SOLVING ILL-CONDITIONED AND SINGULAR LINEAR SYSTEMS: A TUTORIAL ON REGULARIZATION

ARNOLD NEUMAIER *

Abstract. It is shown that the basic regularization procedures for finding meaningful approximate solutions of ill-conditioned or singular linear systems can be phrased and analyzed in terms of classical linear algebra that can be taught in any numerical analysis course. Apart from rewriting many known results in a more elegant form, we also derive a new two-parameter family of merit functions for the determination of the regularization parameter. The traditional merit functions from generalized cross validation (GCV) and generalized maximum likelihood (GML) are recovered as special cases.

Key words. regularization, ill-posed, ill-conditioned, generalized cross validation, generalized maximum likelihood, Tikhonov regularization, error bounds

AMS subject classifications. primary 65F05; secondary 65J20

1. Introduction. In many applications of linear algebra, the need arises to find a good approximation $\hat{x}$ to a vector $x \in \mathbb{R}^n$ satisfying an approximate equation $Ax \approx y$ with ill-conditioned or singular $A \in \mathbb{R}^{m\times n}$, given $y \in \mathbb{R}^m$.

Usually, y is the result of measurements contaminated by small errors (noise). A may be, e.g., a discrete approximation to a compact integral operator with unbounded inverse, the operator relating tomography measurements y to the underlying image x, or a matrix of basis function values at given points relating a vector y of approximate function values to the coefficients of an unknown linear combination of basis functions. Frequently, ill-conditioned or singular systems also arise in the iterative solution of nonlinear systems or optimization problems.

The importance of the problem can be seen from a glance at the following, probably incomplete list of applications: numerical differentiation of noisy data, nonparametric smoothing of curves and surfaces defined by scattered data, image reconstruction, deconvolution of sequences and images (Wiener filtering), shape from shading, computer-assisted tomography (CAT, PET), indirect measurements and nondestructive testing, multivariate approximation by radial basis functions, training of neural networks, inverse scattering, seismic analysis, parameter identification in dynamical systems, analytic continuation, inverse Laplace transforms, calculation of relaxation spectra, air pollution source detection, solution of partial differential equations with nonstandard data (backward heat equation, Cauchy problem for parabolic equations, equations of mixed type, multiphase flow of chemicals, etc.), ... The surveys by Engl [8] and the book by Groetsch [12] contain many pertinent references.

In all such situations, the vector $\tilde{x} = A^{-1}y$ (or in the full rank overdetermined case A^+y, with the pseudo inverse $A^+ = (A^*A)^{-1}A^*$), if it exists at all, is usually a meaningless bad approximation to x. (This can be seen from an analysis in terms of the singular value decomposition; see Section 6.) Moreover, even when some vector $\hat{x}$ is a reasonable approximation to a vector x with $Ax = y$, the usual error estimates $\|x - \hat{x}\| \le \|A^{-1}\|\ \|A\hat{x} - y\|$ in the square case or $\|x - \hat{x}\| \le \|A^+\|\ \|A\hat{x} - y\|$ in the overdetermined case are ridiculously pessimistic.

So-called *regularization techniques* are needed to obtain meaningful solution estimates for such *ill-posed* problems, where some parameters are ill-determined by least

*Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria.
email: neum@cma.univie.ac.at WWW: http://solon.cma.univie.ac.at/~neum/

squares methods, and in particular when the number of parameters is larger than the number of available measurements, so that standard least squares techniques break down.

A typical situation (but by no means the only one) is that most parameters in the state vector are function values of a function at many points of a suitable grid (or coefficients in another discretization of a function). Refining a coarse grid increases the number of parameters, and when a finite amount of (grid-independent) data are given, one cannot use a very fine grid with standard least squares, which requires $m > n$. Of course, one expects the additional parameters introduced by a refinement of the grid to be closely related to the parameters on the coarse grid, since the underlying function is expected to have some regularity properties such as continuity, differentiability, etc.. To get sensible parameter estimates in such a case it is necessary to be able to use this additional qualitative information. (The underlying continuous problem is often an integral equation of the first kind; see, e.g., WING & ZAHRT [37] for an easy-to-read introduction.)

Though frequently needed in applications, the adequate handling of such ill-posed linear problems is hardly ever touched upon in numerical analysis text books. Only in GOLUB & VAN LOAN [11], the topic is briefly discussed under the heading *ridge regression*, the statisticians' name for Tikhonov regularization; and a book by BJÖRCK [4] on least squares problems has a section on regularization.

The main reason for this lack of covering seems to be that the discussion of regularization techniques in the literature (TIKHONOV [31], ENGL et al. [9], HANKE [14], WAHBA [34]) is usually phrased in terms of functional analytic language, geared towards infinite-dimensional problems. (A notable exception is the work by Hansen (see HANSEN [17], HANKE & HANSEN [15], and the recent book HANSEN [18]) that is closer to the spirit of the present paper.) This tends to make the treatments unduly application specific and clutters the simplicity of the arguments with irrelevant details and distracting notation. We summarize the functional analytic approach in Section 2, mainly to give those familiar with the tradition a guide for recognizing what happens in the rest of the paper.

However, those unfamiliar with functional analysis may simply skip Section 2, since the main purpose of the present paper is to show that *regularization can be discussed using elementary but elegant linear algebra,* accessible to numerical analysis students at any level. The linear algebra setting is also much closer to the way in which regularization methods are implemented in practice.

Most results derived in this paper (the major exception is the theory in Section 10 leading to a new family of merit functions) are known in a more or less similar form for problems in function spaces. What is new, however, is the derivation from assumptions that make sense in a finite-dimensional setting, and with proofs based on standard linear algebra.

Section 3 motivates an abstract and general approach to smoothness conditions of the form $x = Sw$ with a vector w of reasonable norm and a suitable smoothness matrix S derivable from the coefficient matrix and some assumed order p of differentiability. Section 4 derives and discusses the basic Theorem 4.1, giving deterministic error bounds that show that regularization is possible. It is also shown that general ill-posed problems behave in a way completely analogous to perhaps the simplest ill-posed problem, numerical differentiation, for which the cure has been understood for a very long time.

Section 5 specializes Theorem 4.1 to obtain some specific regularization tech-

niques. In particular, good approximate inverses for regularization can be derived by modifying the standard least squares formula. The traditional Tikhonov regularization by means of

$$\hat{x} = (A^*A + h^2 I)^{-1} A^* y$$

and an iterated version of it are covered by the basic theorem. Section 6 then invokes the singular value decomposition to compute more flexible approximate inverses, including a familiar one based on the truncated singular value decomposition.

In Section 7 we discuss an inherent limitation of regularization techniques based on deterministic models – there cannot be reliable ways to determine regularization parameters (such as h^2 and p) in the absence of information about the size of the residuals and the degree of smoothness. However, the latter situation is very frequent in practice, and finding valid choices for the regularization parameter is still one of the current frontiers of research.

The remainder of the paper therefore discusses a stochastic framework in which it is possible to rigorously study techniques for the selection of the optimal regularization parameter in the absence of prior information about the error level. We first prove (in Sections 8 and 9) stochastic results that parallel those obtained for the deterministic case, with some significant differences. In particular, Section 8 shows that the expected squared norm of the residual can be minimized explicitly (a result due to Bertero et al. [3]), and leads in the simplest case again to Tikhonov regularization.

Section 9 expresses the optimal estimator in a more general situation in terms of the singular value decomposition and discusses the attainable limit accuracy. The resulting formulas are similar to those arising in deconvolution of sequences and images by *Wiener filters* (Wiener [36]), perhaps the earliest practical use of regularization techniques. A modern treatment from a practical point of view can be found, e.g., in Katsaggelos [19].

In the stochastic approach, the regularization parameter turns out to reappear as a variance quotient, and this permits its estimation through variance component estimation techniques. In Section 10, which is less elementary than the rest of the paper, we derive a family of merit functions whose minimizers give approximations to the ideal regularization parameter; the merit functions contain the generalized cross validation approach and the generalized maximum likelihood approach as special cases.

Finally, Section 11 extends the stochastic approach to the situation where the smoothness condition $x = Sw$ is replaced by the condition that some vector Jx, usually composed of suitably weighted finite differences of function values, is reasonably bounded. This is again a smoothness condition, but by judicious choice of J, the smoothness requirements can be better adapted to particular problems.

2. Regularization in function spaces. In this section (only), we assume the reader to be familiar with concepts from functional analysis; however, for those unfamiliar with these concepts, there is no harm in skipping the section, since the remainder of the paper is completely independent of it.

We shall only give a short sketch of the traditional theory; for more detailed references, in depth treatments and applications we refer to the surveys mentioned in the introduction. In particular, we shall use in this section the notation of the book by Engl, Hanke & Neubauer [9].

The objective (e.g., in computing the inverse Radon transform in tomography) is to solve a linear operator equation of the form $Tx = y$, where $T : \mathcal{X} \to \mathcal{Y}$ is a compact

linear operator between two Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, $y \in \mathcal{Y}$ is given, and $x \in \mathcal{X}$ is wanted. Typically, $\mathcal{X}$ and $\mathcal{Y}$ are closed subsets of Sobolev spaces of functions, and T is an integral operator. In numerical applications, everything needs to be discretized and becomes finite-dimensional. Thus x and y are replaced by discrete versions consisting of function values or coefficients of expansions into basis functions, and T becomes a matrix A with some structure inherited from the continuous problem.

The choice of Hilbert spaces amounts to fixing norms in which to measure x and y, and has in the discretized version an analogue in the choice of suitable scales for the norms to be used to assess accuracies. (In $\mathbb{R}^n$, all norms are equivalent, but for numerical purposes, correct scaling may make a crucial difference in the magnitude of the norms.)

In the function space setting, useful error estimates can be obtained only if x satisfies some smoothness restrictions. These restrictions take two main forms.

In the first form (due to Tikhonov [31], cf. [9], Chapter 3), one assumes that x is in the range $\mathcal{R}(B)$ of a smoothing operator B, a product of p alternating factors T and its adjoint T^*. For problems involving univariate functions, p can be roughly interpreted as the number of weak derivatives bounded in the $\mathcal{L}^2$ norm. To obtain error bounds, one has to assume the knowledge of a bound ω on the norm of some w with $x = Bw$. The number ω can be interpreted as a rough measure of the magnitude of the pth derivative of x.

The choice of the smoothness operator B (in the above setting uniquely determined by p) is an a priori modeling decision, and amounts to selecting the smoothness of the reconstructed solution. Note that even for analytical problems, it is not always appropriate to choose p large, since ω often grows exponentially with p; consider, e.g., $x(t) = \sin(\omega t)$ for large ω. Another a priori modeling decision is the selection of an appropriate level δ for the accuracy of the approximation y to Tx.

The results obtained about particular regularization methods are all of the form that if p is not too large and $\|Tx-y\| \le \delta$ then the regularized solution x_δ has an error $\|x_\delta - x\| = O(\delta^{p/(p+1)}\omega^{1/(p+1)})$. The reduced exponent $p/(p+1) < 1$ reflects the fact that the inverse of a compact operator is unbounded, resulting in the ill-posedness of the problem. Standard results (cf. Proposition 3.15 of [9]) also imply that this is the best exponent of δ achievable.

The construction of an approximation satisfying this error bound depends on the knowledge of p and δ or another constant involving δ such as δ/ω; by a theorem of Bakushinskii [1] (reproved in [9] as Theorem 3.3), any technique for choosing regularization parameters in the absence of information about the error level can be defeated by suitably constructed counterexamples whenever the pseudo inverse of T is unbounded.

However, in practice, there is often not enough information available that provide adequate values for δ and p, which limits the deterministic approach. Some results in a stochastic functional analytic setting are given in the context of smoothing splines by Wahba [34] (Chapter 4.5 and references there).

The second form of the smoothness restrictions (due to Phillips, cf. [28], Natterer [25] or [9], Chapter 8) is given by the assumption that, for some differential operator L and some integer $k > 0$, the $\mathcal{L}^2$ norm of $L^k x$ is finite. The theory gives results and limitations similar to that for the other smoothness restriction.

In the discretized version, the second form of the smoothness restrictions is usually modeled by imposing the condition that the norm of some vector Jx composed of suitable finite differences of the solution vector x is assumed to have moderate size

only; cf. Section 11.

3. Modeling smoothness. At least in the rank-deficient case where the rank of A is smaller than the dimension of x, it is clear that some additional information is needed to find a satisfactory solution of $Ax = y$, since infinitely many solutions exist if there is one at all. From a numerical point of view, ill-conditioned systems behave just like singular ones, and additional information is needed, too. In the applications, the correct one among all (near) solutions is characterized by additional properties, most commonly by requiring the "smoothness" of some function, curve or surface constructed from x.

This qualitative knowledge can be modeled as follows. Given a vector w representing a continuous function f, for example as a list of function values $w_i = f(t_i)$ at the points t_i of some grid, we may apply some form of numerical integration to w to arrive at an (approximate) representation w^1 of a differentiable function f^1 with derivative f; and repeating this p times, we get a representation w^p of a p times differentiable function. Algebraically, we have $w^p = Sw$ for a suitable coefficient matrix S, the *smoothing matrix*. Therefore, we can formulate smoothness of x by requiring that x can be represented in the form $x = Sw$ with some vector w of reasonable norm. The smoothing matrix S characterizes the additional information assumed about x. (More flexible alternative smoothness conditions are discussed in Section 11.)

In practice, S might be a discretized integral operator augmented by boundary conditions. The most convenient way, however, constructs S directly from A and thus works without any detailed knowledge about smoothness requirements.

To motivate the recipe we rewrite the numerical differentiation of a $p+1$ times continuously differentiable function $y : [0,1] \to \mathbb{R}$ as an (infinite-dimensional) regularization problem. The situation is so simple that no functional analysis is needed, and we simplify further by also assuming that y and its first $p+1$ derivatives vanish at $t = 0$ and $t = 1$. Finding the derivative $x(t) = y'(t) = \nabla y(t)$ can be posed as the problem of solving the linear system $Ax = y$ where A is the integral operator defined by

$$Ax(t) := \int_0^t x(\tau)d\tau.$$

The differentiability assumption says that $w = \nabla^{p+1} y = \nabla^p x$ is continuous and bounded, and since $A = \nabla^{-1}$, we may write this in the form $x = A^p w$.

To rewrite this in a form capable of generalization, we note that the adjoint integral operator is defined by the formula $(A^* x_1, x_2) = (x_1, Ax_2)$ in terms of the inner product $(x_1, x_2) = \int_0^1 x_1(t)x_2(t)dt$. With $y_i = Ax_i$, we have $(x_1, Ax_2) = (y_1', y_2) = -(y_1, y_2') = (-Ax_1, x_2)$ by partial integration and our boundary conditions, so that we conclude $A^* = -A$. Up to a sign that may be absorbed into w, we can therefore write the condition $x = A^p w$ as $x = Sw$, where

$$S = \begin{cases} (A^*A)^{p/2} & \text{if } p \text{ is even,} \\ (A^*A)^{(p-1)/2}A^* & \text{if } p \text{ is odd} \end{cases} \tag{1}$$

is a product of $p \geq 1$ alternating factors A^* and A.

This simple and powerful general way of choosing the smoothing matrix S where, in general, A^* is the transposed matrix (the conjugate transposed in the complex case and the adjoint operator in the infinite-dimensional situation), works much more generally and is sufficient for many practical problems.

Indeed, assume that (as in most applications) the operator A is smoothing, in the sense that if $x \in \mathbb{R}^n$ is a discretized version of a k times differentiable function f then $y = Ax \in \mathbb{R}^m$ is a discretized version of a $k+1$ times differentiable function $\bar{f}$. The operator A^* is generally smoothing, too. Thus, if $y \in \mathbb{R}^m$ is a discretized version of a k times differentiable function g, then $A^*y \in \mathbb{R}^n$ is a discretized version of a $k+1$ times differentiable function $\bar{g}$. (Typically, A is related to some integral operator, and A^* to its adjoint integral operator.)

Thus we get smoother and smoother functions by repeating applications of A and A^* to some vector w corresponding to a bounded continuous function only. This suggests that we require that a smooth x is represented as $x = Sw$ with a vector w of reasonable norm. (Since we need $x \in \mathbb{R}^n$, the final matrix applied must be A^*, which is the case for the choice (1).)

Thus, by requiring that x can be represented in the form $x = Sw$ with a smoothing matrix S given by (1), qualitative knowledge on smoothness can be modeled. Note that (1) implies

$$SS^* = (A^*A)^p. \tag{2}$$

While $x = Sw$ with (1) is a frequently used assumption, a discussion of how to find a suitable value of p is virtually missing in the literature. We shall return to this point in Section 10.

4. The error estimate. The key to the treatment of *ill-posed linear systems*, a term we shall use for all linear systems where the standard techniques are inadequate, is a process called *regularization* that replaces A^{-1} by a family C_h ($h > 0$) of approximate inverses of A in such a way that, as $h \to 0$, the product C_hA converges to I in an appropriately restricted sense. The parameter h is called the *regularization parameter*. (More generally, any parameter entering the definition of C_h may be referred to as a regularization parameter.)

In order that the C_h may approximate ill-conditioned (or, in the infinite-dimensional case, even unbounded) inverses we allow that their norm grows towards infinity as $h \to 0$. It is usually possible to choose the C_h such that, for a suitable exponent p (often $p = 1$ or 2), the constants

$$\gamma_1 = \sup_{h>0} h\|C_h\|$$

and

$$\gamma_2 = \sup_{h>0} h^{-p}\|(I - C_hA)S\|$$

are finite and of reasonable size. Then

$$\|C_h\| \le \frac{\gamma_1}{h}, \quad \|(I - C_hA)S\| \le \gamma_2 h^p. \tag{3}$$

The feasibility of regularization techniques is a consequence of the following fundamental deterministic error bounds. The result is valid for *arbitrary* norms, though we shall apply it later only for the Euclidean norm. (However, cf. (18) below, implicit scaling may be needed to adjust the data to the Euclidean norm.)

THEOREM 4.1. *Suppose that*

$$x = Sw, \quad \|Ax - y\| \le \Delta\|w\| \tag{4}$$

for some $\Delta > 0$. *Then* (3) *implies*

$$\|x - C_h y\| \le \left(\gamma_1 \frac{\Delta}{h} + \gamma_2 h^p\right) \|w\|. \tag{5}$$

Proof. We have

$$\begin{aligned} \|x - C_h y\| &= \|(I - C_h A)x + C_h(Ax - y)\| \\ &\le \|(I - C_h A)x\| + \|C_h\| \, \|Ax - y\| \\ &\le \|(I - C_h A)Sw\| + \|C_h\|\Delta\|w\| \\ &\le \|(I - C_h A)S\| \, \|w\| + \|C_h\|\Delta\|w\| \\ &\le \gamma_2 h^p \|w\| + \frac{\gamma_1}{h}\Delta\|w\| \end{aligned}$$

by (4) and (3). This implies (5).□

Note that the two assumptions (3) and (4) have a very different status. (4) is an assumption about the problem to be solved, while (3) is a requirement on C_h that we may satisfy by choosing the latter in a suitable way. If (4) is violated, nothing can be said about the problem, while a violation of (3) only implies that the particular algorithm for choosing C_h is unsuitable for solving the problem.

(The proof shows that one may allow in (4) the more general relation $x = Sw + u$ with arbitrary u satisfying $(I - C_h A)u = 0$, without affecting the conclusion. In some applications, this allows a more flexible treatment of boundary conditions.)

Traditionally, one assumes a residual bound $\|Ax - y\| \le \epsilon$; our form of the bound in (4) is motivated by the fact that the optimal h depends on ϵ and w only through the quotient $\Delta = \epsilon/\|w\|$. Indeed, the bound (5) is smallest when the derivative of the right hand side vanishes, giving

$$\hat{h} = \left(\frac{\gamma_1 \Delta}{\gamma_2 p}\right)^{1/(p+1)} = O(\Delta^{\frac{1}{p+1}}). \tag{6}$$

For this choice, $\gamma_1 \Delta = \gamma_2 p \hat{h}^{p+1}$, and we find an error bound

$$\|x - C_{\hat{h}} y\| \le \gamma_2 (p+1) \hat{h}^p = O(\Delta^{\frac{p}{p+1}}), \tag{7}$$

with reasonable factors hidden in the Landau symbol $O(\cdot)$. Here and later, the Landau symbols refer to the behavior for tiny $\Delta > 0$. However, all nonasymptotic results in this paper are valid for any value of Δ. This is important since regularization is often used when the available data (or the assumed model) are of low accuracy only.

For a well-posed data fitting problem, i.e., one with a well-conditioned normal equation matrix $A^T A$, the least squares estimate has an error of the order of Δ. This follows from (5); indeed, $C_h = A^+$ satisfies (5) for $h^{-1} = \|A^+\| = O(1)$ with $\gamma_1 = 1, \gamma_2 = 0$ independent of p. (One can therefore allow $p \to \infty$ without blowing up the bound, so that the right exponent 1 is approached in (7).)

For an ill-posed problem, the reduced exponent $\frac{p}{p+1} < 1$ in (7) (that can be shown to be optimal under the assumptions made, cf. Proposition 3.15 of [9]) implies a loss of precision due to the ill-posed nature of the problem. Roughly speaking, it means that the number of correct digits that can be expected in an approximate solution is only a fraction of $\frac{p}{p+1}$ of those one would expect in a well-posed problem. In particular, as familiar from numerical differentiation, the accuracy of the approximation for $p = 1$

is only roughly half the number of digits to which the data are accurate. As one can see from (5) and the proof, the error is bounded as a sum of two terms. The first term, divergent as $h \to 0$, is proportional to the error level of the data vector y, and the second term, divergent as $h \to \infty$, reflects the error in the approximative inverse. Thus, the situation is completely analogous to that encountered in numerical differentiation. There (a thorough discussion is in Chapter 8.6 of GILL et al. [10]) $f'(x)$ is approximated using inaccurate function values $\tilde{f}(x+\tau h)$ with absolute errors bounded by Δ. In approximation formulas like

$$\left| f'(x) - \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} \right| \leq \frac{2\Delta}{h} + \frac{h}{2}\|f''\|_\infty$$

for forward differences or

$$\left| f'(x) - \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{h} \right| \leq \frac{2\Delta}{h} + \frac{h^2}{6}\|f''\|_\infty$$

for central differences, two terms with opposite behavior for $h \to 0$ and $h \to \infty$ arise. As in (6)–(7), an optimal choice of $h = O(\Delta^{1/2})$ in the first case and $h = O(\Delta^{1/3})$ in the second case gives the approximation orders $O(\Delta^{1/2})$ for the error of forward differences and $O(\Delta^{2/3})$ for the error of central differences.

In order to evaluate (6) for a specific problem one needs to know γ_1, γ_2, p and Δ. Usually, p is assumed to be fixed, typically at $p = 1$ or $p = 2$; then γ_1, γ_2 can be determined from the formula used to define C_h; cf. Section 5. On the other hand, in practice, one hardly knows precise values for Δ. If we choose instead of the optimal value (6) some value

$$h = \gamma_0 \delta^{1/(p+1)}$$

with a guess δ for Δ and some positive constant γ_0, we find

$$\|x - C_h y\| \leq \text{const.} \max(\Delta, \delta)\delta^{-1/(p+1)}.$$

This immediately shows that, with the same relative error in δ, a choice $\delta > \Delta$ has less influence on the size of the bound than a choice $\delta < \Delta$, at least when $p > 1$. Thus if in doubt, δ should be taken conservatively large. (In the context of solving nonlinear systems or optimization problems, the overestimation is usually made good through very few additional iterations, unless δ is chosen excessively large, while a too small choice of δ may be disastrous.)

Note also that (6) implies that, generally, problems with larger p are solved with a larger value of h, at least as long as $\gamma_1\Delta/\gamma_2 p$ remains much smaller than 1.

5. Choosing approximate inverses. For a well-conditioned approximation problem $Ax \approx y$, the residual $\|A\hat{x} - y\|_2$ becomes smallest for the choice

$$\hat{x} = (A^*A)^{-1}A^*y. \tag{8}$$

When A is rank deficient or becomes increasingly ill-conditioned, this choice is impossible or increasingly useless. We may improve the condition of A^*A by modifying it. The simplest way to achieve this is by adding a small multiple of the identity. Since A^*A is symmetric and positive semidefinite, the matrix $A^*A + h^2 I$ has its eigenvalues

in $[h^2, h^2 + \|A\|^2]$ and hence a condition number $\leq (h^2 + \|A\|^2)/h^2$ that becomes smaller as h increases. With this replacement, the formula (8) turns into

$$\hat{x} = (A^*A + h^2 I)^{-1} A^* y, \tag{9}$$

a formula first derived by TIKHONOV [31] in 1963. He derived it by solving the modified least square problem

$$\|Ax - y\|^2 + h^2 \|x\|^2 = \min!$$

Formula (9) corresponds to the family of approximate inverses defined by

$$C_h = (A^*A + h^2 I)^{-1} A^*. \tag{10}$$

More generally, we may consider using other matrix functions of A^*A as approximate inverses. Indeed, arbitrary functions φ defined on $\mathbb{R}_+$ can be extended to functions of $T = h^{-2}A^*A$. For example, if φ is continuous, we may approximate it by a sequence of polynomials φ_l with $\varphi(t) = \lim_{l\to\infty} \varphi_l(t)$ uniformly for $|t| \leq \|T\|$, we have $\varphi(T) = \lim_{l\to\infty} \varphi_l(T)$.

THEOREM 5.1. *For any smoothing matrix S satisfying* (2)*, the bounds* (3) *hold in the* 2*-norm for the family of approximate inverses*

$$C_h = h^{-2}\varphi(h^{-2}A^*A)A^* \tag{11}$$

for any function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that the constants

$$\gamma_1 = \sup_{t\geq 0} |\varphi(t)t^{1/2}|, \quad \gamma_2 = \sup_{t\geq 0} |1 - \varphi(t)t|t^{p/2}$$

are finite.

Proof. We use the fact that the 2-norm of a matrix B is given by the square root of the supremum of the eigenvalues of the matrix BB^*. If we denote the eigenvalues of $T = h^{-2}A^*A$ by t_i then the t_i are nonnegative. Now the eigenvalues of

$$C_h C_h^* = h^{-4}\varphi(T)A^*A\varphi(T) = h^{-2}\varphi(T)^2 T \tag{12}$$

are $h^{-2}\varphi(t_i)^2 t_i \leq h^{-2}\gamma_1^2$, whence $\|C_h\|_2 \leq h^{-1}\gamma_1$. Similarly, since

$$R := I - C_h A = I - h^{-2}\varphi(T)A^*A = I - \varphi(T)T \tag{13}$$

and $SS^* = (A^*A)^p = (h^2T)^p$, the eigenvalues of

$$(RS)(RS)^* = RSS^*R = (I - \varphi(T)T)^2(h^2T)^p \tag{14}$$

are $(1 - \varphi(t_i)t_i)^2(h^2t_i)^p \leq h^{2p}\gamma_2^2$, whence $\|(I - C_hA)S\|_2 = \|RS\|_2 \leq h^p\gamma_2$. □

In particular, Tikhonov regularization (9) is obtained by choosing

$$\varphi(t) = \frac{1}{t+1}, \tag{15}$$

and for this choice we find

$$\gamma_1 = \sup_{t\geq 0} \frac{t^{1/2}}{t+1} = \frac{1}{2},$$

$$\gamma_2 = \sup_{t\geq 0} \frac{t^{p/2}}{t+1} = \begin{cases} 1/2 & \text{if } p = 1, \\ 1 & \text{if } p = 2, \\ \infty & \text{if } p > 2. \end{cases}$$

Thus Tikhonov's solution formula (9) can exploit smoothness only up to a degree $p \leq 2$, since smoother solutions have to be treated by taking $p = 2$. In particular, Tikhonov regularization cannot approximate better than up to $O(\Delta^{2/3})$, cf. (7).

Iterated Tikhonov regularization. To obtain smoother solutions one may use the formula (9) with iterative refinement (on the unregularized system):

$$x^{(0)} = 0, \quad r^{(0)} = y,$$

and, for $l = 1, 2, \ldots$,

$$x^{(l)} = x^{(l-1)} + (A^*A + h^2 I)^{-1} A^* r^{(l-1)}, \quad r^{(l)} = y - Ax^{(l)}. \tag{16}$$

This is commonly called *iterated Tikhonov regularization*, but it can also be considered as a *preconditioned Landweber iteration.* (The *Landweber iteration* itself is defined by dropping in (16) the factor $(A^*A + h^2 I)^{-1}$ and can be treated in a similar way, but it is too slow to be useful in practice.) Actually, (16) predates Tikhonov's work and was found already by Riley [29] in 1956.

Note that iterated Tikhonov regularization is not the same as iterative refinement for solving the regularized normal equations (9). (The latter is obtained by using in place of $A^*r^{(l-1)}$ the vector $A^*y - (A^*A + h^2 I)x^{(l-1)} = A^*r^{(l-1)} - h^2 x^{(l-1)}$, and has inferior approximation powers.)

Proposition 5.2. *We have*

$$x^{(l)} = h^{-2}\varphi_l\left(h^{-2}A^*A\right)A^*y, \tag{17}$$

where

$$\varphi_l(t) = \frac{1}{t}\left(1 - \frac{1}{(t+1)^l}\right).$$

Proof. This holds for $l = 0$ since $\varphi_0(t) = 0$. Assuming (17) for $l-1$ in place of l, we have, with $T = h^{-2}A^*A$,

$$\begin{aligned} A^*r^{(l-1)} &= A^*y - A^*Ax^{(l-1)} \\ &= (I - T\varphi_{l-1}(T))A^*y = \psi_l(T)A^*y, \end{aligned}$$

where

$$\psi_l(t) = 1 - t\varphi_{l-1}(t) = \frac{1}{(t+1)^{l-1}}.$$

Thus

$$\begin{aligned} x^{(l)} &= x^{(l-1)} + (A^*A + h^2 I)^{-1} A^* r^{(l-1)} \\ &= h^{-2}\varphi_{l-1}(T)A^*y + (h^2 T + h^2 I)^{-1}\psi_l(T)A^*y \\ &= h^{-2}\varphi_l(T)A^*y \end{aligned}$$

since

$$\begin{aligned}\varphi_{l-1}(t) + (t+1)^{-1}\psi_l(t) &= \frac{1}{t}\left(1 - \frac{1}{(t+1)^{l-1}}\right) + \frac{1}{(t+1)^l} \\ &= \frac{1}{t}\left(1 - \frac{t+1}{(t+1)^l} + \frac{t}{(t+1)^l}\right) \\ &= \frac{1}{t}\left(1 - \frac{1}{(t+1)^l}\right) = \varphi_l(t).\end{aligned}$$

Thus (17) holds generally.□

Therefore, Theorem 5.1 applies with $\varphi(t) = \varphi_l(t)$, and Theorem 4.1 gives an error bound for $\hat{x} = x^{(l)}$ defined by (17). Now it is easy to see that this choice leads to $\gamma_1 < \infty$ for all $l > 0$ and

$$\gamma_2 = \sup_{t \geq 0} \frac{t^{p/2}}{(t+1)^l} < \infty \text{ for } p \leq 2l.$$

Thus we are able to exploit smoothness for all orders $p \leq 2l$.

Inspection of the proof of Theorem 5.1 shows that the supremum needs only be taken over all $t \in [0, h^{-2}\|A\|^2]$. This implies that one may get reasonable values for γ_1 and γ_2 also by choosing for φ suitable polynomials. This allows $C_h y$ to be computed iteratively using only matrix-vector multiplications with A and A^*. For example, one can use it in the context of conjugate gradient methods. This makes regularization a viable technique for large-scale problems. For details, see Brakhage [5], Nemirovski [26], Hanke & Raus [16]. A different technique for the large-scale case, allowing more general linear and even nonlinear regularization conditions, is discussed in Kaufman & Neumaier [20, 21].

Scaling. In many applications, such as nonlinear systems and optimization, the linear systems that need to be solved may be badly scaled, and it is advisable to apply the preceding results to scaled variables $\bar{x} = Dx$ related to x by a diagonal transformation matrix D that ensures a common magnitude of all variables. (This is equivalent to using Theorem 4.1 for a scaled Euclidean norm; hence corresponding error estimates are valid.)

If an appropriate scaling matrix D is known, the transformed system to be solved is

$$\bar{A}\bar{x} = y, \quad \bar{A} = AD^{-1}, \tag{18}$$

and the iteration (16) becomes

$$\bar{x}^{(l)} = \bar{x}^{(l-1)} + (\bar{A}^*\bar{A} + h^2 I)^{-1}\bar{A}^* r^{(l-1)}, \quad r^{(l)} = y - \bar{A}\bar{x}^{(l)}.$$

In terms of the original variables $x^{(l)} = D^{-1}\bar{x}^{(l)}$, this can be rewritten as

$$x^{(l)} = x^{(l-1)} + (A^*A + h^2 D^2)^{-1}A^* r^{(l-1)}, \quad r^{(l)} = y - Ax^{(l)}. \tag{19}$$

Note that (19) can be computed efficiently with a single matrix factorization.

If no appropriate scaling matrix D is available, a natural choice is often (but not always) given by the diagonal matrix with diagonal entries

$$D_{kk} = \sqrt{(A^*A)_{kk}} = \|A_{:k}\|_2, \tag{20}$$

where $A_{:k}$ denotes the kth column of A. With this choice, the results are invariant under rescaling of the right hand side and the variables (and corresponding scaling of the rows and columns of A). In this scale, using $A^*A + h^2D^2$ in place of A^*A amounts to multiplying the diagonal entries of A^*A by $\kappa = 1 + h^2$ before doing the factorization.

Since the iteration (19) allows one to use in the bound (5) higher and higher values of p, one can neglect the second term in the parenthesis in (5) if h is not too close to 1, and make the first term small by choosing h largest subject to this requirement. The value $h = 0.1$ corresponding to $\kappa = 1.01$, is a good compromise.

When used in a context where inaccuracies in the solution can be corrected at later stages, a simple stopping criterion for the iteration (19) is to accept $x^{(l)}$ when either $\|r^{(l)}\|$ is sufficiently small or when $\|x^{(l+1)}\|$ exceeds a certain bound. Suitable thresholds are usually available from the context.

6. Using the singular value decomposition. We now assume that we know a singular value decomposition (SVD) of A,

$$A = U\Sigma V^*, \quad U, V \text{ orthogonal}, \ \Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_n). \tag{21}$$

For $A \in \mathbb{R}^{m\times n}$, U is a square orthogonal $m \times m$-matrix, V a square orthogonal $n \times n$-matrix, and $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_n)$ a rectangular diagonal $m \times n$-matrix with diagonal entries $\Sigma_{ii} = \sigma_i$. For details see, e.g., Golub & van Loan [11].

It is well-known that the minimum norm solution of the least squares problem $\|Ax - y\|_2^2 = \min!$ is given by

$$x = V\Sigma^+U^*y = \sum_{\sigma_i\neq 0} \frac{1}{\sigma_i}(U^*y)_i V_{:i}, \tag{22}$$

where the ith column $V_{:i}$ of V is the *singular vector* corresponding to the ith *singular value* σ_i, and

$$\Sigma^+ = \mathrm{Diag}(\sigma_k^+), \quad \sigma_k^+ = \begin{cases} 1/\sigma_k & \text{if } \sigma_k \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Ill-conditioned matrices are characterized by the presence of tiny singular values σ_i, and it is clear from the representation (22) that errors in y that show up in the corresponding components $(U^*y)_i$ are drastically magnified. Therefore, the minimum norm least squares solution is useless for problems with tiny but nonzero singular values.

Using the SVD, one can give meaning to (11) even for irrational functions φ. Indeed, then $A^*A = V\Sigma^*\Sigma V^*$ and $(A^*A)^l = V(\Sigma^*\Sigma)^l V^*$ for $l = 0, 1, 2, \ldots$, so that

$$\varphi\left(h^{-2}A^*A\right) = V\,\mathrm{Diag}\left(\varphi(\frac{\sigma_k^2}{h^2})\right)V^*$$

for all polynomials φ, and by a limiting process also for arbitrary continuous φ. Therefore, (11) becomes

$$C_h y = V\Sigma_h U^* y, \quad \text{with } \Sigma_h = \mathrm{Diag}\left(\frac{\sigma_k}{h^2}\varphi(\frac{\sigma_k^2}{h^2})\right). \tag{23}$$

This differs from (22) in the replacement of the ill-conditioned diagonal matrix Σ^+ by a better behaved diagonal matrix Σ_h. In particular, the choice

$$\varphi(t) = \begin{cases} 1/t & \text{for } t \geq 1, \\ 0 & \text{otherwise} \end{cases}$$

removes the small singular values from the sum (22) and leads to

$$\Sigma_h = \mathrm{Diag}(d_k), \quad d_k = \begin{cases} 1/\sigma_k & \text{for } \sigma_k \geq h, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

The use of (23) and (24) is referred to as regularization by a *truncated singular value decomposition* (TSVD). It has $\gamma_1 = \gamma_2 = 1$ independent of the degree p of smoothness. (However, p is still needed to find the optimal h in (6).)

The alternative choice $\varphi(t) = 1/\max(1, t)$ also works for all p. It has $\gamma_1 = 1$, too, but a smaller constant $\gamma_2 = \dfrac{2}{p+2}(\dfrac{p}{p+2})^{p/2} < 1$, suggesting a slight superiority over TSVD.

In applications where scaling problems may arise, it is again advisable to apply these formulas to the scaled problem (18) with D defined by (20).

7. The difficulty of estimating Δ. So far, we assumed that Δ or an approximation $\delta \approx \Delta$ was known. After having chosen φ, one can compute γ_1 and γ_2 from Theorem 5.1, and then choose an optimal regularization parameter by (6). To complete the discussion it is therefore sufficient to consider ways to estimate Δ from the information in A and y.

At first sight this seems an intractable problem, since when Δ is unknown then the assumption (4) is virtually vacuous. Indeed, by a theorem of Bakushinskii [1] (reproved as Theorem 3.3 in Engl et al. [9]), any technique for choosing regularization parameters in the absence of information about the error level can be defeated by suitably constructed counterexamples, and indeed the techniques in use all fail on a small proportion of problems in simulations where the right hand side is perturbed by random noise.

As a consequence, the literature about choosing regularization parameters (see Hanke & Hansen [15] for a recent review) is surrounded by a sense of mystery and typically based on heuristics and the study of model cases with special distributions of singular values and special right hand sides.

The only arguments of a more rigorous nature, obtained in the context of smoothing splines and summarized in Chapter 4.5 of Wahba [34], are based on stochastic assumptions. But they are discussed using arguments inaccessible to people not trained in statistics, and this makes the approach difficult to understand. As we shall show now, a stochastic approach is feasible in the general situation, and with only elementary machinery.

In view of the above remarks, we can only hope to construct estimates of Δ that work most of the time. We formalize this in a stochastic setting. As a byproduct, we are also able to find useful estimates for the order p of differentiability.

8. Stochastic error analysis. We suppose that we are solving a particular problem

$$y = Ax + \epsilon, \quad x = Sw \tag{25}$$

from a class of problems characterized by stochastic assumptions about ϵ and w. More specifically, we assume that the components ϵ_i and w_k are uncorrelated random variables with mean zero and variance σ^2 and τ^2, respectively. In terms of expectations, this assumption implies that

$$\langle \epsilon_i w_k \rangle = 0 \quad \text{for all } i,k,$$

$$\langle \epsilon_i \epsilon_k \rangle = \langle w_i w_k \rangle = 0 \quad \text{for } i \neq k,$$

and

$$\langle \epsilon_k^2 \rangle = \sigma^2, \quad \langle w_k^2 \rangle = \tau^2 \quad \text{for all } k.$$

Here $\langle z \rangle$ denotes the expectation of a random variable or random vector z. In vector form, these conditions become

$$\langle \epsilon\epsilon^* \rangle = \sigma^2 I, \quad \langle \epsilon w^* \rangle = 0, \quad \langle ww^* \rangle = \tau^2 I. \tag{26}$$

Since the expectation operator $\langle\cdot\rangle$ is linear, we can compute from (25) expectations of arbitrary quadratic forms in ϵ and w. In particular,

$$\begin{aligned} \langle xx^* \rangle &= \langle Sww^*S^* \rangle = S\langle ww^* \rangle S^* = \tau^2 SS^*, \\ \langle xy^* \rangle &= \langle Sw(w^*S^*A^* + \epsilon^*) \rangle = \tau^2 SS^*A^*, \\ \langle yy^* \rangle &= \langle (ASw + \epsilon)(w^*S^*A^* + \epsilon^*) \rangle = \tau^2 ASS^*A^* + \sigma^2 I. \end{aligned} \tag{27}$$

Since σ measures the size of $\varepsilon = y - Ax$ and τ measures the size of w, the quotient

$$\Delta = \sigma/\tau \tag{28}$$

is an appropriate measure for the relative noise level in the stochastic case; cf. Miller [24]. Because of the stochastic assumptions made, Δ can be estimated from an available right hand side; see Section 10. Once Δ is known, the following result (due to Bertero et al. [3]) gives the best one can hope to get. ($\operatorname{tr} B$ denotes the trace of a square matrix B.)

Theorem 8.1. *The error term $\langle \|x - Cy\|^2 \rangle$ is minimal for $C = \hat{C}$, where*

$$\hat{C} = (SS^*A^*A + \Delta^2 I)^{-1} SS^*A^*. \tag{29}$$

For this choice, the optimal estimator $\hat{x} = \hat{C}y$ is computable as $\hat{x} = S\hat{w}$ from the solution of the positive definite symmetric linear system

$$(S^*A^*AS + \Delta^2 I)\hat{w} = S^*A^*y, \tag{30}$$

and we have

$$\langle \|x - \hat{x}\|^2 \rangle = \tau^2 \operatorname{tr}(SS^* - \hat{C}ASS^*). \tag{31}$$

Proof. We look at the behavior of the error term when C is replaced by a perturbed matrix $C + E$. Since

$$\begin{aligned} \|x - (C+E)y\|^2 &= \|x - Cy - Ey\|^2 \\ &= \|x - Cy\|^2 - 2(x - Cy)^*Ey + \|Ey\|^2 \\ &= \|x - Cy\|^2 - 2\operatorname{tr} Ey(x - Cy)^* + \|Ey\|^2, \end{aligned}$$

we have

$$\langle\|x-(C+E)y\|^2\rangle = \langle\|x-Cy\|^2\rangle - 2\,\mathrm{tr}\,E\langle y(x-Cy)^*\rangle + \langle\|Ey\|^2\rangle.$$

The trace term vanishes for the choice

$$\hat{C} = \langle xy^*\rangle\langle yy^*\rangle^{-1} = \tau^2 SS^*A^*(\tau^2 ASS^*A^* + \sigma^2 I)^{-1}, \tag{32}$$

since

$$\langle y(x-\hat{C}y)^*\rangle = \langle yx^*\rangle - \langle yy^*\rangle\hat{C}^* = 0.$$

Hence

$$\langle\|x-(\hat{C}+E)y\|^2\rangle = \langle\|x-\hat{C}y\|^2\rangle + \langle\|Ey\|^2\rangle \geq \langle\|x-\hat{C}y\|^2\rangle.$$

Putting $E = C - \hat{C}$, we see that $\langle\|x-Cy\|^2\rangle$ is minimal for $C = \hat{C}$. Now $K := \tau^2(\tau^2 ASS^*A^* + \sigma^2 I)^{-1}$ is the solution of

$$(\tau^2 ASS^*A^* + \sigma^2 I)K = \tau^2 I.$$

Multiplying by $\tau^{-2}SS^*A^*$, we find $(SS^*A^*ASS^*A^* + \Delta^2 SS^*A^*)K = SS^*A^*$, hence

$$SS^*A^*K = (SS^*A^*A + \Delta^2 I)^{-1}SS^*A^*.$$

By inserting this into (32), we get the formula (29). Moreover,

$$\begin{aligned}\langle\|x-\hat{C}y\|^2\rangle &= \langle\mathrm{tr}(x-\hat{C}y)(x-\hat{C}y)^*\rangle\\ &= \mathrm{tr}\langle xx^*\rangle - 2\,\mathrm{tr}\,\hat{C}\langle yx^*\rangle + \mathrm{tr}\,\hat{C}\langle yy^*\rangle\hat{C}^*\\ &= \mathrm{tr}\langle xx^*\rangle - \mathrm{tr}\,\hat{C}\langle xy^*\rangle^*\\ &= \tau^2\,\mathrm{tr}(SS^* - \hat{C}ASS^*),\end{aligned}$$

giving (31).□

Of course, in practice, we may need to scale the system first, cf. (18)–(20).

In the special case $S = I$ we recover Tikhonov regularization. Thus Tikhonov regularization is the stochastically optimal regularization method under the weak qualitative assumption that $x = w$ is reasonably bounded (corresponding to the case $p = 0$ of Section 2). However, as we shall see below, estimators of good accuracy are possible only when assuming $p > 0$, and then Tikhonov regularization is no longer optimal.

In the applications, σ will be small while τ will be large enough to guarantee that $\Delta \ll \|AS\|_2$. This inequality guarantees that the regularization term in the above optimality result does not dominate the information in the system coefficient matrix.

Note that Δ corresponds only roughly (within a factor $O(1)$) to the Δ used in the deterministic discussion. In particular, while estimates of Δ (such as those found in Section 10 below) may be used to calculate $\hat{x}$ by Theorem 8.1, it appears dubious to use it in Theorem 4.1 with the optimal $\hat{h}$ computed from (6).

For any fixed Δ, we may solve (30) by a direct or iterative method. However, when a singular value decomposition is computationally feasible, one can find explicit formulas that give cheaply the solution for many Δ simultaneously, an important consideration when Δ is unknown and must be found iteratively.

9. Using the singular value decomposition. For the standard choice (1) of S, where $SS^* = (A^*A)^p$, we note that the matrix $\hat{C}$ defined by (29) has the form (11) with $\varphi(t) = t^p/(t^{p+1}+1)$. Therefore we can use the singular value decomposition as before to simplify the formulas.

THEOREM 9.1. *If $SS^* = (A^*A)^p$ and $A = U\Sigma V^*$ is a singular value decomposition of A then the optimal linear estimator $\hat{x} = \hat{C}y$ can be computed from $c = U^*y$ as $\hat{x} = Vz$, where, with $\Delta = \sigma/\tau$,*

$$z_k = \frac{\sigma_k^{2p+1} c_k}{\sigma_k^{2p+2} + \Delta^2},$$

and we have

$$\langle \|x - \hat{x}\|_2^2 \rangle = \sigma^2 \sum_k \frac{\sigma_k^{2p}}{\sigma_k^{2p+2} + \Delta^2}, \tag{33}$$

$$\langle cc^* \rangle = \tau^2 \operatorname{Diag}(\sigma_i^{2p+2} + \Delta^2). \tag{34}$$

Proof. If $A = U\Sigma V^*$ is a singular value decomposition of A, we can write the covariance matrix $\langle yy^* \rangle$ calculated in (27) as

$$\begin{aligned} W &:= \langle yy^* \rangle = \tau^2 A(A^*A)^p A^* + \sigma^2 I = \tau^2 (AA^*)^{p+1} + \sigma^2 I \\ &= U(\tau^2(\Sigma\Sigma^*)^{p+1} + \sigma^2 I)U^*. \end{aligned}$$

The vector

$$c := U^*y \tag{35}$$

has a diagonal covariance matrix,

$$\langle cc^* \rangle = U^* \langle yy^* \rangle U = \tau^2(\Sigma\Sigma^*)^{p+1} + \sigma^2 I. \tag{36}$$

Now (32) takes the form

$$\begin{aligned} \hat{C} &= \tau^2 (A^*A)^p A^* W^{-1} = \tau^2 A^* (AA^*)^p W^{-1} \\ &= \tau^2 V\Sigma^* (\Sigma\Sigma^*)^p (\tau^2(\Sigma\Sigma^*)^{p+1} + \sigma^2 I)^{-1} U^* = VDU^* \end{aligned}$$

with the diagonal matrix

$$D = \Sigma^* (\Sigma\Sigma^*)^p ((\Sigma\Sigma^*)^{p+1} + \Delta^2 I)^{-1}. \tag{37}$$

Therefore

$$\hat{x} = Vz, \tag{38}$$

where $z = DU^*y = Dc$ has the components

$$z_i = \frac{\sigma_i^{2p+1}}{\sigma_i^{2p+2} + \Delta^2} c_i. \tag{39}$$

It remains to express the residual term (31). Since

$$SS^* = (A^*A)^p = V(\Sigma^*\Sigma)^p V^*,$$

$$\hat{C}A = VDU^*U\Sigma V^* = VD\Sigma V^*,$$

we get the desired result from

$$\begin{aligned}\langle \|x - \hat{C}y\|_2^2 \rangle &= \tau^2 \operatorname{tr}(I - \hat{C}A)SS^* \\ &= \tau^2 \operatorname{tr} V(I - D\Sigma)(\Sigma\Sigma^*)^p V^* \\ &= \tau^2 \operatorname{tr}(I - D\Sigma)(\Sigma\Sigma^*)^p V^*V \\ &= \tau^2 \operatorname{tr}(I - D\Sigma)(\Sigma\Sigma^*)^p \\ &= \tau^2 \sum_k (1 - D_{ii}\sigma_k)\sigma_k^{2p} \\ &= \tau^2 \sum_k \frac{\Delta^2 \sigma_k^{2p}}{\sigma_k^{2p+2} + \Delta^2} = \sigma^2 \sum_k \frac{\sigma_k^{2p}}{\sigma_k^{2p+2} + \Delta^2}.\end{aligned}$$

□

The above formulas are related to those known for a long time in the classical technique of *Wiener filtering* (Wiener [36]) for the deconvolution of sequences or images, probably the earliest application of regularization techniques. The singular values σ_k correspond to the absolute values $|H_k|$ of the Fourier coefficients of the point spread function H defining a convolution operator A. The deconvolution of a sequence or image y is done by multiplying its Fourier transform $\tilde{y}$ by the Fourier transform of the filter function, $H^*/(|H|^2 + P_n/P_g)$, where P_n is the noise power and P_g the power in a model sequence or image, and operations are elementwise. If one assumes that $P_g = |H|^{2p}$ (which is an implicit smoothness assumption for the intensities of the true image or density) and writes $c = \arg(H^*)y$ and $P_n = \Delta^2$, the product takes the form (39). (All this finds its natural explanation by noting that the singular value decomposition of a circulant matrix can be written explicitly in terms of the Fourier transforms.) For details see, e.g., Katsaggelos [19].

Order of convergence. The following result about the optimal convergence order reveals a difference between the stochastic and the deterministic situation. We assume that $\tau = O(1)$, and that $\sigma = O(\Delta)$ is small.

Theorem 9.2. *For any $q \in [0, 1]$, we have*

$$\langle \|x - \hat{x}\|_2^2 \rangle \leq \alpha_q \tau^{2-2q} \sigma^{2q} = O(\Delta^{2q}) \tag{40}$$

with

$$\alpha_q = q^q (1-q)^{1-q} \sum \sigma_k^{2p-2q(p+1)}. \tag{41}$$

Proof. We rewrite (33) as

$$\begin{aligned}\langle \|x - \hat{x}\|_2^2 \rangle &= \sigma^2 \sum \frac{(\sigma_k^{p+1}/\Delta)^{2q}}{(\sigma_k^{p+1}/\Delta)^2 + 1} \Delta^{2q-2} \sigma_k^{2p-2q(p+1)} \\ &\leq \sigma^2 \left(\sup_{x \geq 0} \frac{x^{2q}}{x^2 + 1} \right) \Delta^{2q-2} \sum \sigma_k^{2p-2q(p+1)}.\end{aligned}$$

The bound (40) follows since $\sigma^2\Delta^{2q-2} = \sigma^2(\sigma/\tau)^{2q-2} = \tau^{2-2q}\sigma^{2q}$ and

$$\sup_{x\geq 0} \frac{x^{2q}}{x^2+1} = q^q(1-q)^{1-q}$$

(attained at the zero $x = \sqrt{q/(1-q)}$ of the gradient).□

Since in situations that need to be regularized we always have some tiny σ_k, the constant α_q is reasonable only when $q \leq \frac{p}{p+1}$. In particular, since we need $q > 0$ in order that the bound (40) goes to zero as the model error σ goes to zero, p must be strictly positive.

The choice $q = \frac{p}{p+1}$ yields an α_q proportional to the rank of A, the number of nonzero singular values. Thus we recover the deterministic convergence order only when the rank of A is not too large. For typical discretized function space problems, however, A has full rank, the dimension is large, and, usually, the singular values σ_k of A approach the singular values σ_k^* of the continuous problem,

$$\sigma_k \to \sigma_k^* \quad \text{as } n \to \infty;$$

cf. Hansen [18]. To have a bounded α_q in this limit we can only use a reduced exponent

$$q = \frac{p-e}{p+1},$$

where e is a number such that

$$\sum_{k=1}^{\infty} (\sigma_k^*)^{2e} < \infty,$$

and then have

$$\langle \|x - \hat{C}y\|_2^2 \rangle = O(\Delta^{2q}),$$

independent of the dimension.

10. Estimating the regularization parameter. We now return to the problem of estimating Δ. We base our discussion on the SVD; how to adapt the estimation in the case when a SVD is not available is postponed to Section 11 (after (81)). The key is relation (34) that expresses the covariance matrix of the (high-dimensional) vector c in terms of the unknown parameters τ and $\Delta = \sigma/\tau$. From this relation, we get

$$\langle c_k^2 \rangle = \sigma^2 + \tau^2\sigma_k^{2p+2} \quad \text{for } k = 1, \ldots, n. \tag{42}$$

Now in general, relations of the form

$$\langle c_k^2 \rangle = v_k(\theta^*) \quad \text{for } k = 1, \ldots, n \tag{43}$$

allow a low-dimensional parameter vector θ^* to be estimated from a single realization of the high-dimensional random vector c. Indeed, the estimation of variance components of random vectors is a classical problem in statistics (see, e.g., Barndorff-Nielsen & Cox [2]). We shall give a direct and elementary derivation of a family of estimators.

THEOREM 10.1. *Suppose that c is a random vector satisfying* (43). *Let ψ be a strictly concave function, and define*

$$f(c,\theta) = \sum_{k=1}^{n} \alpha_k \left(\psi(v_k(\theta)) + \psi'(v_k(\theta))(c_k^2 - v_k(\theta))\right) \tag{44}$$

with weights $\alpha_k > 0$. Then $\langle f(c,\theta)\rangle$ is bounded below, and any global minimizer $\hat{\theta}$ of $\langle f(c,\theta)\rangle$ satisfies $v_k(\hat{\theta}) = v_k(\theta^*)$ for $k = 1,\dots,n$.

Proof. Write $v_k = v_k(\theta), v_k^* = v_k(\theta^*)$. Then

$$\langle f(c,\theta)\rangle = \sum_{k=1}^{n} \alpha_k(\psi(v_k) + \psi'(v_k)(v_k^* - v_k)),$$

$$\langle f(c,\theta^*)\rangle = \sum_{k=1}^{n} \alpha_k(\psi(v_k^*) + 0),$$

hence

$$\langle f(c,\theta)\rangle - \langle f(c,\theta^*)\rangle = \sum_{k=1}^{n} \alpha_k(\psi(v_k) - \psi(v_k^*) - \psi'(v_k)(v_k - v_k^*)).$$

Since ψ is strictly concave and $\alpha_k > 0$, the kth term in the sum is nonnegative, and vanishes only for $v_k = v_k^*$. Hence $\langle f(c,\theta)\rangle \geq \langle f(c,\theta^*)\rangle$, with equality iff $v_k = v_k^*$ for all k.□

COROLLARY 10.2. *If, in addition, θ^* is determined uniquely by the values of $v_k(\theta^*)$ then θ^* is the unique global minimizer of $\langle f(c,\theta)\rangle$.*

The case of two parameters. For

$$\langle c_k^2\rangle = \sigma^2\mu_k + \tau^2\lambda_k \quad \text{for } k = 1,\dots,n, \tag{45}$$

the case of interest in our application, the estimation problem reduces to a global optimization problem in two variables σ and τ. (We introduce μ_k and λ_k since we shall need it in Section 11.) It is possible to reduce this further to an optimization problem in the single variable

$$t^* = \Delta^2 = \sigma^2/\tau^2 \tag{46}$$

by restricting attention to the family of concave functions defined by

$$\psi_q(v) = \begin{cases} (v^q - 1)/q & \text{if } 0 \neq q < 1, \\ \log v & \text{if } q = 0 \end{cases} \tag{47}$$

for some parameter $q < 1$. In the following, exp denotes the exponential function.

THEOREM 10.3. *Suppose that c is a random vector satisfying* (45). *Define, for given weights $\alpha_k > 0$,*

$$\gamma_q(t) = \sum \alpha_k(\lambda_k + t\mu_k)^{q-1}c_k^2, \tag{48}$$

$$\beta_q(t) = \sum \alpha_k(\lambda_k + t\mu_k)^q, \tag{49}$$

$$f_q(t) = \begin{cases} \log\gamma_q(t) + \frac{1-q}{q}\log\beta_q(t) & \text{if } 0 \neq q < 1, \\ \log\gamma_0(t) + \sum \alpha_k \log(\lambda_k + t\mu_k)/\sum\alpha_k & \text{if } q = 0. \end{cases} \tag{50}$$

Then $\langle \exp f_q(t)\rangle$ *is bounded below, and* $t^* = \sigma^2/\tau^2$ *is the unique global minimizer of* $\langle \exp f_q(t)\rangle$*. Moreover,*

$$\sigma^2 = t^*\langle\gamma_q(t^*)\rangle/\beta_q(t^*). \tag{51}$$

Proof. Condition (45) is equivalent to (43) with

$$v_k(\theta) = \theta_1\mu_k + \theta_2\lambda_k, \quad \theta^* = \begin{pmatrix}\sigma^2\\ \tau^2\end{pmatrix}. \tag{52}$$

We apply Theorem 10.1 with $\psi = \psi_q$ given by (47), and consider the optimal point $\theta = \theta^*$, and the associated $t = t^* = \sigma^2/\tau^2$. Since θ is a minimizer of $\langle f(c,\theta)\rangle$, the gradient

$$\begin{aligned}\nabla\langle f(c,\theta)\rangle &= \sum \alpha_k(\psi'(v_k) + \psi''(v_k)(\langle c_k^2\rangle - v_k) + \psi'(v_k)(-1))\nabla v_k \\ &= \sum \alpha_k\psi''(v_k)(\langle c_k^2\rangle - v_k)\nabla v_k\end{aligned}$$

vanishes. Since $(\sigma^2,\tau^2)\nabla v_k = (\sigma^2,\tau^2)(\mu_k,\lambda_k)^T = v_k$, we find

$$\sum \alpha_k\psi''(v_k)(\langle c_k^2\rangle - v_k)v_k = 0,$$

and since $\psi_q'(v) = v^{q-1}, \psi_q''(v) = (q-1)v^{q-2}$ for the choice (47), we find

$$\sum \alpha_k v_k^{q-1}(\langle c_k^2\rangle - v_k) = 0. \tag{53}$$

Using (52), (48) and (49), this equation becomes $\tau^{2q-2}\langle\gamma_q(t)\rangle - \tau^{2q}\beta_q(t) = 0$, giving

$$\tau^2 = \frac{\langle\gamma_q(t)\rangle}{\beta_q(t)}, \quad \sigma^2 = t\frac{\langle\gamma_q(t)\rangle}{\beta_q(t)}. \tag{54}$$

Thus τ^2 and σ^2 can be expressed in terms of the single parameter t. Now (44) and (53) imply

$$\langle f(c,\theta)\rangle = \sum \alpha_k\psi_q(v_k). \tag{55}$$

For the limiting case $q = 0$, we find

$$\begin{aligned}\langle f(c,\theta)\rangle &= \sum \alpha_k \log v_k = \sum \alpha_k(\log\tau^2 + \log(\lambda_k + t\mu_k)) \\ &= \beta_0(\log\langle\gamma_0(t)\rangle - \log\beta_0) + \sum \alpha_k\log(\lambda_k + t\mu_k) \\ &= \beta_0\log\langle\exp f_0(t)\rangle - \beta_0\log\beta_0,\end{aligned}$$

where $\beta_0 = \sum\alpha_k$ is independent of t. Thus minimizing $\langle f(c,\theta)\rangle$ is equivalent to minimizing $\langle\exp f_0(t)\rangle$. And for $q \neq 0$ we have

$$\begin{aligned}q\langle f(c,\theta)\rangle + \sum\alpha_k &= \sum \alpha_k(q\psi_q(v_k) + 1) = \sum \alpha_k v_k^q \\ &= \tau^{2q}\sum \alpha_k(\lambda_k + t\mu_k)^q = \tau^{2q}\beta_q(t) \\ &= \left(\frac{\langle\gamma_q(t)\rangle}{\beta_q(t)}\right)^q \beta_q(t) = \langle\gamma_q(t)\rangle^q\beta_q(t)^{1-q},\end{aligned}$$

and this expression must therefore be minimized (for $q > 0$) or maximized (for $q < 0$) to get the optimal t. By taking the logarithm and dividing by q, we reverse the monotony behavior for $q < 0$; thus minimizing $\langle f(c,\theta)\rangle$ is again equivalent to minimizing $\langle \exp f_q(t)\rangle$. Thus the theorem follows from the previous result.□

Choice of weights. We now consider the choice of weights in (48)–(49). Since the weights are not random, the α_k must be independent of the c_k. It is desirable that the objective function does not change when condition (45) is rescaled by positive scale factors π_k. Therefore the weights must be chosen in such a way that the transformation $\bar\lambda_k = \pi_k\lambda_k$, $\bar\mu_k = \pi_k\mu_k$, $\bar c_k = \pi_k^{1/2}c_k$ transforms the α_k into $\bar\alpha_k = \pi_k^{-q}\alpha_k$. This is the case for the choices

$$\alpha_k = \lambda_k^r \mu_k^s, \quad q = -r - s;$$

to ensure that tiny values of λ_k and μ_k do not cause trouble we need $r, s \geq 0$. Since q is determined by r and s, it is convenient to use the index pair rs in place of q.

For $r = s = 0$, we find the *generalized maximum likelihood* (GML) merit function

$$f_{00}(t) = \log \sum \frac{c_k^2}{\lambda_k + t\mu_k} + \frac{1}{n}\sum \log(\lambda_k + t\mu_k). \tag{56}$$

(Indeed, under the additional assumption that the c_k are independent Gaussian random variables with zero mean, the global minimizer $\hat t$ of $f_{00}(t)$ is the traditional maximum likelihood estimator for t^*.)

If $r > 0$ or $s > 0$, we find the (r,s)-*generalized cross validation* (GCV) merit function

$$f_{rs}(t) = \log \gamma_{rs}(t) - \frac{1+r+s}{r+s}\log \beta_{rs}(t), \tag{57}$$

where

$$\beta_{rs}(t) = \sum \left(\frac{\lambda_k}{\lambda_k + t\mu_k}\right)^r \left(\frac{\mu_k}{\lambda_k + t\mu_k}\right)^s, \tag{58}$$

$$\gamma_{rs}(t) = \sum \left(\frac{\lambda_k}{\lambda_k + t\mu_k}\right)^r \left(\frac{\mu_k}{\lambda_k + t\mu_k}\right)^s \left(\frac{c_k^2}{\lambda_k + t\mu_k}\right). \tag{59}$$

(Indeed, in an important special case discussed in Section 11 below, the global minimizer of the (0,1)-GCV merit function is the familiar GCV estimate for the regularization parameter.)

Practical use. For the regularization problem, the above results apply with

$$\lambda_k = \sigma_k^{2p+2}, \quad \mu_k = 1 \tag{60}$$

(compare (42) and (45)).

In practice, the expectation $\langle \exp f_{rs}(t)\rangle$ cannot be computed since only a single realization is known. However, the theorem suggests that one can estimate t by finding the global minimizer of $\exp f_{rs}(t)$ or equivalently of $f_{rs}(t)$, where λ_k and μ_k are given

TABLE 1
Failure rates in percent

case	$e_1\backslash e_2$	1	2	4	8	16	32
$n = 50$	1	2	0	0	2	7	14
alg.	2	0	0	0	0	1	1
	4	0	0	0	0	0	0
	8	2	1	0	0	0	0
	16	5	0	0	0	0	0
	32	10	1	0	0	0	0
case	$e_1\backslash e_2$	1	2	4	8	16	32
$n = 50$	1	4	12	16	22	21	16
exp.	2	0	0	3	6	10	17
	4	0	0	0	0	4	16
	8	0	0	0	0	1	2
	16	0	0	0	0	0	0
	32	1	0	0	0	0	0
case	$e_1\backslash e_2$	1	2	4	8	16	32
$n = 500$	1	0	0	0	0	0	1
alg.	2	0	0	0	0	0	0
	4	0	0	0	0	0	0
	8	0	0	0	0	0	0
	16	0	0	0	0	0	0
	32	0	0	0	0	0	0
case	$e_1\backslash e_2$	1	2	4	8	16	32
$n = 500$	1	0	1	10	18	14	17
exp.	2	0	0	0	3	7	17
	4	0	0	0	0	1	1
	8	0	0	0	0	0	0
	16	0	0	0	0	0	0
	32	0	0	0	0	0	0

in terms of p by (60). If p is unknown, one can also minimize over a set of values $p = 1, 2, 3...$ to find the best order of differentiability.

Assuming that the c_k are independent Gaussian random variables with zero mean, and with some technical additional assumptions, it can be proved that the resulting estimate $\hat{t}$ for t^* (and $\hat{p}$ for p) is an asymptotically unbiased estimate as $n \to \infty$; see, e.g., WELSH [35]. In particular, this holds for the maximum likelihood estimator (56), whose asymptotic properties are well-known (see, e.g., BARNDORFF-NIELSEN & COX [2]).

To find the best values for r and s, we note first that the results only depend on the distribution of the quotients

$$q_k = \frac{\sigma^2 \mu_k}{\tau^2 \lambda_k}.$$

We therefore performed a simulation study of the estimation properties for various choices of these quotients. We looked at two different kinds of distributions representative of realistic problems, namely algebraic decay, $q_k = q_1 k^{-a}$, and exponential decay, $q_k = q_1 e^{-ak}$, where q_1 and a are chosen such that the q_k range between 10^{-e_1}

TABLE 2
Number of first, second and third places

case	$2r$	$2s$	gold	silver	bronce
$n = 500$	0	0	1674	433	12
alg.	0	1	1174	340	23
	1	0	1136	351	22
	1	1	1033	256	11
	0	2	884	245	16
	2	1	871	202	11
	1	2	867	198	13
	2	0	831	260	16
	2	2	803	120	7
	1	3	743	182	12
	3	1	737	196	6
	2	3	719	116	5
	3	2	715	113	4
	0	3	706	210	12
	3	3	692	95	2
	3	0	638	233	14
	4	2	632	119	7
	2	4	623	134	7
	1	4	621	204	9
	3	4	620	90	3
	4	1	617	192	9
	4	3	607	89	3
	4	4	588	68	2
	0	4	560	187	10
	4	0	514	199	13

case	$2r$	$2s$	gold	silver	bronce
$n = 500$	0	0	993	394	37
exp.	1	0	785	296	19
	0	1	715	265	24
	1	1	707	217	13
	2	1	658	179	4
	2	0	621	233	19
	1	2	603	171	12
	2	2	584	127	4
	3	2	568	125	5
	0	2	557	228	31
	2	3	540	120	5
	3	1	521	176	5
	3	0	505	207	18
	1	3	505	186	18
	4	3	495	89	2
	3	3	495	86	4
	4	2	476	145	5
	2	4	474	145	8
	1	4	466	196	20
	0	3	458	187	31
	4	1	458	185	14
	3	4	456	89	4
	4	4	437	77	2
	4	0	393	170	22
	0	4	388	173	28

and 10^{e_2}, with constants e_1 and e_2 that were varied independently in $\{1, 2, 4, 8, 16, 32\}$ (if e_1 and e_2 had different signs, estimation was extremely unreliable). Then we chose $\sigma^2 = 10^{-e_1}$, $\tau = 1$, $\lambda_k = \sigma^2/q_k$ and $\mu_k = 1$, and minimized for each case the curves for $f_{rs}(t)$ with $r, s \in \{0, \frac{1}{2}, 1, 1\frac{1}{2}, 2\}$, using samples of various sizes n of independently distributed Gaussian noise c_k with mean zero and variance given by (45).

For the simulation, the minimization was done by evaluation at $t = 10^g t^*$, where $t^* = \sigma^2/\tau^2$ and $g \in \{-2.0, -1.9, ..., 1.9, 2.0\}$. If the smallest function value occured for $|g| > 1$ for *all* values of (r, s) under study, the estimation was classified as failure. Failure rates decrease with increasing dimension n; some are reported in Table 1. As to be expected, when the dimension n is small, there is a larger range of distributions

of the q_k where the estimates are unreliable for all choices of (r, s).

If the minimum occured at some $|g| \leq 1$ for at least one value of (r, s), we ranked the pairs (r, s) by the value of $|g|$, giving gold medals for the smallest $|g|$, and silver and bronce medals to the next smallest. (But if there were three gold medals, no other medals were given, and if there were two gold medals or two silver medals, no bronce medals were given.) The ranking by number of medals is given for $n = 500$ and 3600 test cases in Table 2; the fact that there are much fewer silver medals and very few bronce medals shows that often several choices of (r, s) share the best position. The GML merit function (with $r = s = 0$) is the clear winner. For smaller n or for smaller values of the threshold defining failures, the behavior is similar, though there are fewer gold medals, and GML remains uniformly at the top of the list, both in terms of gold medals and in terms of total number of medals.

Thus one may recommend to estimate $t = \Delta^2$ and p by minimizing the GML merit function (56), and then to use Theorem 8.1 to get $\hat{x}$. A suitable starting point for the minimization is given by

$$t = \text{median}(\lambda_k/\mu_k), \quad p = 1.$$

(With this starting point, the local univariate minimizer I used, based on a bracket search and local parabolic interpolation, cf. Gill et al. [10], usually takes about 7–15 function values, most typically 9, to find a local minimizer t^*.)

To assess more completely the impact of various methods to estimate t and p one also needs to study how sensitive the accuracy of a computed solution $\hat{x}$ depends on the choice of t and p. The only investigation in this direction known to me is a paper by Wahba [34] that shows that there are circumstances under which (0,1)-GCV is more robust than GML when p is misspecified. It is unknown whether this persists when p is estimated together with t from (0,1)-GCV or GML.

11. More flexible smoothness constraints. We now look at the regularization of general linear stochastic models

$$y = Ax + \epsilon, \quad \langle \epsilon \epsilon^* \rangle = \sigma^2 V, \tag{61}$$

where $A \in \mathbb{R}^{m \times n}$, $\sigma^2 V \in \mathbb{R}^{m \times m}$ is a symmetric and positive definite covariance matrix, and σ is typically unknown.

The qualitative constraint is assumed to be given in the form that certain components or linear combinations of the state vector cannot become large, and we can formulate this with a suitable matrix $J \in \mathbb{R}^{r \times n}$ by assuming

$$Jx = w \tag{62}$$

with w of reasonable norm. In the applications, J is usually a matrix of suitably weighted first or second order differences that apply to some part of x containing function values or densities. For problems involving images, x often contains grey values or color intensities, interpreted as a two- or (for movies or multiple slices through an object) three-dimensional function of the position. Then (62) is again a smoothness condition, and by judicious choice of J, the smoothness requirements can be adapted to particular problems.

Modeling smoothness by (62) has the advantage that the smoothness condition may be unrelated to A. In particular, one can use it also for problems where A is well-conditioned but the solution of $Ax = y$ would lead to an unacceptably rough solution x. An important case where A is the identity matrix is that of curve or

image smoothing. Here y contains the function values of a noisy curve or the pixel intensities of a noisy image, and one wants to reconstruct the smooth original x. (For the relaxed requirement that x is only piecewise smooth, one needs nonlinear regularization terms, that must be handled iteratively. See, e.g., [21, 32].)

Our assumptions lead to the conditions

$$\|Jx\| \quad \text{of reasonable size,} \tag{63}$$

$$\langle (Ax-y)^* V^{-1}(Ax-y)\rangle = \sigma^2 m. \tag{64}$$

Traditionally, one interprets (63) by adding a multiple of $\langle \|Jx\|^2\rangle$ as penalty to (64), and solving the associated least squares system

$$\sigma^{-2}(y-Ax)^* V^{-1}(y-Ax) + \tau^{-2}\|Jx\|^2 = \min! \tag{65}$$

With $t = \Delta^2 := \sigma^2/\tau^2$ as regularization parameter, the solution can be found from the normal equations

$$B_t \hat{x} = A^* V^{-1} y, \quad \text{where } B_t = A^* V^{-1} A + tJ^* J, \tag{66}$$

These equations are again a generalization of Tikhonov regularization, which is obtained as the special case $V = I$, $J = I$.

If J is square and nonsingular, we may rewrite (62) as $x = Sw$ with $J = S^{-1}$. If $V = I$ (which can always be enforced by a transformation of ε), we may use the stochastic setting and obtain from Theorem 8.1 the optimal estimator $\hat{x} = \hat{C}y = (SS^*A^*A + \Delta^2 I)^{-1} SS^* A^* y = (A^*A + \Delta^2 J^* J)A^* y$, and this formula agrees with (66). Therefore, (65) is the natural generalization of the optimal situation considered before to the present context.

For large scale problems, (66) can be solved using one Cholesky factorization for each value of t, and we show below how these factorizations can be used to find an appropriate regularization parameter, too. Elden [7] observed that a significant amount of work can be saved by first reducing the problem to bidiagonal form. For full matrices, this can be done with $O(n^3)$ work, and each subsequent function evaluation costs only $O(n)$ operations. Unfortunately, the method is limited to dense problems (banded problems work too, but only for $J = I$) since sparsity in A and/or J is destroyed by bidiagonalization. For problems with banded A and J one can still save a little work by first reducing A and J to banded triangular form; see Section 5.3.4 of Björck [4].

If the sparsity structure of B is such that the computation of several Cholesky factorizations is not practical, iterative methods have to be used. For details see the references given in Section 5.

Using the generalized SVD. We now show that we can choose a special coordinate system where both (63) and (64) become separable and hence easy to analyze.

Theorem 11.1. *(i) There are nonsingular matrices $M \in \mathbb{R}^{n\times n}$ and $N \in \mathbb{R}^{m\times m}$, and nonnegative diagonal matrices $D \in \mathbb{R}^{m\times n}$ and $E \in \mathbb{R}^{n\times n}$ such that*

$$N^*N = V^{-1}, \quad NA = DM, \quad J^*J = M^*E^*EM. \tag{67}$$

(ii) Whenever (67) *holds, the transformed variables*

$$z := Mx\,, \quad c := Ny \tag{68}$$

satisfy

$$\|Jx\|^2 = \|Ez\|^2, \tag{69}$$

$$(Ax-y)^*V^{-1}(Ax-y) = \|Dz-c\|^2. \tag{70}$$

Proof. We give a fully constructive proof that directly translates into an algorithm for computing M, N, D and E. We begin by factoring $V = LL^*$ and consider, for some $\rho \neq 0$, the QR-factorization

$$\begin{pmatrix} L^{-1}A \\ \rho J \end{pmatrix} = QR \qquad \text{with } Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \in \mathbb{R}^{(m+r)\times n}, \quad Q^*Q = I \tag{71}$$

and upper triangular $R \in \mathbb{R}^{n\times n}$. Thus

$$L^{-1}A = Q_1R, \quad \rho J = Q_2R' \quad Q_1^*Q_1 + Q_2^*Q_2 = I. \tag{72}$$

Using the singular value decomposition

$$Q_1 = UDW^* \qquad \text{with orthogonal } U \in \mathbb{R}^{m\times m},\ W \in \mathbb{R}^{n\times n} \tag{73}$$

and a nonnegative diagonal matrix $D \in \mathbb{R}^{m\times n}$ we define

$$M := W^*R, \quad N := U^*L^{-1}. \tag{74}$$

From (73) we find $D^*D = (U^*Q_1W)^*(U^*Q_1W) = W^*Q_1^*Q_1W$, so that the diagonal matrix

$$I - D^*D = W^*W - W^*Q_1^*Q_1W = W^*Q_2^*Q_2W = (Q_2W)^*(Q_2W) \tag{75}$$

is positive semidefinite. Therefore its entries are nonnegative, and we can form the nonnegative diagonal matrix

$$E := \rho^{-1}(I - D^*D)^{1/2}, \tag{76}$$

with component-wise square roots. Now

$$\begin{aligned} N^*N &= L^{-T}UU^*L^{-1} = L^{-T}L^{-1} = (LL^*)^{-1} = V^{-1}, \\ NA &= U^*L^{-1}A = U^*Q_1R = U^*UDW^*R = DW^*R = DM, \\ \rho J &= Q_2R = Q_2WW^*R = Q_2WM, \\ \rho^2J^*J &= (Q_2WM)^*(Q_2WM) = M^*(Q_2W)^*(Q_2W)M = M^*(I-D^*D)M \\ &= \rho^2M^*E^*EM, \end{aligned}$$

since E is a square diagonal matrix. This proves (67). We now conclude from (67) and (68) that

$$\|Jx\|^2 = x^*J^*Jx = x^*(EM)^*(EM)x = \|EMx\|^2 = \|Ez\|^2,$$

$$\begin{aligned} (Ax-y)^*V^{-1}(Ax-y) &= (Ax-y)^*N^*N(Ax-y) \\ &= \|N(Ax-y)\|^2 = \|Dz-c\|^2. \end{aligned}$$

□

When $V = I$, the factorization (67) is generally referred to as a *generalized singular value decomposition*. See VAN LOAN [22, 23], STEWART [30] and PAIGE [27] for further properties of the generalized SVD. The case $V \neq I$ seems not to have been discussed before.

An implementation of the transformation may proceed according to (71), (73) and (76). In (71) one should choose

$$\rho = \|L^{-1}A\|_\infty / \|J\|_\infty \tag{77}$$

or a similar expression in order to ensure that the two matrices on the left hand side have similar magnitude. This guarantees a stable computation when V and hence L are well-conditioned. (See VAN LOAN [23] for the stability of algorithms for computing the GSVD. The counterexamples to the stability of the GSVD computed using (73) given there do not apply if we use the matrices M and N only implicitly, as indicated below.)

Stiff models (61), where V is ill-conditioned, arise in some path tracking applications where there is noise on two different time scales (small system noise and larger measurement noise). In this case, the numerical solution according to (71), (73) and (76) may suffer from instability, and a stable formulation is unknown.

Since by (75), the diagonal entries of D lie in the interval $[0, 1]$, we can replace in (73) any computed singular values > 1 (produced by finite precision calculations) by 1. Then the calculation of E in (76) causes no problem.

Having determined the variances, we can find z by solving the least squares problem

$$\sigma^{-2}\|c - Dz\|^2 + \tau^{-2}\|Ez\|^2 = \min! \tag{78}$$

corresponding to (65). After multiplication by σ^2 we find the solution $\hat{z}$ from the normal equations

$$(D^*D + tE^*E)\hat{z} = D^*c,$$

where $t = \sigma^2/\tau^2$. Since D and E are diagonal, we obtain

$$\hat{z}_k = \begin{cases} \dfrac{D_{kk}c_k}{D_{kk}^2 + tE_{kk}^2} & \text{for } k \leq \bar{n}, \\ 0 & \text{for } k > \bar{n}. \end{cases}$$

Note that c can be computed from y by

$$c = U^*(L^{-1}y); \tag{79}$$

and since $M^{-1} = R^{-1}W^{-T} = R^{-1}W$, the solution estimate $\hat{x}$ can be recovered from $\hat{z}$ by

$$\hat{x} = R^{-1}(W\hat{z}). \tag{80}$$

Therefore, the matrices M and N in (74) need not be formed explicitly. The vectors (79) and (80) can be efficiently computed by triangular solves and matrix-vector multiplications.

Finding the regularization parameter. In the new coordinates, the model equation (61) implies that

$$\bar\epsilon := c - Dz = Ny - DMx = N(y - Ax) = N\epsilon$$

satisfies

$$\langle \bar\epsilon \bar\epsilon^* \rangle = N \langle \epsilon\epsilon^* \rangle N^* = \sigma^2 NVN^* = \sigma^2 I$$

and in particular,

$$\langle \bar\epsilon_k^2 \rangle = \sigma^2 \quad \text{for } k = 1, \ldots, m.$$

On the other hand, the qualitative condition (63) says that $\|Ez\|$ is of reasonable size. In analogy with Section 8, we now consider the class of problems characterized by random vectors z such that $w = Ez$ satisfies

$$\langle w_k^2 \rangle = \tau^2$$

with some $\tau = O(1)$.

If we also make the natural assumption that the w_k and the $\bar\epsilon_k$ are uncorrelated, $\langle w_k \bar\epsilon_k \rangle = 0$, the variances σ^2 and τ^2 can again be determined using Theorem 10.3. Indeed, we have

$$\bar c_k = E_{kk} c_k = D_{kk} E_{kk} z_k + E_{kk} \epsilon_k = D_{kk} w_k + E_{kk} \epsilon_k,$$

so that

$$\langle \bar c_k^2 \rangle = D_{kk}^2 \tau^2 + E_{kk}^2 \sigma^2 \quad \text{for } k = 1, \ldots, \bar n = \min(m, n).$$

Hence Theorem 10.3 applies with

$$\bar c_k = E_{kk} c_k, \quad \bar\lambda_k = D_{kk}^2, \quad \bar\mu_k = E_{kk}^2 \tag{81}$$

in place of c_k, λ_k and μ_k, respectively.

For large scale problems, the generalized SVD is prohibitively expensive to compute or store. Therefore it is important that the above formulas can be rewritten in terms of the original matrices. This allows them to be used for problems with large and sparse A and J, provided that V (and hence N) are diagonal or at least block diagonal. This is the case in a large number of applications.

To rewrite the merit functions $f_{rs}(t)$ for finding the regularization parameter t, we note first that (67) implies

$$B_t = A^* V^{-1} A + t J^* J = M^* (D^* D + t E^* E) M;$$

therefore

$$P_t := NAB_t^{-1} A^* N^* = D(D^* D + t E^* E)^{-1} D^*$$

is a diagonal matrix. Using (81) we get

$$P_t = \operatorname{Diag}\left(\frac{D_{kk}^2}{D_{kk}^2 + t E_{kk}^2} \right) = \operatorname{Diag}\left(\frac{\bar\lambda_k}{\bar\lambda_k + t \bar\mu_k} \right)$$

and

$$I - P_t = t\,\mathrm{Diag}\left(\frac{E_{kk}^2}{D_{kk}^2 + tE_{kk}^2}\right) = t\,\mathrm{Diag}\left(\frac{\bar{\mu}_k}{\bar{\lambda}_k + t\bar{\mu}_k}\right).$$

Using (58) and (59) with the barred quantities, this yields

$$\beta_{rs}(t) = t^{-s}\,\mathrm{tr}\,P_t^r(I - P_t)^s,$$

$$\gamma_{rs}(t) = t^{-s-1}(Ny)^* P_t^r (I - P_t)^{s+1}(Ny).$$

Therefore, if $r + s > 0$, we get

$$f_{rs}(t) = \log(Ny)^* P_t^r (I - P_t)^{s+1}(Ny) - \frac{1 + r + s}{r + s}\log \mathrm{tr}\,P_t^r (I - P_t)^s - \frac{r}{r + s}\log t.$$

In the special case $r = 0$ and $s = 1$, the merit function becomes

$$f_{01}(t) = \log\|(I - P_t)Ny\|^2 - 2\log \mathrm{tr}(I - P_t) = \log(GCV/n),$$

where

$$GCV = \frac{\frac{1}{n}\|(I - P_t)Ny\|^2}{(\frac{1}{n}\,\mathrm{tr}(I - P_t))^2}$$

is the well-known *generalized cross validation* (GCV) formula of Craven & Wahba [6]. The need for P_t, generally a full matrix, causes computational difficulties for sparse matrices. However, the product $P_t Ny$ and the trace $\mathrm{tr}\,P_t$ can be computed at about three times the cost of a factorization of B_t, without forming P_t explicitly, thus making the formula useful as long as B_t has a sparse factorization; details can be found, e.g., in Sections 6.7.4 and 5.3.5 of Björck [4]. The case where $r = 1$ and $s = 0$ seems to be new but can be handled in the same way. It is unknown whether fast methods exist that permit in the sparse case the efficient evaluation of the merit functions with $r + s > 1$ or nonintegral r or s.

On the other hand, the limiting case $r = s = 0$ is even simpler. To write the formula in terms of the untransformed matrices, we assume for simplicity that $\bar{n} = n$, i.e., $m \geq n$. Then

$$\begin{aligned} \log\det B_t &= 2\log\det M + \log\det(D^*D + tE^*E) \\ &= 2\log\det M + \log\prod(D_{kk}^2 + tE_{kk}^2) \\ &= 2\log\det M + \sum\log(\bar{\lambda}_k + t\bar{\mu}_k), \end{aligned} \tag{82}$$

so that, up to an irrelevant constant,

$$f_{00}(t) = \log(Ny)^*(I - P_t)(Ny) - \log t + \frac{1}{n}\log\det B_t. \tag{83}$$

If $m = n$ and E is nonsingular, the expression (82) can also be written as

$$\log\det B_t = \mathrm{const} + n\log t - \log\det(I - P_t),$$

showing that, again up to an irrelevant constant, $f_{00}(t) = \log GML$, where

$$GML = \frac{(Ny)^*(I - P_t)(Ny)}{\sqrt[n]{\det(I - P_t)}}$$

is the generalized maximum likelihood formula of Wahba [33]. This also holds in the general case, but the expression given for GML must then be modified since one must take account of zero eigenvalues of $I - P_t$. No such modification is needed for (83).

Function values for the GML merit function (83) can be computed efficiently from a sparse Cholesky factorization

$$A^*V^{-1}A + tJ^*J = L_tL_t^*$$

as

$$f_{00}(t) = \log(y^*V^{-1}y - \|u_t\|^2) - \log t + \frac{2}{n}\log\det L_t, \tag{84}$$

where u_t is the solution of the triangular linear system

$$L_tu_t = A^*V^{-1}y. \tag{85}$$

Each function evaluation requires a new factorization; therefore an efficient univariate minimizer should be used. A suitable starting value for the minimization is $t = 0.01\,\mathrm{tr}\,A^*V^{-1}A/\,\mathrm{tr}\,J^*J$. Once a good regularization parameter t is determined, the solution $\hat{x}$ of the least squares problem (66) is found by completing (85) with a back substitution, solving the triangular linear system

$$L_t^*\hat{x} = u_t. \tag{86}$$

The evaluation of the GCV formula is more expensive; it requires, in addition to the factorization needed to compute P_tNy, significant additional work for the computation of $\mathrm{tr}\,P_t$. It is gratifying that GML, the computationally simplest formula was found in the previous section to be also the most efficient one.

More general regularization problems. In a number of applications, there are several qualitative constraints of different origin. If one formulates each such constraint as the condition that some linear expression $J_\nu x$ is assumed to be well-scaled and not too large, one may again take account of these constraints as penalty terms in the least squares problem. In analogy to (66), we get the solution as

$$B_t\hat{x} = A^*V^{-1}y, \tag{87}$$

but now

$$B_t = A^*V^{-1}A + \sum_\nu t_\nu^2 J_\nu^* J_\nu$$

contains several regularization parameters t_ν, one for each qualitative constraint. In some cases, one may assume that the t_ν are proportional to some known constants, thereby reducing the problem to one with a single regularization parameter (the proportionality factor). However, more frequently, no information is available about the relative size of the t_ν, and all regularization parameters must be determined simultaneously. The GML criterion generalizes in a natural way; (84)–(86) remain valid, but now L_t is a Cholesky factor of B_t, and the vector t of regularization parameters may be found by a multivariate minimization of $f_{00}(t)$. The more expensive GCV merit function extends in a similar way to this situation (Wahba [34]). See also Gu & Wahba [13].

Acknowledgment. Part of this research was done in 1994, when the author enjoyed a stimulating year at Bell Laboratories, Murray Hill, NJ. I also want to thank Waltraud Huyer for her careful reading of a draft, that lead to significant extensions of the results on merit functions for the regularization parameter.

REFERENCES

[1] A. B. Bakushinskii, *Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion*, USSR Comput. Math. Math. Phys. 24 (1984), pp. 181–182.
[2] O. E. Barndorff-Nielsen and D. R. Cox, *Inference and Asymptotics*, Chapman and Hall, London 1994.
[3] M. Bertero, C. de Mol and G. A. Viano, *The stability of inverse problems*, in Inverse Scattering in Optics, Baltes, ed., Springer 1980, pp. 161–214.
[4] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia 1996.
[5] H. Brakhage, *On ill-posed problems and the method of conjugate gradients*, in Inverse and Ill-Posed Problems, H. W. Engl and C. W. Groetsch, eds., Academic Press, Boston 1987, pp. 165–175.
[6] P. Craven and G. Wahba, *Smoothing noisy data with spline functions*: *Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math. 31 (1979), pp. 377–403.
[7] L. Eldeén, *Algorithms for the regularization of ill-conditioned least squares problems*, BIT 17 (1977), pp. 134–145.
[8] H. W. Engl, *Regularization methods for the stable solution of inverse problems*, Surveys Math. Indust. 3 (1993), pp. 71–143.
[9] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht 1996.
[10] P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, Academic Press, London 1981.
[11] G. H. Golub and C. F. van Loan, *Matrix Computations*, John Hopkins Univ. Press, Baltimore 1989.
[12] C. W. Groetsch, *Generalized Inverses of Linear Operators*, Dekker, New York 1977.
[13] Chong Gu and G. Wahba, *Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method*, SIAM J. Sci. Stat. Comput. 12 91991), pp. 383–398.
[14] M. Hanke, *Conjugate Gradient Type Methods for Ill-Posed Problems*, Pitman Res. Notes Math., Longman, Harlow, Essex, 1995.
[15] M. Hanke and M. P. C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust. 3 (1993), pp. 253–315.
[16] M. Hanke and T. Raus, *A general heuristic for choosing the regularization parameter in ill-posed problems*, SIAM J. Sci. Comput. 17 (1996), pp. 956–972.
[17] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Review 34 (1992), pp. 561–580.
[18] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion*, SIAM, Philadelphia 1997.
[19] A. K. Katsaggelos, *Digital Image Restoration*, Springer, Berlin 1991.
[20] L. Kaufman and A. Neumaier, *PET regularization by envelope guided conjugate gradients*, IEEE Trans. Medical Imag. 15 (1996), pp. 385–389.
[21] L. Kaufman and A. Neumaier, *Regularization of ill-posed problems by envelope guided conjugate gradients*, J. Comput. Graph. Stud., to appear.
[22] C. F. van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal. 13 (1976), pp. 76–83.
[23] C. F. van Loan, *Computing the CS and generalized singular value decomposition*, Numer. Math. 46 (1985), pp. 479–492.
[24] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal. 1 (1970), pp. 52–74.
[25] F. Natterer, *Error bounds for Tikhonov regularization in Hilbert scales*, Appl. Anal. 18 (1984), pp. 29–37.
[26] A. S. Nemirovski, *The regularization properties of the adjoint method in ill-posed problems*, USSR Comput. Math. Math. Phys. 26 (1986), pp. 7–16.
[27] C. C. Paige, *Computing the generalized singular value decomposition*, SIAM J. Sci. Stat. Comput. 7 (1986), pp. 1126–1146.

[28] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comp. Mach. 9 (1962), pp. 84–97.
[29] J. D. Riley, *Solving systems of linear equations with a positive definite symmetric but possibly ill-conditioned matrix*, Math. Tables Aids Comput. 9 (1956), pp. 96–101.
[30] G. W. Stewart, *A method for computing the generalized singular value decomposition*, in Matrix Pencils, B. Kagström and A. Ruhe, eds., Springer, New York 1983, pp. 207–220.
[31] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl. 4 (1963), pp. 1035–1038.
[32] C. R. Vogel and M.E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput. 17 (1996), pp. 227–238.
[33] G. Wahba, *A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem*, Ann. Statist. 13 (1985), pp. 1378–1402.
[34] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia 1990.
[35] A. H. Welsh, *On M-processes and M-estimation*, Ann. Statist. 17 (1990), pp. 337-361. [Correction, Ann. Statist. 18 (1990), p. 1500.]
[36] N. Wiener, *Cybernetics*, MIT Press, Cambridge, MA, 1948.
[37] G. M. Wing and J. D. Zahrt, *A Primer on Integral Equations of the First Kind*, SIAM, Philadelphia 1991.