A Practical Examination of Some Numerical Methods for Linear Discrete Ill-Posed Problems

J. M. Varah

Computer Science Department

The University of British Columbia

Abstract

Four well-known methods for the numerical solution of linear discrete ill-posed problems are investigated from a common point of view: namely, the type of algebraic expansion generated for the solution in each method. A sensitivity analysis of each method is made, and numerical results given for some particular problems. These results are interpreted from this algebraic point of view, and some anomalies explained.

1. Introduction

In this paper we wish to examine some popular numerical methods for linear discrete ill-posed problems; that is, methods for solving the linear system

$$Kf = g \tag{1.1}$$

where K is an ill-conditioned n×n matrix (in fact K could be m×n as well). Because of the ill-conditioning, standard methods give inappropriate results, usually including large oscillations. All the methods we consider modify the problem or solution so these large oscillations do not appear.

Since these discrete problems arise as discretizations of continuous ill-posed problems, such as integral equations of the first kind, it is important to understand the modifications mentioned above from the point of view of the continuous problem.

For the case of an integral equation of the first kind,

$$\int \tilde{K}(s,t)\ \tilde{f}(t)dt = \tilde{g}(s)$$

with $\tilde{K}$ compact, the nature of the problem is in one sense completely specified by the Picard theorem (see Courant and Hilbert [2, pg 159]). There exist adjoint L_2 orthogonal functions $\{\phi_i(s)\}$, $\{\psi_i(t)\}$, and real scalars $\lambda_i \to 0$ so that

$$\int \tilde{K}(s,t)\ \phi_i(s)ds = \lambda_i\ \psi_i(t)$$

and

$$\int \tilde{K}(s,t)\ \psi_i(t)dt = \lambda_i\ \phi_i(s).$$

Thus if $\tilde{g}(s) = \sum_1^\infty \tilde{\beta}_i \psi_i(s)$, then $\tilde{f}(t) = \sum_1^\infty \left(\frac{\tilde{\beta}_i}{\lambda_i}\right) \phi_i(t)$; however $\tilde{f} \in L_2$ only if $\sum_1^\infty (\frac{\tilde{\beta}_i}{\lambda_i})^2 < \infty$, and it is this (Picard) condition which specifies whether a nice solution exists to the problem.

However when the problem is discretized, further difficulties arise: all expansions become finite, and the arithmetic also becomes finite. Now one is interested in whether the problem (Kf = g) has a smooth approximate solution which is relatively insensitive to changes in the data. As well, we are concerned with whether the particular method we are using can find this solution accurately: all the usual methods generate a solution of the form

$$\bar{f} = \sum_1^k c_i y_i \qquad (1.2)$$

where k and the vectors $\{y_i\}_1^k$ depend on the method and on the problem. Although there is no restriction on the vectors $\{y_i\}$, we shall assume they are orthogonal. Using this, we can make the following (qualitative) definition of conditioning:

DEFINITION: The problem Kf = g is well-conditioned with respect to the vectors $\{y_i\}_1^k$ if there is a solution $\bar{f}$ of the form (1.2) with $||K\bar{f} - g||$ small, and where the vectors $\{y_i\}_1^k$ do not reflect the ill-condition of K.

This last statement can be quantified as follows: let Y_k be the n×k matrix with columns $y_1, \ldots, y_k$; then we mean that the condition number $\kappa(KY_k)$ is not large. Here if A is an n×k matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > 0$ $\kappa(A) = \sigma_1/\sigma_k$.

Of course we could describe this in other ways: to find a solution of the form (1.2), we need to solve the n×k system $(KY_k)\underline{c} = \underline{g}$; thus we need $\underline{g}$ to be consistent with the columns of (KY_k) and $\kappa(KY_k)$ not large. However we describe it, it is clear that the condition of the discrete problem depends not only on the problem itself, but the method we use to solve it; thus a particular method may work well on some problems (where a solution like (1.2) exists) and not on others. This will be seen more clearly in the numerical examples.

In the next section we describe the methods, with particular attention to the kind of vectors $\{y_i\}$ generated. We give a sensitivity analysis of the methods in Section 3, and describe the numerical results in Section 4. We should add that a similar investigation for some particular methods was made by Rutishauser [8].

2. The Methods

(a) Truncated QR

The simplest and most general method for solving Kf = g, given the discussion in the first section, is to assume some expansion $\underline{f}^{(k)} = \sum_1^k c_i \underline{y}_i$ for some given set of orthogonal vectors $\{\underline{y}_i\}$, and solve for the $\{c_i\}$ by a least squares technique. If this is done by a QR factorization, we need not prescribe k in advance. Indeed, let Y be the matrix of columns $\{\underline{y}_i\}_1^m$, where $m \le n$ and we know we want to choose $k \le m$. Now perform a QR factorization KY = QR, and solve the first k equations of

$$R\underline{c} = Q^T g.$$

In all these methods there is some free parameter to choose, some trade-off point between smoothness and accuracy of the solution. Here it is k, and it should be chosen as small as possible, consistent with obtaining a good represen-

tation of the solution. This can be measured by the residual $||Kf^{(k)} - g||_2$, which in turn is given by the l_2-norm of the last (n-k) components of $Q^T\underline{g}$. So we pick k large enough that this residual is small, but not so large that the condition number of (KY_k) is too big. Unfortunately (and this is the main problem with this method), although this condition number is also the condition number of R_k = the first k rows and columns of R, it is not always reflected in the size of the diagonal elements of R_k, which is easily monitored. As with all these methods, the free parameter k is best determined by solving the problem interactively, using several choices for k, and choosing that solution which is "best" in the opinion of the user.

Notice that the success of the method is highly dependent on the choice of vectors $\{y_i\}$; a solution giving a small residual with only a few vectors is most desirable, and this depends on the problem, the data $\underline{g}$, and on the ingenuity of the user. See Section 4 for some numerical examples.

(b) Truncated Singular Value Decomposition

This well-known method (see [1], [4], [11]) forms the singular value decomposition of K, $K = UDV^T$, where U and V are orthogonal and $D = \text{diag}(\sigma_1,\sigma_2,\ldots,\sigma_n)$ where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ are the singular values of K. Then the system $Kf = g$ is solved using this decomposition and the relevant orthogonal transformations. Thus if $\{\underline{u}_i\}$ are the columns of U and $\{\underline{v}_i\}$ are the columns of V,

$$\underline{g} = \sum_1^n \beta_i \underline{u}_i \quad \text{and} \quad \underline{f} = \sum_1^n \left(\frac{\beta_i}{\sigma_i}\right) v_i \tag{2.1}$$

Again the expansion for $\underline{f}$ is truncated to $\underline{f}^{(k)} = \sum_1^k \left(\frac{\beta_i}{\sigma_i}\right) \underline{v}_i$

using an interactive approach to obtain the best k.

Of course this is also a special case of the truncated QR method: the vectors $\{y_i\}$ are the singular vectors $\{v_i\}$, and the triangular matrix R is now the diagonal matrix D (and thus incidentally there is no problem in determining the condition number from the diagonal elements!). Thus if there is a good solution in terms of the first few singular vectors, this method works well; however this depends on the problem and the data $\underline{g}$. In particular since the vectors $\{u_i\}$ and $\{v_i\}$ are independent of $\underline{g}$, this method provides a good solution only for those $\underline{g}$ whose components $\beta_i \to 0$ faster than do the σ_i. Although this looks like a discrete analog of the Picard theorem mentioned in the first section, so that one is tempted to say that it is only for such $\underline{g}$ that the problem is well-conditioned, this is completely misleading. For a given $\underline{g}$, even though for the singular vectors $\underline{u}_i$ the β_i never tend to zero (so the SVD gives a poor solution), there may be expansions in other vectors $\{y_i\}$ which give a good approximate solution for $\underline{f}$. We give such an example in Section 4.

(c) Damped Least Squares or Regularization

This method is also very well known (see [7], [9], and [3]). The problem $Kf = g$ is modified to

$$\min_f(||Kf - g||_2^2 + \alpha^2||f||_2^2) \tag{2.2}$$

where α is a free parameter. This is equivalent to finding the least squares solution to the overdetermined linear system

$$\begin{pmatrix} K \\ \alpha I \end{pmatrix} f = \begin{pmatrix} g \\ 0 \end{pmatrix}$$

which is in turn equivalent to the normal equations

$$(K^TK + \alpha^2 I)f = K^Tg. \tag{2.3}$$

Since one normally must solve this using several values of α, the most efficient computational procedure is to again use the singular value decomposition of K. Assume $K = UDV^T$ and $g = \sum_1^n \beta_i u_i$; then the solution (which is easily modified for different α) is

$$f_\alpha = \sum_1^n \left(\frac{\beta_i}{\sigma_i + \alpha^2/\sigma_i} \right) v_i \ . \tag{2.4}$$

Thus the solution can again be expressed in terms of the singular vectors $\{v_i\}$ (compare (2.1)), so the same comments we made in (b) apply here: namely that if there is a good approximate solution in terms of these singular vectors, this method works well (and in fact this method and the truncated singular value decomposition produce very similar results).

Sometimes this problem is expressed in a more geometric way as a constrained least squares problem:

$$\min_f ||Kf - g||_2 \quad \text{subject to} \quad ||f||_2 \leq \gamma$$

where γ is now the free parameter. Normally (see Elden [3]) the solution occurs when $||f||_2 = \gamma$, and the problem with an equality constraint is equivalent to minimizing the quadratic form with Lagrange multiplier α^2:

$$\Phi(f) = (g - Kf)^T(g - Kf) + \alpha^2 f^T f.$$

This leads again to the normal equations (2.3), and the equality constraint gives a nonlinear equation relating the parameters α and γ:

$$f^T f = \gamma^2 = \sum_1^n \frac{\beta_i^2}{(\sigma_i + \alpha^2/\sigma_i)^2} \quad .$$

Thus the problems are equivalent, and although this constrained approach may look more natural, it is more difficult to solve (because of the additional nonlinear equation) and there is still a free parameter γ to choose.

(d) Modified regularization

Here the minimization problem of (2.2) is modified to

$$\min_f(||Kf - g||_2^2 + \alpha^2||Lf||_2^2) \tag{2.5}$$

where L is some matrix, normally a discrete approximation to some derivative operator. For example (and we shall use this specific one in the numerical examples)

$$L = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}_{(n-1)xn} \tag{2.6}$$

which is a discrete approximation to the first derivative, except for a scaling factor. The geometric motivation for (2.5) is that instead of keeping $||f||_2$ small as in (2.2), we keep $||Lf||_2$ small, which should mean we find some non-oscillatory solution f. This is of course rather vague, and we feel that a better understanding of the method (as with the other three we've discussed)

comes from an algebraic examination. As in (c), (2.5) is equivalent to finding the least squares solution to the overdetermined linear system

$$\begin{pmatrix} K \\ \alpha L \end{pmatrix} f = \begin{pmatrix} g \\ 0 \end{pmatrix}$$

which is in turn equivalent to the normal equations

$$(K^T K + \alpha^2 L^T L) f = K^T g \quad . \tag{2.7}$$

Again α is a free parameter, and we would like to be able to solve (2.7) easily for different values of α. This however is not so easy, and cannot be effected using merely the singular value decomposition of K as in (c). Two ways have been devised for coping with this problem:

(i) convert (2.5) back to a standard regularization problem

This could be done easily if L were invertible, but it is not; in fact if L is a discrete approximation to a p-th order derivative, L is normally an $(n-p)\times n$ matrix of rank (n-p). However, this can still be done using the pseudoinverse L^{ψ}, by a technique due to Elden [3]. One can also think of L^{ψ} as the discrete Green's function associated with the differential operator which L approximates. Thus the technique is a discrete version of a technique due to Hilgers[5] who connects problem (d) to problem (c) in the continuous case (i.e. assuming f is a function rather than a vector) using the Green's function.

The discrete technique of Elden goes as follows: assuming L is $n\times(n-p)$ and of rank (n-p), find the QR decomposition of L^T:

$$L^T = (V_1 | V_2)\begin{pmatrix} R \\ \hline 0 \end{pmatrix}$$

so V_2 spans the nullspace of L. Then $L^{\psi} = V_1 R^{-T}$. Now set $x = Lf$; then $f = L^{\psi}x + V_2 y$ for some x and y. Now we want $||Kf - g||_2 = ||Ax - b||_2$ for some A,b to give a problem like (2.2). But

$$Kf - g = KL^{\psi}x - g + KV_2 y \quad ,$$

and if we use the QR decomposition of $KV = (Q_1 | Q_2)\begin{pmatrix} U \\ \hline 0 \end{pmatrix}$, then

$$Q^T(Kf - g) = \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix}(KL^{\psi}x - g + KV_2 y)$$

$$= \begin{pmatrix} Q_1^T(KL^{\psi}x - g) + Uy \\ Q_2^T(KL^{\psi}x - g) \end{pmatrix}$$

As long as KV_2 has full rank (i.e. if K and L have no nullspace in common), U is nonsingular so we can determine y by demanding that the first part of this vector be zero. Thus if we define $A = Q_2^T KL^{\psi}$, $b = Q_2^T g$, then $||Ax - b||_2 = ||Kf - g||_2$ and we have reduced our problem to

$$\min_x(||Ax - b||_2^2 + \alpha^2||x||_2^2).$$

This is certainly a useful computational procedure, and we have verified it computationally; in all the examples of Section 4 it gave the same results as solving (2.7) directly. There is a second way, however, of dealing with (2.7) which is more useful for us here.

(ii) the generalized singular value decomposition (van Loan [10])

Given K and L, this forms the decomposition

$$K = UD_a X^{-1}, \qquad L = VD_b X^{-1}$$

where X is the eigenvector matrix of $(K^TK - \lambda L^TL)$. $U(n\times n)$ and $V((n-p)\times(n-p))$ are orthogonal, $D_a = \text{diag}\,(a_1, \ldots, a_n)$

$$D_b = \begin{pmatrix} b_1 & & & & \\ & \ddots & & & 0 \\ & & b_{n-p} & & \end{pmatrix}_{(n-p)\times n},$$

and the generalized eigenvalues

$$\lambda_i = {}^{a}i/b_i \ , \ i = 1, \ldots, n-p$$

$$= \infty \ , \ i = n-p+1, \ldots, n.$$

Thus the last p columns of X are a basis for the nullspace of L. For convenience we set $b_i = 0$, $i = n-p+1, \ldots, n$ so $\lambda_i = {}^{a}i/b_i$, $i = 1, \ldots, n$. Now consider (2.7) using this decomposition;

$$(K^TK + \alpha^2 L^TL)f_\alpha = K^Tg$$

becomes

$$(D_a^2 + \alpha^2 D_b^T D_b)X^{-1} f_\alpha = D_a U^T g.$$

So if $g = \sum_1^n \beta_i u_i$,

$$f_\alpha = \sum_1^n \left(\frac{\beta_i}{a_i + \alpha^2 b_i^2/a_i} \right) x_i \ , \tag{2.8}$$

and the straightforward solution of $Kf = g$ is $f = \sum_1^n \frac{\beta_i}{a_i} x_i$.

Thus we again have an algebraic expansion of the solution, not in the singular vectors of K (as in (b) and (c)), but in the generalized singular vectors of (K,L). Because of the nature of K and L, a permutation of the $\{a_i\}$ (which are the singular values of KX) tends to zero as do the singular values of K, while the $\{b_i\}$ (singular values of LX) do not, except that the last p are zero. These zero b_i however do not correspond to small a_i unless K and L have "near-nullspaces" in common, which leads to instabilities; this is not the case here because the nullspace of L has very smooth vectors, yet K is nearly singular only for very oscillatory vectors. Thus the nature of the expansion is much like that of (c) in (2.4) except for different vectors $\{x_i\}$. Notice also that the components in (2.8) corresponding to the last p vectors (i.e. the smooth nullspace basis for L) are independent of the free parameter α. Again we come back to the central idea of the first section: if there is a good approximate solution to $Kf = g$ in terms of these vectors $\{x_i\}$, this method works well (using a reasonable α).

3. Perturbation Analysis

The most important requirement of the solutions to any of the techniques of the previous section is that they be insensitive to changes in the data. Here we shall discuss this for each of the methods by perturbing the data g to $\bar{g}$ and

and examining what happens to the generated solution $\bar{f}$. For this to be meaningful, we must assume that the original problem $Kf = g$ has a smooth solution f; that is, in the sense of the definition of conditioning made in Section 1, the exact solution $f = \sum_1^k c_i y_i$, for some $\{y_i\}$ which do not reflect the ill-condition of K.

(a) Truncated QR method

Suppose we use orthogonal vectors $\{z_i\}$; then the exact solution $f = \sum_1^n d_i z_i$, where $(KZ)\underline{d} = \underline{g}$ (we may need all z_i, i=1, ..., n if $\{z_i\} \not\equiv \{y_i\}$). The truncated QR method will give $\bar{f} = \sum_1^k \bar{d}_i z_i$, where $\{\bar{d}_i\}$ is the least squares solution to the overdetermined system $(KZ_k)\underline{\bar{d}} = \bar{g}$. Since the $\{z_i\}$ are orthogonal,

$$||f - \bar{f}||_2^2 = \sum_1^n (d_i - \bar{d}_i)^2 = \sum_1^k (d_i - \bar{d}_i)^2 + \sum_{k+1}^n d_i^2 . \qquad (3.1)$$

Thus we see immediately that unless the $\{z_i\}$ are "like" $\{y_i\}$ in the sense that the components $|d_i|$ are small for $i > k$ (whatever k is chosen), the perturbed solution $\bar{f}$ cannot be close to f, even with no perturbation.

Now perform the QR decomposition:

$$KZ = QR, \quad R = \begin{pmatrix} R_1 & S \\ 0 & R_2 \end{pmatrix},$$

and thus

$$\begin{pmatrix} R_1 & S \\ 0 & R_2 \end{pmatrix} \begin{pmatrix} \underline{d}_1 \\ \underline{d}_2 \end{pmatrix} = \begin{pmatrix} (Q^T g)_1 \\ (Q^T g)_2 \end{pmatrix}, \qquad R_1 \underline{\bar{d}}_1 = (Q^T \underline{\bar{g}})_1$$

So we have

$$R_1(\underline{d}_1 - \overline{\underline{d}}_1) = Q^T(\underline{g} - \overline{\underline{g}}) - S\underline{d}_2$$

and hence if we scale our perturbation so $||g - \overline{g}||_2 = \varepsilon$,

$$||d_1 - \overline{d}_1||_2 \leq ||R_1^{-1}||_2 \, (\varepsilon + ||S||_2 \, ||\underline{d}_2||_2). \tag{3.2}$$

Notice that for k small, the dominant term in (3.2) or (3.1) is $||\underline{d}_2||_2$. As k is increased, this decreases but the term $||R_1^{-1}||_2$ increases as $(KZ)_k$ becomes more ill-conditioned. Thus the situation is like that in the definition of conditioning in Section 1: if the $\{z_i\}$ are such that a solution for some k can be found with k large enough that there is a small residual, yet small enough that the ill-condition of K is not reflected, the generated solution $\overline{f}$ will be insensitive to perturbation in the data g. Notice that there are two sources of error: the perturbation $g \to \overline{g}$, and the different definition of the exact and generated solutions f and $\overline{f}$, even with no perturbation.

(b) Truncated SVD

As in (2.1), let $K = UDV^T$, $g = \Sigma\beta_i u_i$, and $\overline{g} = \Sigma\overline{\beta}_i u_i$. Then the generated solution $\overline{f}_k = \sum_1^k \frac{\overline{\beta}_i}{\sigma_i} v_i$, and

$$f - \overline{f}_k = \sum_1^k \left(\frac{\beta_i - \overline{\beta}_i}{\sigma_i} \right) v_i + \sum_{k+1}^n \frac{\beta_i}{\sigma_i} v_i \, .$$

Thus

$$||f - \overline{f}_k||_2^2 \leq \varepsilon^2 \sum_1^k \frac{1}{\sigma_i^2} + \sum_{k+1}^n \left(\frac{\beta_i}{\sigma_i} \right)^2 \tag{3.3}$$

or

$$E(k) = R(k) + T(k)$$

Here again we have one term T(k) which decreases with k, and one term R(k) which increases with k. Clearly this error bound is minimized if we pick k that $|\beta_i| < \varepsilon$ for $i > k$. In fact on first sight it appears that this rather simple criterion could be used to choose k in practice, with ε being the error in the data. However this does not work well; other errors are involved than just the perturbation in g, so the $|\beta_i|$ rarely fall below ε and the optimal k depends on the $\{\sigma_i\}$ as well.

One can however carry the perturbation analysis further, and estimate the minmum error possible, given a particular distribution of the $\{\beta_i\}$ and $\{\sigma_i\}$.

(i) geometric distribution: $\sigma_i = a^{-i}$, $\beta_i = a^{-2i}$ with $a > 1$ and $\varepsilon << 1$. Then assuming $n >> k$, (3.3) gives

$$E(k) = \frac{\varepsilon^2 a^2 (a^{2k} - 1) + a^{-2k}}{a^2 - 1}$$

which is minimized for $k = k_0$ such that $a^{2k_0} = \frac{1}{\varepsilon a}$, with minimum error

$$E(k_0) = \frac{\varepsilon a(2 - \varepsilon a)}{a^2 - 1} \cong \frac{\varepsilon}{a} .$$

Thus the minimal perturbation error in $\bar{f}$, given this distribution, is $\sqrt{\varepsilon/a}$.

(ii) polynomial distribution: $\sigma_i = i^{-p}$, $\beta_i = i^{-q}$, with $q > p$. Approximating the sums in (3.3) by integrals and assuming $n >> k$ gives

$$E(k) \cong \frac{\varepsilon^2 k^{2p+1}}{2p+1} + \frac{k^{-(2q-2p-1)}}{2q-2p-1}$$

which is minimized for $k = k_0 = \varepsilon^{-1/q}$, giving as minimum error

$$E(k_0) \cong \frac{2q}{(2p+1)(2q-2p-1)} \varepsilon^{\frac{2q-2p-1}{q}}$$

Thus the minimal perturbation error in $\bar{f}$, taking the square root, is roughly $\varepsilon^{1 - \frac{2p+1}{2q}}$.

(c) damped least squares

Again assuming $g \to \bar{g} = \sum_1^n \bar{\beta}_i u_i$, and using the SVD expansion of K so that the exact solution $f = \sum_1 \frac{\beta_i}{\sigma_i} v_i$, the generated solution is, from (2.4),

$$\bar{f}_\alpha = \sum_1^n \left(\frac{\bar{\beta}_i}{\sigma_i + \alpha^2/\sigma_i}\right) v_i \ .$$

Thus the error is

$$f - \bar{f}_\alpha = \sum_1^n \left(\frac{(\beta_i - \bar{\beta}_i)\sigma_i + \alpha^2 \beta_i/\sigma_i}{\sigma_i^2 + \alpha^2} \right) v_i \ . \qquad (3.4)$$

For α close to zero, this error has large components in the high-order v_i since the $\sigma_i \to 0$ and we know only $|\beta_i - \bar{\beta}_i| < \varepsilon$. Moreover as $\alpha \to \infty$, $\bar{f}_\alpha \to 0$ so the error is again large (but finite). Hopefully there is some intermediate value of α where the error is minimized, but this is not clear in general, and depends on

the distribution of $\{\sigma_i\}$ and $\{\beta_i\}$.

(i) geometric distribution: $\sigma_i = a^{-i}$, $\beta_i = a^{-2i}$, with $a > 1$.

For this, (3.4) gives

$$||f - \bar{f}_\alpha||_2^2 \leq (\varepsilon/\alpha + \alpha)^2 \sum_1^n \frac{1}{(\frac{1}{\alpha a^i} + \alpha a^i)^2}$$

This is difficult to evaluate explicitly, but the sum can be bounded independent of α, for any n. Since the denominator is of the form $(x + \frac{1}{x})^2$, the maximum term in the sum occurs when $\alpha a^i = 1$ (call this i_0). Then

$$\sum_{i_0}^{\infty} \leq \sum_{i_0}^{\infty} \frac{1}{(\alpha a^i)^2} \leq \frac{1}{1 - a^{-2}}$$

and

$$\sum_1^{i_0} \leq \sum_1^{i_0} (\alpha a^i)^2 \leq \frac{1}{1 - a^{-2}} \quad .$$

Thus

$$||f - \bar{f}_\alpha||_2 \leq \frac{\sqrt{2}(\varepsilon/\alpha + \alpha)}{\sqrt{1 - a^{-2}}} \quad .$$

So if we choose $\alpha = \sqrt{\varepsilon}$, we again get a perturbation error of the order $\sqrt{\varepsilon}$ (as we did in case (b)).

(ii) polynomial distribution: $\sigma_i = i^{-p}$, $\beta_i = i^{-q}$, with $q > p$.

This gives

$$||f - \bar{f}_\alpha||_2^2 \leq \sum_1^n \left(\frac{\varepsilon i^{-p} + \alpha^2 i^{p-q}}{i^{-2p} + \alpha^2} \right)^2 .$$

For general p and q this is somewhat intractable; however for q = 2p,

$$||f - \bar{f}_\alpha||_2^2 \leq (\varepsilon/\alpha + \alpha)^2 \sum_1^n \frac{1}{(\alpha i^p + \frac{1}{\alpha i^p})^2} ,$$

the same form as in (i). Now choose i_0 such that $\alpha i_0^p = 1$ (i.e. $i_0 = \alpha^{-1/p}$). Then

$$\sum_{i_0}^{\infty} \leq \sum_{i_0}^{\infty} \alpha^{-2} i^{-2p} \cong \frac{\alpha^{-2} i_0^{-2p+1}}{2p - 1} = \frac{i_0}{2p - 1}$$

and

$$\sum_1^{i_0} \leq \sum_1^{i_0} \alpha^2 i^{-2p} \cong \frac{\alpha^2 i_0^{2p+1}}{2p + 1} = \frac{i_0}{2p + 1}$$

where we have approximated the sums by integrals. Thus

$$||f - \bar{f}_\alpha||_2 \leq \sqrt{\frac{4p}{4p^2 - 1}} (\varepsilon/\alpha + \alpha) \alpha^{-1/2p}$$

So if we choose $\alpha = \sqrt{\varepsilon}$, we get a perturbation error in $\bar{f}_\alpha$ of the order $\varepsilon^{\frac{2p-1}{4p}}$, the same as in (b) with the same distribution.

(d) <u>modified regularization</u>

Applying the generalized SVD mentioned in Section 2, with data $g = \sum_1^n \beta_i u_i$ and $\bar{g} = \sum_1^n \bar{\beta}_i u_i$, the exact solution f of Kf = g has the expansion $f = \sum^n \frac{\beta_i}{a_i} x_i$ and the generated solution $\bar{f}_\alpha$ is, from (2.8),

$$\bar{f}_\alpha = \sum_1^n \left(\frac{\bar{\beta}_i}{a_i + \alpha^2 b_{i/a_i}}\right) x_i \ .$$

Thus the perturbation analysis can proceed just as in (c), except that the vectors $\{x_i\}$ are not orthogonal. The minimal perturbation error will now depend on the rates at which $\beta_i \to 0$ and $a_i \to 0$ (rather than β_i and σ_i) and there is the additional factor of $\kappa(X)$.

4. Numerical Results

Here we present the results of applying each of the four methods discussed to three different problems. We give the results for each problem separately, and try to interpret the results in the light of our previous discussion, with particular emphasis to the existence of a good approximate solution expanded in the basis vectors given by each method.

I. Inverse Laplace Transform

Given g(t), we wish to find f(α) so that

$$\int_0^\infty e^{-st} f(s)ds = g(t) \ . \qquad (4.1)$$

This is a common ill-posed problem, occurring frequently in various scientific applications. Normally, g(t) is only measured at certain points; however to test

the numerical methods, we assume g(t) is given analytically, with known transform f(s), so that we can measure the errors in the discrete solutions.

We discretize the problem by applying the Gauss-Laguerre quadrature rule of degree n to (4.1) at n sample points $\{t_i\}_1^n$. Thus the discrete solution is obtained at the abscissae $\{s_i\}$ of the quadrature rule. In our case, the choice of sample points is arbitrary since we know g(t) analytically, but in practice this must be the given data points. (Notice one could use more than n data points; this would give an overdetermined system Kf = g, but each of the methods can be easily modified to handle this.)

The particular problems chosen were

$$g(t) = \frac{1}{t + 0.5}, \quad f(s) = e^{-s/2}$$

and

$$g(t) = \frac{1}{t} - \frac{1}{t + 0.5}, \quad f(s) = 1 - e^{-s/2}.$$

The actual data used was $g(t) + \mu\chi(t)$, where $\chi(t)$ is a normally distributed random variable. This includes noise in the data, with noise level μ (we took $\mu = 0$ and $\mu = .001$). For sample points, we took the equally distributed points $t_i = i$ in the results below. We also tried $t_i = s_i$ (the Gauss-Laguerre abscissae), with comparable results: some results were better, others worse.

soln: exp(-s/2)				
	n=10,μ=0	n=10,μ=.001	n=20,μ=0	n=20,μ=.001
QR	.183	.183	.253	.253
SVD	.099	.099	.065	.065
LS	.053	.053	.071	.071
MLS	.262	.262	.156	.156

soln: 1-exp(-s/2)				
	n=10,μ=0	n=10,μ=.001	n=20,μ=0	n=20,μ=.001
QR	1.0|0.93	1.0|0.93	1.0|0.72	1.0|0.72
SVD	1.0	1.0	1.0	1.0
LS	1.0	1.0	1.0	1.0
MLS	.136	.136	.136	.136

Results quoted are maximum errors at the abscissae $\{s_i\}$ for the best choice of the free parameter k or α. Note first of all that all the results are insensitive to noise. Let us examine the results more closely for each of the methods in turn.

(a) QR: The vectors $\{y_i\}$ chosen for the expansion were simply the unit vectors $\{e_i\}$. This is appropriate for the first example, because of the exponential decay. However (particularly for n=20) the condition number of KY_k became large (giving round off error) before enough vectors could be included in the expansion. The best maximum errors (quoted above) occurred for k = 3. For the second example, these vectors are clearly not appropriate, as all of them are needed to give a good expansion of the solution (thus explaining the errors 1.0). The

vectors were modified to include $e = (1,1,\ldots,1)^T$, giving the second set of errors. Again however KY_k became ill-conditioned very quickly. Better results could probably be achieved with other choices of $\{y_i\}$.

(b) SVD: For this problem, the singular vectors $\{v_i\}$ of K are oscillatory (v_i has (i-1) sign changes), have their maximum at the diagonal component, and damp out very fast. Thus we can expect a good solution using the first few vectors if the expected solution decays to zero. This is the case with the first example (best solution obtained with k = 2 or 3), not with the second.

(c) LS: The same comments apply; for the first example, the best solution was obtained for $\alpha = .04$.

(d) MLS: The results here are very interesting: for the first example, the results are very bad; in fact the maximum error always occurred for s large because the generated solution did not decay to zero, but to some other value (for example to .262 in the case n = 10). For the second example, however, this was the only method which gave a reasonable result, and again the maximum error was asymptotic. We should also mention that this asymptotic error was very dependent on the choice of sample points: for the other choice of sample points ($t_i = s_i$) given earlier, the results were better for the first example, and about the same for the second.

Again, the errors should be interpreted with regard to the particular expansion used (see (2.8)): the L used was that in (2.6), and the $\{x_i\}$ could be explicitly calculated for this K, using the QZ algorithm of [6]. In particular the nullspace of L is spanned by $e = (1,1,\ldots,1)^T$ so this is the last generalized singular vector x_n. The corresponding a_n was in fact the largest $\{a_i\}$, and the

other $\{a_i\}$ or $\{a_i/b_i\}$ tended to zero very fast, much like the singular values $\{\sigma_i\}$ of K. The corresponding vectors $\{x_i\}$ were like the singular vectors $\{v_i\}$, except that they decayed to $x_n = e$ rather than zero. Thus there is a good approximation to the solution of the second example among the dominant $\{x_i\}$, but not for the first example.

II. Finite, Compact Kernel

Here we apply the methods to two examples of compact kernals on finite intervals: $\int_0^1 K(s,t)f(s)ds = g(t)$ with

(i) $K(s,t) = \begin{cases} s(1-t) & \text{for } s \geq t \\ t(1-s) & \text{for } s \geq t \end{cases}$ (Green's function for second derivative)

with the functions $g(t) = t(1-t)e^{-t}$, $f(s) = (4-s)(s-1)e^{-s}$.

(ii) $K(s,t) = \cos(st)$, $g(t) = \frac{\sin t}{t} + \frac{\cos t - 1}{t^2}$, $f(s) = s$.

Now it is appropriate to use the Gauss-Legendre quadrature rule of order n at n sample points $\{s_i\}$, which we take to be equally spaced in the interval [0,1]. Again other choices of sample points gave comparable results, and again we perturbed the data by a discrete normal random variable, giving noise level μ.

Green's function				
	n=10, μ=0	n=10, μ=.001	n=20, μ=0	n=20, μ=.001
QR	.129|.167	.148|.116	.027|.033	.025|.028
SVD	3.6	3.6	3.7	3.7
LS	3.5	3.5	3.9	3.9
MLS	.265	.289	.461	.481

cos(st)				
	n=10, μ=0	n=10, μ=.001	n=20, μ=0	n=20, μ=.001
QR	.00004|.126	.003|.33	.00008|.136	.009|2.3
SVD	.425	.586	.695	.785
LS	.603	.603	.791	.723
MLS	.056	.053	.059	.056

Again the errors given are maximum errors for the best choice of the free parameters. Notice that here the noise does have an effect on the solution generated. Again let us describe each method in turn.

(a) QR: We chose for expansion vectors $\{y_i\}$ discrete versions of two orthogonal sets over [0,1]: orthogonal polynomials and Fourier series. This worked well in the first example, choosing $k \simeq 6$. And orthogonal polynomials worked perfectly in the second example, because the solution is a linear polynomial (notice that when noise was introduced, the error increased to the noise level). The Fourier series expansion did not work well, because of the poor convergence of the Fourier series for $f(s) = s$. Noise swamped the generated solution here, showing that the matrix KY_k was too ill-conditioned.

(b) SVD: For both examples, the singular vectors $\{v_i\}$ of K look like discrete versions of the eigenfunctions of the kernel, namely $\sin(j\Pi s)$, $j = 1,2,\ldots$. Thus the SVD method will work well only for solutions with a rapidly converging sine series. Unfortunately neither of these examples satisfies this, hence the poor results.

(c) LS: Again the same comments apply.

(d) MLS: For the first example (the Green's function), again with L as in (2.6), the $\{a_i/b_i\} \to 0$ like the singular values $\{\sigma_i\}$ of K, and the corresponding $\{x_i\}$ behaved like discrete versions of $\cos(j\Pi s)$, $j = 0,1,2,\ldots$ ($j = 0$ giving $e = x_n$ = nullspace vector of L). This expansion gave a better solution that SVD or LS, but still poor. In the second example, the structure of the $\{x_i\}$ was harder to discern, since the $\{a_i/b_i\}$ became small very quickly (all except two were below 10^{-6}) and thus the vectors contaminated each other. The two "nontrivial" vectors were x_n = e and another vector which behaved nearly linearly; this is why a good solution was found to the second example: the particular solution sought was linear.

References

1. C. T. H. Baker et al, Numerical solution of Fredholm integral equations of the first kind. Comp. J. 7 (1964), 141-148.

2. Courant and Hilbert, Methods of Mathematical Physics I. Interscience, New York, 1953.

3. L. Elden, Numerical methods for the regularization of Fredholm integral equations of the first kind. Tech Rep., Math. Dept., Linköping University, Sweden, 1974.

4. R. J. Hanson, A numerical method for solving Fredholm integral equations of the first kind using singular values. SIAM J. Num. Anal. 8 (1971), 616-622.

5. J. W. Hilgers, On the equivalence of regularization and certain reproducing kernal Hilbert space approaches for solving first kind problems. SIAM J. Num. Anal. 13 (1976), 172-184.

6. C. B. Moler and G. W. Stewart, An algorithm for generalized matrix eigenvalue problems. SIAM J. Num. Anal. 10 (1973), 241-256.

7. D. L. Phillips, A technique for the numerical solution of certain integral equations of the first kind. J. ACM 9 (1962), 84-97.

8. H. Rutishauser, One again: the least squares problem. J. Lin. Alg. Appl. 1 (1968), 479-488.

9. A. N. Tihonov, Solution of incorrectly formulated problems and the regularization method. Dokl. Akad. Nauk. SSR 151 (1963), 501-504. (Soviet Math. Dokl. 4, 1035-1038).

10. C. van Loan, Generalizing the singular value decomposition. Tech. Rep., Math. Dept., Univ. of Manchester, England, 1974.

11. J. M. Varah, On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems. SIAM J. Num. Anal. 10 (1973), 257-267.