On the Numerical Solution of Ill-Conditioned Linear Systems With Applications to Ill-Posed Problems

J. M. Varah
Department of Computer Science
University of British Columbia
Vancouver 8, Canada

1. Introduction

Consider the set of linear equations

$$Ax = b \tag{1.1}$$

where A is an $n \times n$ real nonsingular dense stored matrix. Such a system is commonly solved by Gaussian elimination with pivoting, and it is well-known that the accuracy of the computed solution depends primarily on the condition of the matrix. A convenient measure of the condition is in terms

of the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. These are the nonnegative square roots of the eigenvalues of $A^T A$. If the computed solution $\tilde{x} = x + \delta x$, then

$$(A + \delta A)(x + \delta x) = b, \quad \|\delta A\|_2 \leq K_0(n)\eta_1$$

and (1.2)

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq K_1(n)\left(\frac{\sigma_1}{\sigma_n}\right) \eta_1$$

where η_1 is the machine roundoff level, $\|x\|_2^2 = x^T x$, and $K_0(n)$, $K_1(n)$ depend on the pivotal growth. (See for example Wilkinson [23, Chapter 4] or [24, pg. 106].) We assume A is scaled so that $\sigma_1 = 1$ so the error depends on the smallest singular value σ_n. We emphasize that this bound is independent of b, and in fact is realistic regardless of the right-hand side.

However, when A is ill-conditioned (i.e. σ_n small), we may have prior knowledge that the exact right-hand side and solution depend primarily on the largest few eigenvectors of A (assume A is symmetric for the moment). In this case, the solution will be insensitive to the condition of the full matrix A and if we solve the system by some kind of eigenvector decomposition, much more accuracy can be obtained. The extreme example of this is when b is an eigenvector corresponding to the eigenvalue $\lambda_1 = \sigma_1 = 1$; $x=b$ is a solution no matter how ill-conditioned A is. In what follows, we will develop this idea using as our means of solution the singular value decomposition, and then apply the technique to the solution of some classical ill-posed linear problems, specifically harmonic continuation, inversion of Laplace transforms, and the backwards heat equation.

2. The Singular Value Decomposition

Now we return to the general real matrix A of (1.1). Any such matrix can be expressed as

$$A = UDV^T$$

where U is the orthogonal matrix of eigenvectors of AA^T, V is the same for A^TA, and $D = \text{diag}(\sigma_i)$. For A symmetric positive definite, this is simply the eigenvector decomposition $A = Q \Lambda Q^T$. For a discussion and further references, see Golub and Kahan [12]. More recently, constructive algorithms forming this singular value decomposition have been given (Businger and Golub [4], Golub and Reinsch [13]).

Using this decomposition the linear equation Ax = b becomes

$$D(V^Tx) = U^Tb$$

or

$$Dy = \beta,$$

that is, the system is transformed into diagonal form by rotations of the domain and image space. Put another way, we make a decomposition of b into the basis given by the columns $\{u_i\}$ of U, i.e.

$$b = \sum_1^n \beta_i u_i \tag{2.1}$$

and form

$$y = \sum_1^n \frac{\beta_i}{\sigma_i} e_i = D^{-1}\beta \tag{2.2}$$

and finally x = Vy. Here $\{e_i\}$ denote the Cartesian unit vectors.

We will be interested in cases where the β_i decreases as i increases faster than do the σ_i (i.e., β_i/σ_i decreases). In this case both the data and solution have small components in the high-order u_i, so we can expect an accurate approximation to the solution using only the low-order u_i. We assume in what follows that $||x||_2 = ||y||_2 = 1$ without loss of generality.

Notice that we must be careful in defining the right-hand side b, because the actual $\tilde{b}$ used in the computation will have some data error associated with it. We stipulate that the b used in the analysis is the exact b without any rounding errors, usually obtained from the corresponding continuous problem by

exact discretization. Thus we must consider the right-hand side $\tilde{b}$ used in the computation as having some data error associated with it, say $\|\delta b\|_2 = \|b-\tilde{b}\|_2 \leq d\eta_1$. Since the rounding errors are essentially random, $\tilde{b}$ will always have components of order η_1 in all u_i, so case (iii) would never hold if we were considering $\tilde{b}$ as our given right-hand side.

We could also interpret this another way: given the computational problem $Ax = \tilde{b}$, we choose our b from within $\|b - \tilde{b}\|_2 \leq d\eta_1$ so that its decomposition in the $\{u_i\}$ basis has components β_i which decrease as fast as possible. Then we solve the problem by our technique (which follows) and if the β_i decrease faster than the σ_i, our solution will be close to the exact solution of $Ax = b$ (but not necessarily to $Ax = \tilde{b}$).

The technique we use is a common one: let $D^{(k)} = \text{diag}\,(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. Then form

$$A^{(k)} = U D^{(k)} V^T , \quad b^{(k)} = \sum_1^k \beta_i u_i$$

and solve

$$A^{(k)} x^{(k)} = b^{(k)} \tag{2.3}$$

by the singular value decomposition, i. e.

$$y^{(k)} = \sum_1^k \frac{\beta_i}{\sigma_i} e_i , \quad x^{(k)} = V y^{(k)} . \tag{2.4}$$

This idea of replacing the smaller singular values by zero is used in solving linear least squares problems (Golub [11] and Björck [3]). It is one of many strategies for replacing the n equations given by k (fewer) equations which are better conditioned; the underlying problem of finding the effective rank of a system arises frequently – see Peters and Wilkinson [19] for a good discussion of this.

To decide which k to use, we must examine the error involved. This consists of two parts:

(a) the truncation error $x^{(k)} - x$

(b) the roundoff error in computing $x^{(k)}$.

The error in (a) is easily obtained from the singular value decomposition:

$$\|x^{(k)} - x\|_2 = \|y^{(k)} - y\|_2 = \Big\|\sum_{k+1}^{n} \frac{\beta_i}{\sigma_i} e_i\Big\|_2 = \left(\sum_{k+1}^{n} \left(\frac{\beta_i}{\sigma_i} \right)^2 \right)^{1/2} \qquad (2.5)$$

Now consider the error in computing $x^{(k)}$. Since in (2.4) we ignore σ_i for $i > k$, the exact solution $y^{(k)}$ is the same as that for $D_0^{(k)} y^{(k)} = \beta^{(k)}$, where $D_0^{(k)} = \text{diag}\,(\sigma_1, \ldots, \sigma_k, 1, \ldots, 1)$. Thus $x^{(k)}$ is the exact solution of

$$A_0^{(k)} x^{(k)} = U\, D_0^{(k)} V^T x^{(k)} = b^{(k)} = U \beta^{(k)} . \qquad (2.6)$$

The computed solution is found as follows:

let the computed singular value decomposition be $\tilde{U}\,\tilde{D}\,\tilde{V}^T$ with $\tilde{U} - U = \delta U$, $\tilde{V} - V = \delta V$, $\tilde{D} - D = \delta D$, and assume

$$\max\left(\|\delta U\|_2, \|\delta V\|_2, \|\delta D\|_2\right) \le c\,\eta_1 . \qquad (2.7)$$

Also let $\tilde{D} = \text{diag}\,(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n)$, $\tilde{D}_0^{(k)} = \text{diag}\,(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_k, 1, \ldots, 1)$.

Then we form

$$\tilde{\beta}^{(k)} = \left[fl\,(\tilde{U}^T b) \right]^{(k)} = \left[(\tilde{U}^T + E_1)(b + \delta b) \right]^{(k)}$$

$$\tilde{y}_i^{(k)} = \frac{\tilde{\beta}_i}{\tilde{\sigma}_i}(1 + \epsilon_i) , \text{ so that } (\tilde{D}_0^{(k)} + E_2)\,\tilde{y}^{(k)} = \tilde{\beta}^{(k)}$$

$$\tilde{x}^{(k)} = fl\,(\tilde{V}\,\tilde{y}^{(k)}) = (\tilde{V} + E_3)\,\tilde{y}^{(k)}$$

where $\|E_i\|_2 \le r_i\,\eta_1$. The r_i are small positive numbers and depend

on the machine arithmetic used (see Wilkinson [22, Chapter 1]).

Combining we have

$$(\tilde{D}_0^{(k)} + E_2)(\tilde{V} + E_3)^{-1}\tilde{x}^{(k)} = \tilde{\beta}^{(k)} = \beta^{(k)} + \xi^{(k)} .$$

Then if we set $(\tilde{V} + E_3)^{-1} = V^T + R$, we have

$$U(D_0^{(k)} + \delta D^{(k)} + E_2)(V^T + R)\tilde{x}^{(k)} = U(\beta^{(k)} + \xi^{(k)})$$

or

$$(U D_0^{(k)} V^T + F)\tilde{x}^{(k)} = b^{(k)} + e^{(k)} .$$

Now we proceed to bound $\|F\|_2$ and $\|e^{(k)}\|_2$. First,

$$\xi = \tilde{\beta} - \beta = (\tilde{U}^T + E_1 - U^T)b + (\tilde{U}^T + E_1)\delta b$$

and thus for any k

$$\|e^{(k)}\|_2 = \|\xi^{(k)}\|_2 \leq \|\delta U\|_2 + \|E_1\|_2 + (1 + \|\delta U\|_2 + \|E_1\|_2)\|\delta b\|_2$$

$$\leq (c + r_1 + (1 + c + r_1)\eta_1)\eta_1 \equiv r_4\eta_1 . \qquad (2.8)$$

Then

$$F = U(\delta D^{(k)} + E_2)(V^T + R) + U D_0^{(k)} R$$

and

$$\|R\|_2 \leq \frac{\|\delta V\|_2 + \|E_3\|_2}{1 - (\|\delta V\|_2 + \|E_3\|_2)} \leq \frac{(c + r_3)\eta_1}{1 - (c + r_3)\eta_1} \equiv \rho\eta_1$$

gives

$$\|F\|_2 \leq (1 + \rho\eta_1)(\|\delta D\|_2 + \|E_2\|_2) + \rho\eta_1$$

$$\leq (1 + \rho\eta_1)(c + r_2)\eta_1 + \rho\eta_1 \equiv r_5\eta_1 . \qquad (2.9)$$

Now the standard perturbation results (see for example Wilkinson [23, page 189]) give, assuming $\|(A_0^{(k)})^{-1} F\|_2 < 1$,

$$\|\tilde{x}^{(k)} - x\|_2 \leq \frac{\|(A_0^{(k)})^{-1}\|_2}{1 - \|(A_0^{(k)})^{-1} F\|_2} \left(\|e^{(k)}\|_2 + \|F\|_2 \|x^{(k)}\|_2 \right) .$$

Here, $\|x^{(k)}\|_2 = \|y^{(k)}\|_2 \leq \|y\|_2 = 1$ and from its construction $\|(A_0^{(k)})^{-1}\|_2 = \frac{1}{\sigma_k}$, so that assuming k is chosen so that $\sigma_k > 2 r_5 \eta_1$, we have

$$\|\tilde{x}^{(k)} - x^{(k)}\|_2 \leq 2 (r_4 + r_5) \frac{\eta_1}{\sigma_k} \equiv \frac{K_2 \eta_1}{\sigma_k} \quad . \tag{2.10}$$

For computations in single precision arithmetic, $K_2 = 10n$ is a reasonable estimate. We combine (2.5) and (2.10) as follows.

THEOREM:

Let the solution to (1.1) be approximated by (2.3). Then for k chosen so that $\sigma_k > 2 r_5 \eta_1$, the computed solution $\tilde{x}^{(k)}$ to (2.3) satisfies

$$\|\tilde{x}^{(k)} - x\|_2 \leq \frac{K_2 \eta_1}{\sigma_k} + \sqrt{\sum_{k+1}^{n} \left(\frac{\beta_i}{\sigma_i} \right)^2} \quad . \tag{2.11}$$

Since the first term is increasing in k and the second decreasing, there will be some optimum value of k to choose which will minimize the error bound. For a given problem, it would be of theoretical interest to compute analytically this optimum value; computationally however, the $\{\beta_i\}$ and $\{\sigma_i\}$ are produced in the course of the computation and the optimal number of equations to use can be explicitly calculated.

Before proceeding to the applications, it is of interest to regard the technique from a different point of view. We can consider the eigenvectors $\{u_i\}$, for a fixed n and A as a discrete orthogonal basis for the solution space, and the solution $x^{(k)}$ as the truncated discrete Fourier expansion

in this basis. Moreover, if A is totally positive (see Gantmacher [9 , page 98]) these eigenvectors also have the oscillation property: u_1 has constant sign, u_2 has one sign change, etc., thus replicating the desirable property of continuous bases formed from solutions of Sturm–Liouville problems (Courant and Hilbert [6, page 451ff]).

3. Application I — Harmonic Continuation

All of the applications considered are to classical ill-posed problems, and in fact each example reduces to the solution of a linear integral equation of the first kind:

$$g(s) = \int_a^b K(s,t)f(t)dt.$$

As is well-known, small changes in g(s) cause arbitrarily large changes in the high-order Fourier modes of f(t). Problems of this type have been considered numerically by many different people: for example Phillips [20], Bellman et al [1], and Tihonov [21]. More recently, Hanson [15] has also used the singular value decomposition to solve these problems.

The first application is to harmonic continuation; given a harmonic function $u(r, \theta)$ in the unit circle with known values for some $r < 1$, $u(r, \theta) = g(\theta)$, to find its values $f(\theta)$ for $r = 1$. Now $f(\theta)$ and $g(\theta)$ are related by the Poisson integral:

$$\frac{1}{2\pi} \int_0^{2\pi} \frac{1-r^2}{1-2r\cos(\theta-\varphi)+r^2} f(\varphi)d\varphi = g(\theta) \qquad (3.1)$$

and their Fourier series are also intimately connected: if

$$f(\theta) = f_0 + \sum_{k=1}^{\infty} (f_k \cos k\theta + f_k' \sin k\theta) ,$$

then

$$g(\theta) = f_0 + \sum_{k=1}^{\infty} r^k (f_k \cos k\theta + f_k' \sin k\theta) .$$

Thus a small absolute change in a high-order Fourier coefficient of $g(\theta)$ causes a large change in $f(\theta)$. Clearly the problem is not well-posed unless we consider only small relative perturbations in $g(\theta)$.

This classical problem has been considered analytically by many people, and recently Franklin [8] has considered the numerical solution by solving a stochastic extension of the problem. Since the integrand in (3.1) is periodic, the obvious discretization is via the trapezoidal rule, using points φ_j, $\theta_j = \frac{2\pi j}{n}$, $j = 1, \ldots, n$, giving $Ax = b$ with

$$A_{ij} = \frac{1}{n} \left(\frac{1 - r^2}{1 + r^2 - 2r \cos \frac{2\pi(i-j)}{n}} \right) , \quad b_j = g\left(\frac{2\pi j}{n}\right) .$$

In this case, the eigenvector decomposition is immediate from the Poisson summation formula:

(i) for $0 \leq p \leq [\frac{n}{2}]$, $\lambda = \frac{r^p + r^{n-p}}{1 - r^n}$ is an eigenvalue with

eigenvector $\left(\cos \frac{2\pi jp}{n}, j = 1, \ldots, n \right)$

(ii) for $0 < p \leq \left[\frac{n-1}{2}\right]$, $\lambda = \frac{r^p + r^{n-p}}{1-r^n}$ is an eigenvalue with

$$\text{eigenvector } \left(\sin \frac{2\pi jp}{n} \ , \ j = 1, \ldots, n \right) \quad .$$

Thus for n large, the singular values $\sigma_{2p+1} = \sigma_{2p} \tilde{=} r^p$, $p = 0, 1, 2, \ldots$. Moreover, the $\{\beta_k\}$ are multiples of the trapezoidal rule approximations to the Fourier coefficients of g:

$$\beta_k = \sum_{j=1}^{n} g\left(\frac{2\pi j}{n}\right) \begin{matrix} \sin \\ \cos \end{matrix} \left(\frac{2\pi jk}{n}\right) \quad .$$

So if the Fourier coefficients of $g(\theta)$ decrease faster than r^k, we can get better accuracy using the singular value decomposition as is expected from the continuous problem.

Suppose for example that the analytic function

$$(g + ih)(z) = \sum_{k=0}^{\infty} (f_k - if'_k) z^k$$

formed from $g(\theta)$ has radius of convergence $1/\rho$ ($\rho < r$) so the Fourier coefficients of $g(\theta)$ decrease like ρ^k ; then the total error in solving the first (2k+1) equations using the singular value decomposition is from (2.11)

$$\| \tilde{x}^{(2k+1)} - x \|_2 \leq \frac{1}{r^k} \ (K_2 \eta_1 + \rho^k) \ .$$

From this, the optimal value for k can be found explicitly.

The technique was tried on the example treated by Franklin [8]: $(g+ih)(z) = z^3 - z + \sin z$, with $r = 1/2$, $n = 50$. The best numerical solution occurred for $k \tilde{=} 10$, $\sigma_{2k} \tilde{=} 10^{-3}$ and

gave at least four correct significant figures in all components, comparing favourably with the best results of [8]. Actually for $4 \le k \le 10$, the results were nearly the same, and the optimal k from (2.11) using the computed σ_i and β_i was k=5. (Computations were made on the IBM 360/75 at Caltech and the 360/67 at UBC; we used $\eta_1=10^{-7}$ and $K_2=10n$.)

4. Application II - Inversion of the Laplace Transform

This well-known ill-posed problem has been attacked numerically by many people (see Bellman et al [2] and references therein). We have

$$g(s) = \int_0^\infty e^{-st} f(t)dt \tag{4.1}$$

with $g(s)$ given, either analytically or at given points $\{s_i\}$. If we allow arbitrarily high-order harmonics in $f(t)$, then infinitesimal changes in $g(s)$ (like rounding errors) can cause large changes in $f(t)$. Essentially, our method discretizes the integration and takes $f(t)$ as that linear combination of the first m harmonics which most closely gives $g(s)$ as its transform. And m is chosen so the error in the transform is less than a prescribed tolerance.

The obvious discretization is using Gauss-Laguerre quadrature (we assume f is not periodic):

$$\int_0^\infty e^{-t} h(t)dt \cong \sum_1^n w_k h(t_k)$$

and so at any set of points $\{s_i\}$, we get

$$g(s_i) = \int_0^\infty e^{-s_i t} f(t)dt \cong \sum_{k=1}^n w_k e^{t_k(1-s_i)} f(t_k)$$

or in matrix form, $Af = g$, with

$$a_{ik} = w_k e^{t_k(1-s_i)} \quad . \tag{4.2}$$

Solving this using the singular value decomposition, we will obtain values for f(t) at the zeros $\{t_k\}_1^n$ of the n-th degree Laguerre polynomial. We could then use some kind of interpolation to give f(t) everywhere. Notice that A is totally positive (see Gantmacher and Krein [10, page 89]), so A^TA and AA^T are also, and thus the basis vectors $\{u_i\}_1^m$ over which we take our solution have the oscillation property mentioned previously.

The abscissas and weights are most easily obtained from the eigensystem of the tridiagonal matrix derived from the three-term recurrence relation for the Laguerre polynomials. (See Golub and Welsch [14] for a general discussion of this.) If

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix} \tag{4.3}$$

with $\alpha_i = 2i-1$, $\beta_i = -i$, then the abscissas $\{t_k\}$ are the eigenvalues of T and if the corresponding eigenvectors are $\left\{q^{(k)}\right\}_2$, normalized so $\|q^{(k)}\|_2 = 1$, then $w_k = \left(q_1^{(k)}\right)^2$.

The choice of the sample points $\{s_i\}$ is an interesting problem. They should be chosen so the points $g(s_i)$ adequately represent the function $g(s)$ and so that the expansion of g in the basis $\{u_i\}$, $\sum_1^n \beta_i u_i$, has the $\{\beta_i\}$ decreasing rapidly relative to the singular values of $\{\sigma_i\}$. This seems to be a difficult problem in general. Another factor here is that the weights $\{w_k\}$ decrease rapidly in size (so $w_k e^{t_k}$ is always of moderate size) and even for $n = 20$, $w_n \tilde{=} 10^{-30}$. If these weights are found as described above, the eigenvalue program cannot be expected to find these small w_k to high relative accuracy, and this affects the choice of the $\{s_i\}$. In particular, the $\{s_i\}$ should all be positive so the effect of the inaccurate w_k for large k is not felt. This may be a rather severe restriction and shows why other means of calculating the $\{w_k\}$ may be preferred in some cases, so the products $\left\{ w_k e^{t_k} \right\}$ have high relative accuracy.

Notice also that with little extra effort, we can put in more points $\{s_i\}$ say $m > n$ giving an overdetermined linear system in (4.2). This can be solved in exactly the same manner using the singular value decomposition.

The method was tried on the simple example

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

Using Gauss-Laguerre integration on 10 points, the best results (in the sense of least absolute error) were obtained using 10 equally spaced points $\{s_i\}$ in $(0, 2)$ for which the maximum function error was 1.10^{-3}, neglecting σ_i below 3.10^{-3} (solving 6 equations). Using a 20-point integration scheme, best results were obtained for 20 points equally spaced in $(0, 5)$; maximum error was 5.10^{-4} neglecting σ_i below 5.10^{-3} (solving 6 equations). Using more data points and solving the overdetermined system did not improve the results. It seemed only a function of the s-interval used. The predicted optimal k from (2.11) was also $k = 6$.

5. Application III - The Backwards Heat Equation

Again, this is a well-known ill-posed problem, which has been considered by many people. We will consider only the inverse Cauchy problem on $-\infty < x < \infty$ (no boundaries). Then we can again represent the problem as an integral equation of the first kind:

$$u(x, t) = \int_{-\infty}^{\infty} \frac{e^{-(x-s)^2/4(t-\tau)}}{\sqrt{4\pi(t-\tau)}} \, u(s, \tau)ds \quad . \qquad (5.1)$$

Of course even if $u(s, \tau)$ is discontinuous, the solution for $t > \tau$ is analytic; thus for the backwards problem we have the additional problem of deciding how far back in time we can solve.

In a very revealing paper, John [16] gives a way of doing this. Suppose our time scale is such that $t = 0$ is the final time, let $u(x, 0) = g(x)$, and define for $0 < a < T$,

$$N_u(a) = \sup_{-\infty < x < \infty} |u(x, -a)|$$

$$M_g(a) = \sup_{\substack{-\infty < x < \infty \\ -\infty < y < \infty}} e^{-y^2/4a} |g(x+iy)| \quad .$$

These are both monotone nondecreasing in a, $Mg(a) \leq N_u(a)$, and from the inversion formula $u(x,-t) = \int_{-\infty}^{\infty} \frac{e^{-s^2/4t}}{\sqrt{4\pi t}} g(x+is)ds$, we get

$$N_u(a) \leq \inf_{a<b<T} M_g(b) \sqrt{\frac{b}{b-a}} \quad .$$

Thus $u(x, -t)$ exists and is finite for $0 \leq t < T$ if and only if $M_g(a)$ is finite for $0 \leq a < T$. So $g(z)$ must be an entire function of order ≤ 2 and if of order 2, of type $\left(\frac{1}{4a}\right)$ for the solution to exist back to $-a$. Thus a well-posed version of the problem is assured if we let the final data $\tilde{g}$ vary from g so that $M_{\tilde{g}-g}(a)$ remains bounded.

For the numerical solution, John uses a high-order accurate difference scheme, making one giant step backwards:

$$u(x, -t) = \sum_{j=-m}^{m} c_j^m(t)\, g(x + jh) \qquad (5.2)$$

where h is to be chosen. The coefficients can easily be generated by solving a Vandermonde system and the method works quite well. It is not clear how best to choose h, although John gives some error estimates.

We propose to solve the integral equation (5.1) directly using the singular value decomposition. If we use the trapezoidal rule with large h (which is accurate for this kind of integrand–see Davis and Rabinowitz [7, page 92]), we obtain $Ax = b$ with A a symmetric Toeplitz matrix:

$$a_{ij} = \frac{h}{\sqrt{4\pi(t-\tau)}} \qquad \alpha^{(i-j)^2} \equiv \alpha_{i-j}$$

with $\alpha = e^{-h^2/4(t-\tau)}$. We can obtain estimates of the rate of decrease of the singular values $\sigma_i^{(n)} = \lambda_i^{(n)}$ using the results of Kac, Murdoch, and Szegö [17] (extended by Parter [18]), i.e. for fixed ν

$$\lambda_\nu^{(n)} = m + \frac{c\pi^2\nu^2}{n^2} + 0\left(\frac{1}{n^2}\right)$$

where $p(\theta) = \sum_{-n}^{n} \alpha_k e^{ik\theta}$ is the associated trigonometric polynomial, $p(\theta_0) = m = \min p(\theta)$, $c = \frac{1}{2} f''(\theta_0)$. Also note that $p(\theta)$ is a theta function:

$$p(\theta) = \frac{h}{\sqrt{4\pi(t-\tau)}} \left(1 + 2\sum_{1}^{n} a^{j^2} \cos j\theta\right) = \theta_3(\theta/2;a).$$

However the numerical results using this scheme were poor; instead we approximate (5.1) using Gauss-Hermite quadrature

(as for the Laplace transform problem, this is only appropriate if the solution is non-periodic). This gives at any set of data points $\{s_i\}$

$$u(s_i, t) \simeq \sum_{k=1}^{n} \frac{w_k \, e^{\left(x_k^2 - (s_i - x_k)^2/4(t-\tau) \right)}}{\sqrt{4\pi(t-\tau)}} \; u(x_k, \tau) \qquad (5.3)$$

or $g = Af$. Here the $\{x_k\}$ are the zeros of the n-th degree Hermite polynomial which, together with the weights $\{w_k\}$, can be found from the eigensystem of (4.3), now with $\alpha_i = 0$, $\beta_i = \sqrt{i/2}$. The matrix A is totally positive so the singular vectors $\{u_i\}$ have the oscillation property, and again there is the interesting question of how to choose the data points $\{s_i\}$. There is the same problem of inaccurate matrix entries due to low relative accuracy in $\left\{ w_k e^{x_k^2} \right\}$ for large k if any data points s_i are very large in modulus. For $|s_i| \geq 3$ or so, constant functions won't be integrated closely; however if the solution decays fairly quickly this will not cause problems.

The numerical example given is that of Cannon [5] who used linear programming methods to solve the heat equation backwards. The example consists of two delta function heat sources at $t = -\tau$:

$$u(x, t+\tau) = \frac{5}{\sqrt{\pi (t+\tau)}} \left(e^{-\frac{(x+0.5)^2}{4(t+\tau)}} + e^{-\frac{(x-0.5)^2}{4(t+\tau)}} \right) .$$

The singular value decomposition was used with 20 point Gauss-Hermite quadrature. For $t = \tau = 0.5$, a maximum error of .0008 was obtained for 20 points equally spaced in (-1, 1), neglecting singular values below .003 (solving 5 equations). This compares favourably with Cannon's results; the best results obtained from John's method (see equation (5.2)) were for $m = 9$, $h = 0.75$, and gave a maximum error of .004 . For $t = \tau = 0.1$, best results were obtained for 20 equally spaced points in (-2.5, 2.5); a maximum error of .027 was obtained when the singular values below .03 were neglected (12 equations solved). Again this is roughly the accuracy obtained by Cannon; however John's method with $m= 9$, $h = 0.34$ gave a maximum error of .0024. As in the case of the Laplace transform, no improvement in the results occurred when more data points were used and the overdetermined system solved. In both cases, the predicted optimal k from (2.11) was the same.

REFERENCES

1. R. Bellman, R. Kalaba, and J. Lockett, Dynamic programming and ill-conditioned linear systems. J. Math. Anal. Appl. 10(1965), 206-215.

2. R. Bellman, R. Kalaba, and J. Lockett, Numerical Inversion of the Laplace Transform. Elsevier Press, New York, 1967.

3. A. Björck, Iterative refinement of linear least squares solutions II. BIT 8(1968), 8-30.

4. P. A. Businger and G. H. Golub, Singular value decomposition of a complex matrix. Algorithm 358, Comm. A. C. M. 12(1969), 564-565.

5. J. R. Cannon, Some numerical results for the solution of the heat equation backwards in time. Numerical Solutions of Nonlinear Differential Equations (D. Greenspan ed.), John Wiley & Sons, New York, 1966.

6. R. Courant and D. Hilbert, Methods of Mathematical Physics. Interscience, New York, 1953.

7. P. Davis and P. Rabinowitz, Numerical Integration. Blaisdell, Waltham, Mass., 1967.

8. J. N. Franklin, Well-posed stochastic extensions of ill-posed linear problems. J. Math. Anal. Appl. 31(1970), 682-716.

9. F. R. Gantmacher, The Theory of Matrices, vol. II, Chelsea, New York, 1964.

10. F. R. Gantmacher and M. G. Krein, Oscillating Matrices. Akademie-Verlag, Berlin, 1963.

11. G. H. Golub, Numerical methods for solving linear least squares problems. Numer. Math. 7(1965), 206-216.

12. G. Golub and W. Kahan, Calculating the singular values and pseudoinverse of a matrix. SIAM J. Numer. Anal. 2(1965), 205-224.

13. G. H. Golub and C. Reinsch, Singular value decomposition and least squares solutions. Numer. Math. 14(1970), 403-420.

14. G. H. Golub and J. H. Welsch, Calculation of Gauss quadrature rules. Math. Comp. 23(1969), 221-230.

15. R. J. Hanson, A numerical method for solving Fredholm integral equations of the first kind using singular values. SIAM J. Numer. Anal., to appear.

16. F. John, Numerical solution of the equation of heat conduction for preceding times. Annali di Matematica (1955), 129-142.

17. M. Kac, W. L. Murdoch, and G. Szegö, On the eigenvalues of certain Hermitian forms. J. Rat. Mech. Anal. 2(1953), 767-800.

18. S. Parter, On the extreme eigenvalues of Toeplitz matrices. Trans. AMS 100(1961), 263-276.

19. G. Peters and J. H. Wilkinson, The least squares problem and pseudoinverses. Comp. J. 13(1970), 309-316.

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

21. A. N. Tihonov, Solution of nonlinear integral equations of the first kind. Soviet Math. Dokl. 5(1964), 835.

22. J. H. Wilkinson, Rounding Errors in Algebraic Processes. Prentice-Hall, New York, 1963.

23. J. H. Wilkinson, The Algebraic Eigenvalue Problem. Clarendon Press, Oxford, 1965.

24. G. E. Forsythe and C. B. Moler, Computer Solution of Linear Algebraic Systems. Prentice-Hall, New York, 1967.