous estimation of K and β seems to be a difficult problem. On the other hand, it is possible to fit accurately the experimental data with the direct model.

ESTIMATION OF CONSTITUTIVE PARAMETERS OF A STEEL

Compression tests are analysed to estimated constitutive parameters of a steel. Figure 1 shows the geometry of the samples. The experimental plan is given in Table 5.

Table 5. Experimental plan

	0.02s^{-1}	**0.5s^{-1}**	**5s^{-1}**
950°C		+	+
1050°C	+	+	+
1150°C	+		+

These tests can be analytically analysed assuming that there is no friction between the dies and the samples and that the test is adiabatic or isothermal. The inverse module allows to study the influence of friction or temperature on the measurements. As we plan to estimate constitutive coefficients, the load is measured during the uniaxial compression tests.
The first constitutive model used for the estimation is the Norton-Hoff law.

$$\sigma_{FL} = 2K \exp\left(\beta/T\right)(\bar{\varepsilon})^n \left(\dot{\bar{\varepsilon}}\right)^m \qquad \lambda = (K, m, n, \beta) \tag{23}$$

The result of the estimation is given in Table 6.

Table 6. Results of the parameter estimation

K	m	n	β
575.325	0.10756	0.650770	6842.49

The final cost function value is 7.2 10^{-2}. Figure 6 shows that the difference between measured and computed data at the end of the estimation is still important (the seven curves correspond to the seven experiments defined in Table 5). This is due to the fact that the model error is not negligible. Then the confidence interval model (15) cannot be applied.

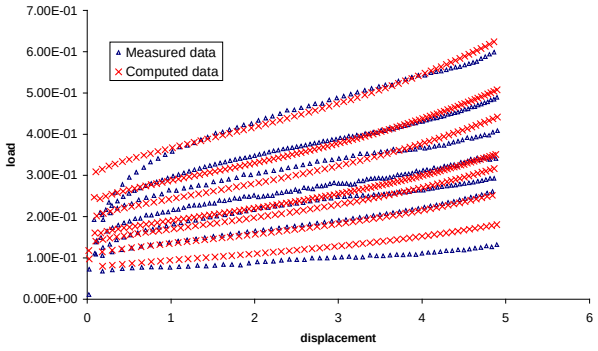


Figure 6. Computed and experimental load

Even though model (15) seems inadequate, confidence intervals give an information about the accuracy of the estimation. If a confidence interval is important, the estimation is probably meaningless. Confidence intervals are used here as a warning. If the parameters are too much correlated one with an other, the confidence intervals will be important.

Table 7. Confidence intervals

$K \in (494.9, 655.7)$
$m \in (0.103, 0.112)$
$n \in (0.055,\ 0.075)$
$\beta \in (6660.9, 7024.1)$

Confidence intervals on K and β are large because of the correlation coefficient between K and β (-0.99). This means that it is difficult to differentiate their effects. In this example :

$$\cos\left(\frac{dD}{dK}, \frac{dD}{d\beta}\right) = 0.9989 .$$

The important model error implies that it is necessary to use a most complex model to be able to represent realistically the behaviour of the material.

$$\begin{cases} \sigma_{FL} = K\dot{\bar{\varepsilon}}^n \\ K = (1-W)K_{ecr} + WK_{sat} \\ K_{ecr} = K_0(\bar{\varepsilon} + \bar{\varepsilon}_0)^n \exp\left(\beta/T\right) \\ K_{sat} = K_{st} \exp\left(\beta_{st}/T\right) \\ W = 1 - \exp(-r\bar{\varepsilon}) \\ m = m_0 + m_1 T \end{cases} \qquad \lambda = (K_0, m_0, m_1, n, \beta, r, K_{st}, \beta_{st}) \tag{24}$$

The use of the semi-analytical sensitivity analysis makes the change of constitutive law easier because it is not necessary to calculate new derivatives. Eight parameters have to be simultaneously estimated. The final cost function value is 4.8.10^{-2}. Using model represented by equation (24) improves the result of the estimation and it is possible to see in Figure 7 that computed data fit better measured data than in Figure 6.

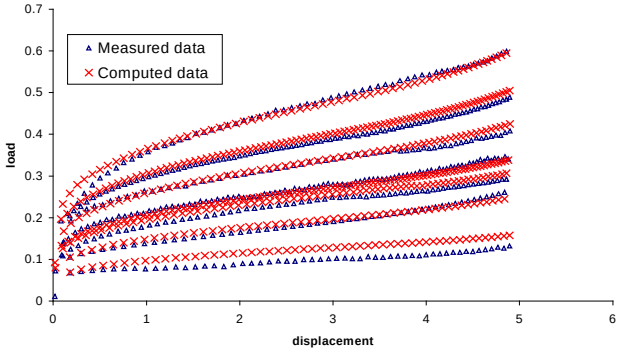


Figure 7. Computed and experimental load

This example shows that the choice of an adequate model is important in mechanical parameter estimation. The visualisation of the residue (Figure 8) indicates that the difference between experimental and computed data is not due to a Gaussian noise but only to model error. Dawning and Blackwell [12] assume that the model error is negligible if there are almost N/2 changes in sign of the residual (N is the number of measurements).

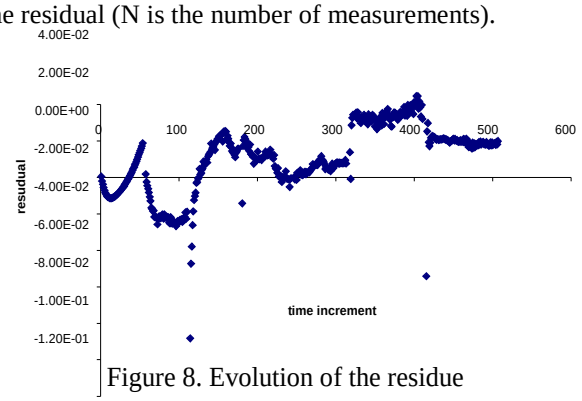


Figure 8. Evolution of the residue

Moreover, confidence intervals for K, β, r, K_{st}, β_{st} are wide (Table 8). This means that it is difficult to decouple their influences on the load. For example : $\cos(\frac{dD}{dK}, \frac{dD}{d\beta}) = 0.9988$ and $\cos(\frac{dD}{dr}, \frac{dD}{d\beta}) = 0.9959$.

Table 8. Confidence intervals

$K \in (1628.5, 273.9)$	$\beta \in (6109.3, 6604.2)$
$m_0 \in (-0.284, -0.1718)$	$r \in (-1.576, -1.356)$
$m_1 \in (2.1.10^{-4}, 2.9.10^{-4})$	$K_{st} \in (2161.9, 3827.9)$
$n \in (0.29, 0.38)$	$\beta_{st} \in (5850.2, 6467.9)$

We can conclude that two indicators are important to estimate the accuracy of a parameter estimation : the residue gives information about the quality of the model and the Gauss-Newton matrix (cosine) gives information about the correlation between the parameters. Confidence intervals can be used as a warning, even if model represented by equation (15) cannot be applied. This example shows that the choice of a realistic model is important to ensure a low value of the residue at the end of the estimation. But the use of too complex a model seems makes the correlation between the parameters impossible since the problem is then ill-posed.

CONCLUSION

An inverse method coupled with a semi-analytical sensitivity module is presented and validated in this paper. It is shown that the semi-analytical method is stable and relatively fast (because it is a linear problem). Moreover, it makes easy the change of constitutive law because it only uses terms calculated for the direct model. It has been shown also that an important correlation between two sensitivity vectors can make the estimation ill-posed : cost function shows 'valleys' and confidence intervals at the end of the estimation can be wide. On the other hand, the inverse method makes possible an efficient fitting of experimental data by computed data.

AKNOWLEDGEMENT

The work has been carried out in the framework of the GROWTH European project TESTIFY, project number GRD1-1999-10714.

REFERENCES

1. Ghouati O., Gelin J.C., Gradint based methods, genetic algorithms and the finite element method for the identification of material parameters, Simulation of Material Processing: Theroy, Methods and Applications, Huétink&Baaijens, pp.157-162, 1998.
2. M. Pietrzyk, Identification of parameters in the history dependent constitutive model of steels, Annals of the CIRP, Vol 50/1/2001, pp.161-164.
3. T.G. Faurholdt, Inverse modelling of constitutive parameters for elastoplastic problems, Journal of strain analysis, 2000,Vol 35, no 6 p471.
4. A. Gavrus, E. Massoni, J.L. Chenot, The rheological parameter identification formulated as an inverse finite element problem, Inverse Problem in Engineering, 1999, Vol 7, pp1-41.
5. J.E. Scheuing, D.A. Tortorelli, Inverse heat conduction problem solutions via second-order design sensitivities and Newton's method, Inverse Problem in Engineering, 1996,Vol 2, pp.227-262.
6. E. Massoni, B. Boyer, R. Forestier, 2001, Inverse Analysis of thermomechanical upsetting tests using gradient method with semi-analytical derivatives, Accepted in the Int. Journal of Thermal Sciences.
7. M. Fortin, D.N. Arnold, F. Brezzi, A stable finite element for the Stokes equations. Calcolo (21) :337 344, 1984.
8. C. Aliaga, E. Massoni, J.L. Treuil, 3D numerical simulation of THEVP behavior using stabilized mixed F.E. formulation: application to 3D heat treatment, Proceedings of the Fourth World Congress on Computational Mechanics, Buenos Aires, Argentina, 29th of June, 2nd of July 1998.
9. T. Coupez,, 1991, Grandes déformations incompressibles-remaillage automatique. PhD thesis, ENSMP, France.
10. J. V. Beck and K. A. Woodbury, Inverse problems and parameter estimation: integration of measurement and analysis. Measurement Sciences and Technologies. Vol 9, pp 839-847, 1998.
11. A. F. Emery and V. N. Aleksey, Optimal experiment design, Measurement Science and Technology. 9 (1998) pp 864-876
12. K. J. Dawning and B. F. Blackwell, Joint experimental/computational techniques to measure thermal properties of solids, Measurement Sciences and Technologies. Vol 9, pp 877-887, 1998.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

GENETIC ALGORITHMS FOR IDENTIFICATION OF ELASTIC CONSTANTS OF COMPOSITE MATERIALS

Mariana Ferreira Teixeira Silva
Lavinia Maria Sanabio Alves Borges
Fernando Alves Rochinha
Department of Mechanical Engineering, EE/COPPE/PEM
Federal University of Rio de Janeiro, UFRJ
Rio de Janeiro, RJ, Brazil
mari@mecsol.ufrj.br
lavinia@serv.com.ufrj.br
faro@serv.com.ufrj.br

Luís Alfredo Vidal de Carvalho
Department of Mechanical Engineering, EE/COPPE/PESC
Federal University of Rio de Janeiro, UFRJ
Rio de Janeiro, RJ, Brazil
okay@iamwaiting.com

ABSTRACT

The aim of this work is to present a technique to identify elastic parameters of composite materials. The identification is based on the adjustment of coefficients in an optimization process in which the objective function is defined by the difference between the analytical natural frequencies and the measured ones. Such analytical natural frequencies are obtained by the finite element method while the experimental ones are determined by ordinary modal tests. The proposed technique is assessed by a number of different tests allows simultaneous identification of several global properties from a single test without damaging the structure. The proposed approach uses genetic algorithm to solve the optimization problem. Since genetic algorithms are not based on the gradient method, they do not require the expensive eigenvectors computations presented in gradient method.

NOMENCLATURE

C_{ijkl} - constitutive matrix
CF - cost function
Dm - extensional stiffness matrix
Df - flexure stiffness matrix
Dmf - coupling stiffness matrix
E_i - elasticity modulus
f_{exp} - experimental frequencies
f_θ- calculated frequencies
G_{ij} - shear modulus
keep - overlapped chromosomes
maxgen - maximum generation number
N - number of layers
nbest - selected individuals for reproduction
pcross - probability of crossover
pmutation - probability of mutation
popsize - number of individuals in a population
Q_{ij} - elastic coefficients for plane stress
$\vec{u}_0$ - displacement in the midplane
$\vec{v}_0$ - normal displacement
ν - Poisson's ratio
ρ - density
θ - elastic constants vector

INTRODUCTION

Recently, composite materials have been used in many structural applications. They are formed by two or more different materials in order to obtain better engineering properties like stiffness, strength, weight reduction and thermal response [1]. In the design of structures, it is of extreme importance to have very precise estimate of the elastic constants which conventional techniques are not fully able to do. In [2] the elastic constants were estimated using simulated ultrasonic phase velocities. Genetic algorithms are used in [3] as a complementary technique to perform the initial estimation of the elastic parameters and then refining the solution by classical updating methods. In the present work, the identification of elastic parameters constitutes an inverse problem, which leads to an optimization formulation. This problem relies on the comparison between experimental natural frequencies and their analytical counterparts. The experimental quantities are obtained by means of modal tests performed on typical composite structures like, for instance, plates. The analytical frequencies are

the eigenfrequencies associated to finite element models of those structures. The novelty of the proposed approach is the use of a genetic algorithm in the numerical solution of the optimization problem. It is worthwhile to remember that the use of this kind of algorithm avoids the expensive computation of gradients involving eigenfrequency values. Moreover, two other advantages are also obtained: no initial guess is required and the optimization process could be more flexible, due to the fact that the search space begin from a set of elastic constants, corresponding to different chromosomes, rather than a single one.

The experimental natural frequencies were obtained by tests performed on aluminium plate [4] and on Glass/Epoxy plate [5]. In order to stress the capacity of the approach simulated experimental frequencies for aluminium, kevlar/epoxy and SCS-6/Ti-15-3 which were obtained by numerical methods were used. The stiffness properties of these materials were obtained from literature [6]. In this manner, the estimated properties from GA were compared with the available data.

CLASSICAL PLATE THEORY

The kinematics of the classical plate theory [7] adopted here is based on the following hypothesis: the vertical displacement v_z of any point of the plate is the same of its projection on the middle surface (Eq. 1) and the straight lines normal to the *xy*-plane before deformation remain straight and normal to the midsurface after deformation – Kirchhoff theory (Eq. 2). These assumptions correspond to neglecting both transverse shear and normal effects, i.e., the deformation is due entirely to bending and in-plane stretching.

$$\frac{\partial v_z}{\partial z} = 0 \quad \rightarrow \quad v_z(x,y,x) = v_z(x,y) \tag{1}$$

$$\frac{\partial \vec{v}_0}{\partial z} + \nabla v_z = 0 \rightarrow \vec{v}_0 = \vec{u}_0 - z\,\nabla v_z \tag{2}$$

where ∇ stands for the spatial gradient.

Thus, the kinematics of the plate's motion relies only on two degrees of freedom defined over the middle surface.

Therefore, the dynamics of the plate is governed by the virtual power principal [7], which lends to the following variational equation:

$$\begin{aligned}
&\int_{\Sigma_0} [D_m \nabla^s \vec{u}_o - D_{mf} \nabla^s(\nabla v_z)] \cdot \nabla^s \vec{u}_0^*\, d\,\Sigma_0 - \\
&\int_{\Sigma_0} [D_{mf} \nabla^s \vec{u}_o - D_f \nabla^s(\nabla v_z)] \cdot \nabla^s(\nabla v_z^*)\, d\,\Sigma_0 \\
&= \int_{\Sigma_0} [I_0\, \ddot{\vec{u}}_o - I_1 \nabla \ddot{v}_z] \cdot \vec{u}_0^*\, d\,\Sigma_0 + \\
&\int_{\Sigma_0} [I_0\, \ddot{v}_z + I_1\, div(\ddot{\vec{u}}_0) - I_2 \nabla^2 \ddot{v}_z]\, v_z^*\, d\,\Sigma_0
\end{aligned} \tag{3}$$

$$[I_0, I_1, I_2] = \sum_{k=1}^{N} \int_{z_k}^{z_{k+1}} r^{(k)} [1, z, z^2] dz \tag{4}$$

$$[D_m, D_{mf}, D_f] = \sum_{k=1}^{N} \int_{z_k}^{z_{k+1}} Q^{(k)} [1, z, z^2] dz \tag{5}$$

$$\begin{aligned}
Q_{11} &= \frac{E_1}{1 - n_{12} n_{21}} \quad Q_{22} = \frac{E_2}{1 - n_{12} n_{21}} \\
Q_{12} &= \frac{n_{12}\, E_2}{1 - n_{12} n_{21}} \quad Q_{66} = 2\, G_{12}
\end{aligned} \tag{6}$$

The composite materials analyzed in this work consist in fibers, which are the principal load-carrying members, in a matrix material, that keeps the fibers together. In each lamina, the fibers are perfectly aligned.

In Eq. 6, E_1 is the elastic modulus in the fibrous direction, E_2 is the elastic modulus in the transverse direction, ν_{12} and ν_{21} are the Poisson ratio and G_{12} is the shear modulus.

Only four out of the five material constant for plane stress of an orthotropic material are independent. In this work, the identified parameters are E_1, E_2, G_{12} and ν_{12}. Therefore, the Poisson ratio ν_{21} is obtained by:

$$n_{21} = \frac{n_{12}\, E_2}{E_1} \tag{7}$$

THE FINITE ELEMENT MODEL

In the finite element model, the plate is discretized by triangular elements with three degrees of freedom per node: two rotations and the transversal displacement. Thus, the associate eigenvalue problem is represented by:

$$(-\nu_i^2 M + K)\, f_i = 0 \tag{8}$$

where ω is the i^{th} natural frequency and f_i is its vibration mode. **M** and **K** are, respectively, the inertia and stiffness matrix of the finite element model.

It is important to remark that the stiffness matrix depends on the elastic parameters that are to be identified, i.e.

$$K = K(E_1, E_2, G_{12}, n_{12}) \tag{9}$$

THE GENETIC ALGORITHM

Genetic algorithms are search algorithms based on the mechanics of natural selection. This technique allows a population composed of many individuals to evolve according to some rules to a state that minimizes a cost function. Comparing with other random search techniques, the GA's are an intelligent way to find the global solution in the search space. These methods should be separated in some categories [8]:

- multiple or single parameter;
- discrete or continuous;
- constrained and unconstrained;

In GA's, a finite number of candidate solution, the chromosomes, are randomly created forming the initial population. In this work, a binary code was used [9]. Each chromosome represents a possible solution, divided in sub-strings that are decoded into their corresponding elastic constants. These chromosomes will create the new generation, by natural selection and reproduction procedures. As the cost function has to be minimized, only a few of the best chromosomes (the members with lower errors) will be kept for breeding.

The natural selection is a procedure that decides which individual should survive, forming the *mating pool*. Individuals with lowest cost reproduce more often than highest cost ones. An overlapping population is permitted. In this case, the offspring will replace the discarded chromosomes. Reproduction procedures consist in crossover and mutation. Two chromosomes, $parent_1$ and $parent_2$, are selected from the mating pool to produce two new offspring, $child_1$ and $child_2$. A crossover point is randomly selected between the first and last bit of the parents, exchanging portions of their strings, in order to form the children. This operation is performed with a probability *pcross*, that is normally a high value. Mutation operation change a bit from ``1" to ``0" or vice versa, with a probability *pmutation*, normally a very low value. Increasing the number of mutations increases the algorithm's search outside the current region of parameter space. It also tends to distract the algorithm from converging on a solution. In order to propagate the best solution unchanged it is usual in GA to keep the fittest chromosome without mutation.
After that, the cost of the new generation is calculated and the described process is repeated, until a stopping criterion. The number of generations depend on whether an acceptable solution is reached or a number of iterations is exceeded (*maxgen*). Figure 1 shows the procedures of the present work.

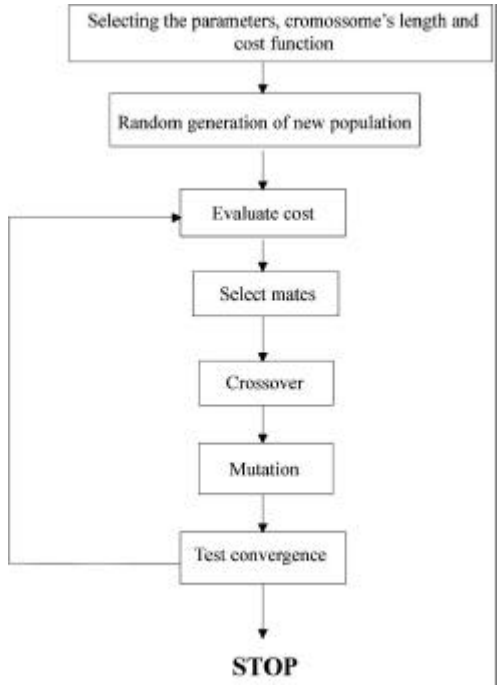


Figure 1: Flow chart

In the application of a GA to the identification problem, the first choice relies in the way of

representing the elastic parameters. A chromosome with 40 bits was used: 16 bits for E_1 and 8 bits for the others. All the elastic parameters are real numbers. The first eight bits of E_1 are decoded to form the integer part and the others are decoded to form the decimal part. The others elastic parameters are calculated as a fractional part of the elasticity modulus E_1. Therefore, the search space for each constant is defined as:

$$0 < E_1 < 256 \qquad 0 < \frac{E_2}{E_1} < 1$$
$$0 < \frac{G_{12}}{E_1} < 1 \qquad 0 < n_{12} < 0.5 \tag{18}$$

In the optimization problem, the objective function is defined by the difference between natural frequencies and analytical ones [10], stated as follows:

$$CF(q) = \sum_{i=1}^{M} \left[\frac{((f_{\exp})_i^2 - f_i^2(q))^2}{(f_{\exp})_i^4} \right] \tag{19}$$

where M is the number of used frequencies and θ is a vector containing the elastic constants. After evaluating the fitness, the chromosomes are ranked from lowest to highest cost. Only the *nbest* chromosomes are kept to form the mating pool, while the others are discarded. The new generations are composed by the *keep* chromosomes, in case overlapping occurs, and completed by the offspring created by crossover and mutation operations.

Parameters' selection is different for each example. Small population size (*popsize*) should lead to premature convergence while a large one is commonly used to increase the variation within a population. However, the increase of the number of function evaluations results in increased computational costs.

APPLICATIONS

The proposed approach is assessed by means of a number of applications. Both, isotropic and orthotropic plates are explored. The former, although it does not constitutes a composite structure, is used as far as it represents a test for the algorithm in which the detection of flaws is simpler.

The elastic constants were identified for each specimen using the data from Tab.1.

Table 1: Parameters of the plate

Sample	a(m)	b(m)	h(m)	ρ(Kg/m^3)
Aluminium	0.6	0.4	0.0063	2700
Kevlar/Epoxy	0.6	0.4	0.004	1380
SCS-6	0.6	0.4	0.004	3860
Glass/Epoxy	0.14	0.14	0.002011	1850

Isotropic Material

In order to enlarge the number of situations that were analyzed, some non-experimental quantities were utilized. They will be referred to as simulated frequencies, which are obtained with an a priori choice of the elastic parameters and the use of a finite element model of the plate. The simulated and experimental natural frequencies are shown in Tab.2 and Fig.2. These two examples are represented by GA_s and GA_e, where the first case was calculated from the simulated frequencies and the second one was calculated from the experimental ones. The experimental frequencies were taken from [4]. In Figure 3, the deviation of the theoretical parameters is presented. The last ones are those used to obtain the simulated frequencies.

For the aluminium plate, the following parameters of the genetic algorithm were utilized.

popsize = 80
keep = 10
nbest = 40
pcross = 0.95
pmutation = 0.03

Table 2: Frequencies on aluminium plate

f_n(Hz)	Simulated	GA_s	Experimental	GA_e
1	88.15	87.98	84.7	83.30
2	93.26	93.28	92.8	93.33
3	204.73	204.56	195.7	197.14
4	216.05	216.06	215.9	217.07
5	247.66	247.48	246.2	244.68
6	287.20	287.57	288.5	286.82
7	377.27	377.33	359.1	369.78
8	427.97	427.75	415.6	415.81
9	498.86	499.81	512.5	502.91
10	574.3	574.21	577	572.79
11	630.31	631.07	623.5	625.12

Table 3-4 show the estimated elastic constants calculated by simulated and experimental

frequencies respectively. The mean value (μ) and the standard deviation (s) were calculated too. In GA_e it is possible to note that the elasticity modulus are better estimated than the shear modulus and the Poisson ratio (Fig. 3), considering the literature values (Tab.3) as the standard ones.

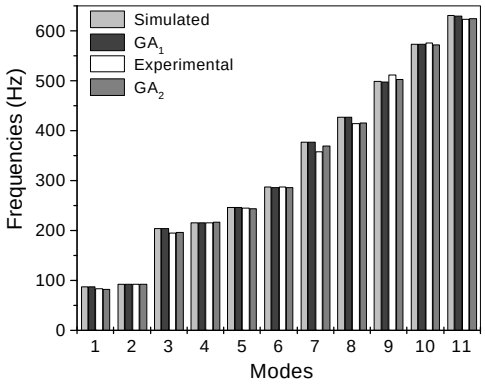


Figure 2: Frequencies on aluminium plate

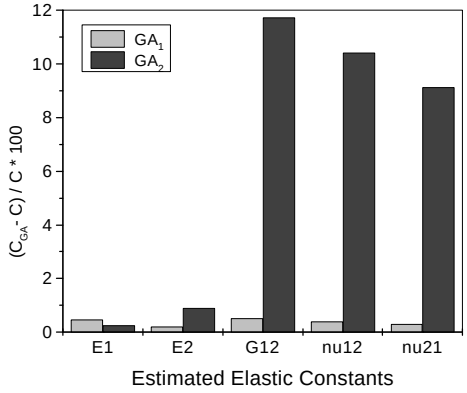


Figure 3: Estimated elastic constants on aluminium plate

Table 3: Estimated elastic constants for an aluminium plate – GA_1

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
Liter.	73	73	28.0789	0.3	0.3
1	72.98	72.69	28.22	0.3027	0.3015
2	73.23	72.66	28.03	0.2988	0.2964
3	73.01	72.73	27.95	0.3066	0.3054
4	74.15	73.28	27.51	0.2968	0.2933
μ	73.34	72.84	27.93	0.3012	0.2991
s	0.22	0.066	0.066	$1.4*10^{-5}$	$2.1*10^{-5}$

Table 4: Estimated elastic constants for an aluminium plate – GA_2

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
Liter.	73	73	28.0789	0.3	0.3
1	73.55	72.11	24.71	0.33	0.32
2	73.01	72.44	24.81	0.33	0.32
3	73.01	72.44	24.81	0.33	0.32
μ	73.19	72.33	24.78	0.33	0.32
s	0.064	0.023	0.0024	0.0	0.0

In Tab. 5, the results obtained by genetic algorithm are compared with the results obtained from the least-squares method in [4]. In the reference, the sample under consideration was modeled as an isotropic plate thus only the elasticity modulus and the Poisson's ratio were calculated.

Table 5: Comparison between GA and least-squares method

	Literature	Estimated Elastic Constants	Reference
E_1	73	73.1953	68.7517
E_2	73	72.3361	68.7517
G_{12}	28.0769	24.7097	26.2616
ν_{12}	0.3	0.3313	0.3090
ν_{21}	0.3	0.3274	0.3090

Orthotropic Material

Kevlar/Epoxy

In this plate, the thickness of each ply is h = 0.001m and the total number of plies is four in order to get the plate thickness. The dimensions and mechanical properties of the plate are shown in Tab.1. The simulated and estimated natural frequencies are shown in Fig.4. It is possible to verify, in the cost function graphic (Fig. 5), that the algorithm escaped a local minimum, looking for the global one. The estimated elastic constants obtained by GA are in good agreement with the literature values (Tab.5).

For the Kevlar/Epoxy plate, the following parameters of the genetic algorithm were utilized.

popsize = 80
keep = 10
nbest = 40
pcross = 0.95
pmutation = 0.04

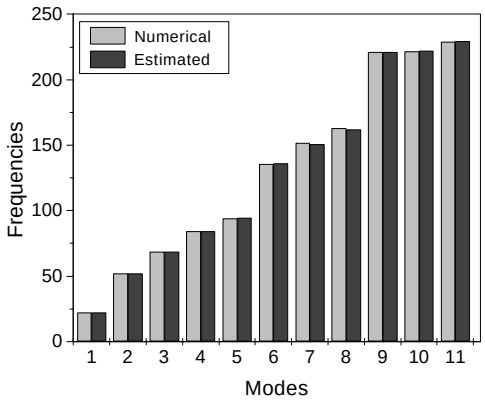


Figure 4: Frequencies on a Kevlar/Epoxy plate

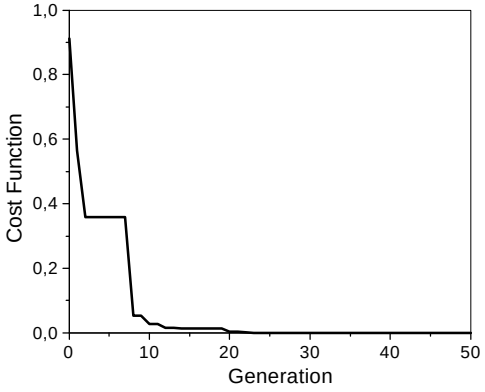


Figure 5: Cost function on a Kevlar/Epoxy plate

Table 5. Estimated elastic constants for Kevlar/Epoxy

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
Liter.	76.8	5.5	2.07	0.34	0.024
1	77.08	5.42	2.1	0.339	0.0238
2	77.09	5.42	2.1	0.333	0.0234
μ	77.085	5.42	2.1	0.336	0.0236
s	$0.25*10^{-4}$	0.0	0.0	$9*10^{-6}$	$4*10^{-8}$

SCS-6/Ti-15-3

In this specimen, the thickness of each ply is h = 0.001m and the total number of plies is four in order to get the plate thickness. The simulated and estimated natural frequencies are shown in Fig.6.

In this example, the algorithm escaped a local minimum too (Fig. 7). In Table 6, the estimated elastic constants are compared to the values of literature. Although the standard deviation has presented high values, the estimated mean values were very close to literature ones.

For the SCS-6/Ti-15-3 plate, the following parameters of the genetic algorithm were utilized.

popsize = 80
keep = 10
nbest = 40
pcross = 0.90
pmutation = 0.03

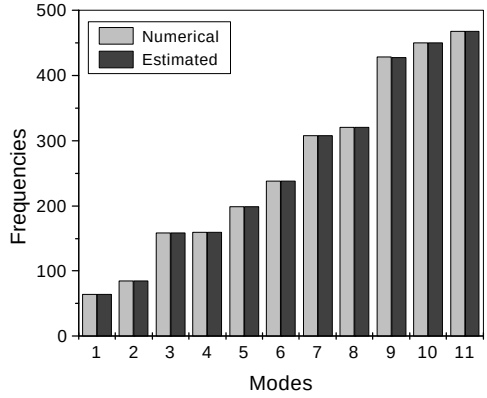


Figure 6: Frequencies on a SCS-6/Ti-15-3 plate

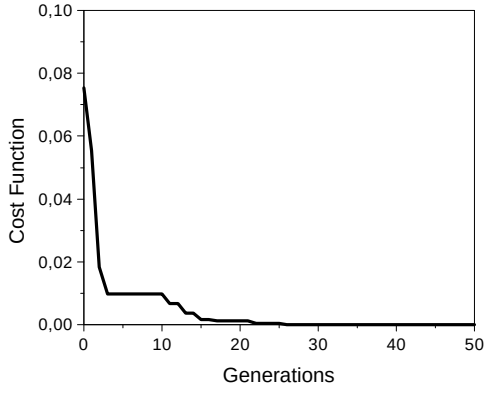


Figure 7: Cost function on a SCS-6/Ti-15-3 plate

Table 6: Estimated elastic constants for SCS-6/Ti-15-3

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
Liter.	221	145	53.2	0.27	0.17
1	224	146.1	53.37	0.24	0.16
2	216.69	143.05	54.17	0.294	0.194
3	220.28	144.55	53.34	0.27	0.18
4	224	145.24	53.37	0.25	0.162
μ	221.243	144.73	53.56	0.263	0.174
s	9.21	1.24	0.123	$4*10^{-4}$	$1.9*10^{-4}$

Glass-Epoxy

In this example, the natural frequencies were obtained by experimental tests and presented in [5] (Tab.7). The plate consists of eight unidirectional layers with a layer stacking sequence $[90/0/90/0]_s$. Table 8 shows the results obtained for the identification procedure applied in [5], comparing the elastic constants of the single layer of the cross-ply laminate (CP) with the properties obtained for the unidirectionally plate from the same material (UD). Table 9 shows the results obtained by the proposed approach. It is seen that the estimated elastic constants E_1, E_2 and G_{12} obtained by genetic algorithm are in good agreement with reference values. However, the Poisson's ratio are not well estimated. Probably, the reason is that the influence of Poisson's ratio on frequencies is considerably smaller than the influences of the others parameters.

Another way to verify the quality of the results is showed in Tab.7. The computed frequencies by using the identified parameters presented in the last column are in good agreement to the experimental ones.

Table 7: Frequencies on a Glass/Epoxy plate

f_n(Hz)	Experimental	FEM	GA
1	166	166.4	162,13632
2	341	344.2	343,76897
3	-	416.2	442,30251
4	484	486.1	497,13064
5	542	545.9	543,87
6	902	895.6	905,45285
7	971	967.5	946,12462
8	1090	1087	1087,01988
9	1155	1165	1158,07626

For the Glass/Epoxy plate, the following parameters of the genetic algorithm were utilized.

popsize = 80
keep = 10
nbest = 40
pcross = 0.95
pmutation = 0.04

Table 8: Elastic constants for Glass/Epoxy [5]

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
CP	38.15	12.44	4.92	0.368	-
UD	38.81	12.12	5.09	0.255	-

Table 9: Estimated elastic constants for Glass/Epoxy

	E_1	E_2	G_{12}	ν_{12}	ν_{21}
1	40.04	10.48	4.06	0.46	0.12
2	41.55	9.90	4.05	0.45	0.10
3	41.3	10.16	4.03	0.44	0.11
4	38.79	11.36	4.09	0.43	0.12
μ	40.42	10.47	4.057	0.445	0.1125
s	1.21	0.30	$4*10^{-4}$	$1.2*10^{-4}$	$6.8*10^{-5}$

The evolution of the cost function durinhg the optimization process is depicted in Fig.8.

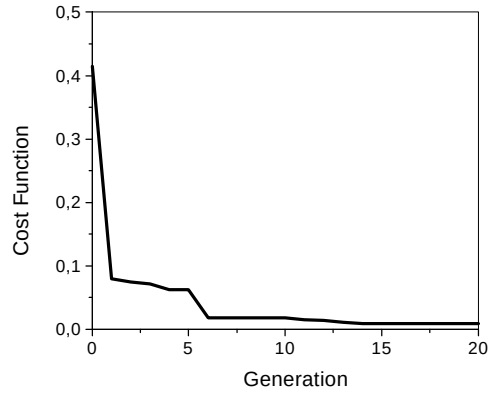


Figure 8: Cost function on a Glass/Epoxy plate

CONCLUSIONS

The proposed method for the identification of elastic constants has shown be effective for the examples that have been presented here. It's clear from the examples that the algorithm was able to skip local minimums, i.e., it can be considered an

efficient method to find the global minimum for the problem under consideration. It should be remarked that the experimental data were obtained out of non-destructive dynamics tests. Therefore, the designer can perform several experiments considering the same composite structure.

REFERENCES

1. Reddy,J.N., *Mechanics of Laminated Composite Plates: theory and analysis* , CRC Press, 1991.

2. Balasubramaniam, K. and Rao, N.S., Inversion of composite material elastic constants from ultrasonic bulk wave phase velocity data using genetic algorithm , *Composites: Part B*, **29B**, 171-180 (1998).

3. Cunha,J. and Cogan,S. and Berthod, C., Application of Genetic Algorithms for the Identification of Elastic Constants of Composite Materials from Dynamic Tests, *International Journal for Numerical Method in Engineering*, **45**, 891-900 (1999).

4. Bastos,S.F., *Identification of elastic parameters by means of modal analysis*, Department of Mechanical Engineering, Federal University of Rio de Janeiro - Brazil, Master Thesis (2001).(in Portuguese)

5. Rikards,R., Chate,A. and Gailis,G., Identification of elastic constants of laminates based on experiment design, *International Journal of Solids and Structures*, **38**, 5097-5115 (2001).

6. Herakovich,C.T., *Mechanics of Fibbrous Composites*, Jonh Wiley and Sons, Inc, 1998.

7. Shames, I.H. and Dym,C.L., *Solid Mechanics: a variational approach*, McGraw-Hill Int. Ed., London, 1973.

8. Haupt,R.L, and Haupt,S.E., *Pratical Genetic Algorithms*, John Wiley and Sons, 1998.

9. Goldberg,D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1998.

10. Bledzki,A.K.,Kessler,A., Rikards,R. and Chate,A., Determination of elastic constants of glass/epoxy unidirectional laminates by the vibration testing of plates, *Composite Science and Technology*, **59** , 2015-2024 (1999).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

AN INVERSE TECHNIQUE FOR IDENTIFICATION OF ELASTIC CONSTANTS OF A GLASS/EPOXY LAMINATED PLATE

X. Han[a] and G.R. Liu[b]
Centre for Advanced Computations in Engineering Science
Department of Mechanical Engineering
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260
http://www.nus.edu.sg/ACES
[a]acehanxu@nus.edu.sg, [b]mpeliugr@nus.edu.sg

ABSTRACT

A computational inverse technique is proposed to identify the material constants of a galss/epoxy laminate plate from dynamic displacement responses obtained at only one receiving point of laminate surfaces. A hybrid numerical method (HNM) is used for forward computation that relates the material constants to the displacement responses. The neural network (NN) is used as the inverse procedure using the surface displacement responses as the inputs and the elastic constants of anisotropic laminated plates as the outputs. The NN model is trained using the results from the simulated results. The NN model would go through a progressive retraining process until the calculated displacement responses using the determined results are sufficiently close to the actual responses. This proposed computational method is verified using one set of elastic constants of glass/epoxy laminated plates. It is found that the present procedure is very robust for reconstruction of the elastic constants of the animated plates.

NOMENCLATURE

$c_{ij}(i,j=1,\cdots 6)$	Elastic constants
d	Distance norm
$f(t)$	External Loading
N	Total number of sample points
u	Displacement response in x direction
$\mathbf{W}=\{w_{ij}^{k}, i=1,\cdots,N_i, j=1,\cdots,N_j; k=1,2,3\}$	Matrix of the weight
$\mathrm{X}=\{x_i, i=1,\cdots,N\}$	Inputs of the NN model
$\mathrm{Y}=\{y_i, i=1,\cdots,M\}$	Outputs of the NN model
σ	Standard deviation

INTRODUCTION

Advanced nondestructive methods for material characterization of composites utilize the complex relationship between the structure behaviors and the material property. This relationship is often represented by a known mathematical model defining the forward problem, which can be analyzed numerically or otherwise. Thus if a set of reasonably accurate experimentally measured structure behavior data is available, the material property of the composite may be identified by solving an inverse problem properly formulated. The material property can often be characterized by minimizing the sum of the squares of the deviations between the experimental and the calculated structure behavior data. Using elastic waves is a promising mean for material characterization of composites. Ultrasonic wave velocity has been used as the structure behavior data for determination of elastic constants of anisotropic composites [1,2]. Liu et al. [3] presented a combined method of genetic algorithm and nonlinear least square method for material determination and applied it to functionally graded material plates and composite laminated plate. First the genetic algorithm was used to locate the initial estimation of the parameter, then the traditional least squares method is applied to determine the material constants. However, it can be generally concluded that these inverse procedures require too many calls for forward solvers.

Neural network (NN) is a novel tool for information processing. It provides a unique computing architecture, which enjoys a massive parallel processing structures. The parallelism of NN enables it to solve many problems that cannot be handled by analytical approaches. NNs provide an effective approach for engineering applications in a very broad spectrum [4,5]. Furthermore, the NN technique is well known for its ability to model nonlinear and complex relationship between the structure parameters and the dynamic characteristics.

In this paper, a novel progressive NN procedure is applied for the identification of elastic constants of anisotropic laminated plates. In the present NN model, the input data are the dynamic displacement responses on the surface of the plate, which can be easily measured using conventional experimental techniques. The NN model is first trained off-line using a set of initial training data that contain various assumed elastic constants and their corresponding displacement responses calculated using the HNM [6] as the forward solver. A modified back-propagation (BP) algorithm is used as the learning process. The NN model is then used to determine the elastic constants of laminated plate by feeding in displacement responses. The determined elastic constants are then used in the HNM to calculate the displacement responses. The NN model would go through a progressive retraining process if the calculated displacement responses deviate un-acceptably from the actual ones. An example of identification of the elastic constants of a glass/epoxy laminated plate is presented to demonstrate the efficience of the proposed inverse technique.

STATEMENT OF THE PROBLEM

Consider a laminated plate with any number of anisotropic layers in the thickness direction, as shown in Figure 1. The thickness of the plate is denoted by *H*. The incident excitation waves to the plate are assumed to be a vertical line load in the *z*-direction acting at $x = 0$ on the upper surface.

The line loads are independent of the *y* axis, but as a function of *t* is dependant as

$$f(t) = \begin{cases} \sin(2\pi t/t_d) & 0 < t < t_d \\ 0 & t \le 0 \text{ and } t \ge t_d \end{cases} \qquad (1)$$

and t_d is the time duration of the incident wave. Equation (1) implies that the time history of the incident wave is one cycle of the sine function.

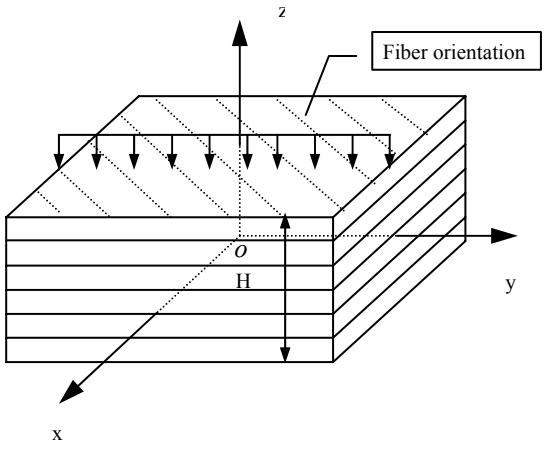


Figure 1. A composite laminate subjected to a line load on the surface.

An NN model is used for the determination of elastic constants of anisotropic laminated plates. The outputs of the NN model are elastic constants. The inputs of the NN model are the time history of displacement responses on the surface of the laminated plate, which can be easily measured using conventional experimental techniques. In this paper, we utilize computer-generated displacement responses calculated using the HNM [6,7] as the forward solver based on actual elastic constants of laminated plates.

Only one receiving point is chosen on the surface of the laminated plates, and the responses in the time domain for displacement components in *z*-direction are selected as the inputs for the NN model.

AN NN PROCESS FOR DETERMING ELASTIC CONSTANTS

An NN model, which consists of a set of nodes arranged into four layers as shown in Figure 2, is used in this work. There are *N* inputs representing the displacement responses on the surface and *M* outputs representing the elastic constants to be determined. Here two hidden layers are used in this work. Mathematically, the NN model represents a nonlinear mapping

between inputs $\mathrm{X} = \{x_i, i = 1, \cdots, N\}$ and outputs $\mathrm{Y} = \{y_i, i = 1, \cdots, M\}$ via the following equation

$$\mathbf{Y} = g(\mathbf{W}, \mathbf{X}) \tag{2}$$

the $\mathbf{W} = \{w_{ij}^k, i = 1, \cdots, N_i, j = 1, \cdots, N_j; k = 1,2,3\}$ is a matrix of weights corresponding to the connections between the layers, and N_i and N_j are the numbers of neurons for the i^{th} and j^{th} layers, respectively. Training of the NN model is referred to as the calculation of the weight matrix **W** using the training data set. Once the training is complete, the NN calculation is very fast regardless of the complexity of the actual physics of the problem. A modified BP learning algorithm [8] with a dynamically adjusted learning rate and an additional jump factor is employed as the learning algorithm. This learning algorithm can overcome the possible saturation of the sigmoid function and speed up the training process of the NN model.

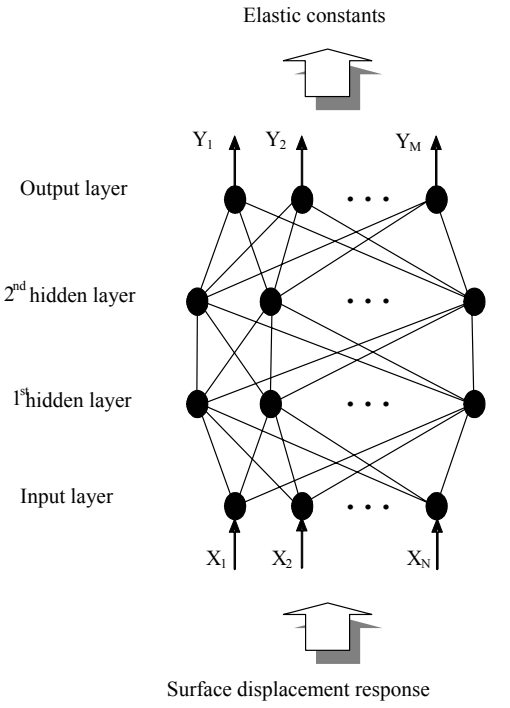


Figure 2 A two-hidden layer NN model

After the initial training of the NN model, the determination of the elastic constants begins by feeding the measured displacement response data $\mathbf{X}_m$ into the NN model. The outputs of the NN model are the determined elastic constants $\mathbf{Y}_I$. These determined elastic constants are then fed into the HNM to produce a set of calculated displacement response data $\mathbf{X}_c$. A comparison between the calculated displacement response $\mathbf{X}_c$ and measured displacement responses $\mathbf{X}_m$ is made based on a given criterion. If these two sets differ significantly such that the criterion is not satisfied, then the NN model will be retrained on-line using adjusted training samples that contain $\mathbf{X}_c$ and $\mathbf{Y}_I$. The retrained NN model is then used to determine the elastic constants again by feeding in the measured displacement esponses $\mathbf{X}_m$. This determination and on-line retraining procedure is repeated until the difference between the calculated and measured displacement responses satisfies the given criterion. At the end of the progression, the final determined elastic constants are guaranteed to produce the displacement responses that are very close to the measured ones when fed into the HNM.

Re-training of the NN model is achieved by adding new samples into the original pool of samples and enforced a more stringent convergence criterion. It has been pointed out that it could be difficult to achieve the same level of convergence while maintaining the same NN architecture when the number of samples increases. To avoid this problem, a dynamic adjustment method for selecting samples for retraining is proposed. While adding the new sample related on the determined elastic constants by the NN model and the displacement responses from the HNM, one sample from the original sample set should be removed so as to maintain the same number of samples. The sample to be removed is the one that has the largest distance norm from the measured displacement responses $\mathbf{X}_m$. The distance norm of the displacement responses between the i^{th} sample $\mathbf{X}_i$ and the measured displacement responses $\mathbf{X}_m$ is defined as

$$d = \left\| \mathbf{X}_m - \mathbf{X}_i \right\|^2 \tag{3}$$

By replacing this remote sample with a new sample, the sample density around the measured displacement responses increase as the process progresses. As a result, the modeling accuracy of the NN model in the neighborhood of the measured displacement responses could be improved.

APPLICATIONS

This NN process for determination of elastic constants of laminated plates is illustrated using one actual laminated plate consisting of six glass/epoxy layers. The stacking sequence of the laminated layers is denoted by $[0/+45/-45]_s$, where the digital numbers stand for the angles of fiber-orientation of each ply to the *x*-axis. The subscript of 's' means that the plate is symmetrically stacked. The glass/epoxy material is the transversely isotropic material; there are only five elastic constants as listed in the 2nd column in Table 1 [9]. Hence there are five parameters, named as $c_{11}, c_{12}, c_{22}, c_{23}$ and c_{55}, needed to be identifieded.

Table 1. The search range for the elastic constants to be identified

Elastic constants	Actual Data (GPa)	Search Range (GPa)
c_{11}	42.02	30-54
c_{12}	6.067	4-8
c_{22}	13.5	10-18
c_{23}	7.277	5-9
c_{55}	3.41	2-4

The NN model used in this paper has two hidden layers, and the neuron numbers of the input, output, 1st and 2nd hidden layers are 10, 5, 30 and 16 respectively. Instead of carrying out actual experiment, the measured displacement responses are simulated using the HNM using the actual elastic constants. In order to simulate the measured displacement responses, noise-contaminated displacement responses are also used for the determination of elastic constants. Noise effects are investigated by adding Gaussian noise directly to the computer generated displacement readings. A Gauss random number generator is used to generate a series of random numbers with the standard deviation as

$$\sigma = 0.01 \times (1/N \sum_{j=1}^{N} (u_j^a)^2)^{1/2} \quad (4)$$

where u_j^a is the measured displacement reading at the *j*th time sample point and *N* is the total number of time sample points.

Table 2. Reconstructed results of elastic constants of laminated plate

Elastic Constants	Original Value(GPa)	Result (deviation) at progressions 1	2	3	4
c_{11}	42.020	41.05(2.3%)	41.18(-2.0%)	41.35(1.6%)	41.05(2.3%)
c_{12}	6.067	5.31(12.5%)	5.40(-11.0%)	5.58(8.0%)	5.90(2.8%)
c_{22}	13.500	12.6(6.6%)	14.18(5.0%)	13.03(3.5%)	13.3(1.5%)
c_{23}	7.277	8.25(13.5%)	7.97(9.5%)	7.97(9.5%)	7.635(4.9%)
c_{55}	3.410	3.70(8.6%)	3.63(6.5%)	3.56(4.5%)	3.57(4.5%)

For this problem, a search range of $\pm$ 30% off from the actual value of elastic constants is used, as shown in Table 1. To formulate the initial training samples, it was assumed that there were 4 levels of change in the search range for these five elastic constants, which correspond to $c_{11}, c_{12}, c_{22}, c_{23}$ and c_{55} of their discrete values. Based on the orthogonal array method, these five

four-level parameters would only require 16 samples to cover the whole domain. In addition, another 21 samples created randomly, were added into the training data set. This combined strategy covers a good cross-section of all possible elastic constants variations.

Table 2 summarizes the reconstructed results of the elastic constants. The results for four progressions are listed. It can be found that the result at the first progression is not accurate as the maximum deviation is high, and the displacement responses corresponding to these reconstructed elastic constants are quite different from the simulated ones using the actual values of elastic constants, the distance norm (Eq. (3)) between them is long. A retraining for NN model is required. A new sample is created from the 1st characterized result and the corresponding displacement responses calculated from the forward sovler. The new sample is added into the original sample pool to replace the sample with large distance norm. The retraining process is repeated until the displacement responses corresponding to the reconstructed elastic constants are sufficiently close to the simulated measurements. The results at stages of progressive training are also listed in Table 2. It can be seen from Table 2 that the accuracy of the determined results increase as the progression number increases, and the determined result is very accurate after 4 progressions. The maximum deviation of the sixth progression elastic constants is as low as 6%. The first training of the presented NN model needs about 1200 seconds, and the retraining time for the following progressions decrease as the progression number increases. However, this decreasion is not very distinct.

The identification with the presence of the noise is also carried out in this paper, and the results are listed in Table 3. It can be found that, the determined result remains stable regardless the presence of the noise, and the required number of progression is not changed, even when the noise is added.

CONCLUSION

A progressive NN technique is proposed for the identification of elastic constants of a glass/epoxy laminated plate, using dynamic displacement responses on the surface as the input data. In this procedure, the HNM is employed as a forward solver to calculate the displacement responses on the surface of the laminated plates. The NN model is trained using the calculated result from the HNM. Once trained, the NN model can be used for on-line identification of elastic constants if the dynamic displacement responses on the surface of the laminated plate can be obtained. The identified elastic constants are then used in the HNM to calculate the displacement responses. The NN model would go through a progressive retraining process until the calculated displacement responses using the determined results are sufficiently close to the actual responses. The accuracy of output from the NN model increases with the increase number of retraining cycles, the required accuracy can be therefore obtained by repeating the retraining process.

Table 3. Reconstructed results with the presence of the noise

Elastic Constants	Original Value(GPa)	Result (deviation) at progressions			
		1	2	3	4
c_{11}	42.020	42.65(1.5%)	42.65(1.5%)	42.78(1.8%)	42.72(1.7%)
c_{12}	6.067	5.18(14.7%)	5.30(12.6%)	5.48(9.7%)	5.81(6.3%)
c_{22}	13.500	14.26(5.6%)	14.02(3.9%)	13.99(3.6%)	13.73(1.6%)
c_{23}	7.277	8.15(12.0%)	8.38(15.1%)	7.74(6.3%)	7.73(6.2%)
c_{55}	3.410	3.71(8.6%)	3.55(4.4%)	3.53(3.7%)	3.55(4.0%)

REFERENCES

1. Y.C Chu and S.I. Rokhlin, Stability of Determination of Composite Moduli from Velocity Data in Planes of Symmetry for Weak and Strong Anisotropies, *J. Acoust. Soc. Am. ,* 95, 213-225(1994).

2. Y.C. Chu, A.D. Degtyar and S.I. Rokhlin, On Determination of Orthoropic Material Moduli from Ultrasonic Velocity Data in Nonsymmetry Planes, *J. Acoust. Soc. Am.* 95, 3191-3203(1994).

3. G. R. Liu, X. Han and K.Y. Lam A combined genetic algorithm and non linear least squares method for material characterization using elastic waves. *Computational Methods in Applied Mechanics and Engineering*, 191, 1909-1921, (2002)

4. Mota Soares CM, Moreira de Freitas M, Araujo Al, Pedersen P. Identification of material properties of composite plate specimens. *Composite Structures* 25: 277-285(1993)

5. X. Wu, J. Ghaboussi and J.H. Garrett. Use of neural networks in detection of structural damage. *Computers & Structures*. 42(4): 649-659(1992).

6. G.R. Liu, J. Tani, T. Ohyoshi and K. Watanabe, Transient waves in anisotropic laminated plates, Part 1: Theory; Part 2: Applications. *Journal of vibration and acoustics.* 113: 230-239(1991)

7. G.R. Liu and Z.C. Xi, Elastic waves in anisotropic laminates, CRC Press, 2001.

8. G. R. Liu, X. Han, Y. G. Xu and K. Y. Lam, Material characterization of functionally graded material using elastic waves and a progressive learning neural network, *Composites Science and Technology*, 61:1401-1411(2001).

9. K. Takahashi and T.W. Chou. Non-linear deformation and failure behavior of carbon/glass hybrid laminates. *Journal of composite materials*. Vol.21 (4), 396-407(1987).

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

A MODIFIED MICRO GENETIC ALGORITHM WITH INTERGENERATIONAL PROJECTION AND INVERSE IDENTIFICATION OF MATERIAL PROPERTIES OF A PRINTED CIRCUIT BOARD

Z. L. YANG
Singapore-MIT Alliance/ Center for Advanced Computations in Engineering Science,National University of Singapore, Singapore 119260
smayzl@nus.edu.sg

G. R. LIU
SMA Fellow, Singapore-MIT Alliance, Department of Mechanical Engineering, National University of Singapore, Singapore 119260
mpeliugr@nus.edu.sg

ABSTRACT

An inverse procedure is introduced to identify the material properties of a printed circuit board (PCB) and components mounted on it. A modified micro genetic algorithm (mGA) with intergenerational projection search technique is presented in this paper to speed up the inverse searching process. The PCB is subjected to an enforced acceleration over a range of excitation frequencies and the frequency response is obtained computationally and experimentally. The frequency response at specific nodes on PCB and components are utilized as the input for the inverse procedure. Material properties that minimise the sum of the squares of the deviations between test and simulation results are inversely determined. The present inverse procedure can be used in a wide range of practical engineering problems.

1. INTRODUCTION

1.1 Problem Description

Modern electronic packages consist of a variety of components. The size, shape, arrangement, and material of the modules vary and are mounted on a PCB using soldered leads or pins. Electronic packages may experience dynamic loads during manufacturing, shipping and service. As a result, the PCB may experience large vibration amplitudes and/or acceleration levels. These vibrations are transmitted throughout the PCB inducing stress in the modules, leads, and solder joints connecting the modules to the PCB. Performance degradation and possible system failure may occur if any component of the system is over-stressed.

Electronic packages are normally subjected to qualification tests. A typical test may consist of subjecting the electronic package to an enforced sinusoidal acceleration and measuring the response, over a range of excitation frequencies, at a few locations on the PCB and components. Making prototypes and conducting physical tests take a long time. From business point of view, it is always preferable to keep the time-to-market as short as possible. Use of computer aided engineering (CAE) tools results in faster evaluation of relative performance of various designs thereby reducing the number of physical prototypes and tests required.

Companies in the manufacturing sector – from small die manufacturers to large automobile manufacturers – are talking CAD/CAM/CAE or virtual prototyping. In the past, engineering analysis was used to evaluate and re-design failed components. However, at present CAE analysis is used to simulate new products, study the effects of design changes on a computer resulting in reduced cost, better quality, and reduced time-to-market. Typical industrial sectors where CAE finds widespread applications are electronic packaging, automotive, aerospace, defence etc.

One of the problems faced by a CAE analyst is the lack of availability of exact material properties. One way of overcoming this difficulty is to work in the opposite way. Instead of looking at the forward problem in which all the material properties, loading, boundary conditions are given and the response is calculated, we look at the inverse problem in which the response is given (from physical tests) and we seek the material properties that result in the given response. Once the material properties are obtained, relative

performance of competing designs using similar components can be evaluated virtually on a computer without the need to make physical prototypes of all the designs and to perform physical tests with all the prototypes. The inverse problem may be regard as an optimization problem in which the objective is to minimize the error between test and simulated response. Many algorithms can be used in the procedure of minimizing the objective function. As the engineering problems are complex and are usually multi-optimum problems, a genetic algorithm is one of the best techniques that can be used to find the global optimum.

1.2 Micro Genetic algorithm

Genetic algorithm was introduced by Holland[1] as a method of searching for global optimum in complex systems. In recently years, many efforts [2-15] are made using inverse procedure to solve the problems in structural engineering. There are several different versions of genetic algorithms. The micro genetic algorithm (mGA) is one of the most widely used GAs. This algorithm produces fewer individuals in each generation and the individuals of each generation are usually created through two operations: selection and crossover. mGA is a very robust algorithm in finding the global optimum rather than local optimum for a given domain. This advantage is especially important in finding the global optimum for multi-minimum or multi-maximum problems. However, as the primitive selection procedure in mGA is random, the time required to find the desired solution is usually very long. The searching time will also increase very rapidly as the number of genes in the individuals increase. It is commonly believed that the mGA is impractical for finding the global optimum for real life problems with large number of parameters, unless measures are taken to speed up the searching process.

One of the methods used for speeding up the search process of GA is the hill climbing technique [16]. This method combines the general GA global search procedure with locally optimized search using hill climbing. This local optimized technique greatly improves the local searching performance of conventional GAs. However, due to the fact that a large number of function evaluations is necessary in the local search, the method encounters difficulties for problems where variables are large and/or a single function evaluation takes considerable computational time. As most of the time is spent on function evaluation, the desired searching method should be the one that requires only a small number of function evaluations. Therefore, measures must be taken to mGA to reduce its number of forward calculations.

1.3 Work in this paper

In this paper, an inverse analysis with modified mGA search algorithm is firstly presented to reduce the required number of forward analyses. The performance of the method is tested through a number of testing functions. The method is then used to determine the material properties that minimize the error between test and simulated frequency response of a PCB with components mounted on it. Conclusions are finally obtained.

2. MODIFIED mGA

To speed up the convergence procedure, in the following part, the conventional mGA is modified using an intergenerational projection searching technique.

2.1 Intergenerational Projection Technique

Intergenerational projection is a search strategy, which can be used to find out the better individual from best individuals of two adjacent generations. Due to the expensive computation of mGA in functional evaluation, the intergenerational projection technique can be better used in mGA than other gradient methods. The investigation of combining mGA with this technique can be found in reference [11]. In using intergenerational projection, two new individuals c_1 and c_2 are generated in each generation using forward and internal interpolations based on the two best individuals in the neighbor generation p_j and p_{j-1}. That is

$$c_1 = p_j + a\,(p_j - p_{j-1}) \qquad (1)$$

$$c_2 = p_j - b\,(p_j - p_{j-1}) \qquad (2)$$

where a and b are two non-negative decimals, these parameters' value can be changed to adjust the distances of these new individuals to original individuals. To get stable convergence, generally, the ranges of these parameters are: $0 \le a \le 1.0$, $0 \le b \le 1.0$. For simplicity, in the following parts of this paper, we fixed the values of a, and b to be 0.2 and 0.5 respectively.

2.2 Performance Test of Using the Intergenerational Projection Technique

In order to compare this method with conventional ones, the selected testing functions are those typical functions used for performance testing. They are listed in table 1.

Table 1 test functions used in performance test

F1	$f(x) = \sin(x) + \sin(10x/3) + \ln(x) - 0.84x + 3,\ 2.7<x<7.5$
F2	$f(x_1, x_2) = \prod_{i=1}^{2} \sin(5.1\pi x_i + 0.5)^6 \exp\frac{-4\log 2(x_i - 0.0667)^2}{0.64},\quad 0 < x_i < 1.0$
F3	$f(x_1, x_2) = \sum_{i=1}^{5} i\cos((i+1)x_1 + i) * \sum_{i=1}^{5} i\cos((i+1)x_2 + i) + \left((x_1 + 1.42513)^2 + (x_2 + 0.80032)^2\right),\quad -10 < x_i < 10$
F4	$f(x_1, x_2, x_3) = \sum_{i=1}^{3}\left(\left(x_1 - x_i^2\right)^2 + \left(x_i - 1\right)^2\right),\quad -5 < x_i < 5$
F5	$f(x_1, x_2, x_3) = \sum_{i=1}^{3}((ax_1 - bx_i^2)^2 + (cx_i - d)^2)$, where, $a = 0.99934,\ b = 1.00056,\ c = 0.99904,\ d = 1.00094$, $-5 < x_i < 5$
F6	$f(x_1, x_2, x_3) = \sum_{i=1}^{10}\left[e^{(-ix_1/10)} - e^{(-ix_2/10)} - \left(e^{(-i/10)} - e^{(-i)}\right)x_3\right]^2,\quad -5 < x_i < 15$
F7	$f(x_1, x_2, x_3, x_4) = \sum_{i=1}^{5}\frac{1}{\sum_{j=1}^{4}(x_j - d(i,j))^2 + c(i)},\quad 0 < x_i < 10$

Table 2 Performance of modified mGA and general mGA

No	**Test Function**	**Modified mGA**		**mGA**		**Ratio N_M/N_m (%)**
	$f_{optimum}$	N_M	f_M	N_m	f_m	
F1	-1.601	7	-1.601	313	-1.601	2.24
F2	1.0	110	1.0	10868	1.0	1.01
F3	-186.7	219	-186.7	11363	-186.7	1.93
F4	0.0	409	-2.235E-8	34618	-2.235E-8	1.18
F5	0.0	306	1.232E-5	26487	1.023E-4	<1.1
F6	0.0	730	-1.459E-8	39997	-.917E-7	<1.8
F7	-10.15	250	-10.15	20576	-5.101	<1.2

Table 2 gives results of the testing functions in table 1, the performance comparison between the mGA with intergenerational projection and the conventional mGA are performed and the rates of number of generations for desirable fitness of present method over conventional method are also listed in the table. Results show the great effectiveness of the proposed method.

Compare to conventional mGA , only a small fraction of generations is needed to get convergence using this intergenerational projection method. The results have shown great success in speeding up the searching procedure.

Table 3 The number of forward calculations between mGAs with and without global modification (GM)

Test Function	Modified mGA			mGA		
	without GM	with GM	saving (%)	without GM	with GM	saving (%)
F1	805	357	55.65	2500	1236	50.56
F2	3980	2116	46.83	100000	56490	43.51
F3	33305	13368	59.86	75000	43147	42.47
F4	42260	19627	53.56	200000	140639	29.68
F5	40220	22075	45.11	150000	105496	29.7
F6	8050	4859	39.64	200000	135878	32.06
F7	8005	5195	35.1	250000	190692	23.72

In order to further improve the search procedure, a pre-treatment procedure is made to ensure random selection of individuals that did not occur previously. This can be done through defining a vector to remember the individuals that are used. Therefore, the domain of candidate individuals for random searching in this method becomes progressively smaller as the searching goes on. The comparisons of number of forward calculation between mGAs with and without this modification are listed in table 3. From table 3, the number of forward calculations is further reduced for modified mGA by using this improvement. The combination use these modified techniques to mGA, the total saving of forward evaluations is significant compare to those using conventional mGA.

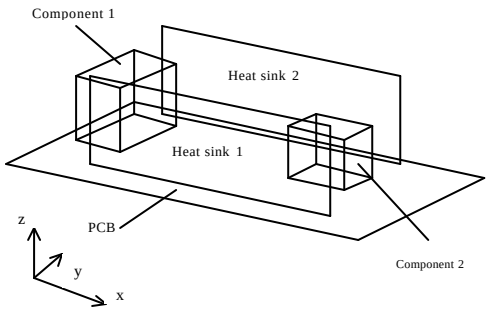


Figure 1. Geometrical model of PCB with heat sinks and components

3. INVERSE IDENTIFICATION OF MATERIAL PROPERTIES

3.1 Problem Definition

For the purpose of illustrating the procedure, a PCB with two heat sinks and two components mounted on it is considered [17]. Figure 1 shows a geometric model of the system. The physical dimensions of the PCB are 200mm×100mm×1.5mm, those of heat sinks are 140mm×50mm×1.4mm, those of component 1 are 50mm×20mm×40mm, and those of component 2 are 40mm×20mm×30mm. The PCB is partially supported along two opposite edges and at a few locations on the other pair of opposite edges. This corresponds to the PCB being housed in a casing. The system is subjected to an enforced acceleration of 1.0g over a frequency range of 20 Hz to 140 Hz. The frequency response (acceleration levels) at a few locations on the PCB and on the components is obtained for a given set of material properties. In the actual case, this response, obtained for a given set of material properties, will correspond to test results. Now, the problem is to determine a set of material properties if exact values are not known but a range of values is available for key parameters. This is done such that the error between test and simulated frequency response is minimized. The formulations are described in the following part.

3.2 Two-Stage Inverse Procedure

The inverse analysis procedure used to determine the material properties in this work is performed by a two-step procedure. First, matching of natural frequencies is performed. The objective function (fitness) for this step is

$$\text{minimise} \quad \sum_i \{(f_t)_i - (f_s)_i\}^2 \quad i = 1, 2, ..., m \qquad (3)$$

where f_t and f_s are the natural frequencies from test and simulation, respectively and i denotes the mode number. After matching the natural frequencies, the design variables used in this step are not allowed to vary in the next step in which matching of frequency response is performed. The remaining design variables are used in this step. The objective function (fitness) used in this step is

$$\text{minimise} \quad \sum_k (T_k - S_k)^2 \quad k = 1, 2, ..., N \qquad (4)$$

where T_k and S_k denote the frequency response from test and simulation, respectively and k denotes the output location number. With this step, the inverse analysis procedure for determining material properties for matching of test-simulation frequency response is completed. The modified mGA is used in the procedure of finding the minimums in (3) and (4). In the forward calculation the finite element representation of the PCB system and the numerical experiments are performed with the finite element model.

3.3 Finite Element Representation

Based on the geometric model shown in Figure 1, a finite element model is created. The PCB and

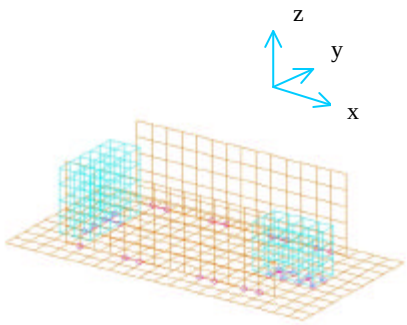


Figure 2. Finite element representation of PCB and components

heat sinks are modeled using four-node plate elements (CQUAD4 in MSC/NASTRAN) while the components are modeled using eight-node brick elements (CHEXA). The heat sinks and the components are connected to the PCB using rigid elements (RBE2 in MSC/NASTRAN). The finite element representation is shown in Figure 2. The model consists of 340 CQUAD4 elements and 64 CHEXA elements. For the purpose of specifying

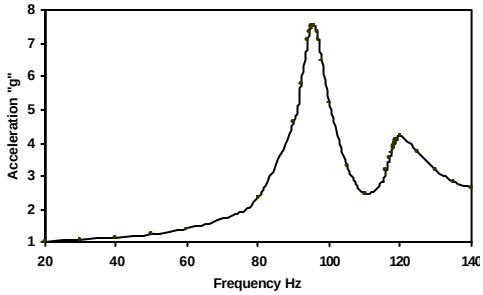


Figure3. Acceleration response at PCB

an enforced acceleration, the large mass method is used. A large mass (1.0×10^7 ton) is connected to the nodal points where the system is supported. An appropriate force on the heavy mass in the required direction will result in an enforced acceleration. The following section describes the numerical experiments that are performed and the results obtained from those experiments.

3.4 Numerical Experiments, Results and Discussion

Before performing a frequency response analysis, knowledge of the natural frequencies of the system is required. Knowing the natural frequencies will help take smaller steps near natural frequencies. So, a normal modal analysis is performed first with the finite element model

Table 4. Material properties for obtaining "test" results

Property	PCB	Heat Sinks 1&2	Component 1 & 2
Young's modulus, E, MPa	13500	71,000	5,000
Mass density, ρ, ton/mm^3	1.5E-9	2.7E-9	1.0E-9
Poisson's ratio	0.3	0.3	0.2
Structural damping coefficient GE	0.08	0.08	0.08

described in the previous section. For the purpose of obtaining "test" results, the set of material properties given in Table 4 is used. The first three natural frequencies are found to be f_1 = 95.39 Hz, f_2 = 118.54 Hz, and f_3 = 173.82 Hz. The fundamental mode is observed to be dominantly a bending mode. Following normal modal analysis, a modal frequency response analysis is performed with the excitation frequencies in the range 20 Hz to 140 Hz. The peak response corresponds to the fundamental frequency. Figure 3 shows the response at PCB center. Table 5 lists the acceleration response at some specific locations.

Table 5. Acceleration response from "test"

No	Location	Acceleration (g)
1	PCB Centre	7.567
2	PCB Front	8.100
3	PCB Back	7.324
4	Below Component 1	5.956
5	Top of Component 1	6.545
6	Below Component 2	5.999
7	Top of Component 2	6.350

After obtaining the test results, attention is now focused on the inverse problem in which a set of material properties, each one of which lies within a given range that minimizes the error between test and simulated response is determined. To begin with, in the inverse analysis step, a set of values that corresponds to the material properties provided by a "client" is chosen. These values are listed in Table 6. A normal modal analysis is performed to determine the natural frequencies. The first three natural frequencies are f_1 = 102.07 Hz, f_2 = 125.84, and f_3 = 185.57 Hz. As the client is not sure about the exact values of the material properties, they are allowed to vary within permissible ranges. Figures 4 and 5 show the variation of the fundamental frequency with Young's modulus and mass density, respectively, of PCB. A value of 1.63×10^{-9} ton/mm^3 is used for mass density of PCB for generating Figure 4 and a value of 13, 500 MPa is use for the Young's modulus of PCB for generating Figure 5. The values of other parameters are E = 74, 586 MPa, r = 2.67×10^{-9} ton/mm^3 for the heat sinks, and E = 5, 392 MPa, r = 1.0×10^{-9} ton/mm^3 for the components. As can be seen from Figures 4 and 5, variation of the Young's modulus or the mass density or both of the PCB results in variation of the natural frequencies. Figure 6 shows the variation of acceleration response at the centre of PCB with variation of the Young's modulus. An increase in the Young's modulus results in an increase in the natural frequencies. As the fundamental natural frequency increases, the acceleration response decreases since they are inversely related. Variation of the acceleration response at the centre of PCB with variation of the structural damping coefficient is shown in Figure 7.

Table 6. Material properties provided by "client"

Material property	PCB	Heat sinks 1&2	Component 1&2
Young's modulus, Mpa	15,000	68,000	5,000
Mass density, ton/mm^3	1.4E-9	2.5E-9	1.0E-9
Poisson's ratio	0.3	0.3	0.2
Structural damping coefficient	0.07	0.07	0.08

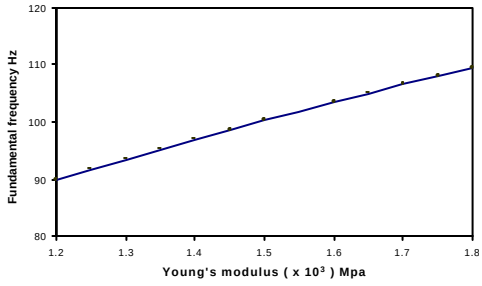


Figure 4. Variation of fundamental frequency with Young's modulus of PCB

Now, the inverse problem of finding a set of material properties that minimizes the error between test-simulation frequency response results is solved. This is done in two steps. In the first step, the natural frequencies that fall in the

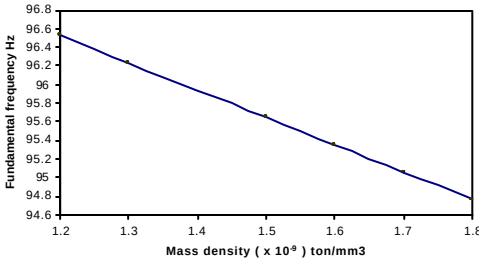


Figure 5. Variation of fundamental frequency with mass density of PCB

excitation range are matched (fitness is minimized). First, a sensitivity analysis is performed to find out those parameters (design variables) that influence the natural frequencies.

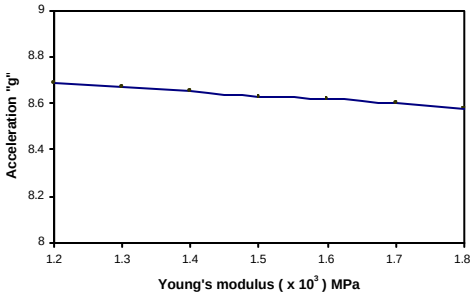


Figure 6. Variation of acceleration response at PCB centre with Young's modulus

The design variables for this step are the Young's modulus and mass density of the PCB, heat sinks, and components. The sensitivity analysis showed that the Young's modulus and density of the PCB material have relatively greater impact on the

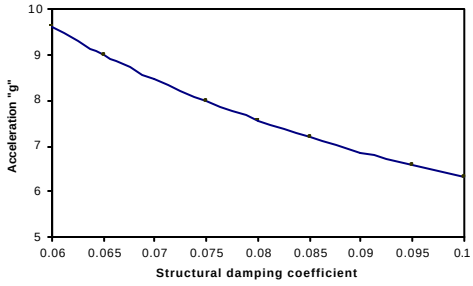


Figure 7. Variation of acceleration at PCB centre with damping coefficient of PCB

natural frequencies. Following sensitivity analysis, frequency matching is performed. The

Table 7. Design variables range for frequency matching

Material property	PCB	Heat sinks 1 & 2	Components 1&2
Young's modulus, E, MPa	$12{,}000 \le E \le 18{,}000$ Initial value: 17,000	$54{,}400 \le E \le 81{,}600$ Initial value: 75,000	$4{,}000 \le E \le 6{,}000$ Initial value: 5,400
Mass density, ρ, ton/mm^3	$1.2\times10^{-9} \le \rho \le 1.8\times10^{-9}$ Initial value: 1.6×10^{-9}	$2.0\times10^{-9} \le \rho \le 3.0\times10^{-9}$ Initial value: 2.5×10^{-9}	$0.8\times10^{-9} \le \rho \le 1.2\times10^{-9}$ Initial value: 0.9×10^{-9}

range of values of the design variables used in this step is given in Table 7. The searching procedure to minimize the fitness using the present modified mGA is shown in figure 8. The desired variables are found in 394 generations. In this example, both components are considered to be having the same material properties though, in practice, there is no such requirement and the procedure can still be followed. The corresponding frequencies, within the range of excitation frequencies, are f_1 = 95.4 Hz and f_2 = 118.5 Hz.

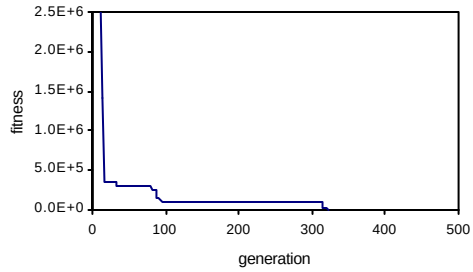


Figure 8 Searching procedure in frequency matching

After matching the first two frequencies, the design variables used in the previous step are kept constant at their respective values obtained at the end of the previous step. A sensitivity analysis is then performed to determine the design variables among the remaining parameters that influence the frequency response (at specified output locations). The structural damping coefficient is found to be the single most significant design variable. So, in matching frequency response, the structural damping coefficients of the PCB, the heat sinks, and the components are made the design variables. After performing a sensitivity

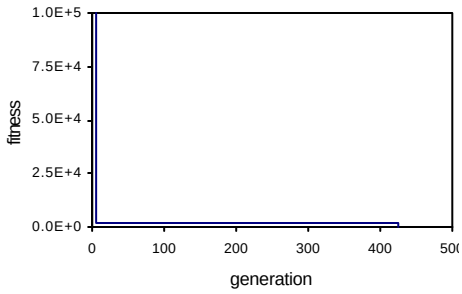


Figure 9. Searching procedure in frequency responds matching

analysis, matching of test-simulation frequency response results is carried out. The objective is to minimize the sum of the squares of the deviations between test and simulated results at the output locations. Figure 9 shows the searching procedure in match frequency responds using modified mGA. The true structural damping coefficients are found in 426 generations. Once material properties are determined this way, they may be used in the evaluation of relative performance of competing designs where the layout of components is different.

4. CONCLUSIONS

A modified mGA with intergenerational projection technique can greatly save forward evaluation and thus shorten the time to get convergence in inverse analysis. This method is efficient enough for finding global optimum in complex multi-optimum engineering problems.

Through sensitivity analysis, the design variables (material properties) that influence the natural frequencies and frequency response cab be determined. A two-step inverse procedure, matching frequency and matching frequency respond, can then be setup to identify material properties of a PCB with heat sinks and electronic components mounted on it by using the finite element model and modified mGA search. It is demonstrated that the two-stage inverse analysis is feasible in real-life applications.

REFERENCES

[1] J. H. Holland, "Adaptation in natural and artificial systems", *The University of Michigan Press*, Ann Arbor, 1975

[2] J. C. Santamarina, D. Fratta, "Introduction to discrete signals and inverse problems in civil engineering". *ASCE Press*, USA, 1998.

[3] B. Krishnan, S. R. Navin, "Inversion of composite material elastic constants from ultrasonic bulk wave phase velocity data using genetic algorithms". *Composites Part B*, 29(1998) pp171-180.

[4] X. Han, Liu G. R. and K. Y. Lam, A Quadratic Layer Element for Analyzing Stress Waves in FGMs and Its Application in Material Characterization. *Journal of Sound and Vibration.* Vol.236(2), 2000, pp.307-321.

[5] Liu G.R. and Chen S.C., Flaw Detection in Sandwich Plates Based on Time-harmonic Response Using Genetic Algorithm, *Computer Methods in Applied Mechanics and Engineering*, Vol. 190, Issue 42, 3 August 2001, pp. 5505-5514.

[6] Ishak, S. I, Liu, G. R. and Lim, S. P., Study on Characterization of Horizontal Cracks in Isotropic Beams. Journal of Sound and Vibration, 238(4), 2000. pp. 661-671.

[7] Ishak S. I, Liu G. R., Lim S. P. and Shang H. M., Locating and Sizing of Delamination in Composite Laminates Using Computational and Experimental Methods. *Composite Part B*. Vol. 32(4), 2001, pp. 287-298.

[8] Xu Y. G., Liu G. R., Wu Z. P., and Hunag X. M., Adaptive Multilayer Perceptron Networks for Detection of Cracks in Anisotropic Laminated Plates, *International Journal of Solids and Structures*, Vol.38, 2001, pp. 5625-5645.

[9] Liu G. R., Han X. and Lam K.Y., Material Characterization of FGM Plates Using Elastic Waves and an Inverse Procedure. *Journal of Composite materials*. Vol.35, No.11, 2001, pp. 954-971.

[10] Liu G. R., Han X., Xu Y. G. and Lam K. Y., Material Characterization of Functionally Graded Material by Means of Elastic Waves and a Progressive-Learning Neural Network, *Composites Science and Technology*, Volume 61, Issu10, August 2001, pp. 1401-1411.

[11] Xu Y. G., Liu G. R., Wu Z. P., A Novel Hybrid Genetic Algorithm Using Local Optimizer Based on Heuristic Pattern Move, *Applied Artificial Intelligence*, Vol. 15(7), 2001, pp.601-631

[12] Liu G. R., Xu Y. G. and Wu Z. P., Total Solution for Structural Mechanics Problems. *Computer Methods in Applied Mechanics and Engineering.* Vol. 191, pp. 989-1012. 2001.

[13] Ishak S. I, Liu G. R., Lim S. P. and Shang H. M., Characterization of Delamination in Beams using Flexural Wave Scattering Analysis. *ASME Journal of Vibration and Acoustics*, 123(4), OCT 2001, pp. 421-427

[14] Liu, G.R., Han, X. and K.Y. Lam, A Combined Genetic Algorithm and Nonlinear Least Squares Method for Material Characterization Using Elastic Waves. *Computer methods in applied mechanics and Engineering*, 191, 1909-1921, 2002.

[15] Z L Yang, G R Liu and K Y Lam, 'An inverse procedure for crack detection using integral strain measured by optical fibres'. *Smart Mater. Struct*. 10, 2001.

[16] Gen, M. and Cheng, R. W., "Genetic Algorithm and Engineering Design", John Wiley & Sons, Inc. New York, 1997

[17] S. H. Venkatasubramanian, G. R. Liu and C. Lu, "PARAMETER IDENTIFICATION AS AN INVERSE PROBLEM IN TEST-SIMULATION CORRELATION", *JSV*, 2002 (to appear)

INVERSE DESIGN AND OPTIMIZATION

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

INVERSE HEAT TRANSFER FOR OPTIMIZATION AND ON-LINE THERMAL PROPERTIES ESTIMATION IN COMPOSITES CURING

Alexandros A. Skordos
Advanced Materials Department
Cranfield University
Cranfield, Bedford, United Kingdom
a.a.skordos@cranfield.ac.uk

Ivana K. Partridge
Advanced Materials Department
Cranfield University
Cranfield, Bedford, United Kingdom

ABSTRACT

This paper presents the development and application of a heat transfer inversion procedure to the cure of thermoset based composites based on genetic algorithms. The procedure is utilized for process optimization applied to the curing of carbon fiber reinforced composites. The optimization objective is the selection of an appropriate cure schedule so that the duration of the curing is minimized subject to constraints related to the thermal gradients developed during the cure.

An alternative use of inversion concerns the integration of monitoring signals with modeling. Inversion is utilized to alter on-line the thermal properties used in the direct model so that monitoring results coincide with simulation predictions. This procedure is applied to the curing of a carbon fiber reinforced thermoset based composite, using thermal conductivity as the variable thermal property.

NOMENCLATURE

c_p	Specific heat capacity
h	Surface heat transfer coefficient
H_{tot}	Total heat of the curing reaction
$\mathbf{K}$	Thermal conductivity tensor
N_i	Interpolation function
$\hat{\mathbf{n}}$	Surface vector
q'	Boundary heat flux
$\hat{\mathbf{r}}$	Spatial coordinate
S_1	Temperature boundary condition surface
S_2	Heat flux boundary condition surface
S_3	Convection boundary condition surface
$\tilde{S}_{ij}$	Interface
T	Temperature
T'	Boundary temperature
T_∞	Ambient temperature
t	Time
t_c	Time to reach conversion 0.84
v_f	Fiber volume fraction
α	Degree of cure
Δt	Time step
θ	Time discretization parameter
ρ	Density of the composite
ρ_r	Resin density

INTRODUCTION

In recent years the need for predictive modelling and for in-situ real time monitoring of composites manufacturing processes has arisen and been met by the development of a family of appropriate techniques. Models representing various aspects of processing have been developed and applied to the majority of processing techniques. Heat transfer models have been implemented in order to simulate the curing phenomena in autoclave processing [1], resin transfer moulding [2], pultrusion [3] and filament winding [4]. Provided that these models are combined with appropriate cure kinetics subroutines [5,6], they offer the ability to calculate the spatial distributions of temperature and of the degree-of-cure and their evolution with time during the curing. Alongside with simulation, process monitoring methods such as thermal [7], dielectric [8,9], fibre optic [10] and acoustic cure monitoring [11], have begun to be implemented in an industrial environment.

Both monitoring and modelling are valuable for optimising the curing process. The predictive ability of the simulation can be used as a part of the process design, while monitoring constitutes a potential tool for on line control. However, both approaches present some inherent drawbacks.

Accurate modelling requires an extensive knowledge of material properties and process characteristics. This may be impossible in some cases due to limited reproducibility of some of the process conditions, or due to the prohibitive costs associated with the knowledge acquisition step. Similarly, monitoring involves insertion of a sensor in some critical area of the component, which composite manufacturers and end users are reluctant to adopt.

A method to overcome these limitations arises from a combination of modelling and monitoring. In the present paper a scheme which combines heat transfer modelling and thermal and dielectric cure monitoring is presented. An inversion of the heat transfer model based on a genetic algorithm is applied to data gathered by monitoring, in order to calculate some of the properties or process characteristics. Thus, an estimation of those modelling parameters that are most difficult to predefine can be performed, in accordance with the results of monitoring. Subsequently the direct model can be solved in order to obtain the global picture of the cure.

The inversion procedure developed here is also applied to purely predictive process design where the minimisation of the cure duration is the objective of the optimisation.

DIRECT MODEL

Heat Transfer Problem

The model concerns the curing of a carbon fiber reinforced composite in a resin transfer mould. In this process dry fabric is placed in a rigid cavity, resin is infused under pressure and vacuum and the curing takes place with further heating of the mould. When forced convection does not occur heat conduction is the only heat transfer mechanism relevant to composites cure. Accordingly, the governing equation is:

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot \mathbf{K} \nabla T + (1 - v_f)\rho_r H_{tot} \frac{d\alpha}{dt} \qquad (1)$$

The second term in the right side of (1) expresses the heat generated by the curing reaction. This equation is accompanied by a set of boundary conditions. In the general case there are three possible boundary conditions: -(i) prescribed temperature:

$$T(\hat{\mathbf{r}}, t) = T'(\hat{\mathbf{r}}, t), \quad \hat{\mathbf{r}} \in S_1 \qquad (2)$$

(ii) prescribed heat flux

$$\hat{\mathbf{n}} \cdot \mathbf{K} \nabla T(\hat{\mathbf{r}}, t) = q'(\hat{\mathbf{r}}, t), \quad \hat{\mathbf{r}} \in S_2 \qquad (3)$$

(iii) and convection

$$\hat{\mathbf{n}} \cdot \mathbf{K} \nabla T(\hat{\mathbf{r}}, t) = h(T(\hat{\mathbf{r}}, t) - T_\infty), \quad \hat{\mathbf{r}} \in S_3 \qquad (4)$$

As in composites curing the heat transfer problem is multimaterial, i. e. thermal properties and especially thermal conductivity present a discontinuity at the tool-composite interface, a separate boundary value problem, of the type expressed by Eqs. (1)-(4), is formed over each subdomain. An additional set of interfacial conditions that ensures temperature and heat flux continuity is defined as follows:

$$T_i(\hat{\mathbf{r}}, t) = T_j(\hat{\mathbf{r}}, t), \quad \hat{\mathbf{r}} \in \tilde{S}_{ij} \qquad (5)$$

$$\hat{\mathbf{n}} \cdot \mathbf{K_i} \nabla T_i(\hat{\mathbf{r}}, t) = \hat{\mathbf{n}} \cdot \mathbf{K_j} \nabla T_j(\hat{\mathbf{r}}, t), \quad \hat{\mathbf{r}} \in \tilde{S}_{ij} \qquad (6)$$

here the indices i and j denote areas of different materials.

Eqs. (1)-(6), accompanied by an appropriate cure kinetics model and a set of thermal properties models expressing thermal conductivity, specific heat capacity and density as functions of degree of cure and temperature, suffice for the complete description of the curing process in a resin transfer mold.

Finite Elements Formulation

In order to solve the problem using finite elements the domain is divided into a number of elements that connect at G nodal points. The unknown variable is approximated as a linear combination of a set of G functions as follows:

$$\overline{T} = \sum_{i=1}^{G} T_i(t) N_i(\hat{\mathbf{r}}), \quad T_i(t) = T'(\mathbf{r^i}, t) \text{ if } \mathbf{r^i} \in S_1 \qquad (7)$$

Function N_i is equal to unity at node i and vanishes at all other nodes and within all the elements to which node i does not connect.

By employing the above approximation, adopting a finite difference scheme to express time derivatives and expressing the heat transfer

problem by its weighted residuals equivalent, the following system of equations is obtained:

$$(\mathbf{M}+\theta\Delta t\mathbf{L})\mathbf{T}^{\mathbf{n+1}}-(\mathbf{M}-(1-\theta)\Delta t\mathbf{L})\mathbf{T}^{\mathbf{n}}-\Delta t\mathbf{F}=0 \qquad (8)$$

where

$$M_{ji}=\int_{\Omega}\rho c_p N_j N_i d\Omega \qquad (9)$$

$$L_{ji}=\int_{\Omega}\nabla N_j\cdot\mathbf{K}\nabla N_i d\Omega+\int_{S_3}N_j h N_i dS \qquad (10)$$

$$F_j=-\int_{S_2}N_j q' dS+\int_{S_3}N_j h T_\infty dS+\int_{\Omega}N_j(1-v_f)\rho_r H_{tot}\frac{d\alpha}{dt}d\Omega \qquad (11)$$

The system of equations described by (8)-(11) can be solved for each time step in order to calculate the distributions of temperature and degree of cure and their evolution with time.

Model Implementation and Validation

A model based on the principles described previously was developed in order to simulate the curing stage of resin transfer molding. The model comprises a core finite elements solver and a set of submodels simulating the cure reaction kinetics, and the changes in thermal properties, i.e. specific heat capacity, thermal conductivity and density, during the cure. The algorithm starts from the initial temperature and degree of cure distributions which are fed into the cure kinetics model. The cure kinetics model produces values for the reaction rate which are fed into the finite elements solver in order to account for heat generation and values for the updated degree of cure which are output to the thermal properties submodels. The three thermal properties submodels use the values of conversion and initial temperature to compute the values of the thermal properties within the different elements of the model. The results are sent to the finite elements solver, which, taking into account the boundary conditions and the initial temperature distributions, computes the resulting temperature distribution. This procedure is repeated for a number of time steps by replacing the initial conditions with the temperature output of the finite elements model and the updated degree of cure distributions as calculated by the cure kinetics and by updating the boundary conditions.

The model was applied to the curing of an RTM6 epoxy resin/ T300 continuous carbon fiber reinforced composite. The material properties submodels were appropriate to the specific materials. The cure kinetics model operates by direct interpolation in the degree of cure-temperature phase space applied to experimental differential scanning calorimetry (DSC) data and is analyzed in detail elsewhere [5].

The specific heat capacity submodel operates in a similar way using experimental data produced by modulated differential scanning calorimetry. The experimental data for the resin in the temperature-fractional conversion phase space are illustrated in Fig. 1. The step change occurring during the cure of the resin marks the glass transition of the thermosetting material. The specific heat capacity of the carbon fiber was found to be a linear function of the temperature as follows:

$$c_p=0.0023\,T+0.765 \qquad (12)$$

In the above equation the units of temprature are ^oC and of specific heat capacity J/g/^oC. Once the values corresponding to the resin and the reinforcement have been computed by the submodel the law of mixtures is employed to calculate the value for the composite.

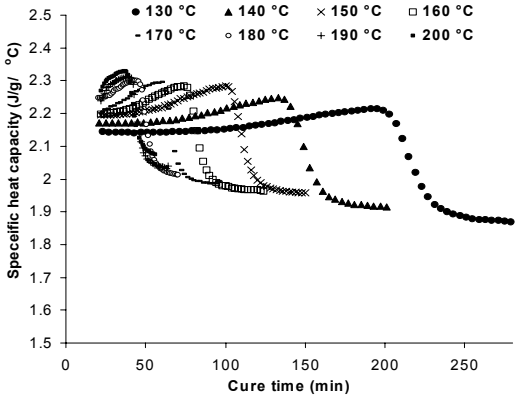


Fig. 1 Specific heat capacity versus cure time during isothermal cures

The thermal conductivity of the anisotropic composite material is computed using an appropriate geometry based model [12] that

combines values for the resin and the carbon fiber. The values for the resin were obtained experimentally using a technique that measures the thermal conductivity of the thermoset while it cures. Details of the method are given elsewhere [13]. The dependence of the resin thermal conductivity to temperature and fractional conversion can be expressed as follows:

$$K = 0.0008\ T\alpha^2 - 0.0011T\alpha - 0.0002T - 0.0937\alpha^2 + 0.22\ \alpha + 0.12 \quad (13)$$

where temperature is given in ^oC and thermal conductivity in W/m/^{o}C. The longitudinal thermal conductivity of T300 carbon fiber can be expressed as follows [14]:

$$K = 4.8 + 0.0074\ T \quad (14)$$

where temperature is given in ^oC and thermal conductivity in W/m/^{o}C. The radial thermal conductivity of T300 carbon fiber is 0.84 W/m/^{o}C.

The density model takes into account thermal expansion of the resin and of the fiber and curing shrinkage. The model assumes that resin shrinkage is proportional to the progress of the curing reaction and uses the law of mixtures in order to calculate the composite density. The density of uncured RTM6 resin at ambient temperature is 1.117 g/cm^3 and the total volumetric chemical shrinkage 4.9 %, while the volumetric thermal expansion coefficient above glass transition is 4.08 10^{-4} $^oC^{-1}$ and below glass transition 1.62 10^{-4} $^oC^{-1}$ [15]. The expansion coefficient of carbon T300 is 5 10^{-6} $^oC^{-1}$ and its density at ambient temperature 1.8 g/cm^3.

The model implementation was tested against experimental data obtained during the cure of a composite. The experimental equipment used is shown in Fig.2. The dimensions of the mold cavity were 800 mm x 340 mm x 3 mm. The sides of the cavity were sealed using silicone rubber while the tool was closed using a glass plate and a set of stiffeners. Heating is achieved by an array of heating elements placed under the mold cavity. The specific experimental configuration was selected in order to reduce the heat transfer problem to one dimension. This enables an easier implementation of the inversion that follows to be made. The carbon fabric used had a surface density of 816 g/m^2 and comprised three layers of unidirectional fiber tows at angles +45^o,-45 ^{o} and 0^o. Four layers of this fabric were used in the cavity to achieve a fiber weight fraction of 0.69. The total sequence of unidirectional tow plies was $[+45/-45/0/0/-45/+45]_{2S}$. Resin filling was carried out at 120 ^oC. After completion of the filling, heating at 1.5 ^oC/min was performed up to 160 ^oC, and then the temperature was kept constant. Two thermocouples (k-type) which measure the temperature at the top of the composite and at the mid-thickness were placed in the center of the curing component.

The modeling domain considered comprised the composite and the glass top plate. The bottom of the composite was considered to follow the thermal profile measured by the tool temperature sensor. The rubber seal was assumed to act as an insulator (zero heat flux) on the sides of the composite component. Natural air convection was considered on the top and sides of the glass plate. The initial condition was considered to be zero fractional conversion and uniform temperature after the end of filling. The thermal properties of the glass plate are given in table 1.

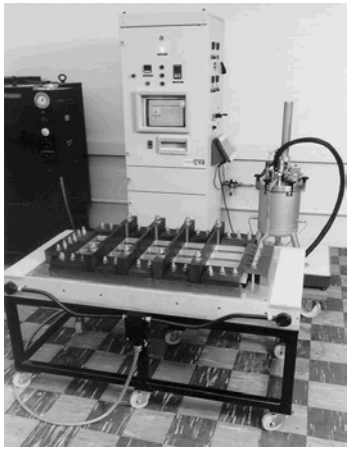


Fig. 2 The resin transfer molding facility

Table 1. Properties of the glass top plate

$c_p\left(J/g/^oC\right)$	0.84
$K\left(W/m/^oC\right)$	0.78
$\rho\left(g/cm^3\right)$	2.7
$h\left(W/m^2\right)$	8.5

The convergence of the three dimensional simulation was investigated. The convergence study indicated an optimum time step of 45 sec,

an optimum element size of 0.05 mm in the thickness direction and 20 mm in the length and width directions of the component. A comparison of the results of the three dimensional case with the results of an one dimensional version of the model where only the thickness direction is considered showed that the 1-D model represents the curing of the specific adequately. The results of the model are set against experimental results in Fig. 3. It can be observed that the agreement achieved is satisfactory.

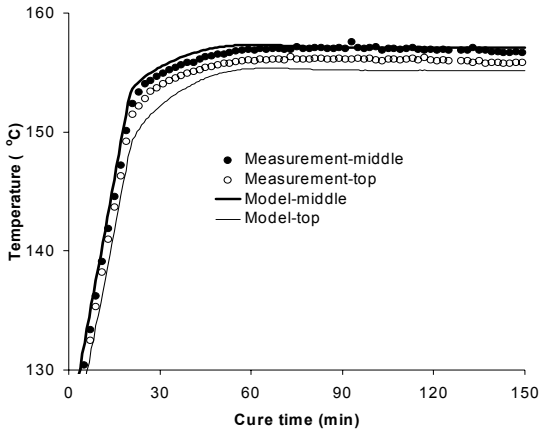


Fig. 3 Experimental and simulation temperature

INVERSION PROCEDURE

The inversion of the heat transfer model was performed using a genetic algorithm. The term describes a family of evolutionary optimization methods which involve a population of points in the search space of solutions (generation) and employ a performance sensitive selection procedure and crossover and mutation operations in order to reproduce a new population. The members of the population (individuals) are usually encoded in bit strings and the algorithm is iterated until some convergence criteria are met. A code which performs these procedures and uses as a direct model the one dimensional heat transfer simulation presented previously has been implemented. The operation of the algorithm is as follows.

A number (N) of initial values for each parameter corresponding to the first generation are created using a random number generator. Then the 1-D simulation is executed N times and the results corresponding to each individual are stored in a file. The fitness of the individuals is calculated subsequently by comparing the output of the direct model with a target, which drives the inversion. The form of the fitness function and the inversion target are specific to the application of the inversion procedure.

The next step of the algorithm is the encoding of the individuals. In this stage, each individual is translated into a unique binary string. The length of the string defines the accuracy of the algorithm. Subsequently, the individuals are sorted according to their fitness. A limited number (m) of individuals with the best fitness are passed directly into the new generation. The rest of the individuals of the new generation are produced with a combination of selection, crossover and mutation. The selection is performed using a standard procedure called "roulette wheel". In this procedure each individual is assigned a slice of a circular wheel, the size of the slice being proportional to the fitness of the individual. Two random numbers between 0 and 360 are generated and the individuals corresponding to them are selected. The application of a uniform crossover operation to the two selected individuals follows. In this operation a predefined probability (exchange probability) is compared with a random number between 0 and 1 at each bit of the binary string. If the number is greater than the exchange probability the two selected individuals exchange their bit values, otherwise the values are preserved. At the end of this operation two new individuals have been produced, each of them containing parts of the old individuals. Subsequently, a mutation operation is applied to the two new individuals. In this stage a very low probability (mutation probability) is compared with a random number for each bit of the two new strings. If the mutation probability is greater than the random number, the bit of the string switches from 1 to 0 or from 0 to 1, otherwise it remains unchanged.

When N-m individuals have been produced, the selection-reproduction procedure stops. These N-m individuals together with the m best individuals of the previous generation form the new generation and the individuals are decoded back to decimal parameters.

At that point the convergence of the algorithm is tested according to a criterion specific to the application which is applied to the best individual. If convergence has been reached, the algorithm outputs the appropriate data and exits. Otherwise the execution of the direct heat transfer model for

the new individuals is performed and the whole procedure of fitness calculation, sorting, encoding, selection, reproduction and decoding is iterated until convergence is achieved.

THERMAL PROFILE OPTIMISATION

A straightforward application of the inversion procedure described previously is the optimization of the thermal profile applied during the cure. The thermal profile comprises a linear heating up and an isothermal segment, thus can be characterised by two parameters:- (i) the ramp up rate and (ii) the isothermal temperature which can be the subject of the optimization. The fitness function was selected so that it rewards the parameter values which reduce the duration of the curing stage, i.e. it increases as the time to reach a fractional conversion of 0.84 in all elements of the component decreases. This is implemented by the function:

$$\text{Fitness} = 1/t_c \tag{15}$$

which is subject to the constraint:

$$\left|\frac{dT}{dz}\right|_{max} < 2.5\ ^oC/mm, \text{ for } \alpha > 0.6 \tag{16}$$

The meaning of the constraint is that for fractional conversions at which the material has reached the rubbery state and residual strain can build up, the maximum thermal gradient must be kept lower than the thermal gradients achieved during conventional cure schedules. The implementation of the constraint is performed by excluding from the selection and reproduction procedures all individuals which violate it.

The values of the optimisation parameters are selected within practically meaningful ranges, i.e. a heat up rate from 0 to 4 ^oC/min and an isothermal temperature from 150 to 190 ^oC. The algorithm is considered to have converged when the individuals of a generation have very small variation, i.e. the average percentage difference between the members of the population and the average is lower than 0.5 %.

Thirteen individuals have been used, each of them represented by a string of one hundred digits. The three best individuals were directly passed to the next generation. The exchange probability used was 40% and the mutation probability 2%.

The convergence of the optimisation is illustrated in Figs. 4 and 5. The problem involves only two parameters, thus convergence occurs very fast within six generations. The optimal values found are a heating rate of 3 ^oC/min and an isothermal temperature of 169 ^oC. Global optimality cannot be ensured, the effectiveness of the optimisation can be evaluated by comparing the resulting cure completion time of 64.5 min with the cure completion time of the conventional thermal profile described in previously which was 87.5 min. A reduction in cure cycle time 26% is achieved, which could have a very significant impact on the total cost of production.

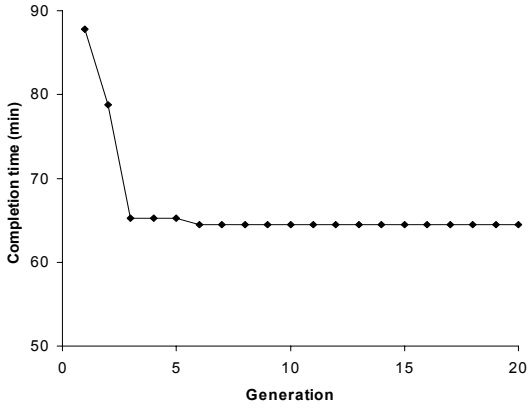


Fig. 4 Cure completion time vs generation number

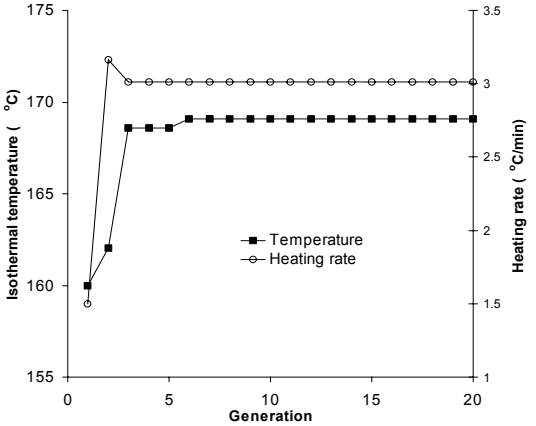


Fig. 5 Isothermal temperature and heating rate vs generation number

MODELLING-MONITORING INTEGRATION

The second application of the inversion procedure examined here concerns the use of on-line temperature measurement results in order to determine the dependence of the composite thermal conductivity on the fractional conversion and temperature. Subsequently, the distributions of the temperature and the degree of cure are calculated.

Temperature measurements performed at the mid-depth of a curing carbon/RTM6 composite are used as the target of the genetic algorithm. The variable parameters are the coefficients of a polynomial, which expresses the dependence of the composite thermal conductivity on fractional conversion and temperature as follows:

$$K = (\alpha^2\,\mathrm{Par1} + \alpha\,\mathrm{Par2} + \mathrm{Par3})(\mathrm{T}\,\mathrm{Par4} + \mathrm{Par5}) \quad (17)$$

The fitness of an individual is calculated as follows:

$$\mathrm{Fitness} = \frac{1}{\sum_{i=1}^{Q} \left| T_i^{mid} - T_M \right|} \quad (18)$$

where T_i^{mid} is the temperature at time step i at the middle node, Q is the number of time steps and T_{M_i} is the temperature measurement at time step i.

The parameters of the finite element model have the values indicated by the convergence study. In the genetic algorithm 26 individuals with a string length of 100 bits were used. Five individuals were passed directly to the next generation, the crossover and mutation probabilities were identical to those used in the thermal profile optimization runs. The range of the five parameters to be estimated was –1 to 1.

The convergence behavior of the algorithm is illustrated in Figs. 6 and 7. The algorithm converges after about 30 generations. Note that the problem in that case involved five parameters and about 900 iterations of the direct model were required for their estimation. A search method equivalent in terms of computational time would have resulted in a accuracy of about 0.5 in the parameter estimation. The solution of the inverse problem is:

$$K = (-0.125\alpha^2 + 0.117\,\alpha - 0.094)(0.358 - 0.034\,\mathrm{T}) \quad (17)$$

Using this model for the calculation of thermal conductivity, the distributions of temperature and degree of cure can be calculated. Their comparison with the results of the direct simulation, which was shown to be in agreement with the thermal monitoring results utilized here for the inversion, is illustrated in Fig. 8. It can be observed that the monitoring-modeling scheme predicts the global distribution of temperature and degree of cure with satisfactory accuracy. The average error in temperature estimation is 0.29 ^{o}C and in fractional conversion determination 0.0019. The error in temperature is lower than the accuracy of the direct model whereas the error in fractional conversion estimation is very low due to the fact that the higher differences in temperature between the inversion results and the direct model occur at low conversions when the reaction rate is very low.

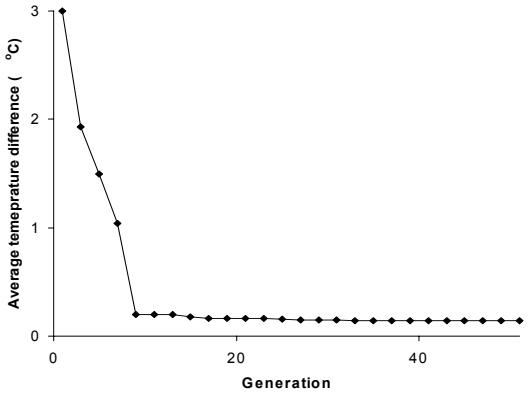


Fig. 6 Difference between measurement and inverse modeling results vs generation

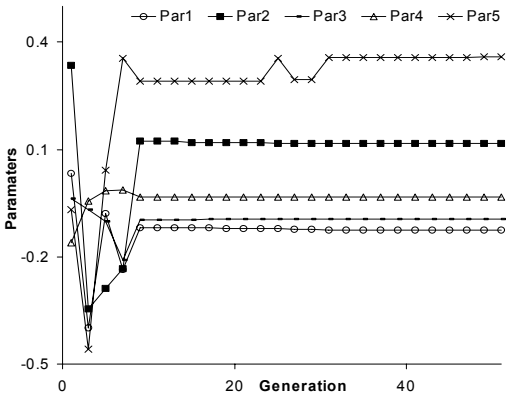


Fig. 7 Parameters of the thermal conductivity polynomial vs generation number

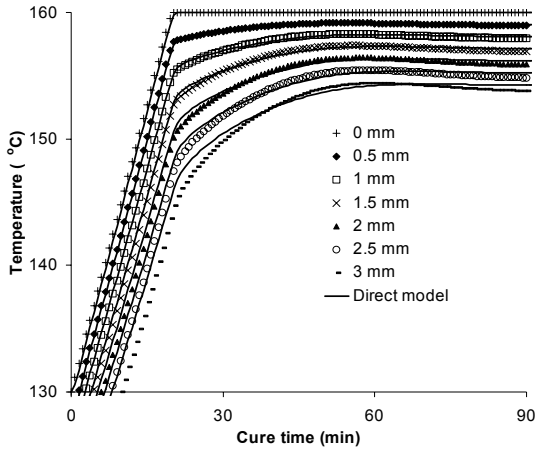


Fig. 8 Temperature vs cure time at different levels as resulting from the inversion procedure and from the direct model

CONCLUSIONS

The inversion procedure based on genetic algorithms presented here can be used for optimization and monitoring-modeling integration in composites manufacturing. Optimal cure schedules with respect to total curing process duration were found for a specific carbon fiber/RTM6 composite component. The monitoring-modeling combined scheme offers the possibility to infer temperature and degree of cure distributions from limited local thermal monitoring signals. Local monitoring results combined with the inversion procedure result in a very accurate estimation of the temperature and degree of cure evolutions during the cure. Both applications of heat transfer inversion can be extended to the case of complex components where two or three-dimensional modeling is required. Application of the monitoring-modeling combined scheme may be limited due to computing time in the case of fast cures, however the majority of advanced composite components are subject to several hours long cure profiles.

REFERENCES

1. A. C. Loos and G. S. Springer, Curing of epoxy matrix composites*, Journal of Composite Materials*, **17**, 135 (1983)

2. D. M. Gao, F. Trochu and R. Gauvin, Heat transfer analysis of non-isothermal resin transfer molding by the finite element method, *Materials and Manufacturing Processes*, **10**, 57 (1995)

3. R. Gorthala, J. A. Roux and J. G. Vaughan, Resin flow, cure and heat transfer analysis for pultrusion process, *Journal of Composite Materials,* **28**, 486 (1994)

4. S. Y. Lee and G. S. Springer, Filament winding cylinders: I. Process model*, Journal of Composite Materials*, **24**, 1275 (1990)

5. P. I. Karkanas and I. K. Partridge, Cure modeling and monitoring of epoxy/amine resin systems. I. Cure kinetics modeling, *Journal of Applied Polymer Science*, **77**, 1419 (2000)

6. A. A. Skordos and I. K. Partridge, Cure kinetics modelling of epoxy resins using a non-parametric numerical procedure, *Polymer Engineering and Science,* **41**, 793 (2001)

7. G. Lebrun, R. Gauvin and K. N. Kendal, Experimental investigation of resin temperature and pressure during filling and curing in a flat steel, *Composites*, *A*, **27**, 347 (1996)

8. D. E. Kranbuehl, P. Kingsley, S. Hart, G. Hasko, B. Dexter and A. C. Loos, In-situ sensor monitoring and intelligent control of the resin transfer molding process, *Polymer Composites*, **15**, 299 (1994)

9. G. M. Maistros and I. K. Partridge, Monitoring autoclave cure in commercial carbon fibre/epoxy composites *Composites, B*, **19**, 245 (1998)

10. D. L. Woederman, J. K. Spoerre, K. M. Flynn and R. S. Parnas, Cure monitoring of the liquid composite molding process using fiber optic sensors, *Polymer Composites*, **18**, 133 (1997)

11. T. M. Whitney and R. E. Green Jr, *Ultrasonics,* Cure monitoring of carbon epoxy composites: An application of resonant ultrasound spectroscopy **34** , 347 (1996)

12. J. D. Farmer and E. E. Covert, Thermal conductivity of a thermosetting advanced composite during its cure, *Journal of Thermophysics and Heat Transfer*, **10**, 467 (1996)

13. A. A. Skordos, PhD Thesis, *Modelling and monitoring of resin transfer moulding*, Cranfield University, UK, 2000

14. T. Yamane, S. Katayama, M. Todoki and I. Hatta, *The measurement of thermal conductivity of carbon fibers*, in Thermal Conductivity 22 by T. W. Tong, p 313, Technomic Publishing, USA, 1994

15. J. A. Holmberg, *Influence of post cure and chemical shrinkage on springback of RTM U-beams*, SICOMP Technical Report 97-004, 1997

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

INVERSE DESIGN OF GAS TURBINE COMPONENTS

Andreas Troxler
Seminar for Applied Mathematics
ETH Zurich
Zurich, Switzerland
atroxler@sam.math.ethz.ch

ABSTRACT

In this paper we present a method for the solution of the target pressure problem of inverse aerodynamic shape design: "Given a static pressure distribution along the sidewalls of a flow device, find the corresponding shape producing this pressure distribution". Two-dimensional and axis-symmetric flows governed by the steady compressible Euler equations are considered, although the method is extendible to three-dimensional configurations. The flow equations are solved on a moving curvilinear body-fitted grid. Grid motion is governed by a parabolic system of differential equations, subject to a novel wall modification procedure. Block structured grids allow for reasonably complex geometries. The capabilities of the method are demonstrated by inverse redesign of a curved annular compressor diffuser with a cooling air bleed-off diffuser. Another simple test case demonstrates the ability of the method to solve transonic problems as well. Remarkable robustness and convergence properties of the method are observed, both in terms of iteration count and execution time.

NOMENCLATURE

a	speed of sound
E	total energy
$\mathbf{F}$	flux tensor
H	total enthalpy
$\mathbf{n} = (n_1, n_2)^T$	outward pointing unit normal
p	static pressure
p^{target}	target pressure
R	residual vector
U	vector of unknowns
$\mathbf{U}$	state vector
$\mathbf{u} = (u, v)^T$	cartesian velocity components
$\mathbf{x} = (x, y)^T$	cartesian coordinates
$\dot{\mathbf{x}}_s$	cell surface velocity
ρ	density
σ, ψ	computational coordinates
Ω	control volume, computational cell

Superscripts in parentheses denote (pseudo-)time indices.

INTRODUCTION

Modern turbo machinery design processes do not merely aim at efficiency increase of the devices, but also at minimizing human interaction and thereby reducing turnaround time. One possible approach is to couple a flow analysis tool to a constrained optimization method [3]. However, constraints on the flow field (eg. given pressure distribution along side walls) are tractable more efficiently by inverse approaches, eventually coupled to an optimization procedure. Therefore, fast, robust, and versatile inverse methods are required. This is the scope of the present paper.

Inverse methods can be divided into two broad categories. In one class we find methods based on stream-line coordinates ([5], [10]). The computational domain remains fixed during the computation, since sidewalls are streamlines. The governing equations assume a simple form, and the coupling between flow and grid is intrinsic. Consequently these methods are usually quite fast. However, streamline coordinates are not always the best choice. For example, difficulties arise near stagnation points, where the distance between adjacent streamlines becomes large. Furthermore, in realistic three-dimensional configurations, secondary flows will produce a rather complicated streamline pattern which is not suitable as a coordinate system.

More flexibility with respect to flow models and geometry is provided by methods based on iterative wall modification and re-gridding procedures (eg. [8], [1], [2]). The basic ingredients of such methods are a scheme for the unsteady flow equations on a (moving) grid, an automatic re-

gridding technique, and a wall modification procedure which updates the wall shape in such a way that in steady state the target pressure distribution is met and the wall is flow-aligned. Often an explicit time marching technique is used to reach the steady state. It has been observed that convergence is impaired if the coupling of fluid and grid motion is neglected [2]. In some cases, the wall update even has to be under-relaxed.

The approach followed here tries to combine the strengths of both classes of methods. Geometric flexibility is provided by general curvilinear body-fitted block-structured moving grids. Grid motion is governed by a system of time dependent parabolic grid update equations, subject to boundary conditions which generate streamline aligned sidewalls in steady state. The coupling between flow and grid is not as strong as within stream-line based methods. Nevertheless, comparable convergence speed can be obtained by an implicit time stepping scheme which drives flow and grid equations simultaneously to steady state (in a strongly coupled manner).

FORMULATION OF THE METHOD

In this section the ingredients of the method are described in detail. The first two subsections are devoted to the flow equations on a moving grid and the equations governing grid motion. The flow boundary conditions as well as the wall modification procedure are presented next. The section is concluded by the description of the spatial and temporal discretization schemes.

Flow Equations

We consider a (quadrilateral) control volume $\Omega(t)$ in a moving curvilinear coordinate system $\mathbf{x}(\sigma,\psi,t)$. The computational coordinates (σ,ψ) remain fixed. For this control volume the compressible Euler equations of gas dynamics are [11]:

$$\partial_t \int_{\Omega(t)} \mathbf{U} d\Omega + \int_{\partial\Omega(t)} (\mathbf{F} - \mathbf{U}\cdot\dot{\mathbf{x}}_s^T)\cdot\mathbf{n} dS = 0. \quad (1)$$

For planar flow the state vector $\mathbf{U}$ and the flux tensor $\mathbf{F}$ are

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{bmatrix} = \begin{bmatrix} \rho u & \rho v \\ \rho uu + p & \rho vu \\ \rho uv & \rho vv + p \\ \rho uH & \rho vH \end{bmatrix}.$$

There are two differences with respect to the Euler equations on a fixed mesh: First, the cell volume $\int_{\Omega(t)} d\Omega$ is time dependent. Second, the advective fluxes depend on the relative flow velocity and therefore the cell surface velocity $\dot{\mathbf{x}}_s$ has to be subtracted. $\dot{\mathbf{x}}_s$ is not specified yet. This is discussed next.

Grid Generation Procedure

The method presented here is quite flexible concerning the choice of a grid generation procedure. In the present paper, we employ a parabolic system of grid update equations similar to standard Poisson-type grid generation equations, see eg. [11]. This ensures that the effect of grid perturbations decays rapidly both in space and in time. Moreover, the velocity of the cell surfaces $\dot{\mathbf{x}}_s$ is easily assessed from the grid point velocity defined by

$$\mathbf{x}_t = g_{22}(\mathbf{x}_{\sigma\sigma} + P\mathbf{x}_\sigma) + g_{11}(\mathbf{x}_{\psi\psi} + Q\mathbf{x}_\psi) - g_{12}\mathbf{x}_{\sigma\psi}, \quad (2)$$

where P and Q are control functions, and

$$\begin{aligned} g_{11} &:= \mathbf{x}_\sigma \cdot \mathbf{x}_\sigma, \\ g_{22} &:= \mathbf{x}_\psi \cdot \mathbf{x}_\psi, \\ g_{12} &:= \mathbf{x}_\sigma \cdot \mathbf{x}_\psi. \end{aligned}$$

If a reasonable initial geometry and good mesh are provided (this might be the case when an existing design has to be improved by inverse redesign) the control functions P and Q may be obtained from the initial mesh by setting $\mathbf{x}_t = 0$ in (2):

$$\begin{bmatrix} g_{22}x_\sigma & g_{11}x_\psi \\ g_{22}y_\sigma & g_{11}y_\psi \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} 2g_{12}x_{\sigma\psi} - g_{22}x_{\sigma\sigma} - g_{11}x_{\psi\psi} \\ 2g_{12}y_{\sigma\psi} - g_{22}y_{\sigma\sigma} - g_{11}y_{\psi\psi} \end{bmatrix}. \quad (3)$$

With this choice of control functions (2) becomes rather an equation for the grid update than a grid generation equation. In fact, for a direct computation (fixed side walls) the mesh remains unchanged.

Depending on how the initial grid has been generated the control functions P and Q may have to

be smoothed by the following scheme which is applied several times [9]:

$$P_{i,j}^{(k+1)} = \omega P_{i,j}^{(k)} + \frac{1-\omega}{2}\left(P_{i,j+1}^{(k)} + P_{i,j-1}^{(k)}\right)$$
$$Q_{i,j}^{(k+1)} = \omega Q_{i,j}^{(k)} + \frac{1-\omega}{2}\left(Q_{i+1,j}^{(k)} + Q_{i-1,j}^{(k)}\right).$$

The smoothing parameter ω should lie between 0 and 1. Note that P is usually only smoothed in ψ direction (index j), since smoothing P in σ direction would lead to a smoothing of the mesh density.

Flow Boundary Conditions

The numerical boundary condition scheme used here is based on spatial characteristic extrapolation [6]. The Jacobian matrix of the surface normal flux $\mathbf{n}\cdot\mathbf{F}(\mathbf{U})$ with respect to the conservative state variables $\mathbf{U}$ is decoupled into characteristic fields. For planar compressible Euler flows the resulting eigenvalues (characteristic speeds) λ_i are given by [6]:

$$\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\end{bmatrix} = \begin{bmatrix} u_n\\ u_n\\ u_n+a\\ u_n-a\end{bmatrix}, \tag{4}$$

and the left eigenvectors $\mathbf{l}_i^T$ correspond to the rows of the following matrix:

$$\begin{bmatrix} a^2-g & \gamma_1 u & \gamma_1 v & -\gamma_1\\ -u_t & n_2 & -n_1 & 0\\ g-au_n & +n_1 a-\gamma_1 u & +n_2 a-\gamma_1 v & \gamma_1\\ g+au_n & -n_1 a-\gamma_1 u & -n_2 a-\gamma_1 v & \gamma_1\end{bmatrix}. \tag{5}$$

Here, the abbreviations $u_n := \mathbf{u}\cdot\mathbf{n}$, $u_t := -n_1 v + n_2 u$, $g := \frac{\gamma-1}{2}(u^2+v^2)$, and $\gamma_1 := \gamma - 1$ have been used.

The number of required boundary conditions equals the number of incoming waves (characteristic speeds $\lambda_i < 0$). For the other waves, a numerical boundary condition has to be used: Non-incoming waves are extrapolated in space by

$$\mathbf{l}_i \cdot (\mathbf{U}_s - \mathbf{U}_{\text{extrapol.}}) = 0, \quad \lambda_i(\mathbf{n}) \geq 0. \tag{6a}$$

Here, $\mathbf{U}_s$ denotes the state vector at the wall surface, and $\mathbf{U}_{\text{extrapol.}}$ is the state vector obtained from the interior cells by linear extrapolation. Incoming waves are replaced by boundary conditions of the general form

$$B_i(\mathbf{U}_s) = 0, \quad \lambda_i(\mathbf{n}) < 0, \tag{6b}$$

where $B_i(\mathbf{U}_s)$ is a nonlinear function expressing the boundary condition, eg. $B_i(\mathbf{U}_s) := p(\mathbf{U}_s) - p^{\text{target}}$.

For subsonic inflow, there are three incoming waves, and therefore three quantities are prescribed, namely the total temperature T_0, the total pressure p_0, and the inflow angle α. In the supersonic case all flow quantities are given. At a subsonic outflow the static pressure p is specified, in the supersonic case all quantities are extrapolated. Along sidewalls one boundary condition is required in any case. If the wall geometry is fixed we impose the slip condition $B_4(\mathbf{U}) := \mathbf{u}\cdot\mathbf{n} = 0$. If instead the static pressure distribution is specified, characteristic extrapolation of the outgoing wave may yield a nonzero normal component of the velocity. The flow tangency condition $\mathbf{u}\cdot\mathbf{n} = 0$ will then be used for the wall modification procedure, as discussed next.

Grid Boundary Conditions

(2) is a system of parabolic second order equations for the two unknowns $\mathbf{x} := (x,y)^T$. Two boundary conditions are required along the boundary of the computational domain. For accuracy reasons an orthogonal grid at the boundaries is desirable. Therefore, one of the boundary conditions is the orthogonality condition

$$\mathbf{x}_\sigma \cdot \mathbf{x}_\psi = 0. \tag{7}$$

For a fixed wall, the other condition may be written in the general form

$$f(\mathbf{x}) = 0. \tag{8}$$

f is any implicit function describing the wall shape, eg.

$$f(\mathbf{x}) = y - \text{polynomial}(x).$$

In the inverse design case application of the characteristic extrapolation scheme (6) with imposed static pressure yields a nonzero normal velocity component. The tangential velocity component is extrapolated by virtue of (6a, $i = 2$). Once a value of the velocity vector $\mathbf{u}$ has been obtained, the geometry condition can be replaced by the flow tangency condition

$$\mathbf{u}\cdot\mathbf{n} = 0 \tag{9}$$

which is regarded as an equation for the wall normal vector $\mathbf{n}$ (ie. for the wall slope). Of course, the starting point of an inverse sidewall has to be fixed.

In case of multi block grids special care has to be taken at block interfaces in order to ensure continuity of the grid. A possible option is to leave the block interfaces floating: Grid vertices at block interfaces are only required to be located in the centre between the four adjacent vertices. Therefore the location of the block boundaries can not be prescribed in advance, but the grid is smooth over block boundaries.

Spatial Discretization

Each block of the computational domain is subdivided into quadrilateral cells in a structured manner. It is quite natural to assign the geometry variables $\mathbf{x} = (x,y)^T$ to the vertices. The grid generation equations (2) are discretized by central finite differences. For block connectivity one layer of ghost vertices is required.

For the flow equations we are free to choose a cell centered finite volume discretization: Flow variables are assigned to cell centers. For simplicity we use the JST scheme [7]: The convective fluxes $\mathbf{F}\cdot\mathbf{n}$ are computed in a central manner, ie. by the mean value of the fluxes evaluated at the neighbouring cell centers. For stabilization second and fourth order differences of the conservative variables are added to each equation. Based on a pressure switch the coefficients of these artificial dissipation terms are chosen in such a way that the second order accuracy of the scheme is retained in smooth flow regions, but still enough stability is provided across shocks. For details the reader is referred to [7]. The fourth order difference requires the values of two neighbouring cells in each direction, therefore two layers of ghost cells are needed.

Time Marching Technique

Spatial discretization of flow (1) and grid (2) equations yields a large system of ODEs

$$U_t + R(U) = 0. \tag{10}$$

The unknown vector U contains discretized flow *and* grid variables. The steady state $R(U) = 0$ can be reached either by explicit or implicit time marching techniques. Here, we use the implicit Euler scheme

$$\frac{1}{\Delta t}\left(U^{(k+1)} - U^{(k)}\right) + R(U^{(k+1)}) = 0. \tag{11}$$

The nonlinear algebraic system for the vector of unknowns at the new time level $U^{(k+1)}$ is solved approximately by one step of Newton's method. $U^{(k)}$ is used as initial guess. The corresponding iteration reads

$$\left(\frac{1}{\Delta t}I + R_U^{(k)}\right)\left(U^{(k+1)} - U^{(k)}\right) + R(U^{(k)}) = 0. \tag{12}$$

I denotes the identity matrix and R_U the Jacobian of $R(U)$. The linear system is solved iteratively by GMRES with ILU preconditioning.

For smooth flows and reasonable initial guess the time step Δt can be chosen very large. With $\Delta t \to \infty$ (12) reduces to the standard Newton iteration for the steady state equations $R(U) = 0$. Hence, quadratic convergence might be expected. In practice, however, the Jacobian matrix R_U is computed with frozen coefficients of artificial dissipation, and the linear system (12) is not solved to machine accuracy. Convergence therefore becomes linear, but the residual is usually reduced by one order of magnitude per iteration.

The time marching scheme (12) requires an initial guess not too far away from the actual solution. As iterations are cheap on a coarse grid it pays to employ a nested iterations technique: First, the problem is solved on a coarse grid. As soon as the iteration error is of the same magnitude as the discretization error, the coarse grid solution is interpolated to a finer grid with twice as many cells in each direction. This procedure is repeated until the target grid has been reached.

The fact that the mesh is time-dependent has to be accounted for in the formulation of the numerical scheme. Although the change of volume of the computational cells $\Omega(t)$ and the cell surface velocity become small as the steady state is approached, this time variations should not be neglected in general, as pointed out in [2]. Indeed, if the grid motion terms are neglected in the flow equations (1), the convergence rate is impaired [8], and the wall or grid motion has to be under-relaxed [1].

However, the situation is quite different when an implicit time marching technique is used. As mentioned before, for large time step (12) reduces to the Newton iteration for the steady state equations $R(U) = 0$. This means $\dot{\mathbf{x}}_s = 0$ (the grid equations are included in the residual vector R), so there is no indication why grid surface motion and cell volume change should be accounted for. However, the flow equations (1) depend on the grid variables through the term $\mathbf{F}\cdot\mathbf{n}$. The corresponding entries in the Jacobian matrix R_U must be accounted for.

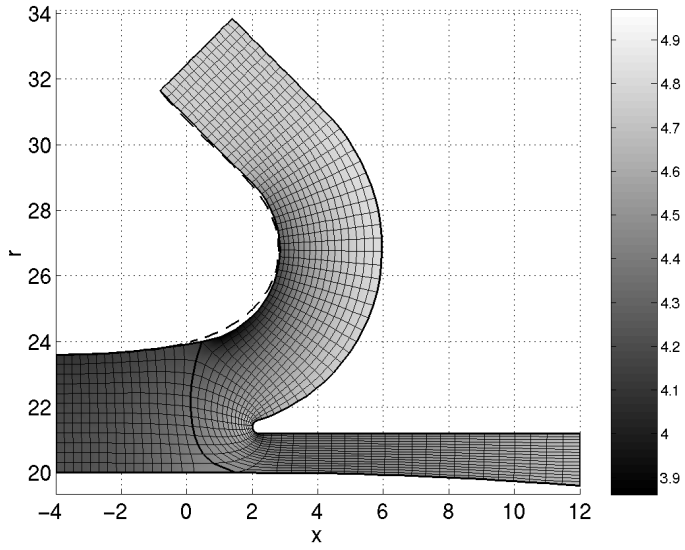


Figure 1: Annular compressor diffuser with cooling air bleed-off diffuser (lower outlet). The original geometry and grid are shown. An H-type mesh is located in the inlet part (left side), and a C-type mesh connecting both outlets is wrapped around the stagnation region.

APPLICATIONS

The method has been implemented for planar and axis-symmetric flows governed by the compressible Euler equations. In this section two application examples are shown. The first one concerns an axis-symmetric problem arising from gas turbine design. The second test case demonstrates the capability of the present method to handle transonic flows as well.

Compressor Diffuser

To demonstrate the capabilities of the inverse method a simple axis-symmetric problem is considered. The design of a compressor diffuser of a gas turbine is often strongly restricted by the space requirements of turbine and combustor. As a consequence there is usually the necessity of choosing a curved diffuser, as shown in Figure 1. The inlet of a cooling air bleed-off diffuser is placed at the beginning of the turn of the main diffuser.

A simple two-block topology has been adopted: An H-type mesh (48×64 cells) is located in the inlet part, and a C-type mesh (64×304 cells) connecting both outlets is wrapped around the stagnation region. This results in a good mesh quality at the stagnation point.

In the first step, the flow is computed for the baseline design (direct mode, all sidewalls are fixed). The resulting static pressure pattern is shown in Figure 1, together with the grid (every fourth grid line is shown). The pressure distribution along the outer diffuser wall (ie. the wall that begins at the larger radius) is indicated by the solid line in Figure 3. Starting from this result it is possible to define a design strategy. In order to satisfy geometric constraints imposed by neighbouring parts the inner wall remains fixed. The outer wall is a better candidate for flow separation and should be designed according to the largest admissible adverse pressure gradient. A possible target pressure distribution is shown in Figure 3 (diamond markers).

The resulting geometry and the new pressure contours are shown in Figure 2. Evidently the small region of acceleration at the beginning of the turn has been removed, although the change in geometry (indicated by the dashed line in Figure 1) is rather minute.

Finally, the convergence history for both computations is shown in Figure 4. It is clearly seen that there is no difference between direct and inverse computation in terms of *number of iterations*. In both cases, the residual is reduced by almost one order of magnitude per iteration (on the finest

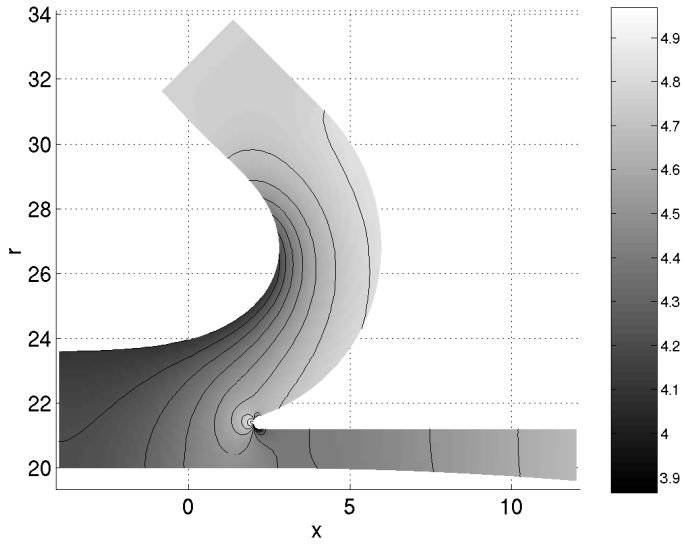


Figure 2: Inverse design of the upper sidewall, resulting geometry and pressure contours. For comparison, the redesigned wall shape is denoted by the dashed line in Figure 1.

grid level). Of course, compared to a purely direct computation with fixed grid, the method presented here consumes more *time*, since the number of unknowns has been increased by two.

Transonic Nozzle

We consider the planar converging-diverging nozzle shown in Figure 5. The inflow conditions are purely subsonic (Mach number $M \approx 0.7$). The flow is accelerated in the converging part of the nozzle, reaches sonic conditions at the throat, and is accelerated further in the diverging part. At the outflow, a constant static pressure value is imposed that would result in subsonic flow there. The back pressure was chosen to give rise to a shock close to the throat. The Mach number and static pressure distributions are shown in Figures 6,7. Solid lines denote the exact solution of the quasi-one-dimensional Euler equations.

In the first run, the flow is computed for fixed walls on a 320×32 grid. The resulting Mach number and static pressure distributions are denoted by the diamond markers in Figures 6 and 7, respectively. The close agreement between exact and approximate solution is seen. In particular, the shock is resolved within a few computational cells.

Next the static pressure distribution obtained from the direct computation is used as boundary condition for an inverse run. The initial shape is a straight channel, and the flow is assumed uniform.

As expected, the original geometry is recovered (up to a small difference in the order of magnitude of the iteration error). This is a clear indication that direct and inverse mode are compatible.

Although the target pressure distribution is discontinuous at the shock, there are no indications of any defect of smoothness of the wall shape. It is thought that this is due to the staggered allocation of flow and grid variables. Of course, it is expected that there arise problems if the prescribed pressure distribution does not satisfy the correct jump relations, see the discussion in [4]. In that case, the wall slope would become discontinuous across the shock, or the method would even fail to converge.

Finally, the observation made for the previous example still holds in the transonic case: there is no difference in the convergence rate between direct and inverse case.

CONCLUSIONS

A method for the two-dimensional and axis-symmetric target pressure problem based on the compressible Euler equations has been presented. Block-structured grids allow for applications relevant to industry. Since the method is very fast, it is feasible to link it to an optimization procedure.

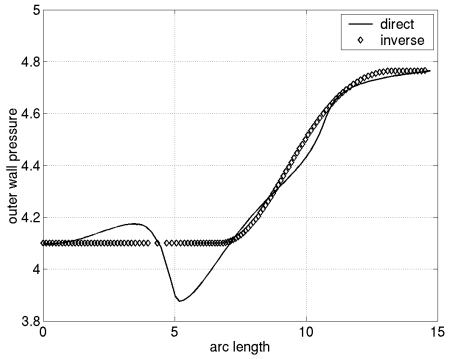


Figure 3: Static pressure distribution along the outer side wall. The target pressure distribution for inverse design is indicated by diamond markers.

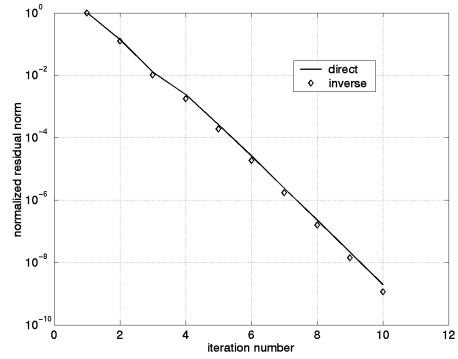


Figure 4: Convergence history for direct and inverse computation: normalized residual norm versus number of time steps (12).

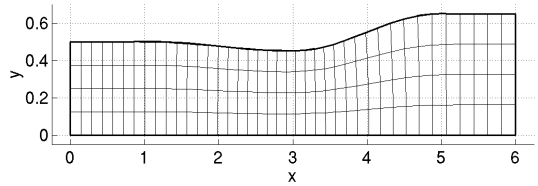


Figure 5: Planar nozzle: geometry and grid (different axis scaling!)

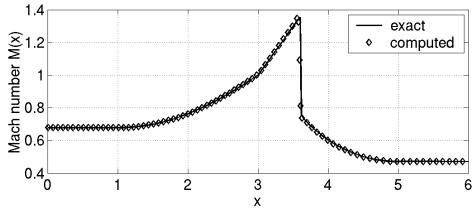


Figure 6: Transonic nozzle flow: Mach number distribution

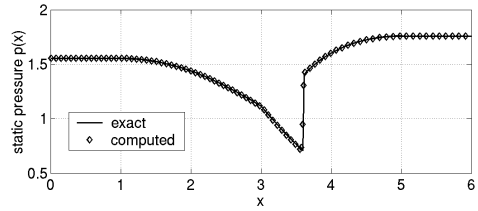


Figure 7: Transonic nozzle flow: static pressure distribution

Such an optimal inverse approach should make it easy to solve inverse problems with additional (geometric) constraints in an efficient manner. Presumably it is worthwhile applying the same nested iterations technique, since generation of a good initial guess on a coarse grid is cheap.

The extension to viscous flows seems rather straightforward, except in one point. While in the inviscid case the flow tangency condition (9) $\mathbf{u} \cdot \mathbf{n} = 0$ can be used for the wall update, this option is prohibited by the no-slip condition $\mathbf{u} = 0$ in the viscous case. However, if viscous effects are restricted to near-wall regions they can be modeled by a zonal viscous/inviscid interaction approach, or by a distributed loss model.

The author acknowledges the support provided by Alstom Power (Switzerland) Ltd. and by the Commission for Technology and Innovation (KTI) under grant No. 4571.1 KTS.

References

1. T. Dang and V. Isgro. Euler-based inverse method for turbomachine blades. Part 1: Two-dimensional cascades. *AIAA Journal*, **33**, 12, 2309–2315 (1995)

2. A. Demeulenaere. An Euler/Navier-Stokes inverse method for compressor and turbine blade design. von Karman Institute for Fluid Dynamics, Lecture Series 1997-05: Inverse Design and Optimisation Methods, April 21-25, 1997

3. W. Egartner and V. H. Schulz. Partially reduced SQP methods for optimal turbine and compressor blade design. In *Bock, Hans Georg (ed.) et al., Proceedings of ENUMATH '97. World Scientific, Singapore, 286-293*, 1998

4. P. D. Frank and G. R. Shubin. A comparison of optimization-based approaches dor a model computational aerodynamics design problem. *J. Comp. Phys.*, **98**, 1, 74–89 (1992)

5. M. B. Giles and M. Drela. Two-dimensional transonic aerodynamic design method. *AIAA Journal*, **25**, 9, 1199–1206 (1987)

6. C. Hirsch. *Numerical Computation of Internal and External Flows*, volume 2. John Wiley & Sons, Chichester, 1988

7. A. Jameson, W. Schmidt, and E. Turkel. Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time stepping schemes. *AIAA paper*, 81-1259 (1981)

8. G. Meauzé. An inverse time marching method for the definition of cascade geometry. *J. Eng. Power*, **104**, 650–656, (1982)

9. P. Mineau. Smoothing of grid discontinuities across block boundaries. In N. P. Weatherill, et. al., (ed.), *Multiblock Grid Generation*, volume 44 of *Notes on Numerical Fluid Mechanics*, pages 139–147. Vieweg, Braunschweig, 1993

10. A. Scascighini. *A Numerical Method for the Design of Internal Flow Configurations Based on the Inverse Euler Equations*. PhD thesis, No. 14440, ETH Zurich, 2001

11. J. F. Thompson, Z. U. A. Warsi, and C. W. Mastin. *Numerical Grid Generation — Foundations and Applications*. Elsevier Science Publishing Co, 1985

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

OPTIMIZATION OF PROCESSES INVOLVING COUPLED ELECTROMAGNETIC AND HEAT TRANSFER ANALYSIS

François Bay
Cemef – Ecole des Mines de Paris
Rue Claude Daunesse – B.P. 207
Sophia Antipolis, 06904, France
Francois.bay@cemef.cma..fr

Valérie Labbé,
Yann Favennec
Cemef – Ecole des Mines de Paris

ABSTRACT

Automatic numerical optimization of processes coupling electromagnetism and heat transfer is a quite intricate problem.

We present here an efficient optimization procedure coupled with a direct finite element model which has been developed and tested successfully in order to deal with prescribed industrial goals such as reaching an homogeneous temperature level in the workpiece to be pre-heated, or achieving a certain level of hardness for the final workpiece.

The sensitivity analysis is carried out either through a finite-difference approach or through the use of an adjoint model which has been specifically designed to include the main features of the algorithm used in the direct numerical model.

Results and discussion on industrial cases are then presented.

INTRODUCTION

Many industrial processes are based on an efficient use of coupling between electromagnetic, thermal and mechanical phenomena. They generally use direct or induced currents to generate heat inside a workpiece in order to get either a prescribed temperature field or some given mechanical or metallurgical properties through an accurate control of temperature evolution with respect to time.

Determining optimal process parameters for these processes in order to reach industrial objectives can be greatly helped by using numerical modeling coupled with optimization techniques.

The objectives have to be formalized using some specific cost functions which can account for time-effects.

Control parameters may include, among others, the electromagnetic source location, frequency, power density, …

We present here the optimization procedure developed and used in our laboratory for these problems.

The two approaches for sensitivity analysis computation – finite differences or use of an adjoint model - will then be presented.

We shall then detail the direct numerical model for the case of an induction heating process and provide some optimization results for it.

THE OPTIMIZATION PROBLEM

A direct numerical model for the analysis of a coupled electromagnetic-heat transfer process can be written in a generic way by:

$$R^E(E,T) = 0 . \quad (1)$$

$$R^T(E,T) = 0 . \quad (2)$$

where R^E and R^T denote the residual vectors for the electromagnetic and heat transfer computations; E and T stand for the vectors of temperatures and electric fields at the nodes of a finite element mesh.

Various industrial objectives can be assigned, such as reaching a temperature as uniform as possible within the part for initial pre-heating,, or prescribing a precise path in space and time for temperature evolution when dealing with heat

treatment applications. We shall assume here that the objective function are always based on temperatures.

A general continuous form of the objective function *J* can therefore be:

$$J(u) = J_1(T(t)) + J_2(T(t_f)) = \int_{t_0}^{t_f} g(T(t))dt + h(T(t_f)) \quad (3)$$

where u denotes the controls. The general optimization problem can therefore be written as follows :

$$\begin{array}{l} \text{Minimize} \quad J(u) \\ \text{under} \quad \text{constraints} \\ R^E(E,T) = 0 \qquad R^T(E,T) = 0 \end{array} . \quad (5)$$

THE GRADIENT-BASED ALGORITHM

In order to solve this optimization problem, we have decided to use a gradient-based algorithm:

$$u^{k+1} = u^k + \alpha^k d^k \quad (6)$$

where d^k denotes the descent direction. Instead of using for d^k the gradient direction at each ietration, we instead use:

$$d^k = -\nabla_{u^k} J + \beta^k d^{k-1} \quad (7)$$

where b^k is computed as in the Polak-Ribière conjugate gradient type method. which can accelerate the convergence of the algorithm.
Once the descent direction d^k has been computed, the a^k descent-step is used to get the value of the cost function as close as possible to 0 through a parabolic interpolation algorithm.

THE SENSITIVITY COMPUTATION

The algorithm for the minimization of the objective function requires the computation of the cost function gradient with respect to the control parameters.

There are at least two different ways of carrying out the computation of this quantity.

The first one is through:

$$\frac{dJ}{du_k} = \frac{\partial J}{\partial u_k} + < \frac{\partial J}{\partial T}, \frac{\partial T}{\partial u_k} > \quad (9)$$

(4)

This approach requires the computation of the sensitivity of the temperatures at all nodes of the mesh with respect to the control parameters. It can be carried out through a finite-difference approach.
The system (1)-(2) is solved for an initial and then a perturbated value of each control parameter.
This approach is the easiest one to implement, but can be quite consuming in terms of computational time. (

The second approach is based on an optimal control approach and avoids the computation of the sensitivities of temperature with respect to the control parameters.

We define the Lagrangian of the problem as:

$$L(u,E,T) = J(u) + < \lambda^E, R^E(E,T > + < \lambda^T, R^T(E,T > \quad (11)$$

If we assume that:

- the thermal residual vector R^T depends only on the temperature field T
- the electromagnetic residual vector R^E depends only on the electric field E and the controls u (8)

we get to solve the following system:

$$\left\langle \frac{\partial R^T}{\partial T}\frac{dT}{du_j}, \lambda^T \right\rangle_{\Omega\times[t_0,t_f]} = -\left\langle \frac{\partial J_1}{\partial T}, \frac{dT}{du_j} \right\rangle_{\Omega\times[t_0,t_f]} - \left\langle \frac{\partial J_2}{\partial T}(t_f), \frac{dT}{du_j}(t_f) \right\rangle_{\Omega} \quad (12)$$

which provides us with values of λ^T

We then solve:

$$\left\langle \frac{\partial R^E}{\partial E}\frac{dE}{du_j}, \lambda^E \right\rangle_{\Omega\times[t_0,t_f]} = -\left\langle \frac{\partial R^T}{\partial E}\frac{dE}{du_j}, \lambda^T \right\rangle_{\Omega\times[t_0,t_f]} \quad (13)$$

from which we get values for λ^E

Finally, the gradient of the cost function is expressed by:

$$\frac{dJ}{du_j} = \left\langle \frac{\partial R^E}{\partial u_j}, l^E \right\rangle_{\Omega\times[t_0,t_f]} \quad (14)$$

APPLICATION TO THE INDUCTION HEATING PROCESS

We shall now see on the induction heating process how the optimization approach can be used.
We shall first presnt the process and derive the coupled system of equations in R^E and R^T which models this process.

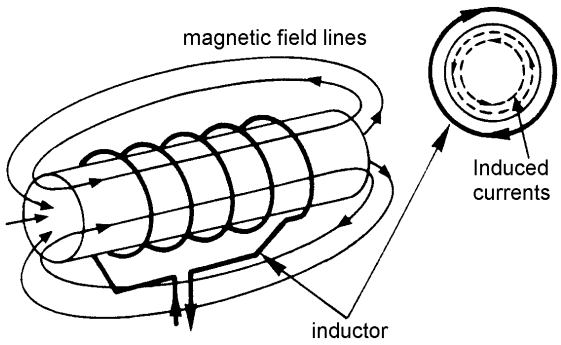


Figure 1 : Induction heating setup

The basic induction setup (see for instance Davies [1990]) consists of one or several inductors and metallic workpieces to be heated (see figure 1). The inductors are supplied with alternating current with frequencies ranging from fifty to several hundred thousand cycles per second. A rapidly oscillating magnetic field is generated and in turn induces eddy currents in the workpiece due to the Joule effect. These currents generate ohmic heat losses inside the workpiece. Moreover, for ferromagnetic materials, alternating magnetization and hysteresis effect also contribute to heat generation.

Most of the heat is produced in a thin layer under the surface of the workpiece; the skin depth d - defined as the depth at which the magnitude of the field drops to a value of e^{-1} of its surface value :

$$\delta = \sqrt{\frac{1}{\pi f \sigma \mu}} \quad (15)$$

where f is the frequency, s the electrical conductivity and m the magnetic permeability. High frequencies are used to achieve surface heating, while low frequencies generate a more uniform heating.

THE DIRECT ELECTRO-THERMAL COMPUTATION

The Electromagnetic Model

The electromagnetic model is classically based on the Maxwell equations:

Magnetic flux equation :

$$\vec{\nabla}.\vec{B} = 0 \quad (16)$$

Maxwell-Gauss equation :

$$\vec{\nabla}.\vec{D} = 0 \quad (17)$$

Maxwell-Faraday equation :

$$\vec{\nabla}\times\vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (18)$$

Maxwell-Ampere equation :

$$\vec{\nabla}\times\vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t} \quad (19)$$

where $\vec{H}$ is the magnetic field, $\vec{B}$ the magnetic induction, $\vec{E}$ the electric field, $\vec{D}$ the electric flux density, and $\vec{J}$ the electric current density associated with free charges.

We also have the following relations which take into account the intrinsic material properties :

$$\vec{D} = e\vec{E} \quad (20)$$

$$\vec{B} = m(|\vec{H}|)\vec{H} \quad (21)$$

$$\vec{J} = s\vec{E} \quad (22)$$

where ε is the dielectric constant, μ the magnetic permeability, and σ the electrical conductivity. They all depend on temperature and the magnetic permeability μ depends also on H.

The range of frequencies dealt with in induction heating (less than 10^6 Hz) enables us to neglect the displacement currents in the Maxwell-Ampere equation (magneto-quasi-static approximation).

A combination of the previous relations leads us to the following equation where the unknown is the electric field.

$$s\,\frac{\partial \vec{E}}{\partial t} + \vec{\nabla} \times \left(\frac{1}{m} \vec{\nabla} \times \vec{E} \right) = -\frac{\partial \vec{J}_S}{\partial t} \quad (23)$$

with $\sigma = \sigma(T)$ and $\mu = \mu(T,H)$

We deal here with axisymmetrical cases, in which the electric field will only have a non-zero component in the θ direction:

$$\vec{E} = (0, E_q(r,z), 0) \quad (24)$$

The Thermal Model

Temperature evolution in the workpiece is governed by the classical heat transfer equation :

$$rC\frac{\partial T}{\partial t} - \mathrm{div}(k\vec{\nabla}T) = \dot{Q}_{em} \quad (25)$$

where r denotes the material density, C and k respectively the specific heat and thermal conductivity, all temperature dependent.

$\dot{Q}_{em}$ denotes the local heat rate, generated by the eddy currents, and integrated over one period:

$$\overline{\dot{Q}}_{em} = \frac{1}{T}\int_0^T \sigma |\vec{E}|^2 dt \quad (26)$$

The boundary conditions can be of various kinds: prescribed heat flux or temperature, convection or radiation.

THE NUMERICAL APPROXIMATION

The finite element space discretization

We define Ω as being a two-dimensional axisymmetrical domain which covers the part to be heated Ω_{part}, the inductor $\Omega_{inductor}$ and a finite volume of air Ω_{air} surrounding the inductor and the part.

We have in fact chosen here to carry out coupling between the part and the inductor for the electromagnetic computations using finite elements rather than boundary elements; the air domain thus needs to be wide enough in order to model accurately electromagnetic wave propagation.

The domain is discretized using second order triangular finite elements (6-nodes triangles). The unknown fields – electric field E_θ for the electromagnetic computations, temperature field T for the thermal computations and velocity field V for the mechanical computations - can thus be approximated over the whole domain by the classical finite element approximation:

$$E(t,r,z) = \sum_{i=1}^{nbnode} E^i(t)\ N^i(r,z) \quad (27)$$

$$T(t,r,z) = \sum_{i=1}^{nbnode} T^i(t)\ N^i(r,z) \quad (28)$$

where $E^i(t)$ denotes the approximated value of the θ-component of the electric field at the node i and at time t, $T^i(t)$ the temperature field, $V^j(t)$ the j-th component (1 or 2) of the velocity field and $N^i(r,z)$ denotes the shape function associated to the node i in the mesh. When the discretized expressions of these fields are introduced in the variational formulation, we get the following equations discretized in space:

$$[C^{em}]\left\{\frac{\partial E}{\partial t}(t)\right\} + [K^{em}]\{E(t)\} = \{B^{em}\} \quad (29)$$

$$[C^{th}]\left\{\frac{\partial T}{\partial t}(t)\right\} + [K^{th}(t)]\{T(t)\} = \{B^{th}\} \quad (30)$$

The time discretization

Numerical models in induction heating often solve a harmonic model. This assumption is quite restrictive when one deals with non-linear

magnetic materials. We have thus chosen to solve the time-dependent model. We need therefore to integrate numerically in time the electromagnetic and thermal equations.

We detail here the selected time integration scheme for the electromagnetic equation. The procedure is the same for the thermal equation. We use a second-order two time step finite difference scheme:

Step 1: the system is solved at time t* such that $t<t^*<t+\delta_2$ with:

$$t^* = a_1(t - dt_1) + a_2 t + a_3(t + dt_2) \quad \text{with} \quad a_1 + a_2 + a_3 = 0 \tag{31}$$

The electric field E* at time t* and its time derivative write:

$$E^* = a_1 E^{t-dt_1} + a_2 E^t + a_3 E^{t+dt_2} \tag{32}$$

$$\frac{\partial E^*}{\partial t} = \gamma \frac{E^{t+\delta_2} - E^t}{\delta t_2} + (\gamma - 1)\frac{E^{t-\delta_2} - E^t}{\delta t_1} \tag{33}$$

The system (20) is written at time t*. E* and its derivative are replaced by expressions (32) and (33). The system is solved for the unknown variable E* :

$$\left(\frac{\gamma}{\alpha_3 \delta t_2}\left[C^{em}\right]^* + \left[K^{em}\right]^*\right)E^* = \left\{B^{em}\right\}^* + c_1\left[C^{em}\right]^* E^t + c_0\left[C^{em}\right]^* E^{t-\delta_1} \quad \text{where} \tag{34}$$

$$c_1 = \frac{\gamma}{\delta t_2} + \frac{\gamma - 1}{\delta t_1} + \frac{\gamma \alpha_2}{\alpha_3 \delta t_2}$$

$$c_2 = \frac{\gamma \alpha_1}{\alpha_3 \delta t_2} - \frac{\gamma - 1}{\delta t_1}$$

The two time steps scheme we have requires the solving of a non-linear equation, as the matrix [C] is dependant on the magnetic field. In order to avoid an additional non-linearity, the matrix is linearized and is approximated using its values at time t and $t-\delta t_1$:

$$[C]^* = (\alpha_1 - \alpha_3 \frac{\delta t_2}{\delta t_1})[C]_{t-\delta t1} + (\alpha_2 + \alpha_3(1 + \frac{\delta t_2}{\delta t_1}))[C]_{t-\delta t1} \tag{35}$$

Second step: computation of :

$$\{E\}^{t+\delta t_2} = \frac{1}{\alpha_3}(\{E\}^* - \alpha_1\{E\}^{t-\delta t1} - \alpha_2\{E\}^t) \tag{36}$$

The electromagnetic/thermal coupling procedure

Physical problems arising from heat transfer and electromagnetism have in common the fact that they are both time-dependent. Their specific time-scales are however very different. The specific time scale of an electromagnetic problem is related to the wave-associated period – typically 10^{-2} s for a 100 Hz frequency down to 10^{-8} s for a 100 MHz frequency – whereas the specific time scale for heat transfer averages normally one second.

A direct model based on finite elements has been developed in our laboratory to cope with these specificities. The model includes a specific coupling procedure for solving:

- the Maxwell equations - in order to access the electromagnetic fields giving the eddy currents dissipated in the material (main source term for the heat transfer equation)
- the heat transfer equation - leading to temperature evolution in the material

The coupling between the electromagnetic and thermal computations relies on a convergence test over the mean heating power and on tests over the variations of the magnetic parameters that determine respectively the passage from an electrical to thermal resolution or inversely from a thermal to an electric one.

Once the electromagnetic field has been calculated, the rate of heat generation $\dot{Q}_{em}$ for the heat equation needs to be evaluated at every integration points. As the electromagnetic time step is far smaller than the thermal one, we do not consider the instantaneous Joule power calculated at a given time at every integration points. We rather consider a mean Joule power averaged over one period of the electromagnetic field:

$$Q_{em}(nT, int) = \frac{1}{T}\int_{(n-1)T}^{nT} \sigma(int, t)\left|E(int, t)\right|^2 dt \quad (37)$$

where *int* is the considered integration point, *T* is the period of the power supply currents, *n* is number of periods considered and $E_q(\text{int}, t)$ is the value at time *t* of the electric field interpolated at the integration point *int*.

At the end of each electromagnetic period, the newly calculated mean power is compared to the one calculated at the previous period until it stabilises. Thermal computations are started with the stabilised thermal source power calculated at (n+1)T if the following convergence test (33) is conducted at every integration points:

$$\frac{\overline{Q}_{em}((n+1)T) - \overline{Q}_{em}(nT)}{\overline{Q}_{em}(nT)} < \varepsilon \quad (38)$$

where ε is the user-supplied convergence parameter.

These thermal computations are valid as long as the variation of the physical magnetic parameters such as the magnetic permeability and the electric conductivity do not exceed 5%. Their variations with temperature are tested after each new thermal computations. The following criteria are tested for every mesh element:

$$\frac{\sigma(T_{max}^{n+1}) - \sigma(T_{max}^{n})}{\sigma(T_{max}^{n})} < 5\% \qquad \frac{\mu(T_{max}^{n+1}) - \mu(T_{max}^{n})}{\mu(T_{max}^{n})} < 5\% \quad (39)$$

where T_{max}^{n+1} is the maximum value of the temperature field in a given element at time $t + dt_{ther}$ and T_{max}^{n} is the maximum value of the temperature field in the same given element at current time t. When the maximum relative variations reach the threshold of 5%, the previously calculated mean heat power is assumed to be irrelevant, and a new electromagnetic calculation is carried out.

For their part, mechanical computations are carried out at the same time steps than thermal computations.

RESULTS

Finite difference approach for sensitivity

We present results on a case where the objective is to reach a homogeneous temperature level in the part with a low-frequency process.

This case is displayed in Figure 2 and is typical of the heating of a billet before forging.

The goal is to obtain a given temperature (1200°C) at several locations in the part (close to the surface), after a 5 seconds heating time.

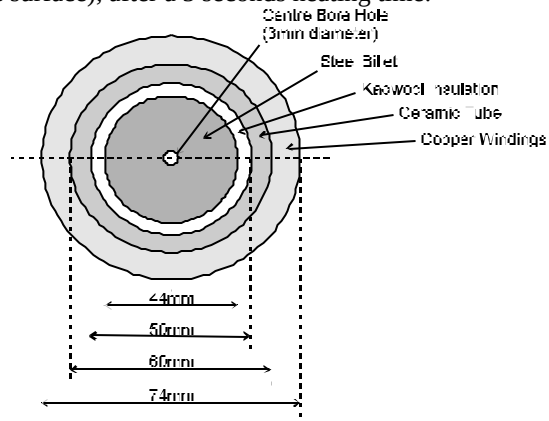


Figure 2 : Induction heating setup

The chosen physical parameters of the billet are the following. The relative magnetic permeability equals 90 at 0 Kelvin, with a temperature sensitivity of 6 (Frohlich-Kenelly). The electrical conductivity is equal to $3.10^6\ \Omega^{-1}m^{-1}$, the thermal conductivity equals 35 $Wm^{-1}K^{-1}$ and the heat capacity equals $4.875.10^6\ JKg^{-1}K^{-1}$.

Figure 3 displays the mesh which has been created for this case.

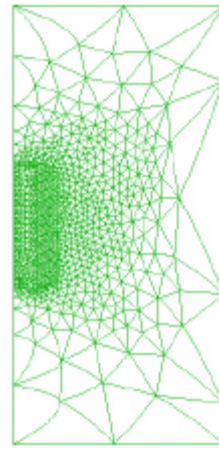


Figure 3 : Finite element mesh

The control parameters here are frequency and current density in the coil. Figure 4 shows how the algorithm has performed in terms of convergence. Initial estimates were 100 Hz and 5.10^9A/m. Convergence towards optimized values has been reached here in 5 iterations.

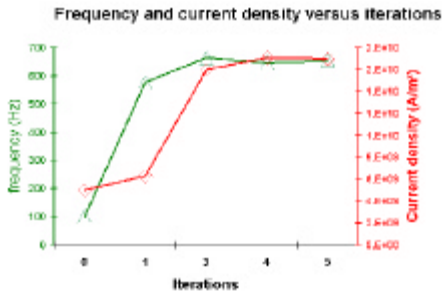


Figure 4: Convergence on frequency (triangles) and current density (squares)

The optimal control approach for sensitivity

We have investigated here two test cases.
The first test case aims at getting on the surface (that is within the radius range [0.018 ; 0.02]) the following time dependent optimal temperature ($T^{opt}(2.5)$=850K), ($T^{opt}(3.5)$=1030K) and ($T^{opt}(5.0)= T^{opt}(t_f)$=1273K).

$$J\left(T\left(t=2.5, t=3.5, t=5\right)\right)= \int_{\Omega_{opt}}\left(T^{cal}\left(2.5\right)-850\right)^2 d\omega + \int_{\Omega_{opt}}\left(T^{cal}\left(3.5\right)-1030\right)^2 d\omega + \int_{\Omega_{opt}}\left(T^{cal}\left(5.0\right)-1273\right)^2 d\omega \quad (40)$$

Figure 5 displays the evolution of input process parameters with respect to iterations as well as the objective function value. Iteration 1 is directly related to the first guessed parameters (frequency = 500 Hz and J0 = 10^9 A/m).

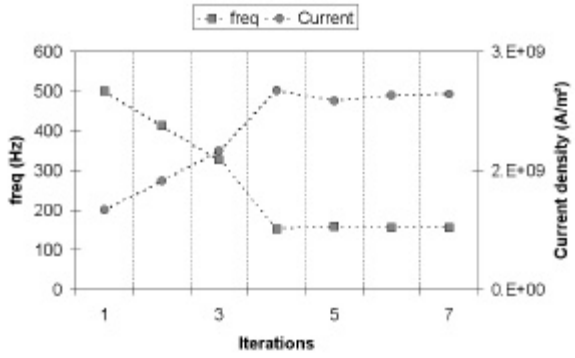


Figure 5 : Process parameter evolutions with respect to optimization iterations

The cost function decrease is displayed in Figure 6. Its value and its gradient are first calculated. Iterations 1, 2, 3, 4 and 5 are then carried out with the uni-dimensional research. The gradient is then calculated once again. Following iterations are related to the second loop. Uni-dimensional algorithm is run again. The cost function value has decreased by a factor 20.

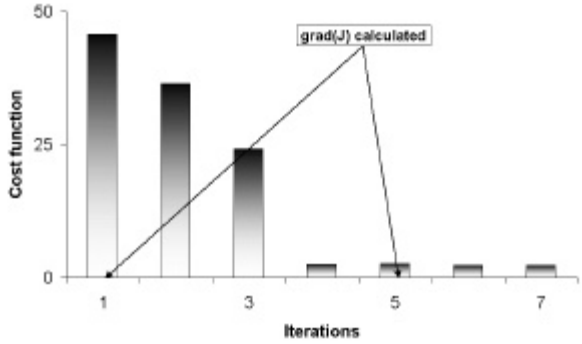


Figure 6: Cost function value with respect to iterations

The second test case uses the same global geometry as for previous test cases except that the inductor is moving along the z-axis as shown in the next figure. The aim here is to improve frequency, current density and coil velocity such that, after ten seconds of heating, the surface between z=22.5cm and z=24.7cm is as close as 850K.

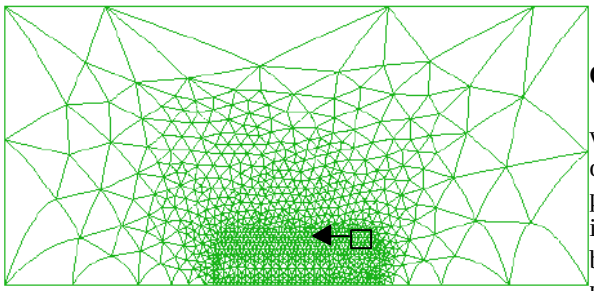


Figure 7: Mesh for the second test case. The inductor is moving at a 10mm/s velocity

Here again, figure 8 presents evolution, of frequency, input current and coil velocity with respect to iterations, while figure 9 presents the evolution of the cost function. Only eight full calculations are needed for decreasing the cost function by a factor 100.

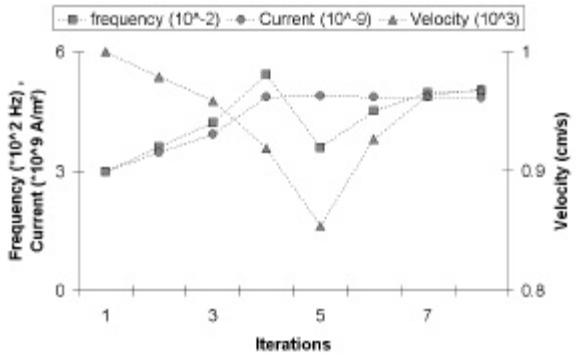


Figure 8 : Process parameter evolutions with respect to optimization iterations

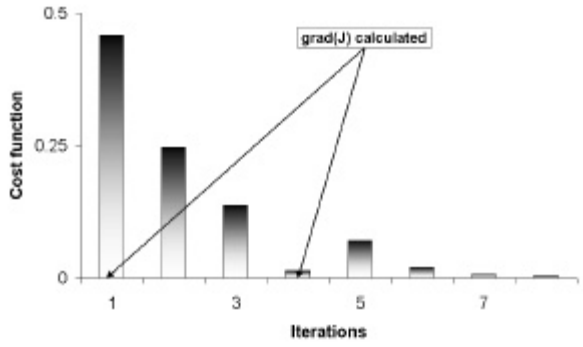


Figure 9 :cost function evolution with respect to optimization iterations

CONCLUSION

We have presented an optimization procedure which can provide a powerful tool for the optimization of coupled electromagnetic-thermal processes. It has been successfully applied for induction heating process optimization, and has been moreover used for identification of physical parameters involved in induction heating (Favennec and al. [2002])

Research is being presently carried out on the parallelization of the model.

REFERENCES

1 F. Bay, V. Labbé, Y. Favennec & J.L. Chenot, "A numerical model for induction heating processes coupling electromagnetism and thermomechanics", Submitted in *Int. J. Num. Meth. Engrg.*, October 2001

2 Davies E.J. *Conduction and Induction Heating*, London : P.Peregrinus Ltd., 1990

3 Favennec Y., Labbe V., Bay F., *Inverse analysis for identification of electromagnetic parameters*, 4th International Conference on Inverse problems in Engineering, Rio De Janeiro, Brazil, May 2002

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

OPTIMAL CONTROL FOR THE ULTRASOUND INDUCED HEATING OF A TUMOR

Matti Malinen
Tomi Huttunen
Department of Applied Physics
University of Kuopio
Kuopio, Finland

Jari P. Kaipio
Department of Applied Physics
University of Kuopio
Kuopio, Finland
Jari.Kaipio@uku.fi

INTRODUCTION

When high intensity ultrasound is directed to a dissipative medium, the energy is partially absorbed and turned into heat. In ultrasound induced bloodless surgery the aim is to direct ultrasound energy to tumors and heat the cancerous tissue so that it is destroyed. The ultrasound transmitters are located outside the body - hence the term bloodless surgery.

In ultrasound surgery the cancerous tissue can be destroyed by rising the temperature to cytotoxic level. The desired temperature in tumor is often 50-60 °C. Although lower temperatures could also be used, the use of high temperatures can reduce the treatment time significantly [1,2,3,4].

The temperature distribution optimization in wave field induced heating problems in medical applications are usually done by optimizing the specific absorption rate (SAR) [5,6,7] or using PID type controllers with pre-focused ultrasound fields [8,9,10]. Steady-state optimization methods have also been used in this type of problems [11] as well as the inverse dynamics approach [12] and fuzzy logic controllers [13] In these controllers the main point is to obtain the desired temperature distribution with pre-focused ultrasound fields. The scanning path of the focus is pre-calculated in these controllers and they alter only the applied power of the transducers, not the phase and the amplitude of the ultrasound waves. These approximations result in linear controller structures, which is implementationally convenient but is usually clearly inferior in performance when compared to more rigidly derived controllers.

Steady-state optimization has been used also to determine the optimal driving parameters for electromagnetic phased array system in [14]. In the study made in [14] the phase and amplitude of the transducers were computed directly from the nonlinear optimization problem.

While some cases such as breast tumors can be treated with relatively simple computational models, the treatment of brain tumors poses significant problems. This is due to the geometrical problems and the high attenuation in the skull. Due to the geometry and the applicable frequencies, typically about 500 kHz, the computation of the ultrasound fields is a major problem. We use the so-called ultra weak variational formulation which enables the use of computationally feasible mesh sizes. For the actual control we employ the Lagrangian approach based on the bioheat equation with the quadratic source control model. We also employ approximate power constraints for the individual sources. We show with simulations that the approach is capable of producing lesions with complex geometries, which enables the treatment of such brain tumors that are near critical brain areas that must not be destroyed.

COMPUTATIONAL MODELS AND APPROXIMATIONS

Wave equation

Linear acoustic wave propagation and scattering in quiescent heterogeneous media is characterized by the wave equation (Helmholtz problem)

$$\nabla \cdot \left(\frac{1}{\rho} \nabla P \right) - \frac{1}{\rho c^2} \frac{\partial^2 P}{\partial t^2} = 0 \qquad (1)$$

where P is acoustic pressure, ρ is density and c is the speed of sound. In the time-harmonic case we have $P(r,t) = p(r)e^{i\omega t}$, where r is the spatial variable, and the space dependent part of the pressure field is the solution of the inhomogeneous Helmholtz equation

$$\nabla \cdot \left(\frac{1}{\rho} \nabla p \right) + \frac{\kappa^2}{\rho} p = 0, \qquad (2)$$

with wave number κ. In dissipative media the wave number is of the form $\kappa = 2\pi f/c + i\alpha$ where f is the frequency of the wave field and α is the absorbtion coefficient [15].

For high wave numbers this requirement results in very large problems with often intolerable computational burden. To avoid this problem ray approximations have been used to compute pressure fields with ultrasound frequencies, see e.g. [16,17,18,19]. However, this approach is feasible only with almost homogeneous media and becomes less accurate in the presence of strongly scattering obstacles.

An alternative approach is to use novel full wave methods which allow the incorporation of a priori information of the solutions to the approximation subspaces. These methods include the partition of unity methods (PUM) [20], the least squares methods [21] and the ultra weak variational formulation (UWVF) [22,23]. Compared with the standard finite elements these methods can reduce the computational burden significantly.

In this paper we use the UWVF to solve the acoustic wave field. Let us partition the domain of interest Ω with disjoint finite elements Ω_j and let ν_j denote the outward unit normal for j'th element. In addition, the boundary between elements Ω_j and Ω_ℓ is denoted by $\Sigma_{j\ell}$. If the element Ω_j is on the boundary of the domain Ω, the coinciding boundary is denoted by $\partial\Omega_j \cap \partial\Omega = \Gamma_j$.

If the material parameters ρ and c are approximated with piecewise constant functions we can decompose the Helmholtz problem for all $1 \le j \le K$ as

$$\Delta p_j + \kappa_j^2 p_j = 0 \quad \text{in } \Omega_j \tag{3}$$

$$\frac{1}{\rho_j}\frac{\partial p_j}{\partial \nu_j} - i\sigma p_j = -\frac{1}{\rho_\ell}\frac{\partial p_\ell}{\partial \nu_\ell} - i\sigma p_\ell \tag{4}$$

$$\frac{1}{\rho_j}\frac{\partial p_j}{\partial \nu_j} - i\sigma p_j = \tau\Big(-\frac{1}{\rho_j}\frac{\partial p_j}{\partial \nu_j} - i\sigma p_j\Big) \tag{5}$$

$$+g \quad \text{on } \Gamma_j \tag{6}$$

where (**??**) is to be fulfilled on $\Sigma_{j\ell}$, $p_j = p|_{\Omega_j}$ $\tau \in \mathbb{C}$, $|\tau| \le 1$, and the coupling parameter $\sigma > 0$, $\sigma \in \mathbb{R}$. The source term is denoted by g.

Define the function f, $f|_{\partial\Omega_j} = f_j$ on the element boundaries as follows

$$f_j = \left(\Big(-\frac{1}{\rho_j}\frac{\partial}{\partial \nu_j} - i\sigma\Big)p_j\right)\Bigg|_{\partial\Omega_j}, \quad 1 \le j \le K. \tag{7}$$

It is shown in [22,23] that f_j satisfies *the ultra weak variational formulation, (UWVF)*

$$\sum_{j=1}^{K}\int_{\partial\Omega_j}\frac{1}{\sigma}f_j\overline{\Big(-\frac{1}{\rho_j}\frac{\partial}{\partial \nu_j} - i\sigma\Big)q_j} \tag{8}$$

$$-\sum_{j=1}^{K}\sum_{\ell=1}^{K}\int_{\Sigma_{j\ell}}\frac{1}{\sigma}f_\ell\overline{\Big(\frac{1}{\rho_j}\frac{\partial}{\partial \nu_j} - i\sigma\Big)q_j}$$

$$+\sum_{j=1}^{K}\int_{\Gamma_j}\frac{1}{\sigma}f_k\overline{\Big(\frac{1}{\rho_j}\frac{\partial}{\partial \nu_j} - i\sigma\Big)q_j}$$

$$=\sum_{j=1}^{K}\int_{\Gamma_j}\frac{1}{\sigma}g\overline{\Big(\frac{1}{\rho_j}\frac{\partial}{\partial \nu_j} - i\sigma\Big)q_j} \tag{9}$$

for all test functions q_j which are the solutions of the adjoint Helmholtz equation

$$\Delta\overline{q}_j + \kappa_j^2\overline{q}_j = 0 \quad \text{in } \Omega_j, \tag{10}$$

where the overbar denotes complex conjugation.

Expressing the solutions in each element as a linear combination of appropriate plane waves (N_k waves in element Ω_k) and using these waves also as test functions as in the more conventional Galerkin approaches, the problem can be written in the form of the matrix equation [22,23]

$$(I - D^{-1}C)X = D^{-1}b. \tag{11}$$

where the unknowns $X = (f_{11}, \ldots, f_{KN_K})^T$ are to be determined. The matrices D and C are sparse and exhibit block structure. To avoid the conditioning problems reported in [22] we allow the number of bases N_j to vary between the elements [24].

Bioheat equation

The temperature evolution in a non-convective medium is governed by the heat equation which is of the form

$$\rho C\frac{\partial T}{\partial t} = \nabla\cdot k\nabla T + \widetilde{Q}, \tag{12}$$

where T is the temperature, ρ is the density of the medium, C is the heat capacity, k is the thermal conductivity and $\widetilde{Q}$ is the distributed heat source or sink [25].

In biological tissues the temperature evolution is usually approximated with the so-called bioheat equation [26]

$$\rho C_t\frac{\partial T}{\partial t} = \nabla\cdot k\nabla T + Q_0 + Q, \tag{13}$$

where $Q_0 = w_{\rm B} C_{\rm B}(T_{\rm A} - T) \leq 0$ is the perfusion (temperature sink) and $Q \geq 0$ is the distributed heat source that is due to the absorbed wave energy. Further, C_t is the heat capacity of tissue, $w_{\rm B}$ is the perfusion due to blood flow, $C_{\rm B}$ is the heat capacity of blood and $T_{\rm A}$ is the arterial blood temperature. The heat source term for the time-harmonic acoustic pressure is [25]

$$Q = \frac{\alpha |p|^2}{\rho c}. \tag{14}$$

Assume that the total field is due to m separate transducers so that $p = \sum_{k=1}^{m} p_k$. The fields p_k are of the form

$$p_k = \tilde{u}_k(t) \tilde{C}_k(r) e^{i\omega t} \,, \tag{15}$$

where $\tilde{u}_k(t) \in \mathbb{C}$ determine the amplitude and phase of the transducer source so that $C_k(r)$ are the time-harmonic solutions of the Helmholtz problems with single point sources of unit source strengths. Thus the heat source is of the form

$$\begin{aligned} Q(r,t) &= \frac{\alpha(r)}{\rho(r)c(r)} |p|^2 \\ &= \frac{\alpha(r)}{\rho(r)c(r)} \left| \sum_{k=1}^{m} \tilde{u}_k(t) \tilde{C}_k(r) \right|^2 \end{aligned} \tag{16}$$

The bioheat equation is discretized according to the usual semidiscrete scheme in which the spatial variable is handled with the Galerkin scheme and the resulting system of ordinary differential equations with appropriate (implicit) schemes such as backward Euler [27].

In the following we parameterize the complex control variables by their real and imaginary parts so that $u \in \mathbb{R}^{2m}$. Then the semidiscrete FEM approximation for the bioheat equation can be written in the form

$$\begin{aligned} M\dot{T} &= (G - w_{\rm B} C_{\rm B} I) T \\ &\quad + w_{\rm B} C_{\rm B} M T_{\rm A} + \tilde{M}_{\rm D} (Bu)^2 \end{aligned} \tag{17}$$

where M is the (ordinary) mass matrix, G is the stiffness matrix, $\dot{T} = dT/dt$, $\tilde{M}_{\rm D}$ is a stack of two inhomogeneous mass matrices corresponding to **(??)** and B is an appropriate real-valued representation of the fields from the transducers (as obtained by the UWVF). In the following this is considered in the form

$$\dot{T} = AT + P + M_{\rm D} (Bu)^2 \tag{18}$$

where we have made the obvious assignments.

CONTROLLER IMPLEMENTATION

Optimality criterion and spatial discretization

For general controller design we refer to [28,29]. Define the quadratic cost function

$$\begin{aligned} \tilde{J}(u) &= \frac{1}{2} \int_0^{t_f} \Bigg\{ \|T(t) - T_{\rm d}(t)\|_\vartheta^2 \\ &\quad + \sum_{k=1}^{2m} s_k \left(\frac{du(t)}{dt} \right)^2 \Bigg\} dt \end{aligned} \tag{19}$$

where $T_{\rm d} = T_{\rm d}(r,t)$ is the desired temperature distribution and s_k are weights for the time derivative of the input thus enforcing smoothness of the control variables. Further, $\|T(t) - T_d(t)\|_\vartheta^2 = \int_\Omega \vartheta(r) \big(T(r,t) - T_{\rm d}(r,t)\big)^2 \, dr$.

In practice the maximum power or pressure amplitude is constrained so that the relevant control problem is of the form

$$\min_u \tilde{J}(u) \qquad \text{subject to } u_k^2 + u_{k+m}^2 \leq \zeta \tag{20}$$

for all $k = 1, \ldots, m$, where u_k and u_{k+m} are the real and imaginary parts for transducer k, respectively, and where we can take $\zeta = 1$ with an appropriate change of variables. This is a quadratic problem with quadratic inequality constraints. Due to the nonlinear constraints and the nonlinearity of the mapping $u \mapsto T$ we have to resort to numerical minimization methods. Furthermore, we approximate the inequality constraint by introducing an additional nonlinear penalty so that the we can define the cost function that is adopted in this paper as

$$J(u) = \tilde{J}(u) + \chi_R(u) \tag{21}$$

where $R \in \mathbb{R}$ and

$$\chi_R(u) = \frac{1}{2} \int_0^{t_f} \sum_{k=1}^{2m} R^{-1} \exp\big(2R|u_k|\big) \, dt \tag{22}$$

The Hamiltonian of the control problem is now

$$\begin{aligned} H &= \frac{1}{2} \Bigg\{ \|T(t) - T_{\rm d}(t)\|_\vartheta^2 \\ &\quad + \sum_{k=1}^{2m} s_k \dot{u}_k^2 + \sum_{k=1}^{2m} R^{-1} \exp\big(2R|u_k|\big) \Bigg\} \\ &\quad + \lambda^{\rm T} \big(AT + P + M_{\rm D} (Bu)^2 - \dot{T} \big) \end{aligned} \tag{23}$$

where $\lambda(t) \in \mathbb{R}^N$, $t \in [0, t_f]$, is the Lagrange undetermined coefficent. In this paper we take $\lambda_k(t_f) = 0$ for all k.

Direct temporal discretization

In this paper we solve the problem by directly discretizing the control and state variabls as well as the Lagrange multiplier with respect to time and using a steepest descent type algorithm for their solution. In the minimization procedure we employ a three-step approach, which is more easily implemented than the straightforward single-step approach involving the gradients of the complete Hamiltonian.

Let the temporal discretization constant be Δt and $N_T = t_f/\Delta t + 1$. In the sequel we denote the temporally discretized variables as $u_t = u(t\tau/\Delta t) \in \mathbb{R}^{2m}$, $\tau \in [0, t_f]$, $t \in [0, \ldots, N_T]$, with other variabes denoted correspondingly. The discretized Hamiltonian is then of the form

$$\begin{aligned} H &= \frac{1}{2}\Bigg(\|T_t - T_{\mathrm{d},t}\|^2_{\vartheta} \\ &\quad + \sum_{k=1}^{2m} s_k \left(u_{k,t} - u_{k,t-1}\right)^2 \\ &\quad + \sum_{k=1}^{2m} R^{-1} \exp\left(2|R|u_{k,t}\right)\Bigg) \\ &\quad + \lambda^{\mathrm{T}}\left(AT_t + P + M_{\mathrm{D}}(Bu)^2 - \dot{T}_t\right) \end{aligned}$$

where Δt^{-2} has been absorbed in s_k.

The time evolution for the system and co-state equations is approximated with the implicit (backward) Euler approach so that we can write

$$\begin{aligned} T_{t+1} &= (I - \Delta t A)^{-1} T_t + \Delta t (I - \Delta t A)^{-1} \\ &\quad \cdot \left(P - M_{\mathrm{D}}(Bu)^2\right) \qquad (24) \\ \lambda_{t+1} &= (I - \Delta t A)^{-1} \lambda_t + \Delta t (I - \Delta t A)^{-1} \\ &\quad \cdot \vartheta \left(T_t + T_{\mathrm{d}}\right) \qquad (25) \end{aligned}$$

The stationary condition is pursued by the Levenberg-Marquardt type stabilized iteration with the search direction

$$\begin{aligned} \nabla_{u_t} H &= \left(F(u_t)^{\mathrm{T}} F(u_t) + \mu I\right)^{-1} \\ &\quad \cdot \left(L(u_t) + F(u_t)^{\mathrm{T}} \lambda_t\right) \qquad (26) \end{aligned}$$

where μ is the stabilization parameter and

$$\begin{aligned} L(u_t) &= \mathrm{sign}\,(u_t) \odot \exp(2R|u_t|) \\ &\quad + S\left(u_t - u_{t-1}\right) \\ F(u_t) &= 2M_{\mathrm{D}}\left(B \odot (Bu, \ldots, Bu)\right) \end{aligned}$$

where $S = \mathrm{diag}\,(s_1, \ldots, s_{2m})$, and $\odot$ denotes the elementwise product of two matrices or vectors. Due to the complexity of the mapping $u \mapsto T$, the algorithm does not necessarily converge to a global minimum.

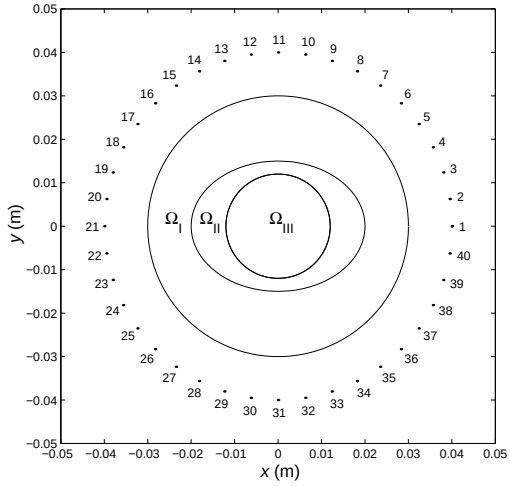


Figure 1. The computing domain. There are 40 point sources located around the computing domain (numbered $1, \ldots, 40$). The acoustic parameters are given in Table 1.

SIMULATION RESULTS

The simulations were carried out in a 2D domain. The computational domain for simulations is shown in Fig. 1. The domain consists of the three subdomains Ω_{I}, Ω_{II} and Ω_{III} with different physical parameters. The physical parameters are given in Table I

TABLE I. The Acoustic Parameters in Different Media for the Control Simulations.

Parameter	Ω_{I}	Ω_{II}	Ω_{III}
Speed of sound c (m/s)	1500	2500	2000
Density ρ (kg/m^3)	1000	2000	1500
Abs. coef. α (Nep/m)	0	4	2

The domain was divided into 840 elements and 445 nodes. The ultrasound fields with frequency of 500 kHz were computed with the UWVF for each point source separately. In this example we consider a system and a specific application in which the maximum pressure amplitude is constrained to less than 1 MPa. Figure 2 shows the mesh and the normed intensity of the UWVF solution of the Helmholtz equation from point source (transducer) 1.

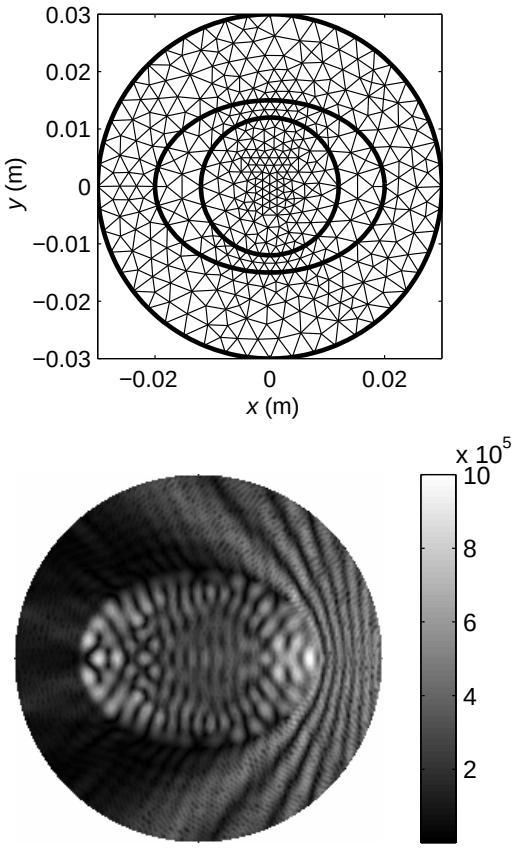


Figure 2. Top: The computing mesh that consists of the 840 elements and 445 nodes. Bottom: The normed intensity $|p|$ from the UWVF solution of the Helmholtz equation from point source (transducer) 1.

The optimal control u_t was computed with the algorithm given in the previous section. The target heat distribution was of the form of the letter "T" in the middle of Ω_{III}. The desired temperature in the target region was 45°C while the desired temperature in other parts of the computing domain was 37°C. The target and the controlled temperature distribution at the final time $t_f = 10$ s are shown in Figure 3. The desired temperature in the target is obtained fairly well.

TABLE II. The Thermal Parameters for Control Simulations.

Heat capacity of tissue C_t (J/kgK)	3700
Thermal cond. of tissue k_t (W/mK)	0.6
Perfusion by blood flow w_b (kg/m^{3}s)	1
Heat capacity of blood C_b (J/kgK)	3800
Arterial blood temperature T_a (°C)	37

The optimal controls for transducers 18 and 20 are shown in Fig. 4. The trajectories are smooth as expected and the amplitudes and thus also the powers of the transducers vanish at the final time. This is due to the chosen final constraint for the Lagrange multipliers. This could also be relaxed which would yield a more homogeneous excitation of the transducers.

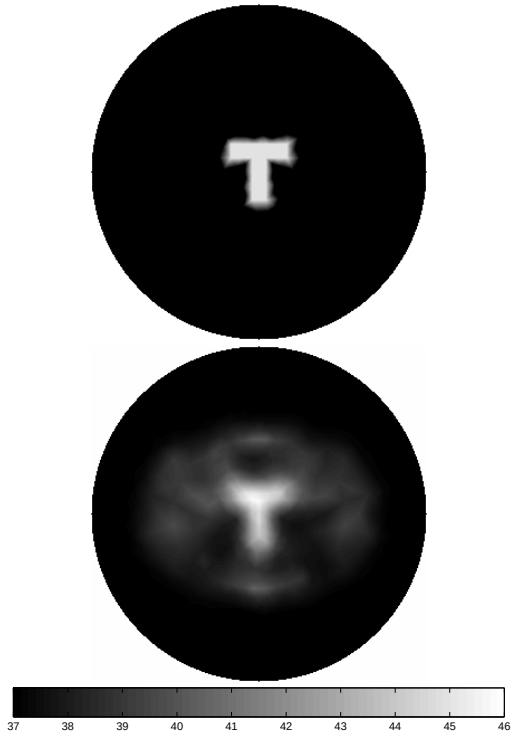


Figure 3. Top: The target temperature distribution with form of the letter T. Middle: The desired temperature in target is 45°C. Bottom: The controlled temperature distribution at the final time t_f =10 s.

CONCLUSIONS

The controller which was proposed in this paper can be applied to ultrasound or microwave induced heating. In microwave induced heating the electrical field is computed from the Maxwell equations. The control algorithm is then applied to the problem in a similar way. The controller proposed here is an approximation for the optimal controller concerning quadratic costs with inequality constraints for the control variables. The simulation indicates that this method is able to produce accurate heat distributions in inhomogeneous media.

The proposed method does not produce a scanning focus type heat source. This may be an important asset in the sense that typical maximum *prefocused* wave field intensities are just on the verge of needin to employ nonlinear wave propagation models. In our case it seems that this is not necessary.

Actual ultrasound surgery is concerned with thermal dose rather than controlled temperature. However, the modification of the proposed temperature control method is the starting point for the thermal dose control problem.

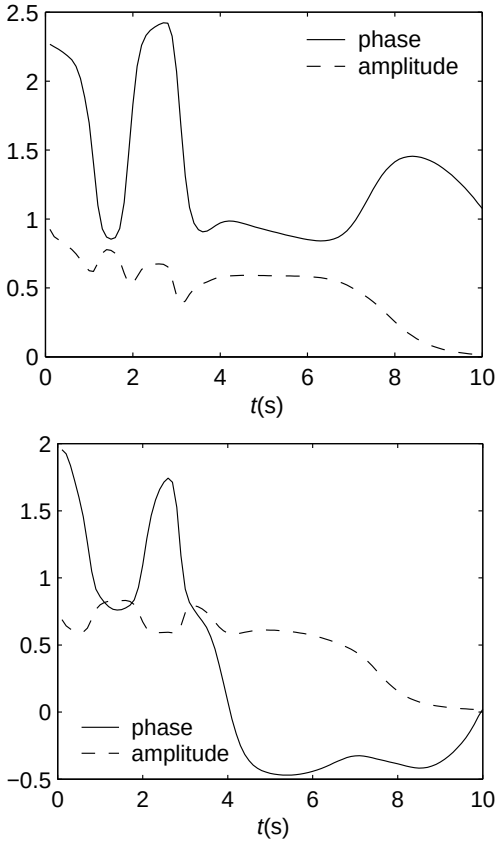


Figure 4. Two optimal amplitude and phase evolutions, transducers 18 (top) and 20 (bottom).

REFERENCES

1. F. A. Duck, A. C. Baker, and H. C. Starrit, *Ultrasound in Medicine*, Institute of Physics Publishing, 1998.

2. K. Hynynen, *Methods of External Hyperthermic Heating*, Springer-Verlag, 1990.

3. G. T. Clement, J. White, and K. Hynynen, "Investigation of a large-area phased array for focused ultrasound surgery through the skull," *Physics in Medicine and Biology*, vol. 45, pp. 1071–1083, 2000.

4. J. Sun and K. Hynynen, "The potential of transskull ultrasound therapy and surgery using the maximum available skull surface area," *Journal of the Acoustical Society of America*, 1999.

5. F. Bardati, A. Borrani, A. Gerardino, and G. A. Lovisolo, "Sar optimization in a phased array radiofrequency hyperthermia system," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 12, pp. 1201–1207, December 1995.

6. J. D. Doss, "Simulation of automatic temperature control in tissue hyperthermia calculations," *Medical Physics*, vol. 12, no. 6, pp. 693–697, November/December 1985.

7. T. Köhler, P. Maass, and P. Wust, *Surveys on Solution Methods for Inverse Problems*, Springer-Verlag, 2000.

8. P. VanBaren and E. S. Ebbini, "Multi-point temperature control during hyperthermia treatments: Theory and simulation," *IEEE Transactions on Biomedical Engineering*, 1995.

9. C. Johnson, R Kress, R Roemer, and K. Hynynen, "Multi-point feedback control system for scanned, focused ultrasound hyperthermia," *Physics in Medicine and Biology*, vol. 35, no. 6, pp. 781–786, 1990.

10. L. Win-Li, R.B. Roemer, and K. Hynynen, "Theoretical and experimental evaluation of a temeperature controller for scanned focused ultrasound hyperthermia," *Medical Physics*, 1990.

11. K. S. Nikita, N. G. Maratos, and N. G. Uzunoglu, "Optimal steady-state temperature distribution for a phased array hyperthermia system," *IEEE Transactions on Biomedical Engineering*, 1993.

12. M. Mattingly, R. B. Roemer, and S. Devasia, "Exact temperature tracking for hyperthermia: A model-based approach," *IEEE Transactions on Control Systems Technology*, 2000.

13. Y. Y. Chen, W. L. Lin, H. L. Liou, J. Y. Yen, and M. J. Shieh, "Self-tuning fuzzy logic control for ultrasound hyperthermia with reference temperature based on objective functions," *Medical Physics*, vol. 26, no. 5, pp. 825–833, 1999.

14. Kowalski M. E. and J. M. Jin, "Determination of electromagnetic phased-array driving signals for hyperthermia based on a steady-state

temperature criterion," *IEEE Transactions on Microwave Theory and Techniques*, vol. 48, no. 11, pp. 1864–1873, 2000.

15. A. B. Bhatia, *Ultrasonic Absorption: An Introduction to the Theory of Sound Absorption and Dispersion in Gases, Liquids and Solids.*, Dover Publications, Inc., 1967.

16. X. Fan and K. Hynynen, "The effect of wave reflection and refraction at soft tissue interfaces during ultrasound hyperthermia treatments," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1727–1736, 1992.

17. E. Kühnicke, "Three-dimensional waves in layered media with nonparallel and curved interfaces: A theoretical approach," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 709–716, 1996.

18. Y. Y. Botros, J. L. Volakis, P. VanBaren, and E. S. Ebbini, "A hybrid computational model for ultrasound phased-array heating in the presence of strongly scattering obstacles," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 11, pp. 1039–1050, 1997.

19. Y. Y. Botros, E. S. Ebbini, and J. L. Volakis, "Two-step hybrid virtual array-ray (VAR) technique for focusing through the rib cage," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 45, no. 4, pp. 989–1000, 1998.

20. I. Babuska and J. M. Melenk, "The partition of unity method," *International Journal for Numerical Methods in Engineering*, vol. 40, pp. 727–758, 1997.

21. P. Monk and D. Wang, "A least squares method for the Helmholtz equation," *Computer Methods in Applied Mechanics and Engineering*, vol. 175, pp. 121–136, 1999.

22. O. Cessenat and B. Despres, "Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem," *SIAM Journal of Numerical Analysis*, vol. 35, no. 1, pp. 255–299, 1998.

23. O. Cessenat, *Application d'une nouvelle formulation variationelle des equations d'ondes harmonigues, Problemes de Helmholtz 2D et de Maxwell 3D*, Ph.D. thesis, Paris IX Dauphine, 1996.

24. T. Huttunen, P. Monk, and J. P. Kaipio, "Computational aspects of the ultra weak variational formulation," *Journal of Computational Physics*, 2001, submitted.

25. A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, Acoustical Society of America, 1991.

26. H. H. Pennes, "Analysis of tissue and arterial blood temperatures in the resting human forearm," *Journal of Applied Physiology*, 1948.

27. C. Johnson, *Numerical Solution of the Partial Differential Equations by the Finite Element Method*, Studentlitteratur, 1987.

28. R. F. Stengel, *Optimal Control and Estimation*, Dover, 1994.

29. J. B. Burl, *Linear Optimal Control.* H_2 *and* H_∞ *Methods*, Addison-Wesley, 1999.

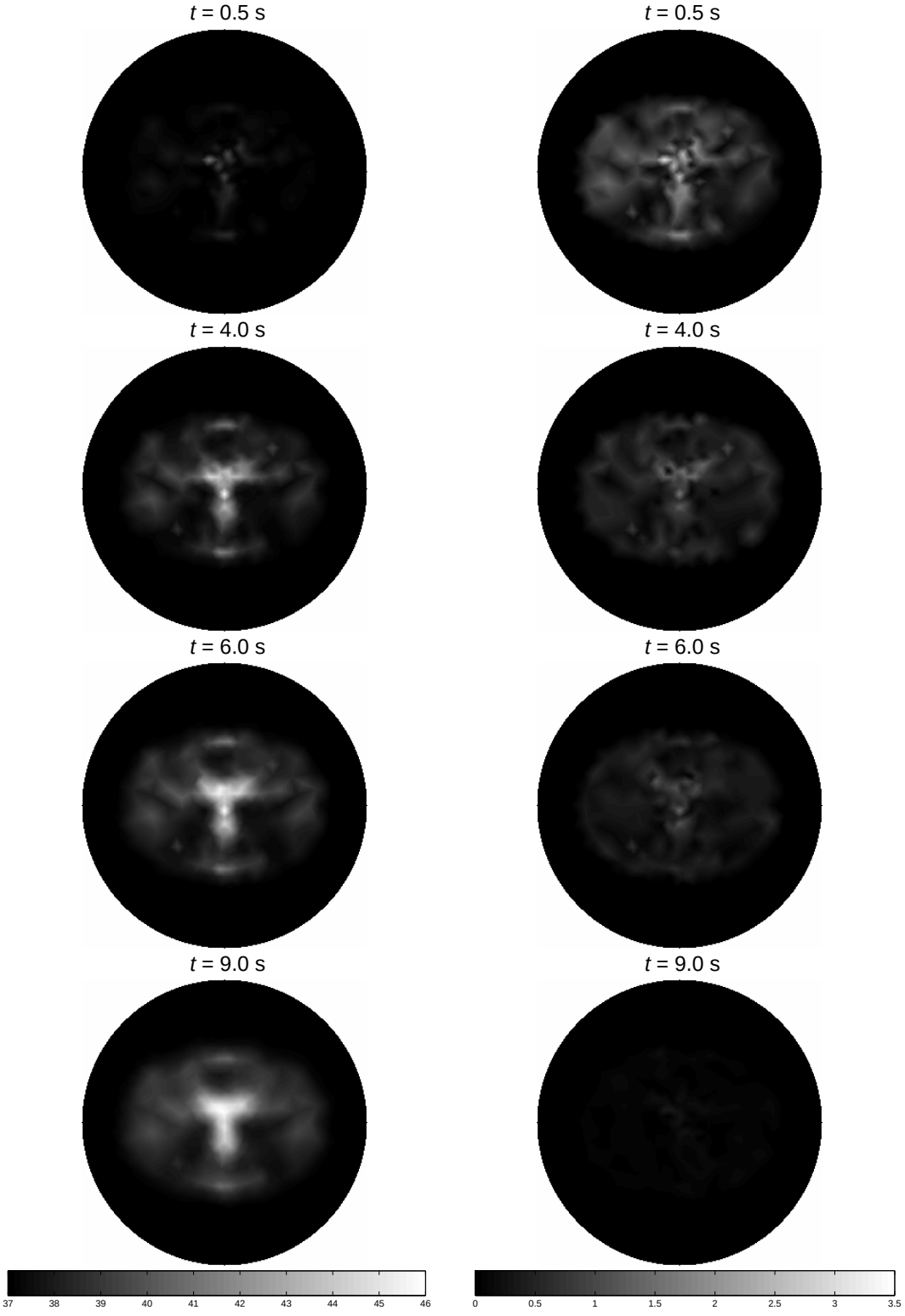


Figure 5: The temperature evolution during the sonication. The left hand column the temperature distribution and the right hand column the square root of the induced distributed heat source, that is, $\sqrt{M_{\mathrm{D}}(Bu)^2}$.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

ROBUST DESIGN OPTIMIZATION STRATEGY OF IOSO TECHNOLOGY

Igor N. Egorov, Gennadiy V. Kretinin, Igor A. Leshchenko
IOSO Technology Center, Moscow, RUSSIA
egorov@iosotech.com

ABSTRACT

The paper presents the main capabilities of the Robust Design Optimization (RDO) strategy of the IOSO (Indirect Optimization based on Self-Organization) Technology. The capabilities of RDO software are demonstrated using examples of solving complex multidimensional (up to 140 design variables) problems. The examples utilize both single and multiobjective optimization problems. Our strategy summarize more than ten years of using RDO for solving real-life problems in various scientific and technical fields. The paper presents the assembly of the newly developed efficient approaches for solving problems requiring RDO. These approaches employ technology of multilevel, multiobjective, and parallel optimization both separately and simultaneously.

INTRODUCTION

Practical application of the numerical optimization results is complicated by the fact that any intricate technical system is a stochastic system and characteristics of this system have a probabilistic nature. We would like to emphasize the point that, speaking of stochastic properties of a technical system within the frame of optimization tasks, we imply a system's essential parameters spread which occurs during the production stage despite the up-to-date level of technology. Random deviations of the system's parameters lead to a random change in system's efficiency.

An efficiency extremum value, obtained during the optimization problem solving in a traditional (deterministic) setting, is simply a maximum attainable value and can be considered as just conventional optimum from the point of view of its practical realization. Thus, one can consider two different types of optimization criteria (fig. 1). One of them is an ideal efficiency which can be achieved under the conditions of absolutely precise practical replication of the preset parameters of the system under consideration (deterministic criterion). Other optimization criteria are of probabilistic nature. For example: mathematical expectation of the efficiency; total probability of assuring the preset constraints; variance of the efficiency and so on

It is evident that the extremum of one of these criteria doesn't guarantee the assurance of the high level of another one. Even more, these criteria may be contradicting each other. Thus, in this case we have a multicriteria optimization problem.

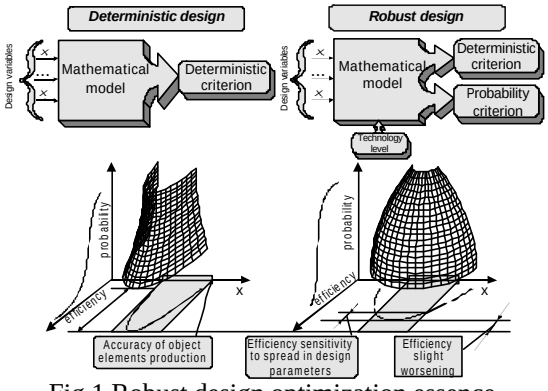


Fig.1 Robust design optimization essence.

Our concept of robust design optimization and robust optimal control allows determining the optimal practical technical solution that could be implemented with the high probability for the given technology level of the production plants [1, 3, 5]. Many current probabilistic approaches either employ estimation of probabilistic efficiency criteria only at the stage of analysis of obtaining deterministic solution, or use significantly simplified estimates of probabilistic criteria during optimization process. The distinctive feature of our approach is that during robust design optimization we solve the optimization problem using direct stochastic

formulation, when estimation of probabilistic criteria is accomplished at each iteration. This procedure reliably produces truly robust optimal solution.

THE MAIN FEATURES OF THE ROBUST DESIGN OPTIMIZATION STRATEGY

IOSO Technology implements the new evolutionary response surface methodology. This methodology differs significantly from both the traditional approaches of nonlinear programming and the traditional response surface approach. Because of that IOSO Technology algorithms have higher efficiency, provide wider range of capabilities, and are practically insensitive with respect to the types of objective function and constraints: smooth, non-differentiable, stochastic, with multiple optima, with the portions of the design space where objective function and constraints could not be evaluated at all, with the objective function and constraints dependent on mixed variables, etc. (see fig. 2).

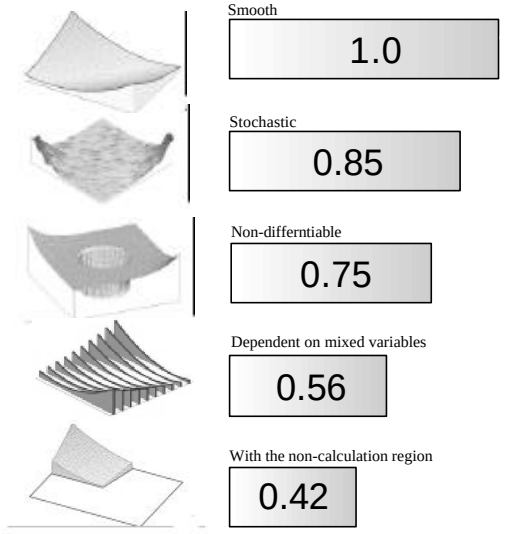


Fig. 2 IOSO algorithms efficiency for different objective functions.

The concepts of Robust Design Optimization and Robust Optimal Control allow finding an optimal technical solution for the particular technology level, accounting that such a technical solution could be realized in practice with high probability. Some other approaches perform evaluation of probability parameters only after the deterministic optimal solution is found or employ very simplified estimates of probability parameters during optimization process. The distinctive feature of our RDO strategy is that optimization problem is solved using stochastic formulation directly, when the evaluation of probability parameters is performed at each iteration. High efficiency of the Robust Design Optimization is provided by the highly efficient capabilities of the developed stochastic optimization algorithms, that reliably work when high level of noise is present in responses. This is confirmed by the thorough testing of the algorithms using well-known test functions.

Using our Robust Design Optimization concept results in considerable (several orders of magnitude) cost and time reduction when developing new highly efficient technical systems. Using Robust Design Optimization concept also provides the considerable (several times) risk reduction when new technical solutions are implemented.

These features were demonstrated during microprocessor control system optimal calibration of the actual automotive engine, when the time reduction of five times was achieved. The methodic of optimal calibration was suggested in [2]. The brief description of this optimization problem is given in the table 1.

Table 1. Brief description of the automobile engine optimal calibrating problem.

Purpose	To insure minimum overthrow of air-fuel ratio (α) during acceleration and throttling processes.
Setting features	3 independent variables; one nonlinear constraint; object under study – actual engine on the experimental bench.
Optimization process features	Two stages of $\Delta\alpha$ minimization: 1. for acceleration when $\Delta\alpha$ of throttling is being constrained; 2. for throttling when $\Delta\alpha$ of acceleration is being constrained.

For solution of this optimization problem we used the 21 experiments only. Objective

improvement history and optimization results are shown on the fig. 3, 4 respectively.

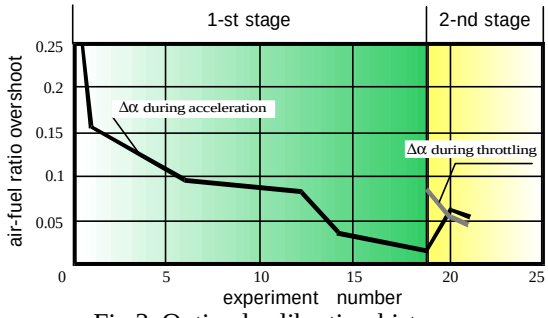


Fig 3. Optimal calibrating history.

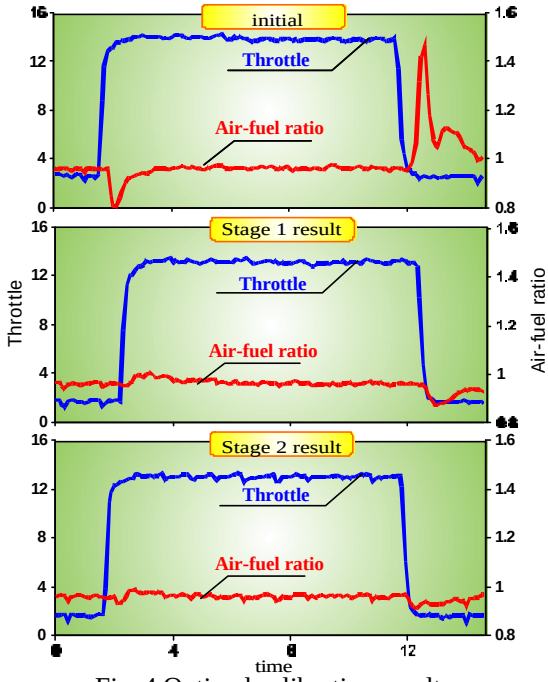


Fig. 4 Optimal calibrating results.

The distinctive feature of our approach is the ability to solve the problems with the large number of design variables (hundreds) and objectives (dozens). These features are available through different original procedures within the frames of robust design optimization strategy. Let us consider some of these procedures.

MULTIOBJECTIVE ROBUST DESIGN OPTIMIZATION

We developed the very effective algorithms of the Multiobjective Robust Design Optimization. The main advantages of the proposed algorithms over traditional mathematical programming approaches are the following [3].

- convolution approaches are not used in solving multiobjective problems;
- the algorithms determine the desired number of Pareto-optimal solutions, so that these solutions are uniformly distributed in the space of objective functions;
- it is possible to solve the optimization problems where the objective functions exhibit complex topology: non-convex, non-differentiable, with many local optima;
- high probability of locating a global optimum in a design space having many local optima;
- relatively small number of mathematical models evaluations;
- it is possible to naturally employ the parallelization of the computational process.

These advantages are the basis for the wide use of the proposed method in the real-life problems.

Let us consider the example of the multiobjective robust design optimization of the multistage axial flow compressor. The brief description of this optimization problem is given in the table 2.

Table 2. Brief description of the compressor robust design optimization problem.

Purpose	To insure the maximum efficiency and maximum implementation probability under preset level of production technology.
Setting features	140 independent variables (flow-path geometry); two objectives; three nonlinear constraints; object under study – quasi-3D mathematical model.
Optimization process features	The set of Pareto-optimal solutions were found.

Fig. 5 shows the main results of this problem. One can see that there is a compromise area between the ideal (deterministic) compressor efficiency and the implementation probability. In general, designer can select any solution from the

obtained set. In this case the design No 4 was selected as the final design.

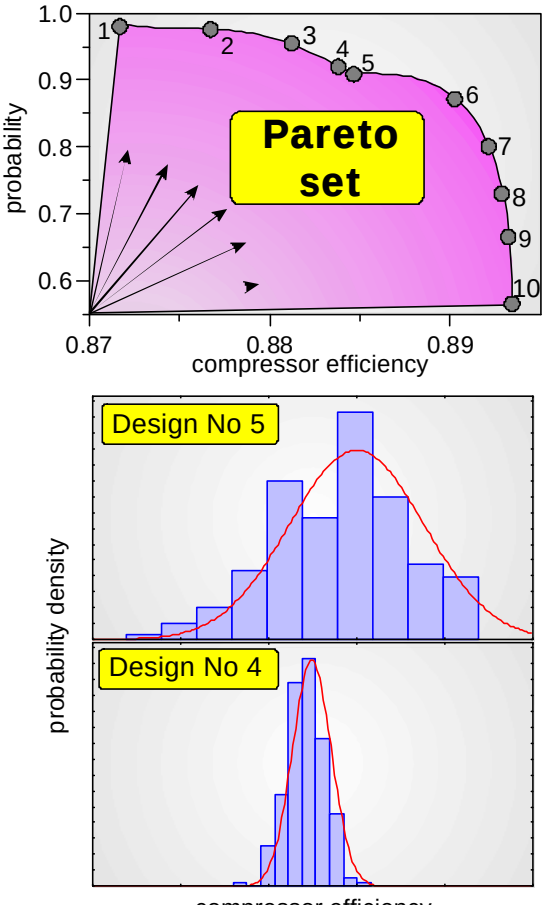


Fig. 5 Results of compressor multicriteria robust design optimization.

MULTILEVEL ROBUST DESIGN OPTIMIZATION

The feature of the Multilevel Robust Design Optimization procedure is the use of mathematical models of various fidelity (from the lowest to the highest) during the solution process and adaptive switching between them [4]. This procedure provides minimization of the number of times the high fidelity models are used without reducing the accuracy of the resulting solution (fig. 6).

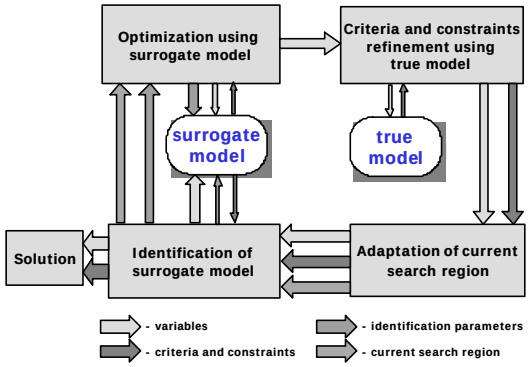


Fig. 6 Multilevel optimization scheme.

The efficiency of this procedure may be demonstrated using the compressor optimization problem with 63 independent variables (fig. 7). For this problem we obtained 10 Pareto-optimal solutions using only 60 direct calls to high-fidelity model. This example shows that it is possible to solve the optimization problem when the number of times the highest fidelity model is involved is less than the number of design variables. This provides considerable (several orders of magnitude) reduction in CPU time required for solution of complex optimization problems.

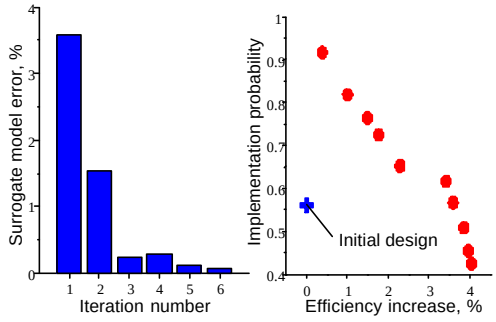


Fig. 7 Results of compressor multilevel robust design optimization.

PARALLEL ROBUST DESIGN OPTIMIZATION

One of the prospective trends in improving optimization process efficiency is the use of computers with multiple processors. In this case, the reduction of elapsed (clock) computing time can be achieved through solution time reduction

by means of parallel computations "inside" the model, as well as by adaptive organization of the optimization process for parallel computations. The first approach implies the use (or development) of mathematical analysis models suitable for using parallel processors. The latter makes it necessary to develop or to modify the corresponding optimization methods.

We have developed the new optimization algorithm, which uses parallel processors (fig. 8). Our algorithm allows us to reach the speed-up parameter value that exceeds the total number of operational CPUs. For example, when using 20 processors we can speed-up the optimization process 40 and more times. Our algorithms allow the most efficient usage of existing computational resources because the number of processors actively involved in solving the problem is independent of the problem dimension. For example, when solving a 10 variable problem one can employ from 1 to 100 and more processors.

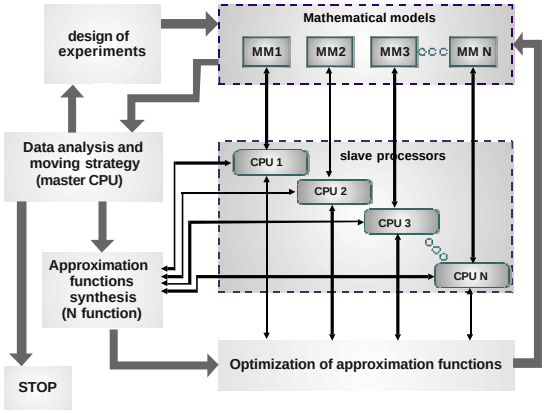


Fig. 8 Parallel optimization scheme

High efficiency of our procedure is due to not only the simple and typical procedure of parallelizing computation of objective function and constraints as it is done in the majority of the known apporaches. In addition to that we parallelize the optimization process as a whole. This is accomplished by obtaining maximum possible information about the topology of the objective function and constraints using our specific response surface methodology at each iteration. Particularly, we construct a set of various approximation function with various properties (local and global accuracy, robustness, good prediction capabilities, etc.). Analysis of the set of these functions and defining new points for analysis at the next iteration is also done in parallel. It is shown that such approach provides considerable (several orders of magnitude) reduction in time required to solve optimization problem. This is makes it realistic to formulate and solve the optimization problems even when many hours are required to the response values for one combination of design variables (for example, 3D CFD codes).

CONCLUSION

Our experience in using the various Robust Design Optimization procedures applied to test problems and real-life problems shows that the total efficiency of the optimization process could be increased 5-10 times. This indicates that combining these procedures significantly broadens the capabilities of the Robust Design Optimization strategy when applied to real-life systems. This strategy is also shown to be a powerful tool for finding new technical solutions, which in turn provide the maximum possible efficiency of complex systems.

REFERENCES

1. I. N. Egorov, Optimization of a Multistage Axial Compressor. Stochastic Approach, ASME, 92-GT-163, 1992.

2. I. N. Egorov, G. V. Kretinin, B. Y. Chernjak, et al., Fast methods of experimental calibrating of microprocessor control system, Proceedings of 5-th International Congress EAEC, SIA9506A36, Strasbourg, France, 1995.

3. I. N. Egorov and G. V. Kretinin, Search for Compromise Solution of the Multistage Axial Compressor's Stochastic Optimization Problem, World Publishing Corporation, *Aerothermodynamics of internal flows III*, pp. 112-120, Beijing, China, 1996.

4. I. N. Egorov, G. V. Kretinin, I. A. Leshchenko, Y. I. Babiy, Optimization of complex engineering systems using variable-fidelity models, MCB University Press, ISBN: 0-86176-650-4, Proceedings of the 1st ASMO UK/ISSMO conference on Engineering Design Optimization, pp.143-149, 1999.

5. I. N. Egorov, G. V. Kretinin, I. A. Leshchenko, Stochastic Optimization of Parameters and Control Laws of the Aircraft Gas-Turbine Engines – a Step to a Robust Design, Elsevier Science Ltd, "Inverse Problem in Engineering Mechanics III", pp.345...353, 2002.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

PARAMETERIZED GEOMETRY FORMULATION FOR INVERSE DESIGN AND OPTIMIZATION

Helmut Sobieczky
German Aerospace Center (DLR)
Bunsenstrasse 10, D-37073 Göttingen, Germany
helmut.sobieczky@dlr.de

George S. Dulikravich
Department of Mechanical and Aerospace Engg.
The University of Texas at Arlington
Arlington, TX 76019, U.S.A.
gsd@mae.uta.edu

Brian H. Dennis
Institute of Environmental Studies,
University of Tokyo
7-3-1 Hongo, Bunko-ku, Tokyo 113-8656, Japan
dennis@garlic.q.t.u-tokyo.ac.jp

ABSTRACT

This contribution focuses on the importance of preprocessing tools for successful design and optimization in practice of turbomachinery engineering. The development of problem-oriented computational geometry generation software is illustrated for the example of aerodynamic inverse design of transonic flow elements which define the compatible boundary conditions (surfaces) in detail. Resulting from learned sensitivity of high speed flows to small changes in airplane wing or turbomachinery blade geometry, preprocessing software is provided to create parametric shapes to be varied for optimization cycles or numerical simulation of mechanical adaptation processes. Supporting the need to design from a multidisciplinary viewpoint, parameterized geometry components for aerodynamic, as well as for thermal and structural considerations are defined. Examples for turbomachinery blade design and optimization are given.

INTRODUCTION

In the past years with rapid expansion of computer speed and storage, and improvements in algorithm speed and accuracy, optimization strategies have become affordable and reliable. Computational analysis and simulation of physical phenomena therefore become valuable design tools to improve technological performance of a product component. Here we focus on the complex technology of coupling the aerodynamics, structural and thermal loading as occurring in turbomachinery component design. In this situation we need realistic and flexible surface modelling to provide boundary conditions produced systematically and in rapid succession, with variations controlled by suitable and efficient sets of parameters.

High speed aircraft design is posing similar coupled problems, as outlined in [1]. Here we use some of the chapters in this book to be adapted and further developed for turbomachinery problems, like aerodynamic blade design with thermal and structural constraints.

With geometry data of a machine component being the common database for desirable aerodynamic, thermodynamic and structural considerations, to name only the most important of disciplines relevant for successful product development in the early engineering phase, we should explain some of the background of these fields as far as they have influenced parametrization of our geometry preprocessor.

GASDYNAMIC PHENOMENA, INVERSE AND DIRECT DESIGN TOOLS

Flow machinery, just like aircraft wings and other free form shapes with a need of refined surface quality is sensitive to the physical phenomena especially in the high speed domain. The knowledge base of transonic and supersonic gasdynamics tells us about regions of influence and

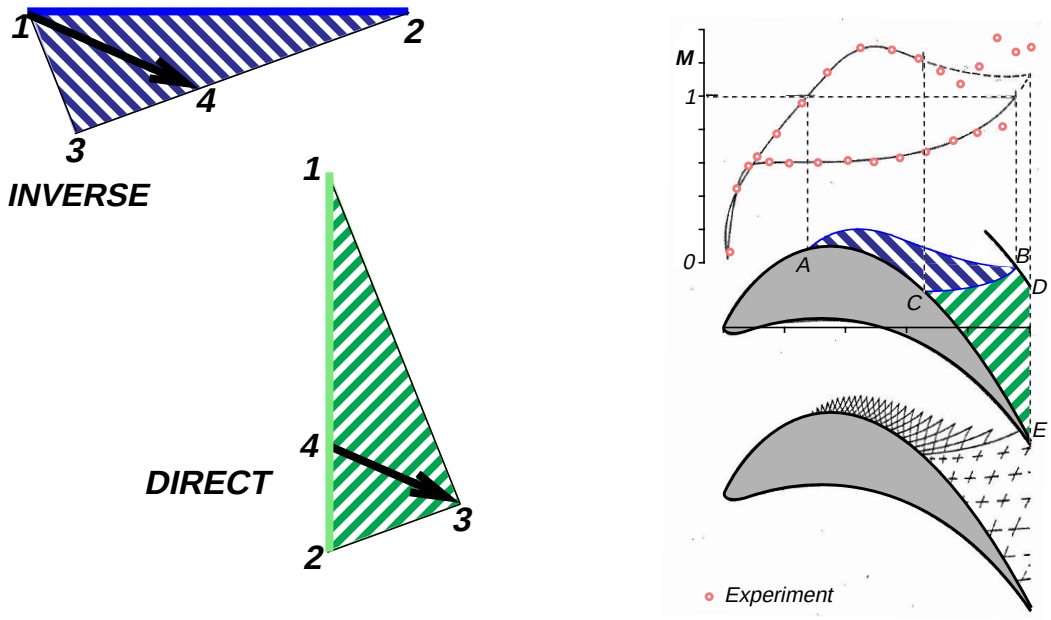


Figure 1: Principle of inverse and direct supersonic design; application of both inverse and direct methods to redesigning parts of the contour for a turbine blade test case.

dependence, this way suggesting the definition of section geometry definition observing the lack of upstream influence of any shape changes in certain regions.

Applied to the design of turbine blades we use such phenomena to perform flow computations both in a direct (downstream) or an inverse (cross-flow) marching procedure (Fig. 1). The latter allows to use certain given starting data to compute the flow along with a compatible boundary condition: the shape results from this inverse approach which, in practice, means that we may control the flow quality to avoid or delay negative effects like separation and obtain design hints how to shape a blade to actually observe this desired flow quality.

Figure 1 shows the principle, infinitesimally starting from known data along ($\overline{12}$), finding the solution within a triangle ($\overline{123}$), for potential flow computation, or with entropy updates along streamlines ($\overline{14}$) for an Euler accurate CFD simulation. Also shown is the application to an experimentally tested turbine blade design: the supersonic domain is re-constructed by starting from the given sonic line ($\overline{AB}$) in the inverse mode, then continuing downstream of ($\overline{BC}$) in the direct mode.

This transonic design method has been used for turbomachinery cascades [2] and many airplane wing design examples [3].

In the following we may not need to use these methods but we use geometry preprocessors which use parametric airfoils and other component functions which have been tailored using experience with this design concept. So we ensure to be close to desirable conditions in the aerodynamic part of the many needed optimization steps in design practice.

GEOMETRY MODELS WITH PARAMETRIC SHAPE CONTROL

Results from the above cited inverse approach in aerodynamics have taught us about shape sensitivities [4] and consequently about the needed refined parameter definition for the following more recent and future optimization efforts. In practical design, there will be a more multidisciplinary approach trying to optimize aerodynamics, structure, thermal properties etc. in a synchronized way. Here we call for a setup of parameters for controlling the complete set of boundaries to vary the shape for each discipline effectively. Restricting our illustrations to turbine blade technology, which might be resulting from the above illustrated

design process, now needs to be created including its structure of coolant passages to allow for a design optimization including structural and thermal loads.

Without knowledge of some physical properties of aerodynamics leading to a suitable parametrization, the size and shape of the mathematical space that contains all the design variables (for example, coordinates of all blade surface points) is very large and complex in a realistic cooled blade geometry. Only when it is possible to use fast flow-field analysis codes could it be affordable to have an ideal optimization situation where each surface grid point on the optimized configuration is allowed to move independently. Otherwise, the designer is forced to somewhat restrict the design space by working with a relatively small number of the design variables by performing parameterization - if not by a specialized software like the one introduced here, for example, by fitting polynomials - of either the 3-D surface geometry or the 3-D surface pressure. The optimization code then needs to identify the coefficients in these polynomials. Since it is often necessary to constrain and sometimes not allow motion of certain parts of the 3-D surface, the most promising choices for the 3-D parameterization appear to be different types of Bezier functions [5] and the geometry preprocessing tools used here which is based on a library of suitable analytical functions and successive manipulations and integrations in 3D cartesian coordinates ([1], pp 123-136).

This approach allows to vary the airfoil parameters as found suitable from 2D design (Fig. 1) into the third dimension, to compose a 3D blade with drastically changing sections as occurring between the root and tip sections of a realistic turbine blade. Moreover, mathematical description of every surface point without any interpolation and iteration to approximate given data, allows for an easy construction of parallel surfaces as needed to meet wall thickness constraints. These are crucial when the inner structure of the blade needs to house a coolant flow passage reducing the heat load on the blade and still maintain structural stiffness to support the forces produced by the flow and through structure transferred to yield shaft torque.

A starting geometry for subsequent simulations and optimization is illustrated in Fig 2.

In the following chapter, some of our first results on optimization, will be commented, obtained prior to the availability of the fully parameterized blade geometry introduced here. The goal is, to learn from bi-disciplinary (aerodynamic-thermal, aerodynamic-structural, thermal-structural) optimization, before a truly multidisciplinary, automated optimization will be feasible.

Finally, the fully parameterized geometry of basic blade with coolant flow passages serves as a test bed for varying the parameters following the suggestions of a structural optimization strategy.

MULTIDISCIPLINARY DESIGN TASKS IN TURBOMACHINERY TECHNOLOGY

With presently available materials such as nickel-based alloys, gas turbine blades cannot withstand metal temperatures in excess of approximately 1300 K. Internal coolant flow passages augmented with heat transfer enhancements, such as trip strips or turbulators, impingement cooling, banks of pin fins and miniature heat exchangers can provide significant enhancements of convection heat transfer. For example, when needed in the initial turbine stages, cooling air can be made to impinge on the leading and trailing edge internal cooling passage surfaces in order to enhance convection. Impingement cooling schemes demand large leading and trailing edge diameters, but this creates thicker blades that can substantially increase aerodynamic losses. Complex heat exchangers have two major drawbacks. First, they induce early transition to turbulence and greatly increase the coolant passage effective friction, while moderately increasing the convective heat transfer. Second, manufacture of such complex internal configurations requires special machining processes.

The design variable set defines the geometry of the turbine blade including the external turbine airfoil shape definition, thermal barrier coating thickness, blade wall thickness distribution, and blade internal strut configurations. The blade stacking axis, twist, and taper are incorporated into the design variable set for three-dimensional blades. With the execution of this geometry generation program, a set of optimization design variables (the parametric model) is used to represent a virtual (electronic) prototype of the turbine blade or vane. The optimization design variable set controlled the internal coolant passage configuration, thickness variation of the coolant passage wall, positions and thicknesses of the internal ribs, and

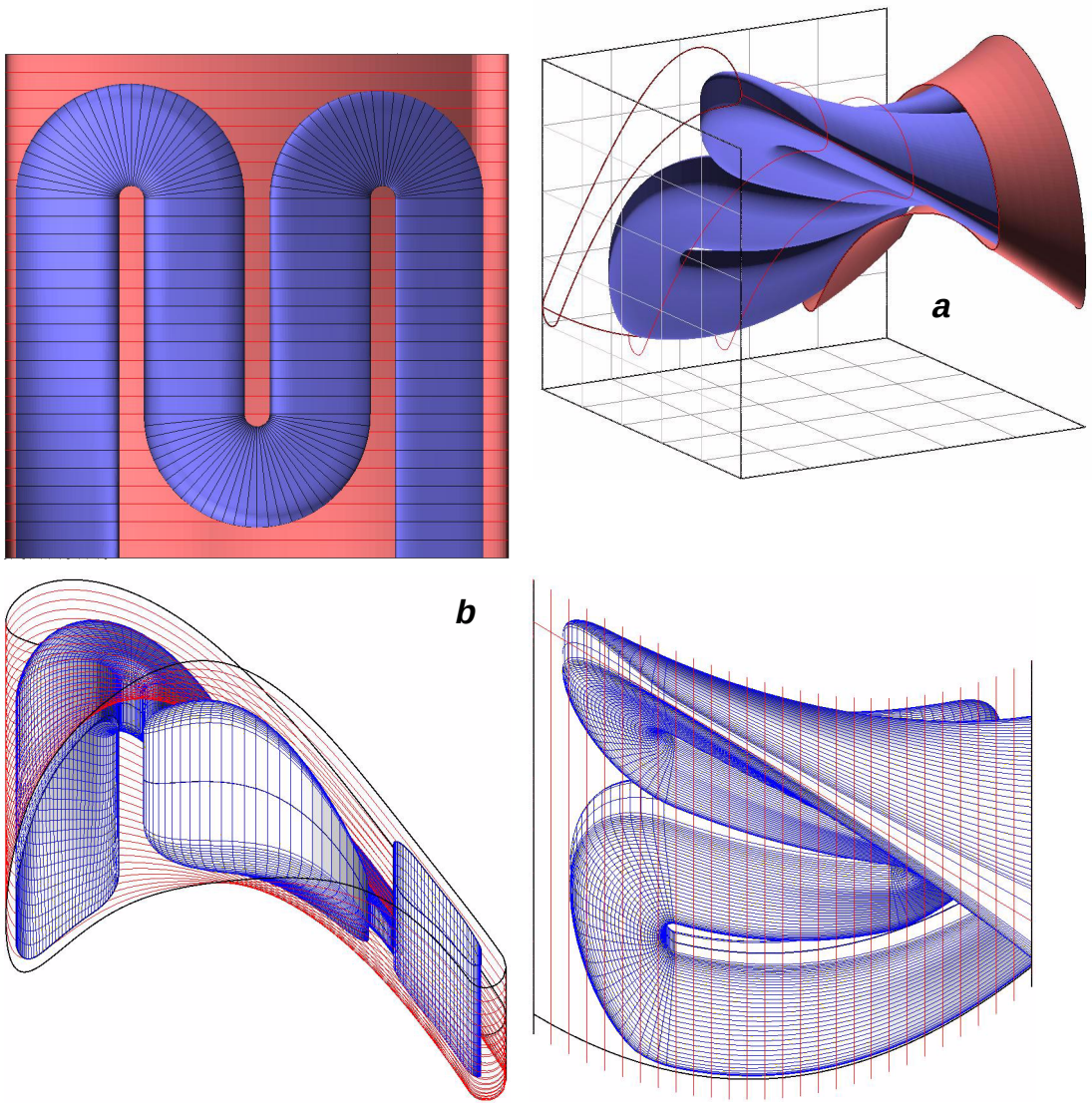


Figure 2: Turbine blade with coolant flow duct: Parametric outer shape definition plus meandering duct within the blade observing local shape control of duct cross section shape and wall thickness. Blade surface partly removed (a), three-view (b)

die pull angles of the ribs [6].

In our first exercise in multidisciplinary design optimization of internally cooled gas turbine blades, a turbulent compressible flow Navier-Stokes solver was used to predict the hot gas flowfield outside of the blade subject to specified realistic hot surface temperature distribution. As a byproduct, this analysis provides hot surface normal temperature gradients thus defining the hot surface convection heat transfer coefficient distribution. This and the guessed coolant bulk temperature and the coolant passage wall convection heat transfer coefficients create boundary conditions for the steady temperature field prediction in the blade and thermal barrier coating materials using fast boundary element technique. The quasi-one-dimensional flow analysis (with heat addition and friction) of the coolant fluid dynamics is coupled to the detailed steady heat conduction analysis in the turbine blade material. By perturbing the design variables (especially the variables defining the internal blade geometry) the predicted thermal boundary conditions on the interior of the blade will be changing together with the coolant flow parameters. As the optimization algorithm runs, it also modifies the turbine inlet temperature. Once

the turbine inlet temperature changes significantly, the entire iterative procedure between the thermal field analysis in the blade material and the computational fluid dynamic analysis of the external hot gas flow-field will be performed again to find a better estimate for thermal boundary conditions on the blade hot surface. This global coupling process, so far, was performed only a small number of times during the course of the entire optimization. This semi-conjugate optimization uses sectional 2-D blade hot flow-field analysis and a simple quasi 1-D coolant flow-field analysis (Fig. 3).

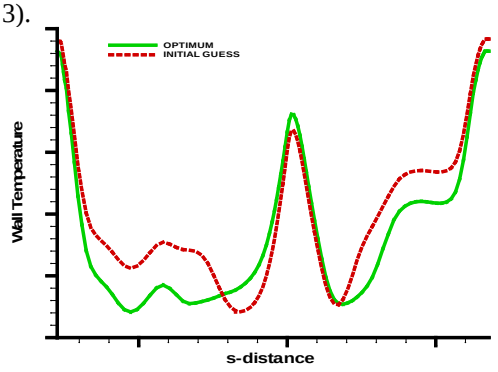


Figure 3. Comparison of external wall temperature variations computed at the quarter-root span of the second HPT blade of the F100 engine

This design methodology was successful at generating a wide range of realistic internally cooled turbine blades and vanes, while the surface meshing, grid generation, and boundary conditions were automatically mapped between the interfacial surfaces. This information was transferred between the various design, optimization, and numerical analysis tools without user intervention. A constrained hybrid optimization algorithm [7] controls the overall operation of the system and guides the multidisciplinary internal turbine cooling design process towards the objectives of cooling effectiveness and turbine blade durability. Design variable sets which had generated an infeasible or impossible geometry, were restored to a feasible shape automatically using a constraint sub-minimization.

There are also possibilities for further improvement in the design of cooled turbine blades. The external turbine blade shape could be modified in an effort to make the external aero-thermodynamics reduce the amount of heat absorbed by the blade. Each new design of the external airfoil would require a fully conjugate viscous three-dimensional steady-state CFD analysis of the hot gas flow field and the temperature field inside the blade [8]. This CFD solution would then be used to predict new external heat transfer coefficients, as well as provide an aerodynamic constraint function so that the efficiency and work of the turbine row could be fixed [9], [10]

RESULTS ON TURBINE BLADE STRUCTURAL ANALYSIS

The geometry preprocessing tool based on analytical functions was already used to model boundary conditions for the automatic structural analysis of internally cooled turbine blades. The preprocessing tool can quickly generate realistic coolant passage shapes within a specified outer blade. The passage shapes are controlled by a set of parameters that the users provide as input. When combined with automatic grid generation and finite element analysis tools, the system is ideal for automatic parameter studies as well as for design optimization. In the current structural analysis system, the geometry preprocessing tool generates a multiblock structured grid that represents the turbine blade geometry. Another program then automatically generates a surface triangulation [11] and then another code makes a volume grid composed of tetrahedrons [12]. A typical surface mesh is shown in Fig. 4. Once a mesh is generated, a structural analysis is performed. The current structural analysis system uses a parallel finite element analysis (FEA) code that can do both linear and nonlinear structural analysis [13]. This code also has the capability of doing automatic partitioning of the mesh as well as automatic FEA. Figure 5 shows an example finite element linear stress analysis result for a turbine blade with coolant passages spinning at 3000 RPM. In this case, the number of degrees of freedom was around 100,000. Two Pentium II 333 MHz processors were used to compute the solution in roughly 15 minutes. With this system, the user only needs to input the parameters that govern the shape of the blade and start the system. Once completed, the system provides the detailed stress and displacement field for the turbine blade without any further interaction with the user. It is hoped that when combined with optimization this automatic geometry generation/analysis system will be a powerful tool for turbine blade design.

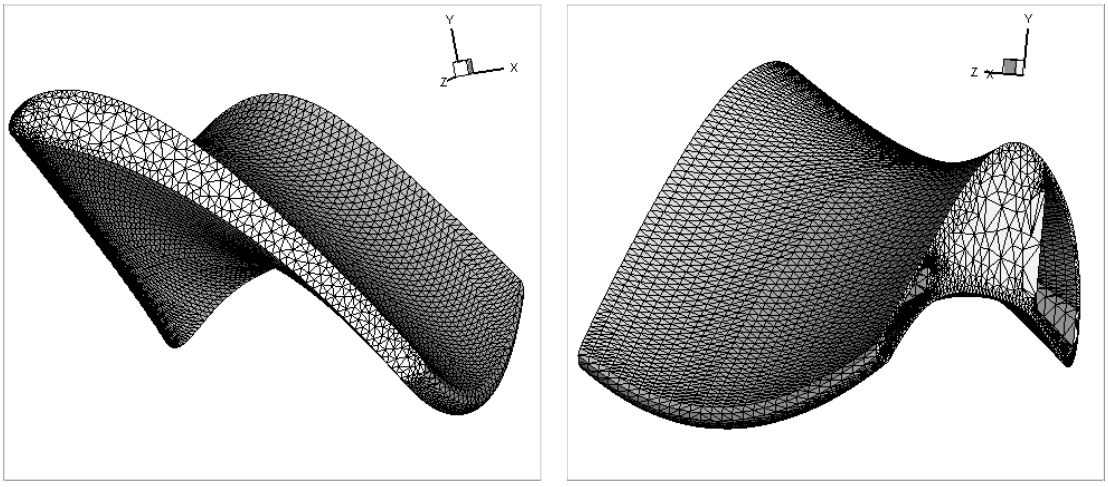


Fig. 4. View of triangular surface mesh from blade tip and from blade root

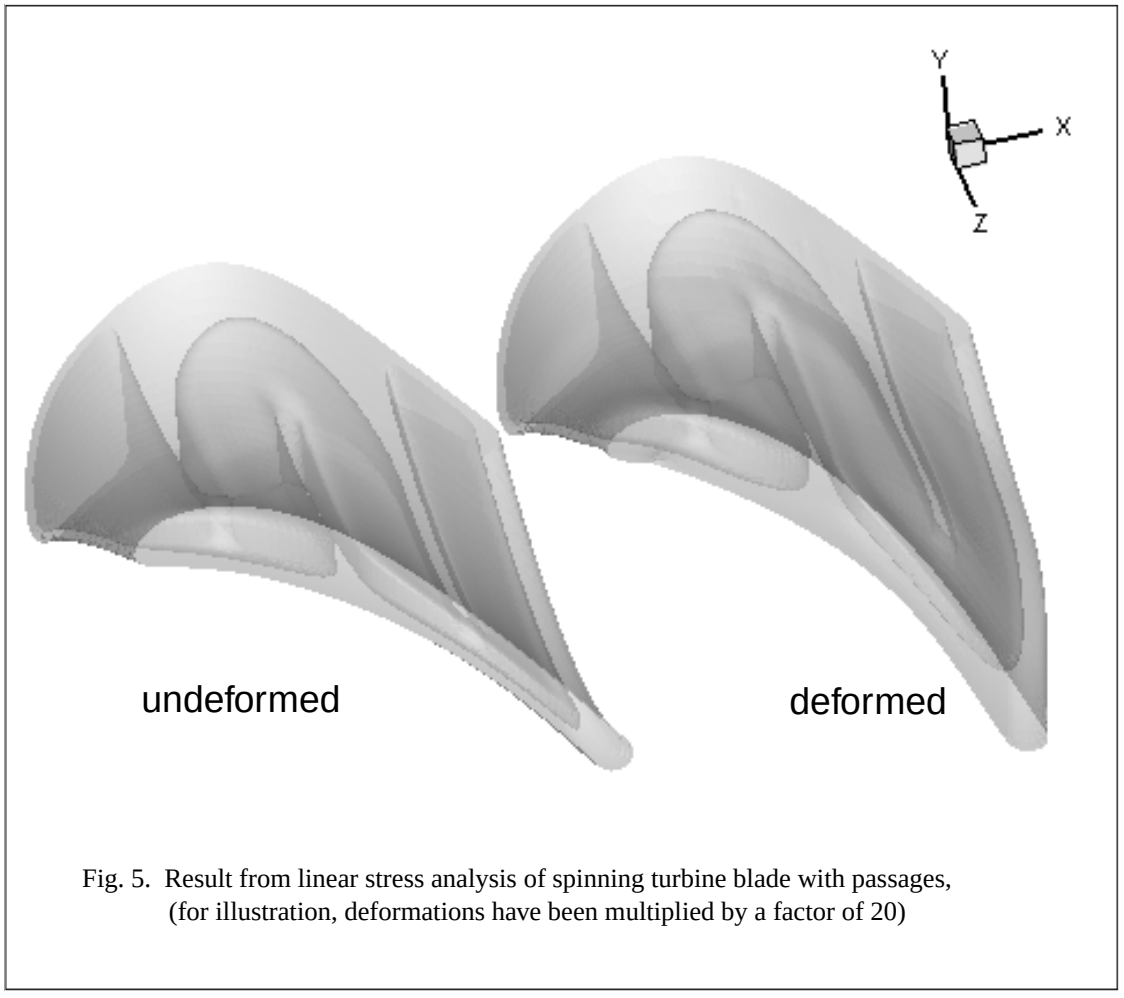


Fig. 5. Result from linear stress analysis of spinning turbine blade with passages,
(for illustration, deformations have been multiplied by a factor of 20)

CONCLUSION

We have shown some first results of what is going to be a software system for multidisciplinary optimization for turbomachinery components like cascades, stators and rotors. Other applications for aerospace and ground vehicle design seem straightforward and rather may be less complicated: A very close coupling of high speed areodynamics, thermal and extreme structural loading may occur only in high speed aircraft design. While showing several results for monodisciplinary design and first results of bidisciplinary optimization, we come to the conclusion and have stressed the fact that fast, flexible and realistic surface modelling for practical components is effectively supporting any future multidisciplinary approach to optimize product components observing advantages and constraints of all mayor disciplines involved in the operation of the component. Optimization of turbomachinery blades poses first, but strong test cases challenging all aspects of the simulation software.

REFERENCES

1. Sobieczky, H., (editor): New Design Concepts for High Speed Air Transport. CISM Courses and Lectures Vol. 366. Wien, New York: Springer (1997)
2. Sobieczky, H., Dulikravich, D. S.: A Computational Design Method for Transonic Turbomachinery Cascades. ASME paper 82-GT-117, (1982)
3. Sobieczky, H., Seebass, A. R.: Supercritical Airfoil and Wing Design.Ann. Rev. Fluid Mech. 16, pp. 337-63 (1984)
4. Klein, M., Sobieczky, H.: Sensitivity of aerodynamic optimization to parameterized target functions. In: M. Tanaka, G.S. Dulikravich, (Eds.), Inverse Problems in Engineering Mechanics, Proc. Int. Symp. on Inverse Problems in Engineering Mechanics (ISIP2001), Nagano, Japan (2001)
5. Farin, G., *Curves and Surfaces for Computer Aided Geometric Design*, Second Edition, Academic Press, 1990.
6. Dennis, B. H., Dulikravich, G. S. and Han, Z.-X., 2001, "Constrained Optimization of Turbomachinery Airfoil Shapes Using a Navier-Stokes Solver and a Genetic/SQP Algorithm", *AIAA Journal of Propulsion and Power*, Vol. 17, No. 5, 2001, pp. 1123-1128.
7. Dulikravich, G. S., Martin, T. J., Dennis, B. H. and Foster, N. F., 1999, "Multidisciplinary Hybrid Constrained GA Optimization," Chapter 12 in *EUROGEN'99 - Evolutionary Algorithms in Engineering and Computer Science: Recent Advances and Industrial Applications*, (editors: K. Miettinen, M. M. Makela, P. Neittaanmaki and J. Periaux), John Wiley & Sons, Ltd., Jyvaskyla, Finland, May 30 - June 3, 1999, pp. 231-260, 1999...
8. Han, Z.-X., Dennis, B. H. and Dulikravich, G. S., 2001, "Simultaneous Prediction of External Flow-Field and Temperature in Internally Cooled 3-D Turbine Blade Material," *International Journal of Turbo & Jet-Engines*, Vol. 18, No. 1, pp. 47-58.
9. Martin, T. J. and Dulikravich, G. S., "Aero-Thermo-Elastic Concurrent Design Optimization of Internally Cooled Turbine Blades", Chapter 5 in *Coupled Field Problems, Series on Advances in Boundary Elements* (eds: Kassab, A. J. and Aliabadi, M. H.), WIT Press, Boston, MA, 2001, pp. 137-184.
10. Martin, T. J. and Dulikravich, G. S., Analysis and Multi-disciplinary Optimization of Internal Coolant Networks in Turbine Blades, ASME IMECE'01, New York, November 11-16, 2001.
11. Shewchuk, J.R., "Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator," First Workshop on Applied Computational Geometry, Philadelphia, Pa, 1996
12. Marcum, D. L. and Weatherhill, N. P., "Unstructured Grid Generation Using Iterative Point Insertion and Local Reconnection," *AIAA Journal*, Vol. 33, No. 9, 1995, pp. 1619-1625.
13. Yoshimura, S. "Development of Computational Mechanics System for Large Scale Analysis and Design," Annual Report of JSPS RFTF CS&E Project, 1999, pp.45-56 (in Japanese).http://adventure.q.t.u-tokyo.ac.jp/.

4th International Conference on Inverse Problems in Engineering
Rio de Janeiro, Brazil, 2002

DEVELOPMENT AND APPLICATIONS OF OPTIMISATION AND INVERSE MODELLING IN THE MANUFACTURING OF ALUMINIUM ALLOY PRODUCTS

Dr. Darius P.K. Singh
Research and Development
Argent Metals Technology Ltd
New Zealand
dariuss@amtnz.com

A/Prof. Gordon Mallinson
Department of Mechanical Engineering
University of Auckland
Auckland
New Zealand

ABSTRACT

Although many foundries use specialized software packages to simulate filling and solidification of castings for process designs, a number of the required input parameters (such as material properties and boundary conditions) are seldom available for valued analysis. A developed and recently patented virtual casting design methodology uses optimization and inverse modeling techniques to firstly calibrate computer models to plant conditions. These models are then used in a second phase of optimisation that improves the operation of the casting plant, thereby playing a major role in reducing costs and improving productivity and quality of cast products. This paper describes an application of this design technology in a low pressure permanent mold casting operation in NZ.

NOMENCLATURE

$f(\boldsymbol{x})$	objective function
$g(\boldsymbol{x})$	constraint function
h	heat transfer coefficient
M	number of time steps
n	number of constraint functions
N	number of thermocouples
p	number of points in $h(T)$ graph.
q	heat flux density
t	time
T	temperature
$\boldsymbol{x}$	vector of design variables.

Subscripts / superscripts

casting	location in casting
experimental	experimental measurement
i	thermocouple index
j	time step index
model	model estimate
n_model	*n*'th model time
n_target	*n*'th target time
surrounding	location in surrounds
t,s,b	mould constraint indices.

INTRODUCTION

A major thrust of cast aluminium research is the development of computer-aided engineering methods to reduce cycle time and cost for producing high quality cast aluminium automotive components. The goal is to provide tools that simulate casting solidification and predict microstructure, mechanical properties and durability of a cast component. This paper describes applications of a methodology that uses numerical simulation and optimisation to enhance a low-pressure permanent mould casting process for aluminium alloy wheels [1].

A common approach [2-4] for casting plant design is to build numerical models that represent all the relevant physical processes as accurately as possible. If appropriate boundary conditions and physical parameters that represent plant conditions can be found, these models produce accurate predictions of plant performance and can be used as design tools. The approach is to find the results (e.g. temperature distribution) from a known cause (e.g. boundary and initial conditions). Unfortunately, it is often impossible to provide input data commensurate with the capabilities of the model thereby reducing its effectiveness as a design tool.

It is possible that the unknown quantities may be determined using extra conditions, which may come from physical measurements elsewhere in the problem domain. Such a problem is termed an *inverse problem* and the process of recovering the boundary conditions is referred to here as *reverse engineering*. Thus, an alternative modelling strategy is to use inverse modelling

with numerical optimisation to adjust a computational model to match measured plant conditions and better understand and predict the stages of the casting process. Using this approach the model may not necessarily be as sophisticated as one used in a conventional analysis. However, by being more closely matched to the real world process, it becomes a more effective design tool. Although the approach does, of course, rely on the availability of detailed plant measurements, its effectiveness for predicting directions for plant improvements can outweigh the expense and difficulty of the measurements.

The design methodology described here uses a finite element model of a casting process that is embedded in an optimisation procedure. Initial stages of optimisation adjust the model's boundary conditions so that it more closely emulates measured temperature-time histories throughout the cast. A second optimisation stage provides the design tool by adjusting die material properties to improve casting performance (i.e. productivity and quality). The modified material properties are then mapped onto suitable adjustments of the casting equipment to effect the improvements.

Figure 1 illustrates the traditional (direct) and reverse engineering (inverse) solution strategies, the latter of which has been employed by a number of researchers [5-8] to design casting mould geometry for optimum casting performance.

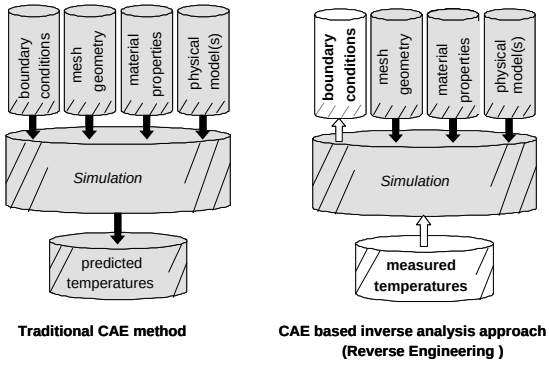


Figure 1 Direct and inverse approach for a computational model.

The approach is extended here to adjust flow and thermal boundary conditions for a specified geometry and is demonstrated by application to a low pressure die-casting manufacturing plant for aluminium alloy automobile wheels.

NUMERICAL METHODS

The die filling process for the wheel illustrated in Figure 2(a) was represented by a 2D axi-symmetric finite element model (Figure 2(b)) aligned along the plane of symmetry through a spoke.

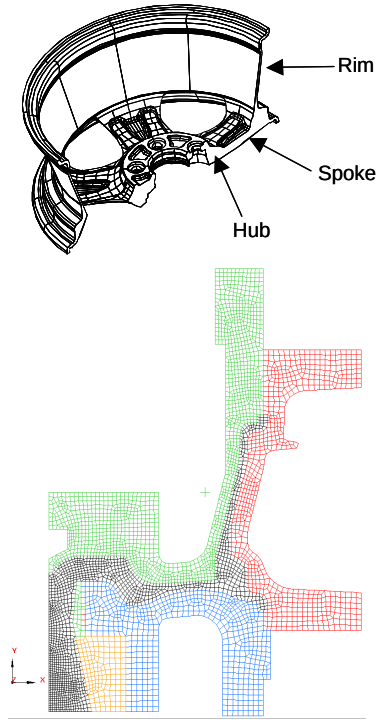


Figure 2 (a) Cross-sectioned solid model of an aluminium alloy wheel, (b) A two-dimensional model of a wheel and tool steel (H13) die, assuming axi-symmetry about the central plane of a spoke.

A commercial package (ProCAST™) was used and, following a mesh convergence study, the model contained 4483 nodes and 3963 linear tetrahedral elements. The maximum allowable time step in the simulations was 0.1 seconds (time of fill was between 8 and 17 seconds) and each unsteady fill simulation took 33 CPU minutes using a SUN ULTRA-1 workstation. The speed of solution is important since the simulations must be completed many times in response to the adjustments made by the optimisation algorithms.

The finite element model was linked via a purpose written user interface to optimisation packages such as DOT™ [9] and SNOPT™ [10]. This interface generates a design file that specifies all the relevant data for optimisation, such as design variables, objective and constraint functions, etc. The design file is used as input to the solution process. The architecture of the optimisation/simulation algorithm is illustrated in Figure 3.

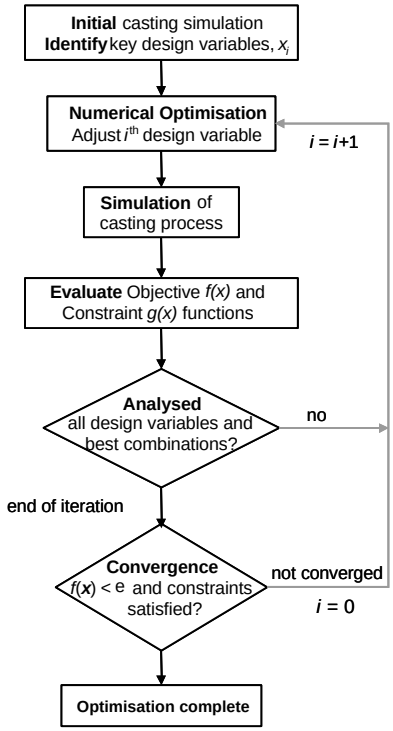


Figure 3 Architechture of the solution control module of the overall design methodology.

Recognising that a wheel is not really axi-symmetric, a 3D model of one fifth of the wheel was also constructed. This model contained 19,755 nodes and 90,683 elements and was used to confirm that the inverse modelling based on the 2D model produced parameters that were relevant to the fully 3D situation.

EXPERIMENTAL MEASUREMENTS

A critical aspect of obtaining data from a casting process under production conditions is to measure temperatures for multiple and consecutive casting shots without causing thermocouple breakage or freezing the thermocouples into the castings. If the thermocouples were to remain inside the solidified casting it would be extremely difficult to open the surrounding die at the end of a cycle and be equally difficult to keep the thermocouples intact. There would also be a high likelihood of damaging parts of the die and its mechanisms. After experimenting with several options, exposed thermocouples were coated with a lubricating graphite die coat during the usual die coating procedure that allows easy extraction of the cast after solidification. Although the coating decreased the temporal responsiveness of the thermocouples slightly it proved to be the only practical way to record multiple and consecutive cycles during warm up and operating stages of production.

A section of the die corresponding to a wheel spoke was instrumented with thermocouples distributed as shown in

Figure 4.

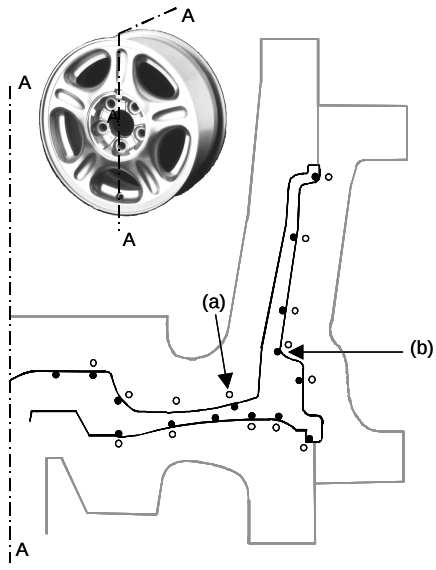


Figure 4 Thermocouple locations in the wheel and die (solid dots indicate thermocouples in the wheel). Cooling curves for the thermocouples marked (a) and (b) are presented in Figure 8.

There were 15 thermocouples that protruded 5 mm into the cavity to measure temperatures in the

molten and solidifying metal. A separate pilot study indicated that a coated thermocouple measured the cooling rate within 1% of that measured by an uncoated thermocouple. Temperature measurement errors from all sources were estimated to be less than 1.3% of the measured temperature. The thermocouples were sampled once a second which was sufficiently rapid to capture the cooling histories while allowing the data collection equipment to store the results from several consecutive cycles.

MODEL ADJUSTMENT USING OPTIMISED INVERSE MODELLING

Modelling of the die casting process is usually divided into a two-part problem. The first stage of the process involves simulating the fluid dynamics during filling of molten metal into the cavity. The second stage involves modelling the heat transfer during solidification.

Generally, the filling sequence is determined by a prescribed velocity boundary condition at the cavity entrance and the solidification profile is controlled by heat transfer boundary conditions across the metal/mould interfaces. The latter is a more complex situation since the boundary conditions comprise a multitude of transient factors ranging from convection in the molten metal during filling, conduction from the solidifying casting to the mould and radiation across isolated air gaps between the casting and mould. The relative importance of these processes depends on experimental or manufacturing conditions and can possibly change as solidification proceeds. The cumulative effect of these heat transport phenomena is often represented by a single heat transfer coefficient, h, embedded in the heat flux condition prescribed at the casting/mould interface

$$q=h(T_{casting}-T_{surrounding}) \quad (1)$$

where q is the heat flux through the interface. In the case of metallic moulds, h can control the solidification rate more than any other single parameter [11]. Hence an accurate calculation of h is essential for an accurate representation of the process.

The matching of the model to plant conditions was done using two stages of inverse modelling. The first stage estimated the inlet velocity during filling. The second estimated the heat transfer boundary conditions during solidification.

Inlet Boundary Condition for a Low Pressure Filling Sequence

The objective function, $f(\boldsymbol{x})$, for optimisation was expressed as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (t_i^{model}-t_i^{experimental}) \quad (2)$$

where $t_i^{experimental}$ and t_i^{model} are the times when the i^{th} thermocouple and its respective node in the model first respond to molten metal contact. The summation is over the total number of cooling curves measured by the 15 thermocouples that protruded into the cast volume. The only design variable in the optimisation was the vertical component of velocity of the metal entering from the riser tube. Unconstrained optimisation using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was used.

Previous filling models for wheels have relied on estimates ranging from 50 to 120 mm/s for the inlet velocity boundary condition [3]. Convergence of the objective function (Equation 2) was achieved in 16 iterations and produced a tuned inlet velocity of 185 mm/s. As shown by the visualisation snap shots in Figure 6, the solution that corresponded to the optimum match had an unsatisfactory flow pattern. Recirculation occurred, causing colder metal to swirl over and mix with hotter incoming metal, increasing the likelihood of trapped air/gas in the cavity.

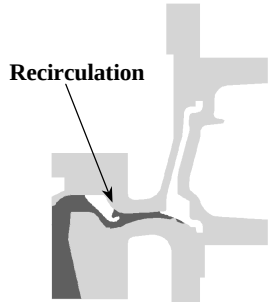


Figure 5 Snapshot of the filling sequence a wheel as predicted by a calibrated 2D finite element model (ProCAST™)

Although previous filling models, based on the lower estimates of inlet velocity, had not predicted the recirculation, that region of the cast had been known to have porosity problems.

The prediction of air/gas entrainment indicated by the solution in Figure 5 has since been validated using a full-scale water analogue model. Figure 6 shows bubbles being generated at the predicted location and propagating throughout the cavity under the influence of the fluid momentum. The optimised calibration has shifted the model to a more representative condition and ultimately led to the identification and understanding of a problematic aspect of this industrial casting process.

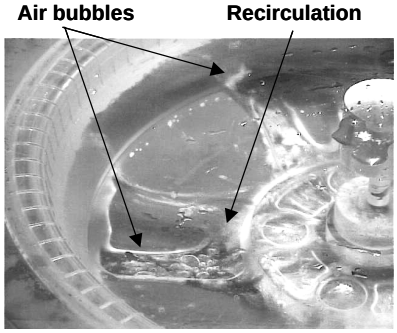


Figure 6 Plan view of filling sequence in a full scalewater analogue model.

Boundary Heat Transfer Coefficients

In this section, inverse engineering is applied to the solidification phase of the same casting process, the objective being to find a distribution of temperature dependent heat transfer coefficients so that the computed and experimental cooling curves closely match. Although the heat transfer during solidification between the casting and die is a function of several variables, temperature was assumed to be the dominant variable. The objective function can be expressed as:

$$f(\boldsymbol{x}) = \sum_{j=1}^{M} \sum_{i=1}^{N} (T_{i,j}^{model} - T_{i,j}^{experimental})^2 \qquad (3)$$

where $T_{i,j}^{model}$ and $T_{i,j}^{experimental}$ are the model and experimental temperatures at the j^{th} time step for the i^{th} thermocouple and M is the total number of time steps over which the optimisation was applied. The second summation is over all the thermocouples, those protruding into the melt and those located in the die. A constraint in this optimisation problem was to maintain decreasing heat transfer coefficients with decreasing temperature to represent the formation of air gaps between the casting and mould, due to casting contraction and mould distortion during solidification. There were three sets of constraint functions, representing the number of die components and interfaces with the casting. These were represented as:

$$g(\boldsymbol{x})_t^{top} = h(T)_t - h(T)_{t+1} \quad , \quad t = (1,p) \qquad (4)$$
$$g(\boldsymbol{x})_b^{bottom} = h(T)_b - h(T)_{b+1} \quad , \quad b = (1,p) \qquad (5)$$
$$g(\boldsymbol{x})_s^{side} = h(T)_s - h(T)_{s+1} \quad , \quad s = (1,p) \qquad (6)$$

where t, b, and s refer to discrete points on each $h(T)$ curve. The Sequential Quadratic Programming (SQP) algorithm was used in the optimisation.

Starting with heat transfer coefficients that were based on previous models and engineering experience, the optimisation produced a 76% improvement in the objective function relative to this initial estimate (Figure 7).

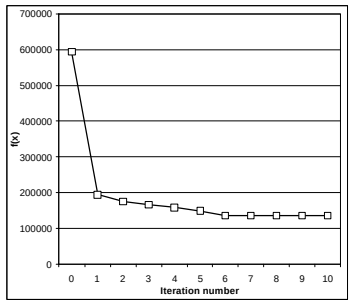
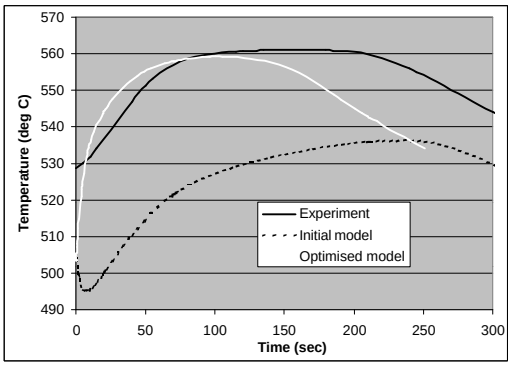


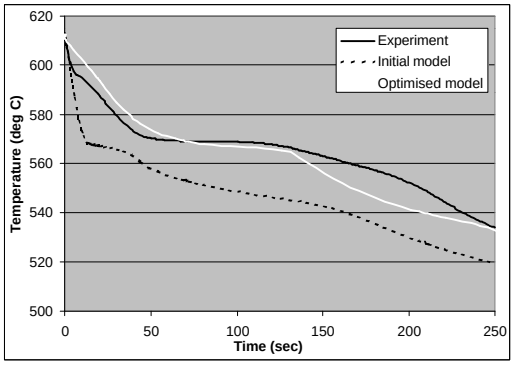
Figure 7 History of the objective function (Equation 3) used to estimate heat transfer coefficients.

An independent analysis has also been conducted to determine the sensitivity of the optimum solution to the initial values of variables in the design space. The results have produced similar optimised solutions for all initial guesses. Sample cooling curves are summarised in Figure 9.

Although an exact match was not produced, an improvement in the thermal predictions for the casting and die was achieved. A reason for the unmatched cooling curves in the die is that some of the blind thermocouples may not have been in complete contact with the internal mould surface, leaving a small air gap between the thermocouple tip and the die.



a)



b)

Figure 8 (a) T-t profiles of the point in the wheel marked "a" in Figure 4, (b) T-t profiles of the point in the casting, marked "b" in Figure 4.

As well, during the experiment there were effects and occurrences in the actual process that are reflected in the experimental data, but not modelled (e.g. variations in die open and close times, breaks in the cooling cycle, metal refills in the furnace and misruns). The effect of any combination of these events can contribute significantly to a source of difference between the plant conditions and the predictions of the model. However, despite these issues, Figure 8 shows that the optimum cooling curves (white lines) show more realistic solidification characteristics than the initial model (dotted lines). The calibrated heat transfer coefficients were used in the 3D wedge model and a resulting isochron plot (reflecting times taken to cool to specified temperatures) was compared with an actual cast piece. The distribution of times taken for the cast to cool to 570°C is shown in Figure 9(a) and indicates a hot spot in the rim / spoke junction. The cast in Figure 9(b) exhibits a corresponding shrinkage defect.

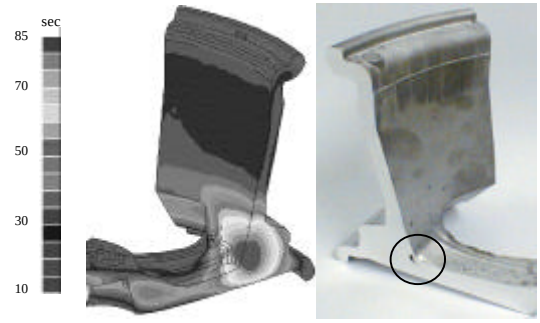


Figure 9 (Left) A calibrated model showing solidification times that indicate a hot spot at the spoke-rim junction. (Right) Photograph of a spoke shrinkage defect at the same location.

CASTING PERFORMANCE OPTIMISATION

The calibrated model was used to optimise the performance of the casting process by modifying thermophysical properties in the die. The objective was to reduce casting defects and achieve a shorter casting cycle time. Constraint functions, $g(\boldsymbol{x})_i$, in the optimisation analysis which were designed to achieve a uni-directional solidification profile are represented by

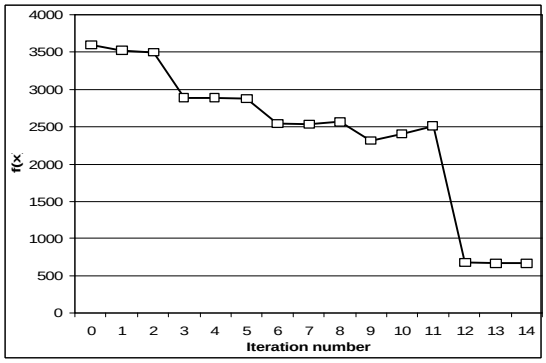
$$g(\boldsymbol{x})_i = T_j - T_k \quad 0 \qquad i = 1,n \tag{7}$$

where the subscripts for temperature denote selected nodes in the model and n denotes the total number of constraint functions. The objective function for this analysis can be expressed by

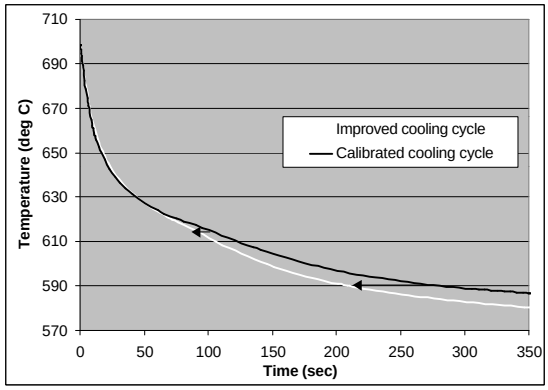
$$f(\boldsymbol{x}) = (t_{1\text{-}model} - t_{1\text{-}target})^2 + (t_{2\text{-}model} - t_{2\text{-}target})^2 \tag{8}$$

where $t_{n\text{-}model}$ and $t_{n\text{-}target}$ denote model and target times, respectively, of the cooling cycle. For the results reported here, the two points in the arbitrarily chosen target cooling curve for a node located in the sprue were 615°C at $t_{1\text{-}target}$ = 75 seconds and 590°C at $t_{2\text{-}target}$ = 160 seconds. A node in the sprue was chosen since it is the last part of the casting to solidify, and is hence a good indicator for the end of a cycle.

Figure 10(a) indicates that a 78% improvement from the initial value of the objective function was achieved and Figure 10(b) illustrates the corresponding reduction in cycle time.



a)



b)

Figure 10 (a) Optimisation history (Equation 8). (b) Initial and desired cooling curves for a selected node.

In a separate analysis the activation periods of four cooling circuits were also optimised using the same objective and constraint functions, producing further refinement in virtual casting performance.

Regions of optimum thermal properties in the die have suggested ideal placements for cooling and insulation and the results have been implemented into an existing low-pressure die cast process that manufactures aluminium alloy wheels. The direct outcome for the industrial plant has been an 80% increase in production capacity (10 to 18 wheels per hour) and a 15% reduction in the design lead time.

CONCLUSIONS

The results presented here have shown how optimisation and inverse modelling can be used initially to tune a computational model so that its predictions match more closely data measured in the industrial casting equipment. The tuned model can then be used in a subsequent optimisation to predict changes that could be made to the casting plant to increase productivity.

In the case of molten metal filling, the tuning produced a model that identified problematic areas in the casting due to recirculations that had not been predicted by previous models. The solidification phase of the casting process was then calibrated and the resulting cooling profiles accurately reflected typical defects in the casting. Both observations were indications that the optimisation had produced better estimates of boundary conditions than had previously been used.

The use of numerical optimisation and modelling has been demonstrated to predict casting phenomenon at macroscopic scales, with very successful predictions for directions of improvement. The inverse methodology encapsulated as a design tool has since been directly incorporated into several vehicle component programmes in the industry.

ACKNOWLEDGEMENT

The research was supported by the New Zealand Foundation for Research Science and Technology (Grant FMCX9601) and Ford Motor Company Scientific Research Labs.

REFERENCES

1. D.P.K. Singh, G.D. Mallinson, S.M. Panton and N. Palle, Die Design Strategy for Improved Productivity and Quality in Die Casting, *AFS Transactions*, 99-25, pp 127-133 (1999).

2. N. Nanda, K. Smith, V. Voller and K. Haberel, A Heuristic Based Practical Tool for Casting Process Design, Modelling, *Casting and Advanced Solidification Processes* VII, pp 381-398 (1995).

3. Y. Otsuka, G. Trapaga and J. Szekely, Optimal Casting Design by Application of Mold Filling and Solidification Simulation, *Light Metals, Minerals, Metals and Materials Society*, pp. 897-905 (1994).

4. S.I. Kang, I.J. Lee and S.D. Shin, Optimisation of Casting Conditions by the measurement of Mold Wall Temperature and Pohang Works, *Proceedings of the Steelmaking Conference*, pp 347-356 (1994).

5. J.M. Drezet and M. Rappaz, Direct Chill Casting of Aluminum Alloys: Ingots Distortions and Mold Design Optimisation, *Light Metals, The*

Minerals, Metals and Materials Society, pp. 1071-1080 (1997).

6. J.V. Beck, Determination of Optimum, Transient Experiments for Thermal Contact Conductance, *Int. J. Heat Mass Transfer*, 12, pp. 621-633 (1980).

7. S.A. Ebrahimi, D.A. Tortorelli and J.A. Dantzig, Sensitivity Analysis and Nonlinear Programming Applied to Investment Casting Design, *Applied Mathematical Modelling*, Vol. 21, pp. 113-123 (1997).

8. R. McDavid and J. Dantzig, Experimental and Numerical Investigation of Mold Filling, *Proc. Int. Modelling, Casting, Welding and Advanced Solidification Processes*, Vol. VIII, B. Thomas and C. Beckermann (eds.), pp. 59-66 (1998).

9. Vanderplaats Research and Development, Inc., *DOT Users Manual*, Version 4.20 (1995).

10. P.E. Gill, W. Murray and A. Saunders, *SNOPT 5.3 User Manual*, Department of Mathematics, UC San Diego (1994).

11. K. Ho and R.D. Pehlke, Metal-Mold Interfacial Heat Transfer, *Metall. Trans. B*, Vol. 16, pp. 585-594 (1985).