Proceedings of 3icipe
1999 ASME Inverse Problems in Engineering
June 13-18, 1999, Port Ludlow, Washington, USA

3ICIPIE/0000

OPTIMIZATION TOOLS FOR INVERSE PROBLEMS USING THE NONLINEAR L- AND A-CURVE

M. E. Gulliksson*
Department of Industrial Technology
Mid Sweden University
S-891 18 Örnsköldsvik
Sweden
Email: marten@ind.mh.se

P.-Å. Wedin
Department of Computing Science
Umeå University
S-901 87 Umeå
Sweden

ABSTRACT

We consider a new idea for solving Tikhonov regularized discretized ill-posed problems. The optimization problem is formulated as a nonlinear least squares problems containing the Tikhonov regularization parameter λ. In order to find the size of the regularization parameter and attain good convergence in the optimization method we use the nonlinear L- and a-curve. The nonlinear L-curve is a direct generalization of the linear L-curve and can be used to find a good regularized solution. The a-curve is the Tikhonov function as a function of the regularization parameter and is most useful in monitoring the global convergence of the method.

Our model algorithm for solving the Tikhonov problem is to use a linearization around the best attained point x_k (possibly given by the nonlinear L-curve) giving a linear L- and *a*-curve. Following the trajectory of the solution to this linear problem the new point chosen is the one that gives sufficient decrease in the size of the residual.

NOMENCLATURE

λ The regularization parameter.
α Step length in optimization method.
x_c The center for the regularization.
x_k Approximation of the Tikhonov problem at iteration k.
$t(x), y(x)$ Size of the residual and solution.
$J(x)$ The Jacobian $\partial f/\partial x$.
t_k, y_k, f_k, J_k Abbreviations for $t(x_k), y(x_k)$, $f(x_k)$ and $\frac{\partial f}{\partial x}(x_k)$.
$\underline{\lambda}_k$ The regularization parameter used as an upper limit for the choice of regularization parameter in step k.
$\bar{t}_k, \bar{y}_k$ The point on the linear L-curve minimizing determining $\bar{\lambda}_k$.

INTRODUCTION

We consider nonlinear equations of the form

$$f(x) = 0, \quad f : \mathbb{R}^n \to \mathbb{R}^m. \tag{1}$$

In our case (1) is a discrete version of an ill-posed infinite dimensional problem. Characteristic for such ill-posed problems are that the singular values of the Jacobian $J = \partial f/\partial x$ decrease rapidly to zero without any useful gap. This fact prevents the efficient use of standard methods such as the Gauss-Newton method.

Therefore, we will use the *Tikhonov problem*

$$\min_x T(x,\lambda), \quad T(x,\lambda) = t(x) + \lambda y(x) \quad \lambda \geq 0 \tag{2}$$

where

$$t(x) = \nu(f(x)) \geq 0, \;\; y(x) \geq 0 \tag{3}$$

---

*Address all correspondence to this author.

are convex functions that attain their minima for $f=0$ and $x=x_c$, respectively. The n-vector x_c is called the center and is chosen a priori (or just zero). The difficulty is to choose the regularization parameter $\lambda \geq 0$ giving both a reasonably small $t(x)$ as well as $\nu(f(x))$ small.

Obviously, the choice of λ is of great importance. In the linear case where $f(x)=Ax+b$ there are many different well analyzed strategies such as the discrepancy principle (10; 6), generalized cross validation (12), and the linear L-curve (9; 7; 11; 8).

For the nonlinear case treated here we propose to use the nonlinear L-curve to find a suitable regularization parameter (1; 5; 4). We make the following definition of the L-curve that is a generalization of the linear L-curve in (8) to nonlinear problems.

Definition 0.1. *Let $x(\lambda)$ solve problem (2), i.e.,*

$$x(\lambda) = \arg\left\{\min_x \; t(x)+\lambda y(x)\right\}, \quad \lambda \geq 0.$$

The L-curve is defined as the curve $(t(x(\lambda)), y(x(\lambda)))$.

The L-curve is monotonically decreasing and convex as shown in (5).

To construct the L-curve in an efficient way and find a good solution there is a need for a robust and efficient method to solve the Tikhonov problem for several λ. However, close to a corner of the L-curve the Tikhonov function varies much and there may be a need for a second order optimization method. We will not consider this aspect here and refer to (2) for special quasi-Newton methods.

Another important curve useful for monitoring the convergence of methods for solving the Tikhonov problem is the a-curve.

Definition 0.2. *The a-curve is defined as the curve $(\lambda, a(\lambda))$ where*

$$a(\lambda) = \min_x \; t(x)+\lambda y(x), \quad \lambda \geq 0. \tag{4}$$

The a-curve is monotonically increasing and concave as shown in (5).

In our earlier implementations, see (1; 3), we used a Gauss-Newton method directly on the Tikhonov problem choosing λ adaptively and monotonically decreasing depending on the size of the step length. This approach seems inefficient since it is quite difficult to safely choose a small λ in each step. Further, we do not use the global information attainable from the L-curve. This is also true for a trust-region method applied on the Tikhonov problem.

Therefore, we propose a special variant of the Gauss-Newton method used on a linearization of the Tikhonov problem (2). The main feature of the method is that the regularization is made relative the center x_c but the linearization is made around the current iteration point x_k (possibly chosen as the best attainable point given by the nonlinear L-curve). This idea combines the regularization effect restricting the size of x_k with the minimization of the size of the residual $\nu(f(x))$. The method is more efficient than a Gauss-Newton or trust-region directly on the Tikhonov problem since we use more information from the linear subproblem of the Gauss-Newton method and we have the possibility to safely choose the smallest possible λ in each step.

A LOCAL TRUST-REGION METHOD

Geometrical motivation

For simplicity, but without loss of generality, we will in this section assume that $\nu(f) = 1/2\,\|f\|^2$ and $t(x) = 1/2\,\|x-x_c\|^2$ where $\|\cdot\|$ is the 2-norm. Thus, the Tikhonov problem (2) can be written as

$$\min_x \; \frac{1}{2}\,\|f(x)\|^2 + \frac{1}{2}\lambda\|x-x_c\|^2. \tag{5}$$

We start by describing the general idea in the k'th step of the algorithm. If we linearize the Tikhonov function in (5) around x_k we get the linear least squares problem

$$\min_p \; \frac{1}{2}\,\|f_k + J_k p\|^2 + \frac{1}{2}\lambda\|p+x_k-x_c\|^2. \tag{6}$$

Using the normal equations we easily attain the solution to (6) as

$$p(\lambda) = -(J_k^T J_k + \lambda I)^{-1}\left(J_k^T, \lambda^{1/2} I\right)\begin{pmatrix} f_k \\ x_k - x_c \end{pmatrix}. \tag{7}$$

The trajectory $x_k(\lambda) = x_k + p(\lambda)$ is seen in Figure 1 and as λ is decreasing $x_k(\lambda)$ is moving from x_c with a decreasing residual $\|J_k p(\lambda) + f_k\|$.

To show more of the implications and possibilities of our idea we reformulate the linear problem (6) using $x = x_k + p$ to get

$$\min_x \; t_{x_k}(x) + \lambda y_{x_k}(x) \tag{8}$$

where we define the functions

$$t_{x_k}(x) = \frac{1}{2}\,\|J_k(x-x_k)+f_k\|^2, \;\; y_{x_k}(x) = \frac{1}{2}\|x-x_c\|^2. \tag{9}$$

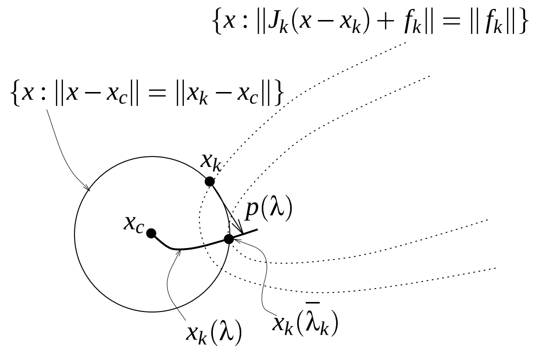

Figure 1. LEVEL CURVES FOR THE LINEAR PROBLEM.

The solution to (8) is $x_k(\lambda) = x_k + p(\lambda)$ and the linear L-curve associated to (8) is $(t_{x_k}, y_{x_k}(t_{x_k}))$ as seen in the left part of Figure 2. Following the L-curve from (t_k, y_k) minimizing the residual

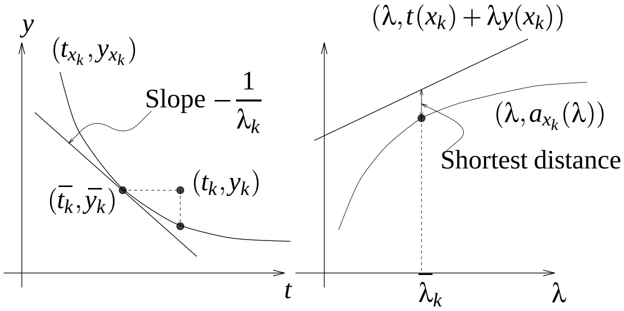

Figure 2. THE LINEAR L- AND a-CURVES CORRESPONDING TO THE LINEAR PROBLEM (6.

$y(x)$ we find the point

$$(\bar{t}_k, \bar{y}_k), \;\; \bar{t}_k = t_{x_k}(x_k(\bar{\lambda}_k)), \;\; \bar{y}_k = y_{x_k}(x_k(\bar{\lambda}_k))$$

where the L-curve has slope $-1/\bar{\lambda}_k$. Apart from the fact that this is the point closest to (t_k, y_k) keeping $t(x) = t_k$ constant we will see that this point has some interesting and useful properties.

The trust-region idea

Following More' (10) we accept the point $x_k(\lambda)$ as our new approximation of the solution to the Tikhonov problem if the inequality

$$\|f(x_k)\|^2 - \|f(x_k(\lambda))\|^2 < \delta\{\|f(x_k)\|^2 - \|J_k p(\lambda) + f_k\|^2\} \tag{10}$$

is satisfied for $\lambda < \bar{\lambda}_k$ and $0 < \delta < 1$. Combining this condition with a strategy for choosing λ gives the following model algorithm where we assume that x_k is a given point.

1. Solve the linear problem (6) with $\lambda < \bar{\lambda}_k$.
2. **while** The condition in (10) is not satisfied **and** $\lambda < \bar{\lambda}_k$.
   (a) Set $\lambda = \mu\lambda$, $\mu > 1$.
   (b) Solve the linear problem (6) getting the solution $p(\lambda)$.
   (c) Update $x_k(\lambda) = x + p(\lambda)$.
3. **if** $\lambda > \bar{\lambda}_k$
   (a) Solve the nonlinear problem (2) with $\lambda = \bar{\lambda}_k$ to a certain accuracy.

The last step in the algorithm is to find a point closer to the solution of the Tikhonov problem for $\lambda = \bar{\lambda}_k$. As we will see later this will make x_k closer to the linear approximation of the Tikhonov problem and make it possible to find a smaller residual.

The choice of $\bar{\lambda}_k$.

Our aim is to find the next point $x_k(\bar{\lambda}_k)$ not very much further from x_c but with a smaller residual. Imagining x_k on the linear L-curve (t_{x_k}, y_{x_k}) we have that $\nabla_x T = \nabla_x t + \lambda_x y = 0$ and the level curves $t(x) = t_k, y(x) = y_k$ are tangential, see Figure 1. Thus, there will be no decrease in the residual for λ greater than the $\bar{\lambda}_k$ defined by the relation $t(x_k(\lambda)) = t(x_k)$ or

$$\|x_k(\lambda) - x_c\| = \|x_k - x_c\|. \tag{11}$$

If we have x_k not on the linear L-curve it seems reasonable to have the same criterion for choosing $\bar{\lambda}_k$. From Figure 1 it is seen that $p(\lambda)$ will always be a descent direction to $t(x)$.

Further, we define the linear a-curve

$$a_{x_k}(\lambda) = \min_x \; t_{x_k}(x) + \lambda y_{x_k}(x)$$

shown in the right part of Figure 2. The linear a-curve can be used to find $\bar{\lambda}_k$ but first we present a useful lemma.

Lemma 0.3. *Assume that $(\tilde{t}, y(\tilde{y})$ is a point on or above the L-curve. Then the solution of*

$$\min_\lambda \; \tilde{t} + \lambda\tilde{y} - a(\lambda)$$

is given by

$$\tilde{\lambda} = -\frac{1}{\frac{dy}{dt}(\tilde{t})}. \tag{12}$$

Moreover, the slope at $\tilde{\lambda}$ is given by

$$\frac{da}{d\lambda}(\tilde{\lambda}) = \tilde{y}.$$

Proof. Define $F(\lambda) = \tilde{t} + \lambda\tilde{y} - a(\lambda)$. We have

$$\frac{dF}{d\lambda} = \tilde{y} - \frac{da}{d\lambda} \text{ and } \frac{d^2F}{d\lambda^2} = -\frac{d^2a}{d\lambda^2} > 0.$$

Hence, $y = da/d\lambda = \tilde{y}$ at the closest point, i.e., the tangent of $a(\lambda)$ has the same direction as the line $\tilde{t} + \lambda\tilde{y}$. Obviously,

$$\frac{da}{d\lambda}(\tilde{\lambda}) = \tilde{y}$$

if $y(\tilde{t}) = \tilde{y}$ and hence

$$\frac{dy}{dt}(\tilde{t}) = -\frac{1}{\tilde{\lambda}}.$$

From Lemma 0.3 we get that the line $(\lambda, t(x_k) + \lambda y(x_k))$ above the linear a_{x_k}-curve is as close as possible to the a-curve at $\bar{\lambda}_k$ suggesting a way to find $\bar{\lambda}$ if we can approximate the a-curve efficiently.

Approximating the L- and a-curve

The convexity of the L-curve and the concavity of the a-curve are direct consequences of the fact that the curves describe the solution of a sequence of optimization problems. It is natural to try to keep these properties when information at a finite point set M in $\mathbb{R}^n$ is used to approximate the functions $y(t)$ and $a(\lambda)$. We define the function $y_{pol}(t)$ as a polygon approximation of $y(t)$ if $y_{pol}(t)$ is a strictly decreasing convex function. To every curve $y_{pol}(t)$ there should also exist a concave, strictly increasing polygon approximation $a_{pol}(\lambda)$ of $a(\lambda)$.

The first step towards smooth approximating curves is to find a subset $\{x_i\}_{i=1}^{p}$ in M such that $0 \le t_1 < t_2 < \ldots < t_p, t_i = t(x_i)$ and the function

$$y_{pol}(t) = y_i \frac{t_{i+1} - t}{t_{i+1} - t_i} + y_{i+1} \frac{t - t_i}{t_{i+1} - t_i}, \tag{13}$$

$t_i \le t \le t_{i+1}$, $y_i = y(x_i)$, $i = 1, \ldots, p-1$ is a strictly decreasing convex function for $t_1 \le t \le t_p$. If we add the points $(t_1, y), y \ge y_1$ to the points defined by $(t, y_{pol}(t))$ the set M defines points $((t(x_i), y(x_i))$ that are inside the convex set defined by (13) as shown in Figure 3. For a given finite set M the polygon curve

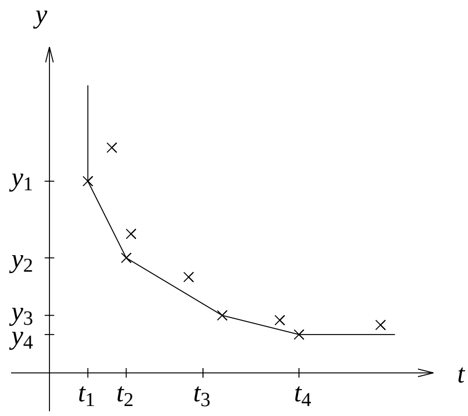

Figure 3. THE SHAPE OF THE y_{pol}-CURVE.

constructed in this way is unique. But the point set $\{x_i\}_{i=1}^{p}$ need not be unique, since there may exist a point $\tilde{x} \in M, \tilde{x} \ne x_i$ such that $t(\tilde{x}) = t(x_i)$ and $y(\tilde{x}) = y(x_i)$.

There is also a polygon approximating curve $a_{pol}(\lambda)$ of $a(\lambda)$ that corresponds to the polygon curve $y_{pol}(t)$ constructed in (13). Define $\lambda_{i,i+1}$ as the point where the two straight lines $t_i + \lambda y_i$ and $t_{i+1} + \lambda y_{i+1}$ intersect. Hence, $t_i + \lambda_{i,i+1} y_i = t_{i+1} + \lambda_{i,i+1} y_{i+1}$ and

$$\lambda_{i,i+1} = -\frac{t_{i+1} - t_i}{y_{i+1} - y_i}. \tag{14}$$

The definition (13) implies that $\lambda_{1,2} < \lambda_{2,3} < \ldots < \lambda_{p-1,p}$. Also define $\lambda_{p,\infty}$ as the point where the straight line $t_p + \lambda y_p$ cuts the asymptote $a = t(x_c)$, i.e., $\lambda_{p,\infty} = t(x_c) - t_p/y_p$. The definition of a_{pol} is now

$$a_{pol}(\lambda) = \begin{cases} t_1 + \lambda y_1, & 0 \le \lambda \le \lambda_{1,2} \\ t_i + \lambda y_i, & \lambda_{i-1,i} \le \lambda \le \lambda_{i,i+1} \\ t_p + \lambda y_p, & \lambda_{p-1,p} \le \lambda \le \lambda_{p,\infty} \end{cases}$$

The function $a_{pol}(\lambda)$ is the unique strictly increasing concave function such that for all points $\tilde{x} \in M$ the straight lines $(\lambda, t(\tilde{x}) + \lambda y(\tilde{x}))$ lie above the curve $(\lambda, a_{pol}(\lambda))$.

Now there is a simple task to construct a smooth approximating L- and a-curve. Let a polygon curve $(t, y_{pol}(t))$ be known and let $(t, y_{sm}(t))$ be a convex decreasing spline function that interpolates the polygon curve at $(t_i, y_i), i = 1, \ldots, p$. The function y_{sm} is our smooth function and by definition it is twice differentiable. Define λ at a given point $(t, y_{sm}(t))$ from the derivative $dy_{sm}/dt = -1/\lambda$ and set $a_{sm} = t + \lambda y_{sm}$ as our smooth function corresponding to the a-curve. As before the differential

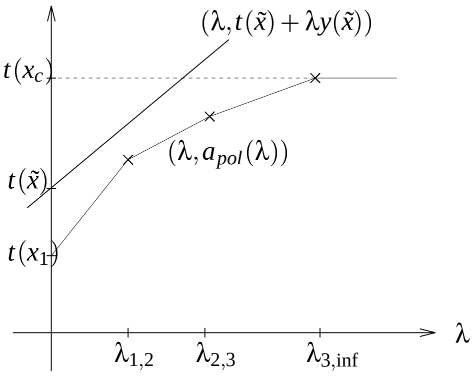

Figure 4. THE SHAPE OF THE a_{pol}-CURVE.

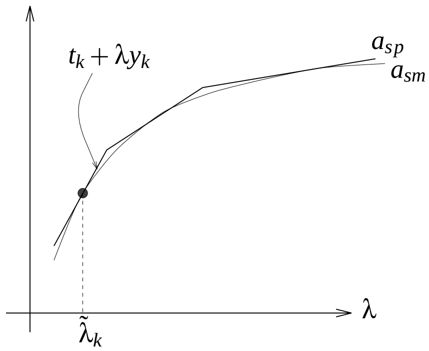

Figure 5. APPROXIMATING $\bar{\lambda}_k$.

$da_{sm} = dt + \lambda dy_{sm} + d\lambda y_{sm} = d\lambda y_{sm}$ giving

$$\frac{da_{sm}}{d\lambda} = y_{sm} > 0$$

and a_{sm} is strictly increasing. Further,

$$\frac{d^2 a_{sm}}{d\lambda^2} = \frac{dy_{sm}}{d\lambda} = \frac{dy_{sm}}{dt}\frac{dt}{d\lambda} = -\lambda^{-1}\frac{dt}{d\lambda}$$

where by definition we have

$$\frac{d\lambda}{dt} = \frac{d}{dt}\frac{1}{\frac{dy_{sm}}{d\lambda}} = \left[\frac{dy_{sm}}{dt}\right]^{-2}\frac{d^2 y_{sm}}{dt^2} = \lambda^2 \frac{d^2 y_{sm}}{dt^2}$$

and thus

$$\frac{d^2 a_{sm}}{d\lambda^2} = -(\lambda^3 \frac{d^2 y_{sm}}{dt^2})^{-1} < 0,$$

making a_{sm} concave.

Approximating $\bar{\lambda}$ using the smooth approximating a-curve.

If we have computed the smooth approximating a-curve as well as the polygon approximating we can use these curves to approximate $\bar{\lambda}$. In Figure 5 this idea is clearly seen where $\tilde{\lambda}_k$ is an approximation of $\bar{\lambda}_k$.

The special case $x_k = x_k(\lambda)$.

In this section we show that the search direction $p(\lambda)$ is well defined even if x_k is very close to $x(\bar{\lambda}_k)$ with $\bar{\lambda}_k$ defined by (11).

By definition we have $p(\bar{\lambda}_k) = 0$ but the interesting quantity to be investigated is

$$\lim_{\lambda \to \bar{\lambda}_k} \frac{p(\lambda)}{||p(\lambda)||}$$

since this is the attainable search direction. We will need the following lemma.

Lemma 0.4. *Assume that $x_k = x(\bar{\lambda}_k)$. Then*

$$\frac{dp}{d\lambda}(\bar{\lambda}_k) = -(J_k^T J_k + \bar{\lambda}_k I)^{-1}(x_k - x_c). \tag{15}$$

Proof. For (15) we have

$$\frac{dp}{d\lambda} = -\left(\frac{d}{d\lambda}J_k^{\#} f_k + \frac{d}{d\lambda}P_N\,(x_k - x_c)\right)$$

where

$$J_k^{\#} = (J_k^T J_k + \lambda I)^{-1} J_k^T, \;\; P_N = \lambda(J_k^T J_k + \lambda I)^{-1}.$$

Further,

$$\frac{d}{d\lambda}J_k^{\#} = -(J_k^T J_k + \lambda I)^{-2} J_k^T$$

and after some algebra

$$\frac{d}{d\lambda}P_N = J_k^T J_k (J_k^T J_k + \lambda I)^{-2} J_k^T.$$

Thus,

$$\frac{dp}{d\lambda} = -(J_k^T J_k + \lambda I)^{-1}(-J_k^{\#} f_k + J_k^{\#} J_k)(x_k - x_c). \qquad (16)$$

By using that x_k lies on the trajectory $x_k(\lambda)$ we have

$$J_k^T f_k + \bar{\lambda}_k (x_k - x_c) = 0$$

or by premultiplying with $(J_k^T J_k + \lambda I)^{-1}$

$$J_k^{\#} f_k + \bar{\lambda}_k (J_k^T J_k + \lambda I)^{-1}(x_k - x_c) = 0. \qquad (17)$$

Inserting (17) into (16) and using that

$$\bar{\lambda}_k (J_k^T J_k + \lambda I)^{-1} + J_k^{\#} J_k = 0$$

we get

$$\frac{dp}{d\lambda}(\bar{\lambda}_k) = -(J_k^T J_k + \lambda I)^{-1}(x_k - x_c).$$

The following theorem proves that $p(\lambda)$ is well defined and a descent direction to $\|f(x)\|$ at $\bar{\lambda}_k$.

Theorem 0.5. *Assume that $p(\lambda)$ is defined by (7) then*

$$q_k = \lim_{\lambda \to \bar{\lambda}_k} \frac{p(\lambda)}{\|p(\lambda)\|} = -\frac{(J_k^T J_k + \bar{\lambda}_k I)^{-1}(x_k - x_c)}{\|(J_k^T J_k + \bar{\lambda}_k I)^{-1}(x_k - x_c)\|} \qquad (18)$$

and

$$q_k^T J_k^T f_k = -\frac{\bar{\lambda}_k (x_k - x_c)^T (J_k^T J_k + \lambda I)^{-1}(x_k - x_c)}{\|(J_k^T J_k + \lambda_k I)^{-1}(x_k - x_c)\|} < 0. \qquad (19)$$

Proof. Using that $p(\bar{\lambda}_k) = 0$ we have

$$\frac{p(\lambda)}{\|p(\lambda)\|} = \frac{p(\lambda) - p(\bar{\lambda}_k)}{\lambda - \lambda_k} \frac{\lambda - \bar{\lambda}_k}{\|p(\lambda) - p(\lambda_k)\|}$$

and if we assume that $\bar{\lambda}_k > \lambda$ we get

$$\frac{p(\lambda)}{\|p(\lambda)\|} = \frac{p(\lambda) - p(\bar{\lambda}_k)}{\lambda - \lambda_k} \frac{1}{\frac{\|p(\lambda) - p(\lambda_k)\|}{\|\lambda - \lambda_k\|}}.$$

Letting $\lambda \to \bar{\lambda}_k$ and using Lemma 0.4 we get the first statement (18) in the theorem.

The second statement (19) is attained directly from (18).

REFERENCES

J. Eriksson. Optimization and regularization of nonlinear least squares problems. Technical Report UMINF 96.09 (Ph.D. Thesis), Dept. of Comp. Science, Umeå University, Umeå, Sweden, 1996.

J. Eriksson. Quasi-Newton methods for nonlinear least squares. Technical Report Accepted for publication in BIT, Dept. of Comp. Science, Umeå University, Umeå, Sweden, 1997.

J. Eriksson, M. E. Gulliksson, P. Lindström, and P.-Å. Wedin. Regularization tools for training feed-forward neural networks. *J. Opt. Meth. Soft.*, 10:49–69, 1998.

M. E. Gulliksson and P.-Å Wedin. Algorithms for using the nonlinear L-curve. Technical Report Sumbitted to SIAM J. Optim., Dept. of Comp. Science, Umeå University, Umeå, Sweden, 1999.

M. E. Gulliksson and P.-Å Wedin. The nonlinear L-curve. Technical Report Sumbitted to SIAM J. Optim., Dept. of Comp. Science, Umeå University, Umeå, Sweden, 1999.

A. Neubauer H. Engl, M. Hanke. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992.

P.C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems. Numerical aspects of linear inversion.* SIAM, Philadelphia, 1997.

R. J. Hanson and C. L. Lawson. *Solving least squares problems*. Prentice Hall, Englewood Cliffs, N. J., 1974.

J. J. More. The Levenberg-Marquardt algorithm: implementation and theory. In G. A. Watson, editor, *Proceedings of the 1977 Dundee conference on numerical analysis*, Lecture notes in mathematics 630, pages 105–116, Berlin, Heidelberg, New York, Tokyo, 1978. Springer Verlag.

T. Reginska. A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.*, 17(3):223–228, 1996.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.