Regularization of Least Squares Problems

Heinrich Voss

voss@tu-harburg.de

Hamburg University of Technology
Institute of Numerical Simulation

TUHH
Technische Universität Hamburg-Harburg

Outline

Well-posed / ill-posed problems

Back in 1923 Hadamard introduced the concept of well-posed and ill-posed problems.

A problem is well-posed, if

- it is solvable
- its solution is unique
- its solution depends continuously on system parameters
  (i.e. arbitrary small perturbation of the data can not cause arbitrary large perturbation of the solution)

Otherwise it is ill-posed.

According to Hadamard's philosophy, ill-posed problems are actually ill-posed, in the sense that the underlying model is wrong.

Ill-posed problems

Ill-posed problems often arise in the form of inverse problems in many areas of science and engineering.

Ill-posed problems arise quite naturally if one is interested in determining the internal structure of a physical system from the system's measured behavior, or in determining the unknown input that gives rise to a measured output signal.

Examples are

- computerized tomography, where the density inside a body is reconstructed from the loss of intensity at detectors when scanning the body with relatively thin X-ray beams, and thus tumors or other anomalies are detected.
- solving diffusion equations in negative time direction to detect the source of pollution from measurements

Further examples appear in acoustics, astrometry, electromagnetic scattering, geophysics, optics, image restoration, signal processing, and others.

Least Squares Problems

Let

$$\|Ax - b\| = \min! \quad \text{where } A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m, m \geq n. \tag{1}$$

Differentiating

$$\varphi(x) = \|Ax - b\|_2^2 = (Ax - b)^T(Ax - b) \tag{2}$$

yields the necessary condition

$$A^T Ax = A^T b. \tag{3}$$

called normal equations.

If the columns of A are linearly independent, then $A^T A$ is positive definite, i.e. φ is strictly convex and the solution is unique.

Geometrically, x^* is a solution of (1) if and only if the residual $r := b - Ax$ at x^* is orthogonal to the range of A,

$$b - Ax^* \perp \mathcal{R}(A). \tag{4}$$

Solving LS problems

If the columns of A are linearly independent, the solution x^* can be obtained solving the normal equation by the Cholesky factorization of $A^TA > 0$.

However, A^TA may be badly conditioned, and then the solution obtained this way can be useless.

In finite arithmetic the QR-decomposition of A is a more stable approach.

If $A = QR$, where $Q \in \mathbb{R}^{m\times m}$ is orthogonal, $R = \begin{bmatrix}\tilde{R}\\ 0\end{bmatrix}$, $\tilde{R} \in \mathbb{R}^{n\times n}$ upper triangular, then

$$\|Ax-b\|_2 = \|Q(Rx-Q^Tb)\|_2 = \left\|\begin{bmatrix}\tilde{R}x-\beta_1\\ -\beta_2\end{bmatrix}\right\|_2, \quad Q^Tb = \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix},$$

and the unique solution of (1) is

$$x^* = \tilde{R}^{-1}\beta_1.$$

Singular value decomposition

A powerful tool for the analysis of the least squares problem is the singular value decomposition (SVD) of A:

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^T \tag{5}$$

with orthogonal matrices $\tilde{U} \in \mathbb{R}^{m\times m}$, $\tilde{V} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\tilde{\Sigma} \in \mathbb{R}^{m\times n}$.

A more compact form of the SVD is

$$A = U\Sigma V^T \tag{6}$$

with the matrix $U \in \mathbb{R}^{m\times n}$ having orthonormal columns, an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Sigma \in \mathbb{R}^{n\times n} = diag(\sigma_1, \ldots, \sigma_n)$.

It is common understanding that the columns of U and V are ordered and scaled such that $\sigma_j \geq 0$ are nonnegative and are ordered by magnitude:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

σ_i, $i = 1, \ldots, n$ are the singular values of A, the columns of U are the left singular vectors and the columns of V are the right singular vectors of A.

Solving LS problems cnt.

With $y := V^T x$ and $c := U^T b$ it holds

$$\|Ax - b\|_2 = \|U\Sigma V^T x - b\|_2 = \|\Sigma y - c\|_2.$$

For $\text{rank}(A) = r$ it follows

$$y_j = \frac{c_j}{\sigma_j},\ j = 1, \ldots, r \quad \text{and} \quad y_j \in \mathbb{R} \text{ arbitrary for } j > r.$$

Hence,

$$x = \sum_{j=1}^{r} \frac{u_j^T b}{\sigma_j} v_j + \sum_{j=r+1}^{n} \gamma_j v_j,\ \gamma_j \in \mathbb{R}.$$

Since $v_{r+1}, \ldots, v_n$ span the kernel $\mathcal{N}(A)$ of A, the solution set of (1) is

$$L = x_{LS} + \mathcal{N}(A) \tag{7}$$

where

$$x_{LS} := \sum_{j=1}^{r} \frac{u_j^T b}{\sigma_j} v_j$$

is the solution with minimal norm called minimum norm or pseudo normal solution of (1).

Pseudoinverse

For fixed $A \in \mathbb{R}^{m\times n}$ the mapping that maps a vector $b \in \mathbb{R}^m$ to the minimum norm solution x_{LS} of $\|Ax - b\| = \min!$ obviously is linear, and therefore is represented by a matrix $A^\dagger \in \mathbb{R}^{n\times m}$.

$A^\dagger$ is called pseudo inverse or generalized inverse or Moore-Penrose inverse of A.

If A has full rank n, then $A^\dagger = (A^TA)^{-1}A^T$ (follows from the normal equations), and if A is quadratic and nonsingular then $A^\dagger = A^{-1}$.

For general $A = U\Sigma V^T$ it follows from the representation of x_{LS} that

$$A^\dagger = V\Sigma^\dagger U^T, \quad \Sigma^\dagger = \mathrm{diag}\{\tau_i\},\ \tau_i = \begin{cases} 1/\sigma_i & \text{if} \quad \sigma_i > 0 \\ 0 & \text{if} \quad \sigma_i = 0 \end{cases}$$

Perturbation Theorem

Let the matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$ have full rank, let x be the unique solution of the least squares problem (1), and let $\tilde{x}$ be the solution of a perturbed least squares problem

$$\|(A+\delta A)x - (b+\delta b)\| = \min! \tag{8}$$

where the perturbation is not too large in the sense

$$\epsilon := \max\left(\frac{\|\delta A\|}{\|A\|}, \frac{\|\delta b\|}{\|b\|}\right) < \frac{1}{\kappa_2(A)} \tag{9}$$

where $\kappa_2(A) := \sigma_1/\sigma_n$ denotes the condition number of A.

Then it holds that

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leq \epsilon\left(\frac{2\kappa_2(A)}{\cos(\theta)} + \tan(\theta)\cdot\kappa_2^2(A)\right) + \mathcal{O}(\epsilon^2) \tag{10}$$

where θ is the angle between b and its projection onto $\mathcal{R}(A)$.

For a proof see the book of J. Demmel, Applied Linear Algebra.

Ill-conditioned problems

In this talk we consider ill-conditioned problems (with large condition numbers), where small perturbations in the data A and b lead to large changes of the least squares solution x_{LS}.

When the system is not consistent, i.e. it holds that $r = b - Ax_{LS} \neq 0$, then in equation (10) it holds that $\tan(\theta) \neq 0$ which means that the relative error of the least squares solution is roughly proportional to the square of the condition number $\kappa_2(A)$.

When doing calculations in finite precision arithmetic the meaning of 'large' is with respect to the reciprocal of the machine precision.
A large $\kappa_2(A)$ then leads to an unstable behavior of the computed least squares solution, i.e. in this case the solution $\tilde{x}$ typically is physically meaningless.

A toy problem

Consider the problem to determine the orthogonal projection of a given function $f : [0,1] \to \mathbb{R}$ to the space Π_{n-1} of polynomials of degree $n-1$ with respect to the scalar product

$$\langle f, g \rangle := \int_0^1 f(x)g(x)\,dx.$$

Choosing the (unfeasible) monomial basis $\{1, x, \ldots, x^{n-1}\}$ this leads to the linear system

$$Ay = b \qquad (1)$$

where

$$A = (a_{ij})_{i,j=1,\ldots,n},\ a_{ij} := \frac{1}{i+j-1}, \qquad (2)$$

is the so called Hilbert matrix, and $b \in \mathbb{R}^n$, $b_i := \langle f, x^{i-1} \rangle$.

A toy problem cnt.

For dimensions $n = 10$, $n = 20$ and $n = 40$ we choose the right hand side b such that $y = (1, \ldots, 1)^T$ is the unique solution.

Solving the problem with LU-factorization (in MATLAB $A\backslash b$), the Cholesky decomposition, the QR factorization of A and the singular value decomposition of A we obtain the following errors in Euclidean norm:

	$n = 10$	$n = 20$	$n = 40$
LU factorization	5.24 E-4	8.25 E+1	3.78 E+2
Cholesky	7.07 E-4	numer. not pos. def.	
QR decomposition	1.79 E-3	1.84 E+2	7.48 E+3
SVD	1.23 E-5	9.60 E+1	1.05 E+3
$\kappa(A)$	1.6 E+13	1.8 E+18	9.8 E+18 (?)

A toy problem cnt.

A similar behavior is observed for the least squares problem. For $n = 10$, $n = 20$ and $n = 40$ and $m = n + 10$ consider the least squares problem

$$\|Ax - b\|_2 = \min!$$

where $A \in \mathbb{R}^{m \times n}$ is the Hilbert matrix, and b is chosen such hat $x = (1, \ldots, 1)^T$ is the solution with residual $b - Ax = 0$.

The following table contains the errors in Euclidean norm for the solution of the normal equations solved with LU factorization (Cholesky yields the message 'matrix numerically not positive definite' already $n = 10$), the solution with QR factorization of A, and the singular value decomposition of A.

	$n = 10$	$n = 20$	$n = 40$
Normalgleichungen	7.02 E-1	2.83 E+1	7.88 E+1
QR Zerlegung	1.79 E-5	5.04 E+0	1.08 E+1
SVD	2.78 E-5	2.93 E-3	7.78 E-4
$\kappa(A)$	2.6 E+11	5.7 E+17	1.2 E+18 (?)

Rank deficient problems

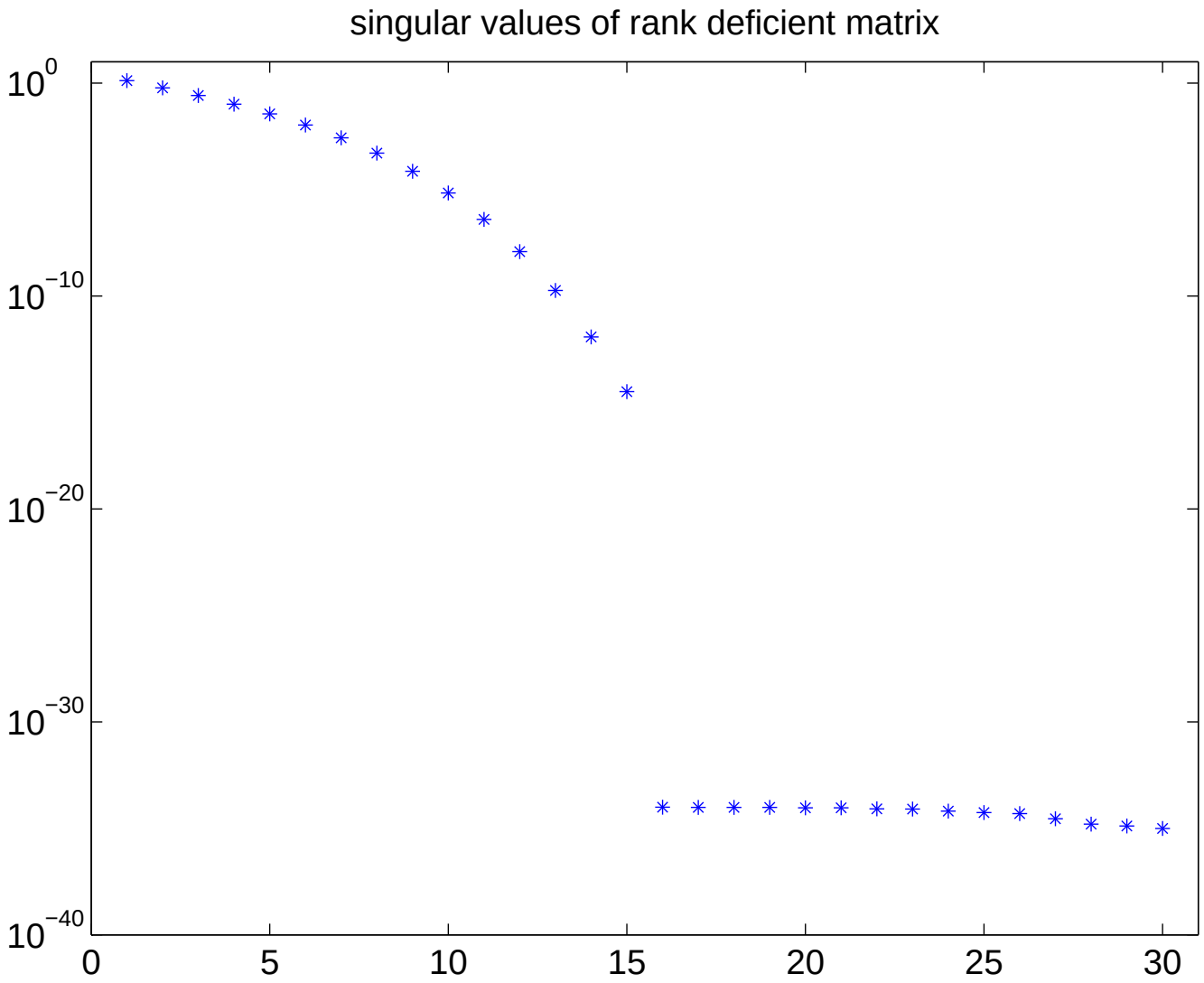
In rank-deficient problems there exist a well-determined gap between large and small singular values of A.

Often the exact rank of the matrix $A \in \mathbb{R}^{m\times n}$ is equal to n, with a cluster of very small singular values that are almost zero. This cluster can simply result from an overdetermined linear system that is nearly consistent. The number of singular values that do not correspond to the cluster at zero is denoted as the numerical rank of A.

Methods for solving rank-deficient problems try to determine the numerical rank of A, which itself is a well-conditioned problem. If the numerical rank is known, it is usually possible to eliminate the ill-conditioning. The problem is then reduced to a well-conditioned LS problem of smaller size.

In this talk we concentrate on discrete ill-posed problems

Rank deficient problems cnt.



Fredholm integral equation of the first kind

Famous representatives of ill-posed problems are Fredholm integral equations of the first kind that are almost always ill-posed.

$$\int_\Omega K(s,t)f(t)dt = g(s), \quad s \in \Omega \tag{11}$$

with a given *kernel* function $K \in L^2(\Omega^2)$ and right-hand side function $g \in L^2(\Omega)$.

Then with the singular value expansion

$$K(s,t) = \sum_{j=1}^{\infty} \mu_j u_j(s) v_j(s), \ \mu_1 \geq \mu_2 \geq \cdots \geq 0$$

a solution of (11) can be expressed as

$$f(t) = \sum_{j=1}^{\infty} \frac{\langle u_j, g\rangle}{\mu_j} v_j(t), \quad \langle u_j, g\rangle = \int_\Omega u_j(s)g(s)\,ds.$$

Fredholm integral equation of the first kind cnt.

The solution f is square integrable if the right hand side g satisfies the Picard condition

$$\sum_{j=1}^{\infty} \left(\frac{\langle u_j, g \rangle}{\mu_j} \right)^2 < \infty.$$

The Picard condition says that from some index j on the absolute value of the coefficients $\langle u_j, g \rangle$ must decay faster than the corresponding singular values μ_j in order that a square integrable solution exists.

For g to be square integrable the coefficients $\langle u_j, g \rangle$ must decay faster than $1/\sqrt{j}$, but the Picard condition puts a stronger requirement on g: the coefficients must decay faster than $\mu_j/\sqrt{j}$.

Discrete ill-posed problems

Discretizing a Fredholm integral equation results in discrete ill-posed problems

$$Ax = b$$

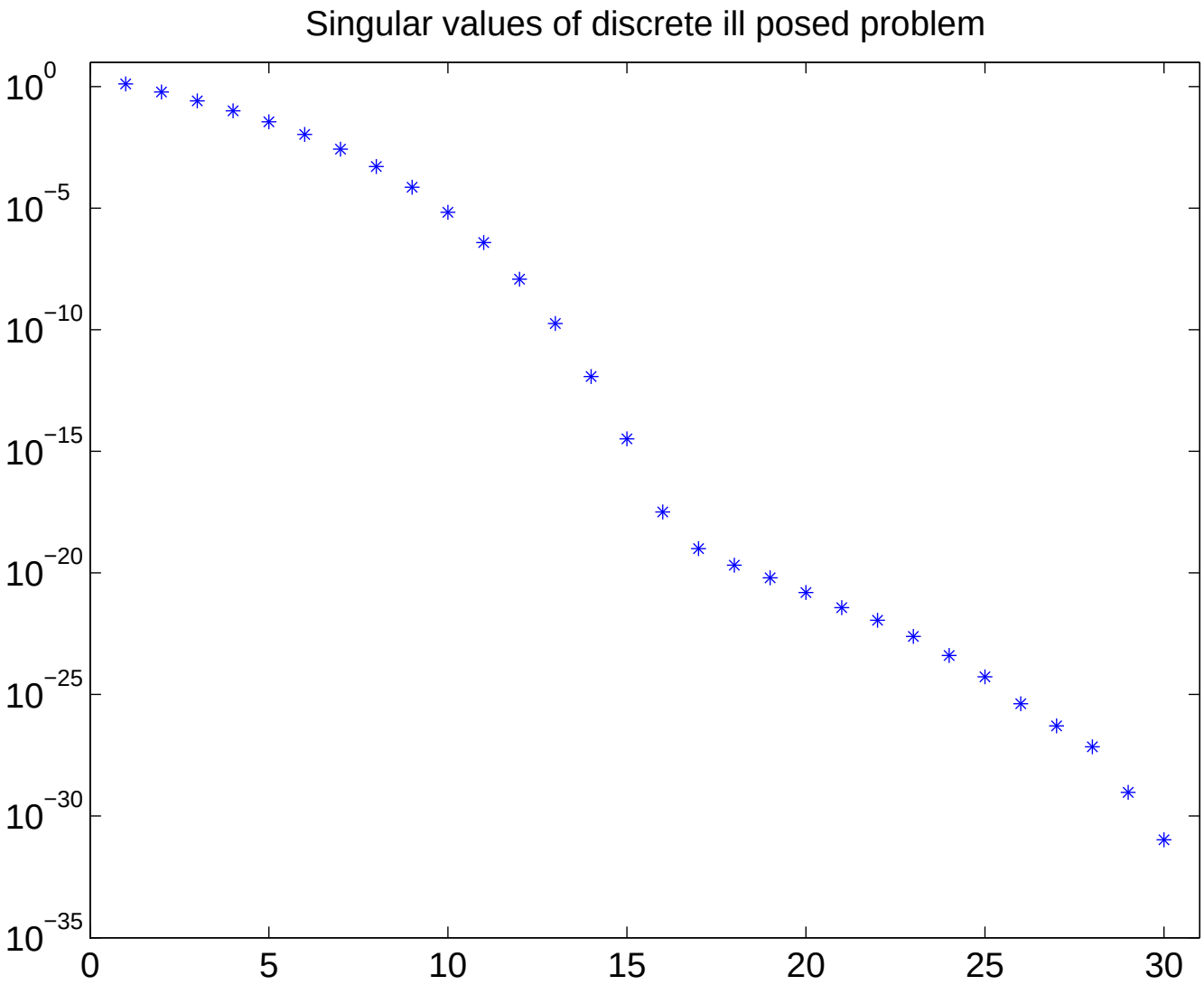
.

The matrix A inherits the following properties from the continuous problem (11): it is ill-conditioned with singular values gradually decaying to zero.

This is the main difference to rank-deficient problems.

> *Discrete ill-posed problems have an ill-determined rank, i.e. there does not exist a gap in the singular values that could be used as a natural threshold.*

Discrete ill-posed problems cnt.



Discrete ill-posed problems cnt.

When the continuous problem satisfies the Picard condition, then the absolute values of the Fourier coefficients $u_i^T b$ decay gradually to zero with increasing i, where u_i is the ith left singular vector obtained from the SVD of A.

Typically the number of sign changes of the components of the singular vectors u_i and v_i increases with the index i, this means that low-frequency components correspond to large singular values and the smaller singular values correspond to singular vectors with many oscillations.

The Picard condition translates to the following discrete Picard condition:

> *With increasing index i, the coefficients* $|u_i^T b|$ *on average decay faster to zero than* σ_i.

Discrete ill-posed problems cnt.

The typical situation in least squares problems is the following:
Instead of the exact right-hand side b a vector $\tilde{b} = b + \varepsilon s$ with small $\varepsilon > 0$ and random noise vector s is given. The perturbation results from measurement or discretization errors.

The goal is to recover the solution x_{true} of the underlying consistent system

$$Ax_{true} = b \tag{12}$$

from the system $Ax \approx \tilde{b}$, i.e. by solving the least squares problem

$$\|\Delta b\| = \text{min!} \quad \text{subject to } Ax = \tilde{b} + \Delta b. \tag{13}$$

For the solution it holds

$$\tilde{x}_{LS} = A^{\dagger}\tilde{b} = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^{r} \frac{u_i^T s}{\sigma_i} v_i \tag{14}$$

where r is the rank of A.

Discrete ill-posed problems cnt.

The solution consists of two terms, the first one is the true solution x_{true} and the second term is the contribution from the noise.

If the vector s consists of uncorrelated noise, the parts of s into the directions of the left singular vectors stay roughly constant, i.e. $u_i^T s$ will not vary much for all i. Hence the second term $u_i^T s/\sigma_i$ blows up with increasing i.

The first term contains the parts of the exact right-hand side b developed into the directions of the left singular vectors, i.e. the Fourier coefficients $u_i^T b$.

If the discrete Picard condition is satisfied, then $\tilde{x}_{LS}$ is dominated by the influence of the noise, i.e. the solution will mainly consist of a linear combination of right singular vectors corresponding to the smallest singular values of A.

Regularization

Assume A has full rank . Then a regularized solution can be written in the form

$$x_{reg} = V\Theta\Sigma^{\dagger}U^T\tilde{b} = \sum_{i=1}^{n} f_i \frac{u_i^T\tilde{b}}{\sigma_i} v_i = \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^{n} f_i \frac{u_i^T s}{\sigma_i} v_i. \qquad (15)$$

Here the matrix $\Theta \in \mathbb{R}^{n\times n}$ is a diagonal matrix, with the so called filter factors f_i on its diagonal.

A suitable regularization method adjusts the filter factors in such a way that the unwanted components of the SVD are damped whereas the wanted components remain essentially unchanged.

Most regularization methods are much more efficient when the discrete Picard condition is satisfied. But also when this condition does not hold the methods perform well in general.

Truncated SVD

One of the simplest regularization methods is the truncated singular value decomposition (TSVD). In the TSVD method the matrix A is replaced by its best rank-k approximation, measured in the 2-norm or the Frobenius norm

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \quad \text{with } \|A - A_k\|_2 = \sigma_{k+1}. \tag{16}$$

The approximate solution x_k for problem (13) is then given by

$$x_k = A_k^\dagger \tilde{b} = \sum_{i=1}^{k} \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^{k} \frac{u_i^T s}{\sigma_i} v_i \tag{17}$$

or in terms of the filter coefficients we simply have the regularized solution (15) with

$$f_i = \begin{cases} 1 & \text{for } i \le k \\ 0 & \text{for } i > k \end{cases} \tag{18}$$

Truncated SVD cnt.

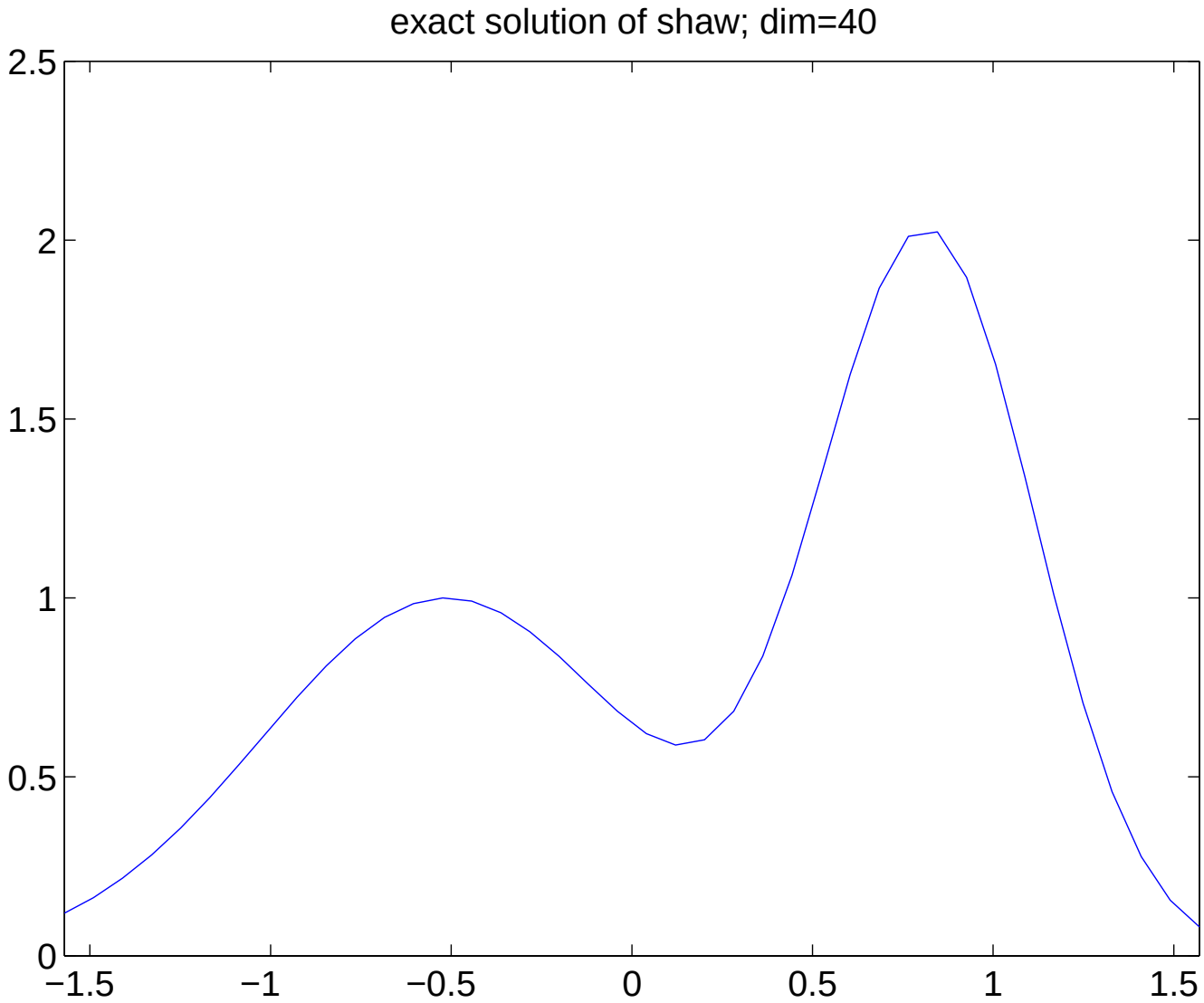
The solution x_k does not contain any high frequency components, i.e. all singular values starting from the index $k+1$ are set to zero and the corresponding singular vectors are disregarded in the solution. So the term $u_i^T s/\sigma_i$ in equation (17) corresponding to the noise s is prevented from blowing up.

The TSVD method is particularly suitable for rank-deficient problems. When k reaches the numerical rank r of A the ideal approximation x_r is found.

For discrete ill-posed problems the TSVD method can be applied as well, although the cut off filtering strategy is not the best choice when facing gradually decaying singular values of A.

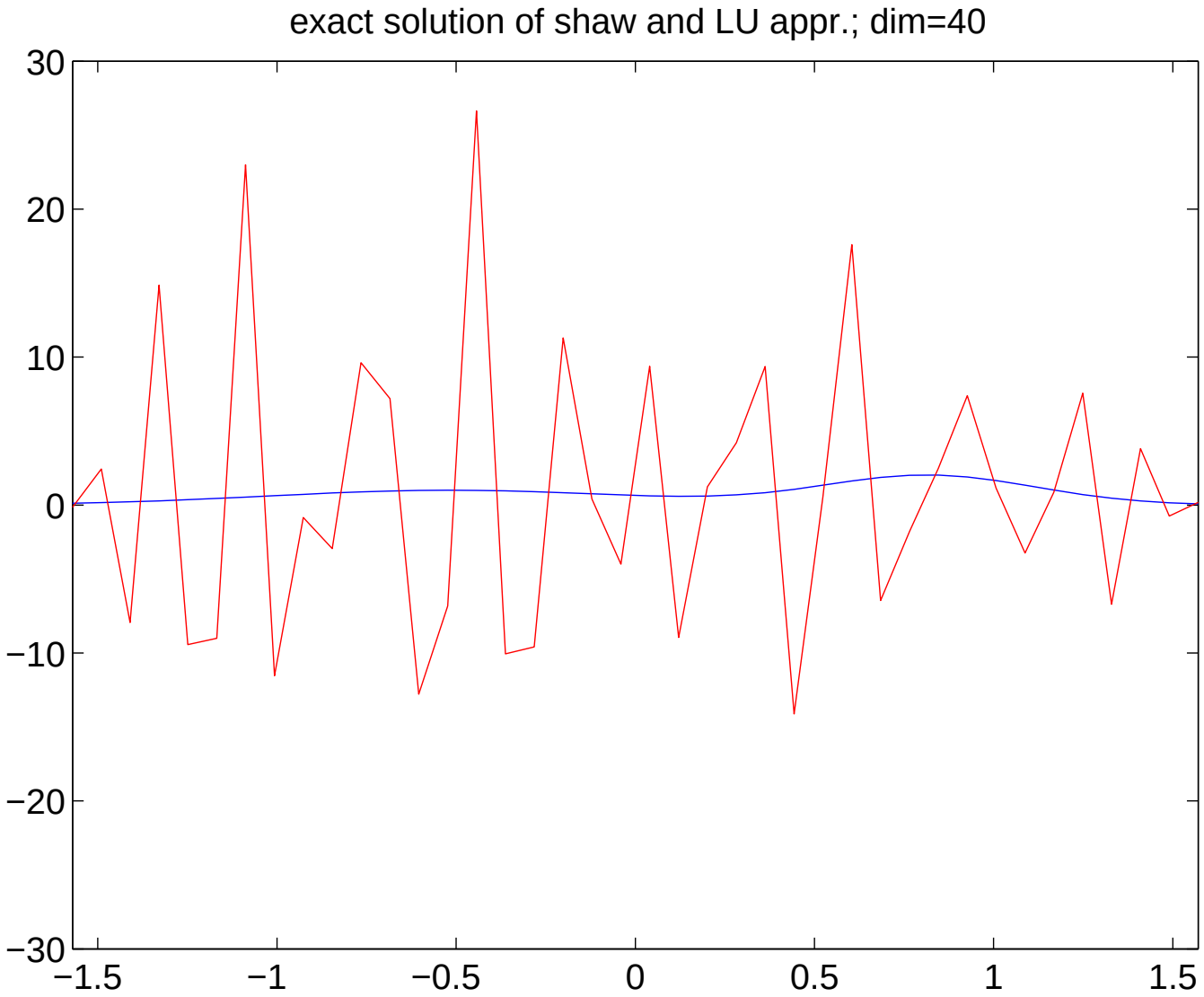
Example

Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40



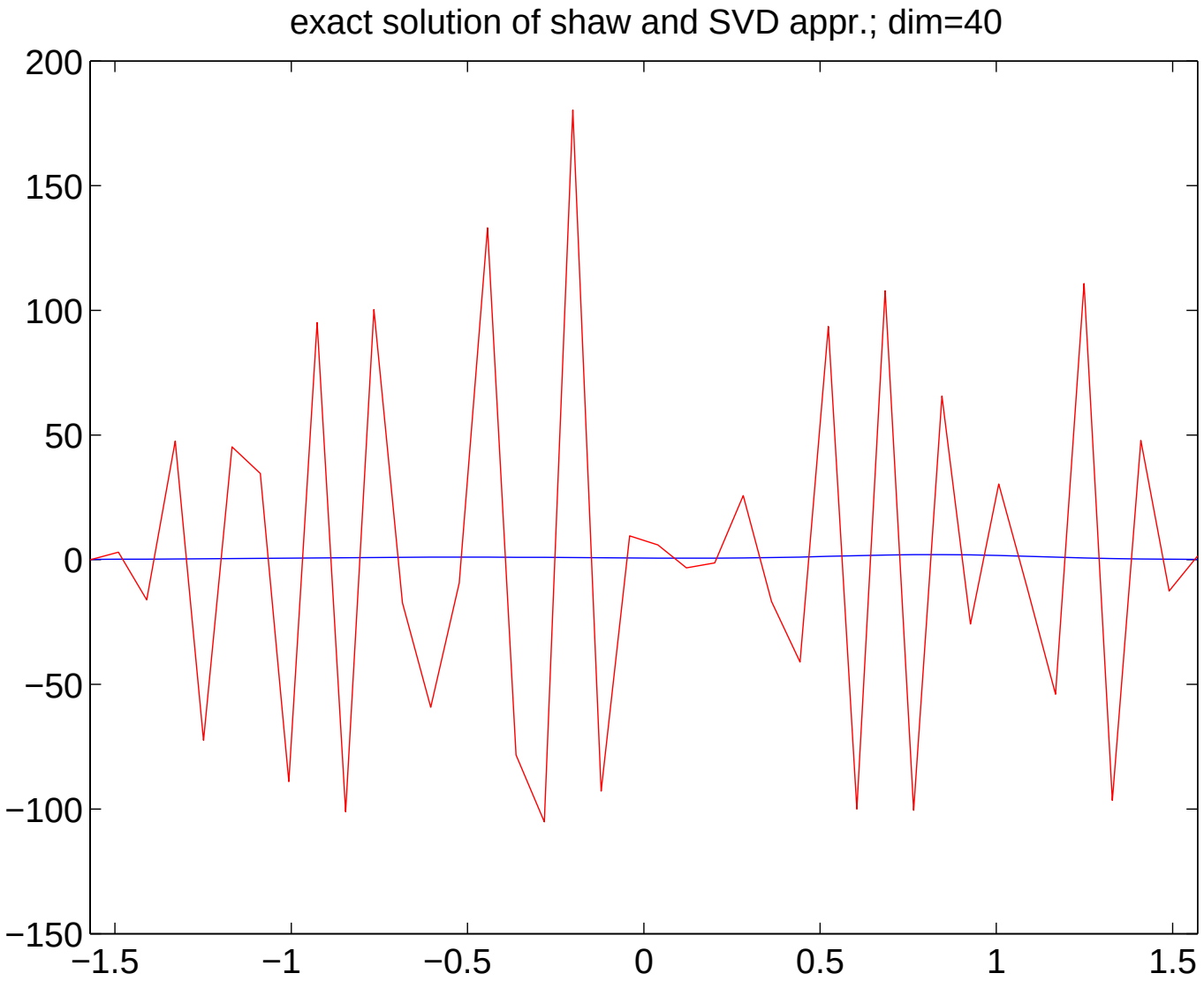
Example

Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40 and its approximation via LU factorization



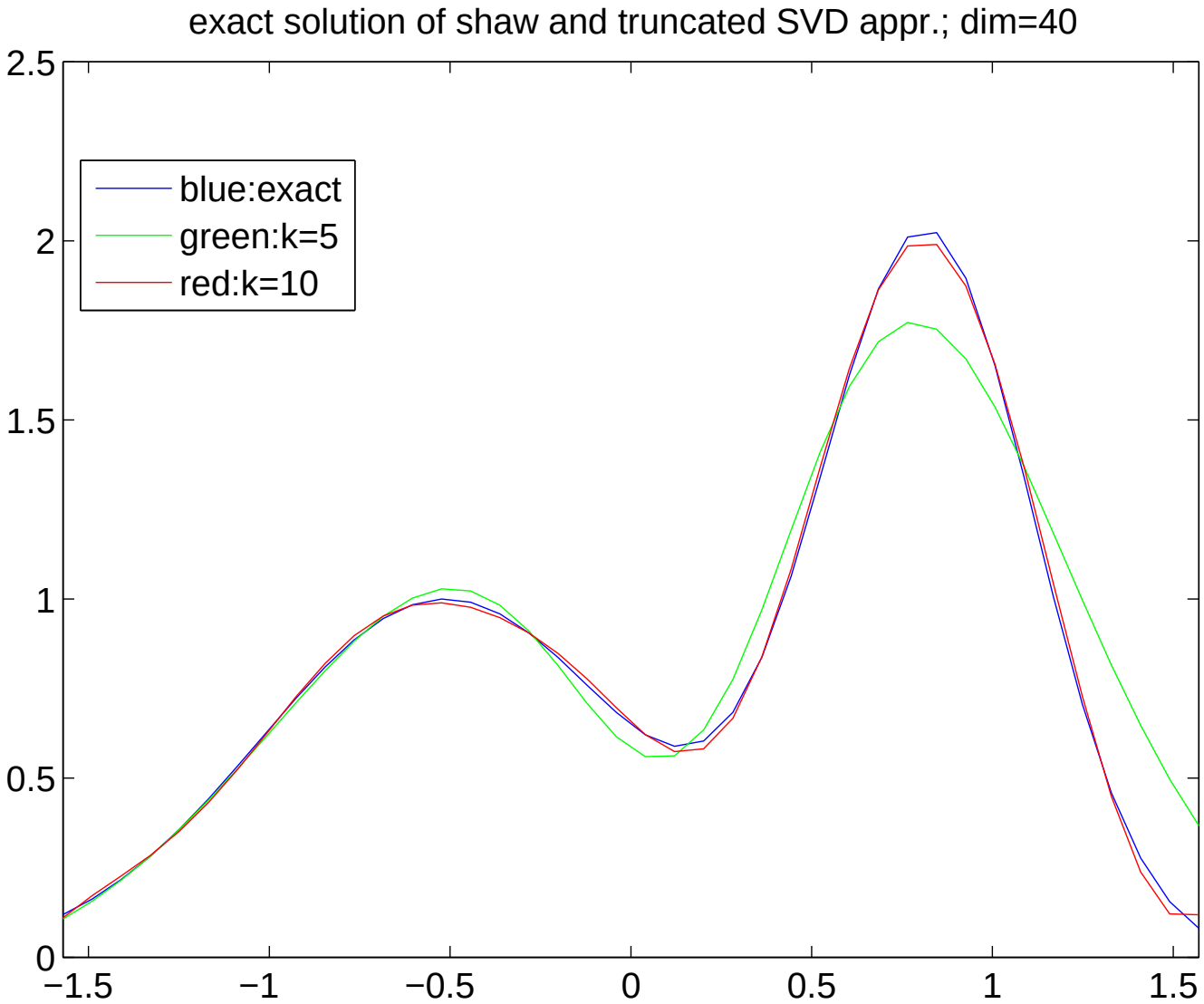
Example

Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40 and its approximation via complete SVD



Example

Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40 and its approximation via truncated SVD



Tikhonov regularization

In Tikhonov regularization (introduced independently by Tikhonov (1963) and Phillips (1962)) the approximate solution x_λ is defined as minimizer of the quadratic functional

$$\|Ax - \tilde{b}\|^2 + \lambda\|Lx\|^2 = \min! \tag{19}$$

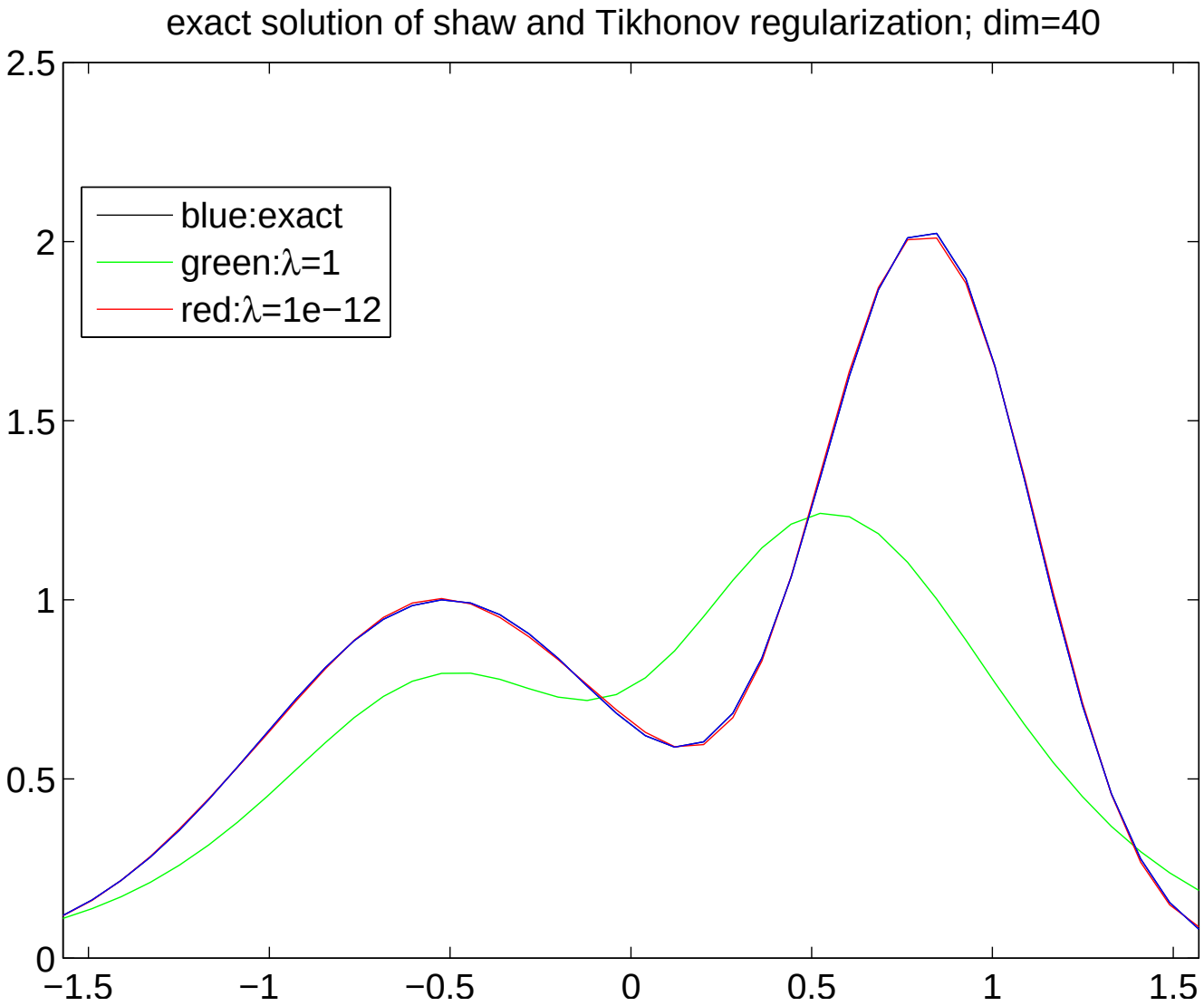
The basic idea of Tikhonov regularization is the following: Minimizing the functional in (19) means to search for some x_λ, providing at the same time a small residual $\|Ax_\lambda - \tilde{b}\|$ and a moderate value of the penalty function $\|Lx_\lambda\|$.

If the regularization parameter λ is chosen too small, (19) is too close to the original problem and instabilities have to be expected.

If λ is chosen too large, the problem we solve has only little connection with the original problem. Finding the optimal parameter is a tough problem.

Example

Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40 and its approximation via Tikhonov regularization



Tikhonov regularization cnt.

Tikhonov regularization has an important equivalent formulation as

$$\min \|Ax - b\| \quad \text{subject to } \|Lx\| \leq \delta \tag{20}$$

where δ is a positive constant.

(20) is a linear LS problem with a quadratic constraint, and using the Lagrange multiplier formulation

$$\mathcal{L}(x, \lambda) = \|Ax - b\|^2 + \lambda(\|Lx\|^2 - \delta^2), \tag{21}$$

it can be shown that if $\delta \leq \|x_{LS}\|$, where x_{LS} denotes the LS solution of $\|Ax - b\| = \min!$, then the solution x_δ of (20) is identical to the solution x_λ of (19) for an appropriately chosen λ, and there is a monotonic relation between the parameters δ and λ.

Tikhonov regularization cnt.

Problem (19) can be also expressed as an ordinary least squares problem:

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} L \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \tag{22}$$

with the normal equations

$$(A^T A + \lambda L^T L)x = A^T \tilde{b}. \tag{23}$$

Let the matrix $A_\lambda := [A^T, \sqrt{\lambda} L^T]^T$ have full rank, then a unique solution exists.

For $L = I$ (which is called the standard case) the solution $x_\lambda = x_{reg}$ of (23) is

$$x_\lambda = V \Theta \Sigma^\dagger U^T \tilde{b} = \sum_{i=1}^{n} f_i \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^{n} \frac{\sigma_i (u_i^T \tilde{b})}{\sigma_i^2 + \lambda} v_i \tag{24}$$

where $A = U \Sigma V^T$ is the SVD of A. Hence, the filter factors are

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \quad \text{for} \quad L = I. \tag{25}$$

For $L \neq I$ a similar representation holds with the generalized SVD of (A, L).

Tikhonov regularization cnt.

For singular values much larger than λ the filter factors are $f_i \approx 1$ whereas for singular values much smaller than λ it holds that $f_i \approx \sigma_i^2/\lambda \approx 0$.

The same holds for $L \neq I$ with replacing σ_i by the generalized singular values γ_i.

Hence, Tikhonov regularization is damping the influence of the singular vectors corresponding to small singular values (i.e. the influence of highly oscillating singular vectors).

Tikhonov regularization exhibits much smoother filter factors than truncated SVD which is favorable for discrete ill-posed problems.

Generalized Singular Value Decomposition

The generalized singular value decomposition (GSVD for short) is a generalization of the SVD for matrix pairs (A, L).

The GSVD was introduced by Van Loan (1976) for analysis of matrix pencils $A^TA + \lambda L^TL$, $\lambda \in \mathbb{R}$, and he showed that (similar to the ordinary case) the generalized singular values are the square roots of the eigenvalues of (A^TA, L^TL). The use of the GSVD in the analysis of discrete ill-posed problems goes back to Varah (1979).

For the GSVD the number of columns of both matrices have to be identical, and typically it holds that $A \in \mathbb{R}^{m\times n}$ and $L \in \mathbb{R}^{p\times n}$ with $m \geq n \geq p$.

It is assumed that the regularization matrix L has full rank and that the kernels do not intersect, i.e. $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$.

Under these conditions the following decomposition exists:

Generalized Singular Value Decomposition cnt.

With orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$ and a nonsingular matrix $X \in \mathbb{R}^{n\times n}$

$$A = U\Sigma_L X^{-1} \quad \text{and} \quad L = V[M, 0]X^{-1} \tag{26}$$

where

$$\Sigma_L = diag(\sigma_1, \ldots, \sigma_p, 1, \ldots, 1) \in \mathbb{R}^{m\times n} \quad \text{and} \quad M = diag(\mu_1, \ldots, \mu_p) \in \mathbb{R}^{p\times p} \tag{27}$$

and it holds that

$$0 \le \sigma_1 \le \cdots \le \sigma_p \le 1 \quad \text{and} \quad 1 \ge \mu_1 \ge \cdots \ge \mu_p > 0 \tag{28}$$

$$\sigma_i^2 + \mu_i^2 = 1 \quad \text{for } i = 1, \ldots, p. \tag{29}$$

The ratios $\gamma_i = \sigma_i/\mu_i, i = 1, \ldots, p$ are termed generalized singular values of (A, L).

For $L = I_n$ the generalized singular values γ_i are identical to the usual singular values of A, but one should be aware of the different ordering, i.e. the γ_i are monotonically increasing with their index.

CS Decomposition

For $m \geq n \geq p$ let $Q_1 \in \mathbb{R}^{m\times n}$ and $Q_2 \in \mathbb{R}^{p\times n}$ such that $Q := \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$ has orthonormal columns.

Then there exist orthogonal matrices $U_1 \in \mathbb{R}^{m\times m}$, $U_2 \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{n\times n}$ such that

$$\begin{pmatrix} U_1^T & O \\ O & U_2^T \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V = \begin{pmatrix} U_1^T Q_1 V \\ U_2^T Q_2 V \end{pmatrix} = \begin{pmatrix} \Sigma_1 \\ \Sigma_2 \end{pmatrix}$$

and

$$\begin{pmatrix} \Sigma_1 \\ \Sigma_2 \end{pmatrix} = \begin{pmatrix} C & O \\ O & I_{n-p,n-p} \\ O & O \\ S & O \end{pmatrix}.$$

C and S are diagonal $p \times p$-matrices with non-negative entries such that $C^2 + S^2 = I$.

Proof

Consider first the case $m = n = p$. Let $Q_1 = U_1 C V^T$ the SVD of Q_1. Then

$$\begin{pmatrix} C \\ \tilde{Q}_2 \end{pmatrix} = \begin{pmatrix} U_1^T & O \\ O & I \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V$$

has orthonormal columns.

Hence, $C^2 + \tilde{Q}_2^T \tilde{Q}_2 = I$ which implies that

$$\tilde{Q}_2^T \tilde{Q}_2 = I - C^2$$

is diagonal.

This means that the columns $\tilde{q}_j^{(2)}$ are orthogonal, and the desired matrix U_2 can be constructed as follows: For each j such that $\tilde{q}_j^{(2)} \neq 0$ set $u_j^{(2)} = \tilde{q}_j^{(2)} / \|\tilde{q}_j^{(2)}\|$, and fill the remaining columns of U_2 with an orthonormal basis of the orthogonal complement of the column space of $\tilde{Q}_2$.

Proof cnt.

From the orthogonality of the columns of Q we have that U_2 is orthogonal, and

$$U_2^T \tilde{Q} U_2 = S \quad \text{where } S = \text{diag}\{s_1, \ldots, s_p\}$$

with $s_j = \|\tilde{q}_j^{(2)}\|$, if $\tilde{q}_j^{(2)} \neq 0$ and $s_j = 0$ otherwise.

This completes the proof for $m = n = p$.

For $m \geq n \geq p$ let

$$Q_2 V_1 = \begin{pmatrix} \tilde{Q}_{21} & O \end{pmatrix}, \quad \tilde{Q}_{21} \in \mathbb{R}^{p \times p}$$

be the RQ factorization of Q_2. Then

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V_1 = \begin{pmatrix} \hat{Q}_{11} & \hat{Q}_{12} \\ \tilde{Q}_{21} & O \end{pmatrix},$$

and the by the orthogonality

$$\hat{Q}_{11}^T \hat{Q}_{12} = O, \quad \hat{Q}_{12}^T \hat{Q}_{12} = I.$$

Proof cnt.

Hence, if

$$\hat{U}_1 = \begin{pmatrix} \hat{U}_{11} & \hat{Q}_{12} \end{pmatrix}$$

is orthogonal, then

$$\begin{pmatrix} \hat{U}_1^T & O \\ O & I \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} V_1 = \begin{pmatrix} \hat{U}_{11}^T \hat{Q}_{11} & O \\ O & I \\ \tilde{Q}_{21} & 0 \end{pmatrix}.$$

Since $m + p - n \geq p$ we may apply the QR factorization to $\hat{U}_{11}^T \hat{Q}_{12}$ to obtain

$$\tilde{U}_1^T (\hat{U}_{11}^T \hat{Q}_{12}) = \begin{pmatrix} \tilde{Q}_{11} \\ O \end{pmatrix}.$$

Thus, the problem is reduced to computing the CS decomposition of the square $p \times p$-matrices $\tilde{Q}_{11}$ and $\tilde{Q}_{21}$, which completes the proof.

Existence of GSVD

Let

$$\begin{pmatrix} A \\ L \end{pmatrix} = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} R$$

be the QR factorization with $Q_1 \in \mathbb{R}^{m\times n}$, $Q_2 \in \mathbb{R}^{p\times n}$ and $R \in \mathbb{R}^{n\times n}$.

The CS factorization of $\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$ yields the existence of orthogonal matrices U, V and W such that

$$Q_1 = U\Sigma_L W^T \quad \text{and} \quad Q_2 = V \begin{pmatrix} M & O \end{pmatrix} W^T$$

where $M = \text{diag}\{\mu_1, \ldots, \mu_p\}$ and the columns of W^T can be ordered such that $\mu_1 \geq \mu_2 \geq \cdots \geq, \mu_p \geq 0$, and $\Sigma_L = \text{diag}\{\sigma_1, \ldots, \sigma_p, 1, \ldots, 1\}$. From $\sigma_j^2 + \mu_j^2 = 1$ it follows that $0 \leq \sigma_1 \leq \cdots \leq \sigma_p$.

Hence, $A = Q_1 R = U\Sigma_L W^T R$, $L = V \begin{pmatrix} M & O \end{pmatrix} W^T R$, and $X = (W^T R)^{-1}$.

The invertibility of R follows from our assumption that $\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}$.

GSVD for discrete ill-posed problems

For general matrix pairs (A, L) little can be said about the appearance of the GSVD. However, for discrete ill-posed problems the following three characteristic features are often found:

- The generalized singular values σ_i decay towards 0 (for $i = p, p-1, \ldots, 1$). There is no typical gap in the spectrum separating large generalized singular values from small ones.

- The singular vectors u_i, v_i, x_i often have more sign changes in their elements as the corresponding σ_i decrease. Hence, singular vectors corresponding to tiny singular values represent highly oscillating functions, whereas those corresponding to large singular values represent smooth functions. The last $n-p$ columns of X usually have very few sign changes.

- The Fourier coefficient of the $u_i^T b$ of the right hand side on the average decay faster to zero than the generalized singular values.

Generalized Singular Value Decomposition cnt.

Let $U = [U_p, U_{m-p}]$ and $X = [X_p, X_{n-p}]$ where the index indicates the number of columns of the corresponding matrix.

Then the columns of X_{n-p} span the kernel of L, i.e. $\mathcal{N}(L) = \text{span}\{X_{n-p}\}$.

The pseudoinverse of A can be written by means of the GSVD of (A, L) as

$$A_L^\dagger := [X_p, X_{n-p}] \begin{bmatrix} \Sigma_L^\dagger & 0 \\ 0 & I_{n-p} \end{bmatrix} [U_p, U_{n-p}]^T = X_p \Sigma_L^\dagger U_p^T + X_{n-p} U_{n-p}^T. \tag{30}$$

The matrix $\Sigma_L^\dagger$ is determined analogously to $\Sigma^\dagger$.

The solution of the least squares problem (1) is

$$x_{LS} = A^\dagger b = A_L^\dagger b = X_p \Sigma_L^\dagger U_p^T b + X_{n-p} U_{n-p}^T b =: x_L^1 + x_L^2.$$

Hence the solution x_{LS} consists of two parts, the second of which x_L^2 lies in the nullspace of L

Truncated GSVD

The idea of the truncated SVD approach can be extended to the GSVD case.

In the TGSVD method only the largest k (with $k \leq p$) components of $X_p \Sigma_L^\dagger U_p^T$ are taken into account whereas the x_L^2 remains unchanged. The solution $x_{L,k}$ is given by $x_{L,k} = A_{L,k}^\dagger b$ with

$$A_{L,k}^\dagger := [X_p, X_{n-p}] \begin{bmatrix} \Sigma_{L,k}^\dagger & 0 \\ 0 & I_{n-p} \end{bmatrix} [U_p, U_{n-p}]^T = X_p \Sigma_{L,k}^\dagger U_p^T + X_{n-p} U_{n-p}^T \qquad (31)$$

where $\Sigma_{L,k}^\dagger = diag(0, \ldots, 0, \sigma_{p-k+1}^{-1}, \ldots, \sigma_p^{-1})$.

The main part of the solution $x_{L,k}$ now consists of k vectors from X_p, i.e. generalized singular vectors that exhibit properties of the regularization matrix L as well.

Since the GSVD is expensive to compute it should be regarded as an analytical tool for regularizing least squares problems with a regularization matrix L.

Tikhonov regularization via GSVD

A particularly nice feature of the GSVD is that it immediately lets us write down a formula for the solution of the regularized solution of (19).

Introducing the filter function

$$f_\lambda(\gamma) := \frac{\gamma}{\gamma + \lambda} \tag{32}$$

the Tikhonov approximation can be written as

$$x_\lambda = \sum_{j=1}^{p} f_\alpha(\gamma_j^2) \frac{u_j^T b}{\sigma_j} x_j + \sum_{j=p+1}^{n} u_j^T b x_j. \tag{33}$$

Again the second (usually smooth) term x_L is untouched by the regularization whereas the terms in the first sum are damped for $\gamma_j \ll \lambda$ (oscillatory components, and due to the discrete Picard condition dominated by noise) and the other ones are only slightly changed.

Implementation of Tikhonov regularization

Consider the standard form of regularization

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \tag{34}$$

Multiplying A from the left and right by orthogonal matrices (which do not change Euclidean norms) it can be transformed to bidiagonal form

$$A = U \begin{bmatrix} J \\ O \end{bmatrix} V^T, \quad U \in \mathbb{R}^{m \times m},\ J \in \mathbb{R}^{n \times n},\ V \in \mathbb{R}^{n \times n}$$

where U and V are orthogonal (which are not computed explicitly but are represented by a sequence of Householder transformation).

With these transformations the new right hand side is

$$c = U^T b, \quad c =: (c_1^T, c_2^T)^T,\ c_1 \in \mathbb{R}^n,\ c_2 \in \mathbb{R}^{m-n}$$

and the variable is transformed according to

$$x = V\xi.$$

Implementation of Tikhonov regularization cnt.

The transformed problem reads

$$\left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \xi - \begin{bmatrix} \tilde{c}_1 \\ 0 \end{bmatrix} \right\|^2 = \text{min!} \tag{35}$$

Thanks to the bidiagonal form of J, (35) can be solved very efficiently using Givens transformations with only $\mathcal{O}(n)$ operations. Only these $\mathcal{O}(n)$ operations depend on the actual regularization parameter λ.

We considered only the standard case. If $L \neq I$ problem (19) the problem is transformed first to standard form.

If L is square and invertible, then the standard form

$$\|\bar{A}\bar{x} - \bar{b}\|^2 + \lambda \|\bar{x}\|^2 = \text{min!}$$

can be derived easily from $\bar{x} := Lx$, $\bar{A} = AL^{-1}$ and $\bar{b} = b$, such that the back transformation simply is $x_\lambda = L^{-1}\bar{x}_\lambda$.

Transformation into Standard Form

In the general case when $L \in \mathbb{R}^{p\times n}$ is rectangular with full row rank p the transformation can be performed via the GSVD.

The key step is to define the *A*-weighted generalized inverse of L as follows

$$L_A^\dagger = \left(I_n - (A(I_n - L^\dagger L))^\dagger A\right) L^\dagger. \tag{36}$$

If $p = n$ then $L^\dagger L = I$, and $L_A^\dagger = L^\dagger = L^{-1}$. For $p < n$ however, $L_A^\dagger$ is in general different from the Moore-Penrose inverse $L^\dagger$.

We further need the component x_0 of the regularized solution in the kernel $\mathcal{N}(L)$. Since $I - L^\dagger L$ is the orthogonal projector onto $\mathcal{N}(L)$, it holds that

$$x_0 = (A(I_n - L^\dagger L))^\dagger b.$$

Transformation into Standard Form ct.

Given the GSVD of (A, L), $L_A^\dagger$ can be expressed as (cf. Eldén 1982))

$$L_A^\dagger = X \begin{bmatrix} \text{diag}\{\mu_j^{-1}\} \\ O \end{bmatrix} V^T. \tag{37}$$

and x_0 as

$$x_0 = \sum_{i=p+1}^{n} u_i^T b x_i.$$

Then the standard-form quantities $\bar{A}$ and $\bar{b}$ obtain the form $\bar{A} := AL_A^\dagger$ and $\bar{b} := b - Ax_0$, while the transformation back to general-form setting becomes

$$x_\lambda = L_A^\dagger \bar{x}_\lambda + x_0$$

where $\bar{x}_\lambda$ denotes the solution of the transformed LS problem in standard form.

Choice of Regularization Matrix

In Tikhonov regularization one tries to balance the norm of the residual $\|Ax - b\|$ and the quantity $\|Lx\|$ where L is chosen such that known additional information about the solution can be implement.

Often some information about the smoothness of the solution x_{true} is known, e.g. if the underlying continuous problem is known to have a smooth solution then this should hold true for the discrete solution x_{true} as well. In that case the matrix L can be chosen as a discrete derivative operator.

The simples (easiest to implement) regularization matrix is $L = I$, which is known as the standard form. When nothing is known about the solution of the unperturbed system this is a sound choice.

From equation (14) it can be observed that the norm of x_{LS} blows up for ill-conditioned problems. Hence it is a reasonable choice simply to keep the norm of the solution under control.

Choice of Regularization Matrix cnt.

A common regularization matrix imposing some smoothness of the solution is the scaled one-dimensional first-order discrete derivative operator

$$L_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}. \tag{38}$$

The bilinear form

$$\langle x, y \rangle_{L^T L} := x^T L^T L y \tag{39}$$

does not induce a norm, but $\|x\|_L := \sqrt{\langle x, x \rangle_{L^T L}}$ is only a seminorm.

Since the null space of L is given by $\mathcal{N}(L) = \text{span}\{(1, \ldots, 1)^T\}$ a constant component of the solution is not affected by the Tikhonov regularization.

Singular vectors corresponding to $\sigma_j = 2 - 2\cos(j\pi/n)$, $j = 0, \ldots, n-1$ are $u_j = (\cos((2i-1)j\pi/(2n)))_{i=1,\ldots,n}$, and the influence of highly oscillating components are damped.

Choice of Regularization Matrix cnt.

Since nonsingular regularization matrices are easier to handle than singular ones a common approach is to use small perturbations.

If the perturbation is small enough the smoothing property is not deteriorated significantly. With a small diagonal element $\varepsilon > 0$

$$\tilde{L}_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{or} \quad \tilde{L}_{1D} = \begin{bmatrix} \varepsilon & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \tag{40}$$

are approximations to L_{1D}.

Which one of these modifications is appropriate depends on the behavior of the solution close to the boundary. The additional element ε forces either the first or last element to have small magnitude.

Choice of Regularization Matrix cnt.

A further common regularization matrix is the discrete second-order derivative operator

$$L_{1D}^{2nd} = \begin{bmatrix} -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n} \tag{41}$$

which does not affect constant and linear vectors.

A nonsingular approximation of L_{1D}^{2nd} is for example given by

$$\tilde{L}_{1D}^{2nd} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{42}$$

which is obtained by adding one row at the top an one row at the bottom of $L_{1D}^{2nd} \in \mathbb{R}^{(n-2)\times n}$. In this version Dirichlet boundary conditions are assumed at both ends of the solution

Choice of Regularization Matrix cnt.

The invertible approximations

$$\tilde{L}_{1D}^{2nd} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \quad \text{or} \quad \tilde{L}_{1D}^{2nd} = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

assume Dirichlet conditions on one side and Neumann boundary conditions on the other.

Choice of regularization parameter

According to Hansen and Hanke (1993): “No black-box procedures for choosing the regularization parameter λ are available, and most likely will never exist”

However, there exist numerous heuristics for choosing λ. We discuss three of them. The goal of the parameter choice is a reasonable balancing between the regularization error and perturbation error.

Let

$$x_\lambda = \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^{n} f_i \frac{u_i^T s}{\sigma_i} v_i \tag{43}$$

be the regularized solution of $\|Ax - \tilde{b}\| = \min!$ where $\tilde{b} = b + \varepsilon s$ and b is the exact right-hand side from $Ax_{true} = b$.

Choice of regularization parameter cnt.

The regularization error is defined as the distance of the first term in (43) to x_{true}, i.e.

$$\left\| \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i - x_{true} \right\| = \left\| \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i \right\| \tag{44}$$

and the perturbation error is defined as the norm of the second term in (43), i.e.

$$\varepsilon \left\| \sum_{i=1}^{n} f_i \frac{u_i^T s}{\sigma_i} v_i \right\| . \tag{45}$$

If all filter factors f_i are chosen equal to one, the unregularized solution x_{LS} is obtained with zero regularization error but large perturbation error, and choosing all filter factors equal to zero leads to a large regularization error but zero perturbation error – which corresponds to the solution $x = 0$.

Increasing the regularization parameter λ reduces the regularization error and increases the perturbation error. Methods are needed to balance these two quantities.

Discrepancy principle

The discrepancy principle assumes knowledge about the size of the error:

$$\|e\| = \varepsilon\|s\| \approx \delta_e.$$

The solution x_λ is said to satisfy the discrepancy principle if the discrepancy $d_\lambda := b - Ax_\lambda$ satisfies

$$\|d_\lambda\| = \|e\|.$$

If the perturbation e is known to have zero mean and a covariance matrix $\sigma_0^2 I$ (for instance if $\tilde{b}$ is obtained from independent measurements) the value of δ_e can be chosen close to the expected value $\sigma_0\sqrt{m}$.

The idea of the discrepancy principle is that we can not expect to obtain a more accurate solution once the norm of the discrepancy has dropped below the approximate error bound δ_e.

Generalized Cross Validation

The generalized cross validation is a popular method when no information about the size of the error $\|e\|$ is available.

The idea behind GCV is that a good parameter should predict missing data values.

Assume that a data point $\tilde{b}_i$ is excluded from the right-hand side $\tilde{b} \in \mathbb{R}^m$ and a regularized solution is computed from the remaining system

$$A([1 : i-1, i+1 : m], :)x_{\lambda,i} \approx \tilde{b}(1 : i-1, i+1 : m).$$

Then the value $A(i, :) \cdot x_{\lambda,i}$ should be close to the excluded value $\tilde{b}_i$ if a reasonable parameter λ has been chosen.

Generalized Cross Validation

In ordinary cross validation the ordering of the data plays a role, while generalized cross validation is invariant to orthogonal transformations on $\tilde{b}$.

The GCV method seeks to minimize the predictive mean square error $\|Ax_{reg} - b\|$ with the exact right-hand side b. Since b is unknown the parameter is chosen as the minimizer of the GCV function

$$\mathcal{G}(\lambda) = \frac{\|Ax_\lambda - \tilde{b}\|^2}{trace(I - AA^\sharp(\lambda))^2} \tag{46}$$

where the $A^\sharp(\lambda)$ denotes the matrix that maps the vector $\tilde{b}$ onto the regularized solution x_λ .

Depending on the regularization method the matrix $A^\sharp$ may not be explicitly available or sometimes not unique. Typically the GCV function displays a flat behavior around the minimum which makes its numerical computation difficult.

L-curve criterion

The L-curve criterion is a heuristic approach. No convergence results are available.

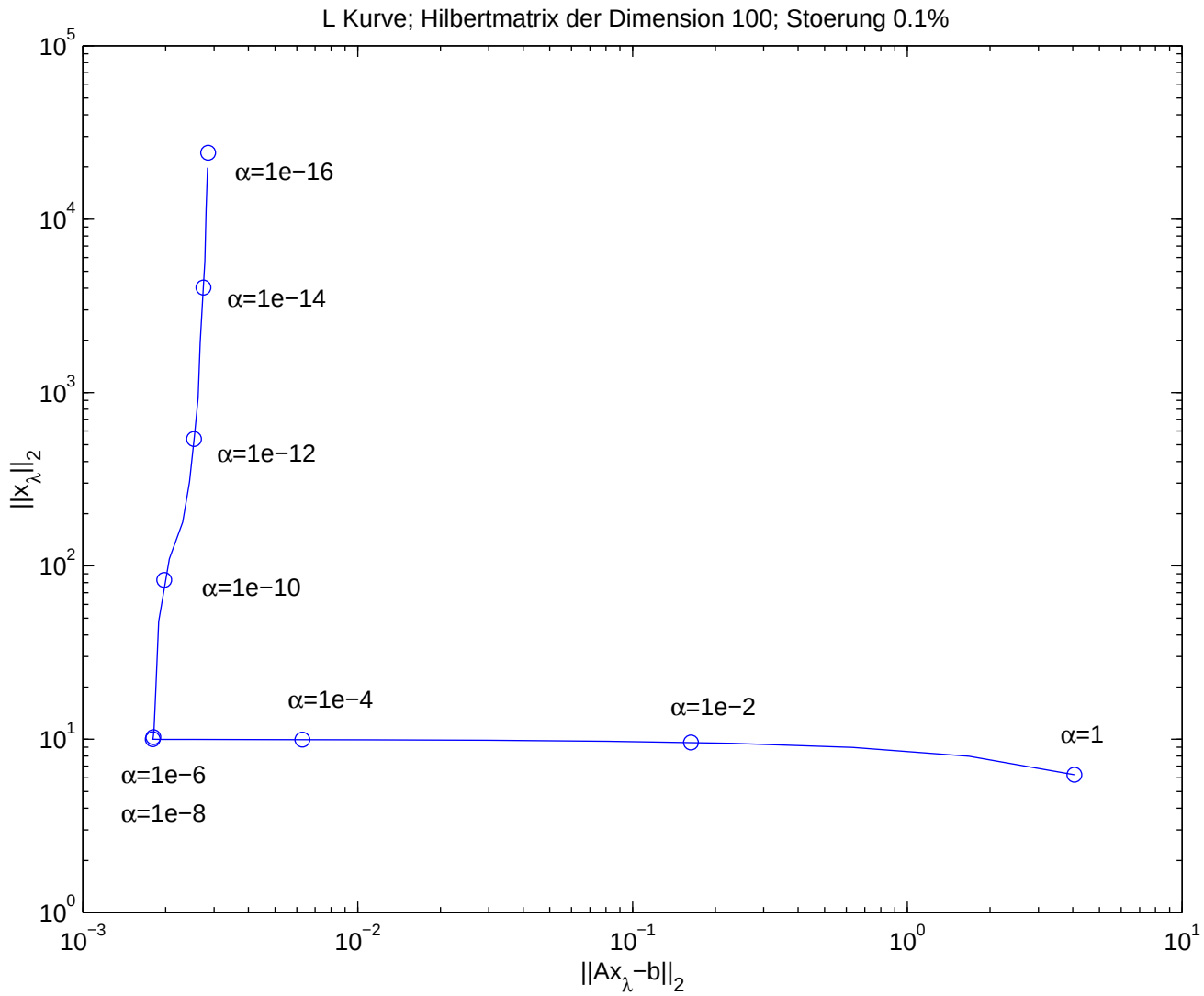
It is based on a graph of the penalty term $\|Lx_\lambda\|$ versus the discrepancy norm $\|\tilde{b} - Ax_\lambda\|$. It is observed that when plotted in log-log scale this curve often has a steep part, a flat part, and a distinct corner seperating these two parts. This explains the name L-curve.

The only assumptions that are needed to show this, is that the unperturbed component of the right-hand side satisfies the discrete Picard condition and that the perturbation does not dominate the right-hand side.

The flat part then corresponds to Lx_λ where x_λ is dominated by perturbation errors, i.e. λ is chosen too large and not all the information in $\tilde{b}$ is extracted. Moreover, the plateau of this part of the L-curve is at $\|Lx_\lambda\| \approx \|Lx_{\text{true}}\|$.

The vertical part corresponds to a solution that is dominated by perturbation errors.

L-curve; Hilbert matrix n=100



Toy problem

The following table contains the errors for the linear system $Ax = b$ where A is the Hilbert matrix, and b is such that $x = ones(n, 1)$ is the solution. The regularization matrix is $L = I$ and the regularization parameter is determined by the L-curve strategy. The normal equations were solved by the Cholesky factorization, QR factorization and SVD.

	$n = 10$	$n = 20$	$n = 40$
Tikhonov Cholesky	1.41 E-3	2.03 E-3	3.51 E-3
Tikhonov QR	3.50 E-6	5.99 E-6	7.54 E-6
Tikhonov SVD	3.43 E-6	6.33 E-6	9.66 E-6

The following table contains the results for the LS problems (m=n+20).

	$n = 10$	$n = 20$	$n = 40$
Tikhonov Cholesky	3.85 E-4	1.19 E-3	2.27 E-3
Tikhonov QR	2.24 E-7	1.79 E-6	6.24 E-6
Tikhonov SVD	8.51 E-7	1.61 E-6	3.45 E-6

Large ill-posed problems

Until now we assumed that the SVD of A or TSVD of (A, L) is available which is only reasonable for not too large dimensions.

We now assume that the matrix A is so large as to make the decomposition of its singular value decomposition undesirable or infeasible.

The first method introduced by Björck (1988) combines the truncated SVD with the bidiagonalization of Golub-Kahan-Lanczos.

With the starting vector b, put

$$\beta_1 u_1 = b, \ \alpha_1 v_1 = A^T u_1, \tag{47}$$

and for $i = 1, 2, \ldots$ compute

$$\left\{ \begin{array}{rcl} \beta_{i+1} u_{i+1} & = & A v_i - \alpha_i u_i \\ \alpha_{i+1} v_{i+1} & = & A^T u_{i+1} - \beta_{i+1} v_i \end{array} \right. \tag{48}$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$, $i = 0, 1, 2, \ldots$ are chosen so that $\|u_i\| = \|v_i\| = 1$.

Bidiagonalization & TSVD

With

$$U_k = (u_1, \ldots, u_k),\ V_k = (v_1, \ldots, v_k),\ \bar{B}_k = \begin{bmatrix} \alpha_1 & & & \\ \beta_1 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_k \end{bmatrix} \tag{49}$$

The recurrence relation can be rewritten as

$$U_{k+1}(\beta_1 e_1) = b, \tag{50}$$

$$AV_k = U_{k+1}\bar{B}_k,\ A^T U_{k+1} = V_k \bar{B}_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T. \tag{51}$$

We are looking for an approximate solution $x_k \in \text{span}\{V_k\}$ and write $x_k = V_k y_k$.

Bidiagonalization & TSVD cnt.

From the first equation of (51) it follows $Ax_k = AV_k y_k = U_{k+1}\bar{B}_k y_k$, and since (in exact arithmetic) U_{k+1} can be shown to be orthogonal, it follows that

$$\|b - Ax_k\| = \|U_{k+1}(\beta_1 e_1 - \bar{B}_k y_k)\| = \|\beta_1 e_1 - \bar{B}_k y_k\|.$$

Hence, $\|b - Ax_k\|$ is minimized over span(V_k) if y_k solves the least squares problem

$$\|\beta_1 e_1 - \bar{B}_k y_k\| = \text{min!}. \tag{52}$$

With $d_k := \beta_1 e_1 - \bar{B}_k y_k$ it holds

$$A^T(Ax_k - b) = A^T U_{k+1} d_k = V_k \bar{B}_k^T d_k + \alpha_{k+1} v_{k+1} e_{k+1}^T d_k,$$

and if y_k solves (52) we get from $b_k^T d_k = 0$

$$A^T(b - Ax_k) = \alpha_{k+1} v_{k+1} e_{k+1}^T d_k. \tag{53}$$

Bidiagonalization & TSVD cnt.

Assume that $\alpha_i \neq 0$ and $\beta_i \neq 0$ for $i = 1, \ldots, k$. If in the next step $\beta_{i+1} = 0$, then $\bar{B}_k$ has rank k and it follows that $d_k = 0$, i.e. $Ax_k = b$. If $\beta_{k+1} \neq 0$ but $\alpha_{k+1} = 0$, then by (53) x_k is a least squares solution of $Ax = b$.

Thus, the recurrence (51) cannot break down before solution of $\|Ax - b\| = \min!$ is obtained.

The bidiagonalization algorithm (47), (48) is closely related to the Lanczos process applied to the symmetric matrix $A^T A$. If the starting vector v_1 is used, then it follows from (48)

$$(A^T A) V_k = A^T U_{k+1} \bar{B}_k = V_k (\bar{B}_k^T \bar{B}_k) + \alpha_{k+1} \beta_k v_{k+1} e_k^T .$$

Bidiagonalization & TSVD cnt.

To obtain a regulrized solution of $\|Ax - b\| = \min!$ consider the (full) SVD of the matrix $\bar{B}_k \in \mathbb{R}^{k+1\times k}$

$$\bar{B}_k = P_k \begin{bmatrix} \Omega_k \\ 0 \end{bmatrix} Q_k^T = \sum_{i=1}^{k} \omega_i p_i q_i^T \tag{54}$$

where P_k and Q_k are square orthogonal matrices and

$$\omega_1 \geq \omega_2 \geq \cdots \geq \omega_k > 0. \tag{55}$$

Then the solution y_k to the least squares problem (52) can be written

$$y_k = \beta_1 Q_k(\Omega_k^{-1}\ ,\ 0)P_k^T e_1 = \beta_1 \sum_{i=1}^{k} \omega_i^{-1} \kappa_{1i} q_i \tag{56}$$

with $\kappa_{ij} = (P_k)_{ij}$.

Bidiagonalization & TSVD cnt.

The corresponding residual vector is

$$d_k = \beta_1 P_k e_{k+1} e_{k+1}^T P_k^T e_1 = \beta_1 \kappa_{1,k+1} p_{k+1},$$

from which we get

$$\|b - Ax_k\| = \|d_k\| = \beta_1 |\kappa_{1,k+1}|.$$

For a given threshold $\delta > 0$ we define regularized solution by the TSVD method:

$$x_k(\delta) = V_k y_k(\delta), \quad y_k(\delta) = \beta_1 \sum_{\omega_i > \delta} \omega_i^{-1} \kappa_{1i} q_i. \tag{57}$$

Notice, that the norm of $x_k(\delta)$ and the residual norm

$$\|x_k(\delta)\|^2 = \beta_1^2 \sum_{\omega_i > \delta} (\kappa_{1i})^2 \quad \text{and} \quad \|r_k(\delta)\| r = \beta_1^2 \sum_{\omega_i \le \delta} \kappa_{1i}^2$$

can be obtained for fixed k and δ without forming $x_k(\delta)$ or $y_k(\delta)$ explicitly. Hence, an "optimal" δ can be obtained by generalized cross validation or an L-curve method, and only after δ has been found $x_k(\delta)$ is determined.

Tikhonov regularization for large problems

Consider the Tikhonov regularization in standard form ($L = I$) and its normal equations

$$(A^T A + \mu^{-1} I)x = A^T b \tag{58}$$

where the regularization parameter μ is positive and finite.

Usually $\lambda = \mu^{-1}$ is chosen, but for the method to develop (58) is more convenient.

The solution of (58) is

$$x_\mu = (A^T A + \mu^{-1} I)^{-1} A^T b \tag{59}$$

and the discrepancy

$$d_\mu = b - A x_\mu. \tag{60}$$

We assume that an estimate for the error ε of b is explicitly known. We seek for a parameter $\hat{\mu}$ such that

$$\varepsilon \leq \|d_{\hat{\mu}}\| \leq \eta\varepsilon \tag{61}$$

where the choice of η depends on the accuracy of the estimate ε.

Tikhonov regularization for large problems cnt.

Let

$$\phi(\mu) := \|b - Ax_\mu\|^2. \tag{62}$$

Substituting the expression (59) into (62) and applying the identity

$$I - A(A^TA + \mu^{-1}I)^{-1}A^T = (\mu AA^T + I)^{-1}$$

one gets that

$$\phi(\mu) = b^T(\mu AA^T + I)^{-2}b. \tag{63}$$

Hence, $\phi'(\mu) \leq 0$ and $\phi''(\mu) \geq 0$ for $\mu \geq 0$, i.e. ϕ is monotonically decreasing and convex, and for $\tau \in (0, \|b\|)$ the equation $\phi(\mu) = \tau$ has a unique solution.

Notice, that the function $\nu \mapsto \phi(1/\nu)$ cannot be guaranteed to be convex. This is the reason why the regularization parameter μ is used instead of $\lambda = 1/\mu$.

Tikhonov regularization for large problems cnt.

Newton's method converges globally to a solution of $\phi(\mu) - \varepsilon^2 = 0$. However, for large dimension the evaluation of $\phi(\mu)$ and $\phi'(\mu)$ for fixed μ is much too expensive (or even impossible due to the high condition number of A).

Calvetti and Reichel (2003) took advantage of partial bidiagonalization of A to construct bounds to $\phi(\mu)$ and thus determine a $\hat{\mu}$ satisfying (61).

Application of $\ell \leq \min\{m, n\}$ bidiagonalization steps (cf. (48)) gives the decomposition

$$AV_\ell = U_\ell B_\ell + \beta_{\ell+1} u_{\ell+1} e_\ell^T, \; A^T U_\ell = V_\ell B_\ell^T, \; b = \beta_1 U_\ell e_1 \tag{64}$$

where $V_\ell \in \mathbb{R}^{n\times\ell}$ and $U_\ell \in \mathbb{R}^{n\times\ell}$ have orthonormal columns and $U_\ell^T u_{\ell+1} = 0$. $\beta_{\ell+1}$ is a nonnegative scalar, and $\|u_{\ell+1}\| = 1$ when $\beta_{\ell+1} > 0$.

$B_\ell \in \mathbb{R}^{\ell\times\ell}$ is the bidiagonal matrix with diagonal elements α_j and (positive) subdiagonal elements β_j (B_ℓ is the submatrix of $\bar{B}_\ell$ in (49) containing its first ℓ rows).

Tikhonov regularization for large problems cnt.

Combining the equations in (64) yields

$$AA^T U_\ell = U_\ell B_\ell B_\ell^T + \alpha_\ell \beta_{\ell+1} u_{\ell+1} e_\ell^T \tag{65}$$

which shows that the columns of U_ℓ are Lanczos vectors, and the matrix $T_\ell := B_\ell B_\ell^T$ is the tridiagonal matrix that would be obtained by applying ℓ Lanczos steps to the symmetric matrix AA^T with initial vector b.

Introduce the functions

$$\phi_\ell(\mu) := \|b\|^2 e_1^T (\mu B_\ell B_\ell^T + I_\ell)^{-2} e_1, \tag{66}$$

$$\bar{\phi}_\ell(\mu) := \|b\|^2 e_1^T (\mu \bar{B}_\ell \bar{B}_\ell^T + I_{\ell+1})^{-2} e_1, \tag{67}$$

Tikhonov regularization for large problems cnt.

Substituting the spectral factorization $AA^T = W\Lambda W^T$ with $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$ and $W^T W = I$ into ϕ one obtains

$$\phi(\mu) = b^T W(\mu\Lambda + I)^{-2} W^T b = \sum_{j=1}^{m} \frac{\tilde{b}_j^2}{(\mu\lambda_j + 1)^2} =: \int_0^\infty \frac{1}{(\mu\lambda + 1)^2}\, d\omega(\lambda) \qquad (68)$$

where $\tilde{b} = W^T b$, and ω chosen to be a piecewise constant function with jump discontinuities of height $\tilde{b}_j^2$ at the eigenvalues λ_j.

Thus, the integral in the right-hand side is a Stieltjes integral defined by the spectral factorization of AA^T and by the vector b.

Tikhonov regularization for large problems cnt.

Golub and Meurant (1993) proved that ϕ_ℓ defined in (66) is the ℓ-point Gauß quadrature rule associated with the distribution function ω applied to the function

$$\psi(t) := (\mu t + 1)^{-2}. \tag{69}$$

Similarly, the function $\bar{\phi}_\ell$ in (67) is the $(\ell + 1)$-point Gauß-Radau quadrature rule with an additional node at the origin associated with the distribution function ω applied to the function ψ.

Since for any fixed positive value of μ the derivatives of ψ with respect to t of odd order are strictly negative and the derivatives of even order are strictly positive for $t \geq 0$, the remainder terms for Gauß and Gauß-Radau quadrature rules show

$$\phi_\ell(\mu) < \phi(\mu) < \bar{\phi}_\ell(\mu). \tag{70}$$

Instead of computing a value μ that satisfies (61), one seeks to determine values ℓ and μ such that

$$\varepsilon^2 \leq \phi_\ell(\mu), \quad \bar{\phi}_\ell(\mu) \leq \varepsilon^2 \eta^2. \tag{71}$$

Tikhonov regularization for large problems cnt.

A method for computing a suitable μ can be based on the fact that for every $\mu > 0$ it holds that

$$\phi_k(\mu) < \phi_\ell(\mu) < \phi(\mu) < \bar{\phi}_\ell(\mu) < \bar{\phi}_k(\mu) \quad \text{for } 1 \le k < \ell \tag{72}$$

which was proved by Hanke (2003).

In general the value ℓ in pairs (ℓ, μ) that satisfy (71) can be chosen quite small (cf. the examples in the paper of Calvetti and Reichel).

Once a suitable value $\hat{\mu}$ of the regularization parameter is available, the regularized solution

$$x_{\hat{\mu},\ell} = V_\ell y_{\hat{\mu},\ell} \tag{73}$$

wher $y_{\hat{\mu},\ell} \in \mathbb{R}^\ell$ satisfies the Galerkin equation

$$V_\ell^T (A^T A + \hat{\mu}^{-1} I) V_\ell y_{\hat{\mu},\ell} = V_\ell A^T b. \tag{74}$$

Tikhonov regularization for large problems cnt.

Using the recurrence relation (64) system (74) can be simplified to

$$(\bar{B}_\ell^T \bar{B}_\ell + \hat{\mu}^{-1} I_\ell) y_{\hat{\mu},\ell} = \beta_1 B_\ell^T e_1. \tag{75}$$

These are the normal equations associated with the least squares problem

$$\left\| \begin{bmatrix} \hat{\mu}^{1/2} \bar{B}_\ell \\ I_\ell \end{bmatrix} y - \beta_1 \hat{\mu}^{1/2} e_1 \right\| = \text{min!} \tag{76}$$

which can be solved in $\mathcal{O}(\ell)$ arithmetic floating operations by using Givens rotations.

Methods based on bidiagonalization for standard regularization problems can also be applied to a Tikhonov regularization problem with general $L \neq I$, provided that it can be transformed to standard form without too much effort (with the A-weighted pseudoinverse of L). For instance, when L is banded with small bandwidth and a known null space, the transformation is attractive.

Iterative projection methods for Tikhonov regularization with general L is an active field of research.