

# Regularization of Least Squares Problems

Heinrich Voss

voss@tu-harburg.de

Hamburg University of Technology  
Institute of Numerical Simulation



- 1 Introduction
- 2 Least Squares Problems
- 3 Ill-conditioned problems
- 4 Regularization

# Outline

- 1 Introduction
- 2 Least Squares Problems
- 3 Ill-conditioned problems
- 4 Regularization

# 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.

# 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.

# 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.

# 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.



# 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.

# 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.

# 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.

# 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.

# Outline

- 1 Introduction
- 2 Least Squares Problems
- 3 Ill-conditioned problems
- 4 Regularization

# Least Squares Problems

Let

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

Differentiating

$$\varphi(x) = \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) \quad (2)$$

yields the necessary condition

$$A^T Ax = A^T b. \quad (3)$$

called **normal equations**.

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

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

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

# Least Squares Problems

Let

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

Differentiating

$$\varphi(x) = \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) \quad (2)$$

yields the necessary condition

$$A^T Ax = A^T b. \quad (3)$$

called **normal equations**.

If the columns of  $A$  are linearly independent, then  $A^T A$  is positive definite, i.e.  $\varphi$  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  $A$ ,

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

# Least Squares Problems

Let

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

Differentiating

$$\varphi(x) = \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) \quad (2)$$

yields the necessary condition

$$A^T Ax = A^T b. \quad (3)$$

called **normal equations**.

If the columns of  $A$  are linearly independent, then  $A^T A$  is positive definite, i.e.  $\varphi$  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  $A$ ,

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



# Least Squares Problems

Let

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

Differentiating

$$\varphi(x) = \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) \quad (2)$$

yields the necessary condition

$$A^T Ax = A^T b. \quad (3)$$

called **normal equations**.

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

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

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

# 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  $A^T A > 0$ .

However,  $A^T A$  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

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

and the unique solution of (1) is

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

# 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  $A^T A > 0$ .

However,  $A^T A$  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

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

and the unique solution of (1) is

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

# 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  $A^T A > 0$ .

However,  $A^T A$  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

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

and the unique solution of (1) is

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

# 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  $A^T A > 0$ .

However,  $A^T A$  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

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

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  $A$ :

$$A = \tilde{U} \tilde{\Sigma} \tilde{V}^T \quad (5)$$

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

$$A = U \Sigma V^T \quad (6)$$

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  $\Sigma \in \mathbb{R}^{n \times n} = \text{diag}(\sigma_1, \dots, \sigma_n)$ .

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:

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

$\sigma_i$ ,  $i = 1, \dots, n$  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  $A$ .

# Singular value decomposition

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

$$A = \tilde{U} \tilde{\Sigma} \tilde{V}^T \quad (5)$$

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

$$A = U \Sigma V^T \quad (6)$$

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  $\Sigma \in \mathbb{R}^{n \times n} = \text{diag}(\sigma_1, \dots, \sigma_n)$ .

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:

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

$\sigma_i$ ,  $i = 1, \dots, n$  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  $A$ .

# Singular value decomposition

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

$$A = \tilde{U} \tilde{\Sigma} \tilde{V}^T \quad (5)$$

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

$$A = U \Sigma V^T \quad (6)$$

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  $\Sigma \in \mathbb{R}^{n \times n} = \text{diag}(\sigma_1, \dots, \sigma_n)$ .

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:

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

$\sigma_i$ ,  $i = 1, \dots, n$  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  $A$ .



# Singular value decomposition

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

$$A = \tilde{U} \tilde{\Sigma} \tilde{V}^T \quad (5)$$

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

$$A = U \Sigma V^T \quad (6)$$

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  $\Sigma \in \mathbb{R}^{n \times n} = \text{diag}(\sigma_1, \dots, \sigma_n)$ .

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:

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

$\sigma_i$ ,  $i = 1, \dots, n$  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

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

Hence,

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

Since  $v_{r+1}, \dots, v_n$  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)

# 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

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

Hence,

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

Since  $v_{r+1}, \dots, v_n$  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).

# 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

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

Hence,

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

Since  $v_{r+1}, \dots, v_n$  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  $A^\dagger = (A^T A)^{-1} A^T$  (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

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

# 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  $A^\dagger = (A^T A)^{-1} A^T$  (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

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

# 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  $A^\dagger = (A^T A)^{-1} A^T$  (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

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

# 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  $A^\dagger = (A^T A)^{-1} A^T$  (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

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



# 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

$$\|(A + \delta A)x - (b + \delta b)\| = \min! \quad (8)$$

where the perturbation is not too large in the sense

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

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

Then it holds that

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

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

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

# 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

$$\|(A + \delta A)x - (b + \delta b)\| = \min! \quad (8)$$

where the perturbation is not too large in the sense

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

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

Then it holds that

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

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

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

# 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

$$\|(A + \delta A)x - (b + \delta b)\| = \min! \quad (8)$$

where the perturbation is not too large in the sense

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

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

Then it holds that

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

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

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

# Outline

- 1 Introduction
- 2 Least Squares Problems
- 3 III-conditioned problems**
- 4 Regularization

# 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.

# 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.

# 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  $f : [0, 1] \rightarrow \mathbb{R}$  to the space  $\Pi_{n-1}$  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  $\{1, x, \dots, x^{n-1}\}$  this leads to the linear system

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

where

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

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



# A toy problem

Consider the problem to determine the orthogonal projection of a given function  $f : [0, 1] \rightarrow \mathbb{R}$  to the space  $\Pi_{n-1}$  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  $\{1, x, \dots, x^{n-1}\}$  this leads to the linear system

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

where

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

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  $y = (1, \dots, 1)^T$  is the unique solution.

Solving the problem with LU-factorization (in MATLAB  $A \setminus b$ ), 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.

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

For dimensions  $n = 10$ ,  $n = 20$  and  $n = 40$  we choose the right hand side  $b$  such that  $y = (1, \dots, 1)^T$  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 that  $x = (1, \dots, 1)^T$  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 (?)

# 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 that  $x = (1, \dots, 1)^T$  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 (?)

# 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 that  $x = (1, \dots, 1)^T$  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 (?)

# 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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  $\mu_j$  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  $\mu_j/\sqrt{j}$ .

# 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  $\mu_j$  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  $\mu_j/\sqrt{j}$ .

# 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  $\mu_j$  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

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

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

$$Ax = b$$

.

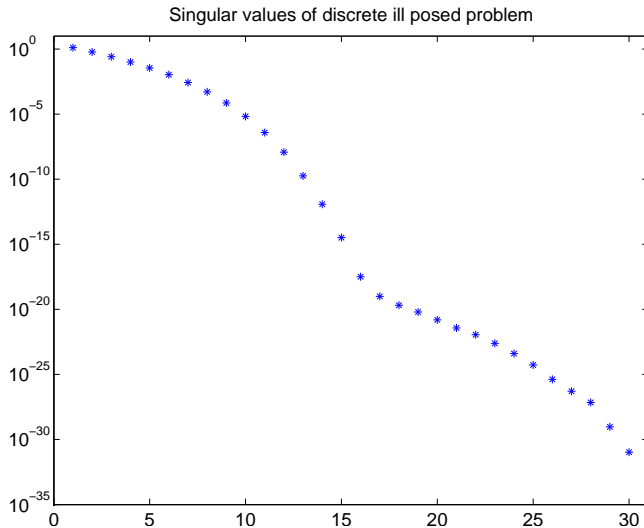
The matrix  $A$  inherits the following properties from the continuous problem (11): it is ill-conditioned with singular values gradually decaying to zero.

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  $i$ th 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  $\sigma_i$ .*

# 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  $i$ th 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  $\sigma_i$ .*

# 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  $i$ th 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  $\sigma_i$ .*

# 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

$$Ax_{true} = b \quad (12)$$

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

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

For the solution it holds

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

where  $r$  is the rank of  $A$ .

# 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

$$Ax_{true} = b \quad (12)$$

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

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

For the solution it holds

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

where  $r$  is the rank of  $A$ .

# 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

$$Ax_{true} = b \quad (12)$$

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

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

For the solution it holds

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

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  $A$ .



# 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  $A$ .

# 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  $A$ .

# 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  $A$ .

# Outline

- 1 Introduction
- 2 Least Squares Problems
- 3 Ill-conditioned problems
- 4 Regularization**

# Regularization

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

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

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.

# Regularization

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

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

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.

# Regularization

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

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

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

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

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

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

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

$$f_i = \begin{cases} 1 & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases} \quad (18)$$



# 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

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

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

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

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

$$f_i = \begin{cases} 1 & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases} \quad (18)$$

# 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

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

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

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

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

$$f_i = \begin{cases} 1 & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases} \quad (18)$$

# 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  $A$ .

# 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  $A$ .

# Truncated SVD cnt.

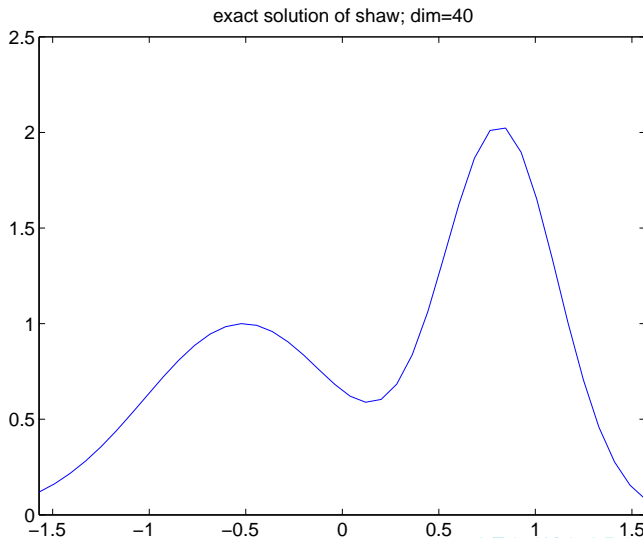
The solution  $x_k$  does not contain any high frequency components, i.e. all singular values starting from the index  $k + 1$  are set to zero and the corresponding singular vectors are disregarded in the solution. So the term  $u_i^T s / \sigma_i$  in equation (17) corresponding to the noise  $s$ 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

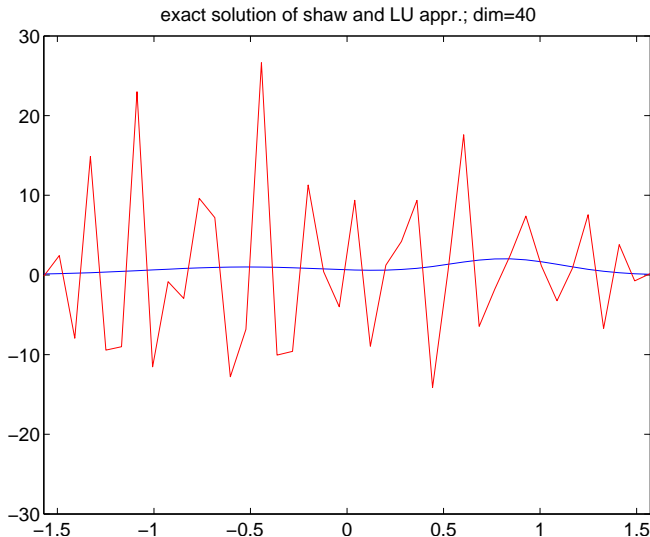
# Example

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



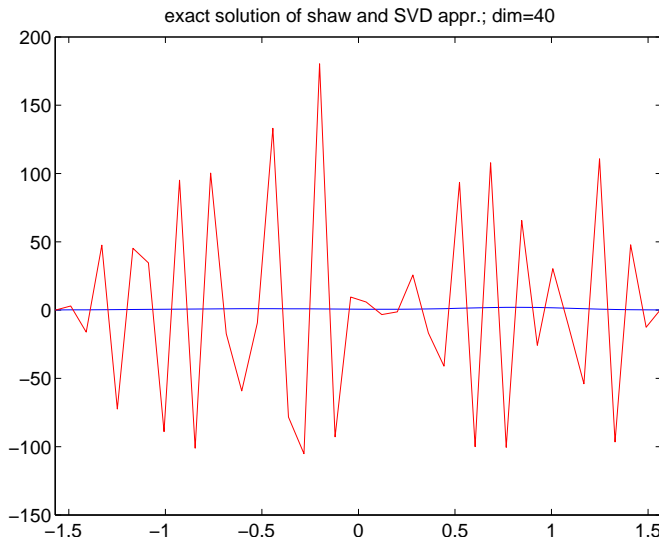
# Example

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



# Example

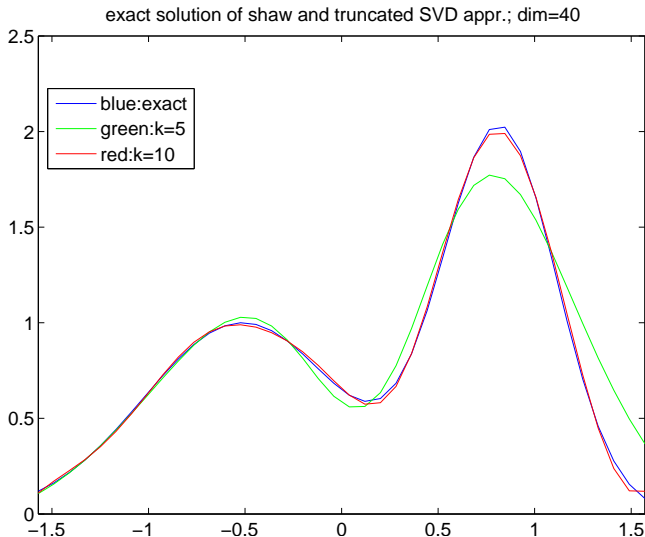
Solution of the Fredholm integral equation `shaw` from Hansen's regularization tool of dimension 40 and its approximation via complete SVD





# 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  $x_\lambda$  is defined as minimizer of the quadratic functional

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

The basic idea of Tikhonov regularization is the following: Minimizing the functional in (19) means to search for some  $x_\lambda$ , 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  $\lambda$  is chosen too small, (19) is too close to the original problem and instabilities have to be expected.

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

# Tikhonov regularization

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

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

The basic idea of Tikhonov regularization is the following: Minimizing the functional in (19) means to search for some  $x_\lambda$ , 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  $\lambda$  is chosen too small, (19) is too close to the original problem and instabilities have to be expected.

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

# Tikhonov regularization

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

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

The basic idea of Tikhonov regularization is the following: Minimizing the functional in (19) means to search for some  $x_\lambda$ , 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  $\lambda$  is chosen too small, (19) is too close to the original problem and instabilities have to be expected.

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

# Tikhonov regularization

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

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

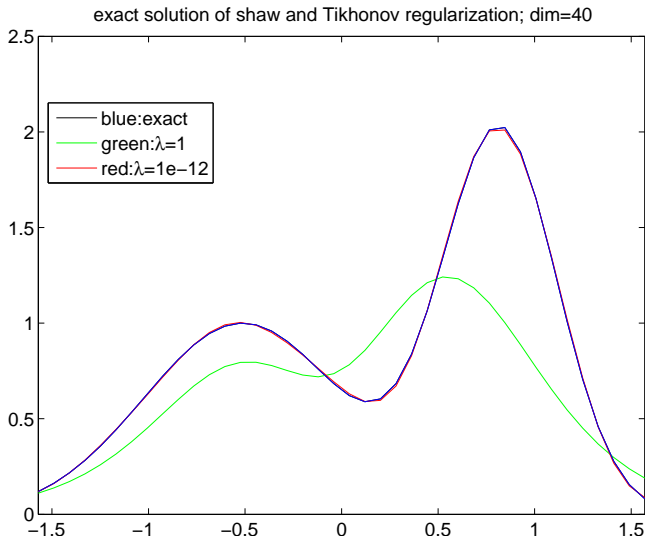
The basic idea of Tikhonov regularization is the following: Minimizing the functional in (19) means to search for some  $x_\lambda$ , providing at the same time a small residual  $\|Ax_\lambda - \tilde{b}\|$  and a moderate value of the penalty function  $\|Lx_\lambda\|$ .

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

If  $\lambda$  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.

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

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} L \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (20)$$

with the normal equations

$$(A^T A + \lambda L^T L) x = A^T \tilde{b}. \quad (21)$$

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  $x_\lambda = x_{\text{reg}}$  of (21) is

$$x_\lambda = V \Theta \Sigma^\dagger U^T \tilde{b} = \sum_{i=1}^n f_i \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^n \frac{\sigma_i (u_i^T \tilde{b})}{\sigma_i^2 + \lambda} v_i \quad (22)$$

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

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \quad \text{for } L = I. \quad (23)$$

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

# Tikhonov regularization cnt.

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

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} L \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (20)$$

with the normal equations

$$(A^T A + \lambda L^T L) x = A^T \tilde{b}. \quad (21)$$

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 (21) is

$$x_\lambda = V \Theta \Sigma^\dagger U^T \tilde{b} = \sum_{i=1}^n f_i \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^n \frac{\sigma_i (u_i^T \tilde{b})}{\sigma_i^2 + \lambda} v_i \quad (22)$$

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

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \quad \text{for } L = I. \quad (23)$$

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



# Tikhonov regularization cnt.

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

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (20)$$

with the normal equations

$$(A^T A + \lambda L^T L)x = A^T \tilde{b}. \quad (21)$$

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 (21) is

$$x_\lambda = V\Theta\Sigma^\dagger U^T \tilde{b} = \sum_{i=1}^n f_i \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^n \frac{\sigma_i(u_i^T \tilde{b})}{\sigma_i^2 + \lambda} v_i \quad (22)$$

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

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \quad \text{for } L = I. \quad (23)$$

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

# Tikhonov regularization cnt.

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

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (20)$$

with the normal equations

$$(A^T A + \lambda L^T L)x = A^T \tilde{b}. \quad (21)$$


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 (21) is

$$x_\lambda = V\Theta\Sigma^\dagger U^T \tilde{b} = \sum_{i=1}^n f_i \frac{u_i^T \tilde{b}}{\sigma_i} v_i = \sum_{i=1}^n \frac{\sigma_i(u_i^T \tilde{b})}{\sigma_i^2 + \lambda} v_i \quad (22)$$

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

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \quad \text{for } L = I. \quad (23)$$

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

# Tikhonov regularization cnt.

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

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

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.

# Tikhonov regularization cnt.

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

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

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.

# Tikhonov regularization cnt.

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

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

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.

# Tikhonov regularization cnt.

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

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

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.

# Implementation of Tikhonov regularization

Consider the standard form of regularization

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (24)$$

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

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

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

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

and the variable is transformed according to

$$x = V\xi.$$

# Implementation of Tikhonov regularization

Consider the standard form of regularization

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (24)$$

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

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

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

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

and the variable is transformed according to

$$x = V\xi.$$



# Implementation of Tikhonov regularization

Consider the standard form of regularization

$$\left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (24)$$

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

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

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

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

and the variable is transformed according to

$$x = V\xi.$$

# Implementation of Tikhonov regularization cnt.

The transformed problem reads

$$\left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \xi - \begin{bmatrix} \tilde{c}_1 \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (25)$$

Thanks to the bidiagonal form of  $J$ , (25) 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  $\lambda$ .

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

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

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

# Implementation of Tikhonov regularization cnt.

The transformed problem reads

$$\left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \xi - \begin{bmatrix} \tilde{c}_1 \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (25)$$

Thanks to the bidiagonal form of  $J$ , (25) 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  $\lambda$ .

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

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

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

# Implementation of Tikhonov regularization cnt.

The transformed problem reads

$$\left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \xi - \begin{bmatrix} \tilde{c}_1 \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (25)$$

Thanks to the bidiagonal form of  $J$ , (25) 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  $\lambda$ .

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

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

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

# Implementation of Tikhonov regularization cnt.

The transformed problem reads

$$\left\| \begin{bmatrix} J \\ \sqrt{\lambda} I \end{bmatrix} \xi - \begin{bmatrix} \tilde{c}_1 \\ 0 \end{bmatrix} \right\|^2 = \min! \quad (25)$$

Thanks to the bidiagonal form of  $J$ , (25) 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  $\lambda$ .

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

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

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

# 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 implemented.

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 simplest (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

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 implemented.

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 simplest (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

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 implemented.

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 simplest (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

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 implemented.

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 simplest (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

$$L_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}. \quad (26)$$

The bilinear form

$$\langle x, y \rangle_{L^T L} := x^T L^T L y \quad (27)$$

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

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

# 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

$$L_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}. \quad (26)$$

The bilinear form

$$\langle x, y \rangle_{L^T L} := x^T L^T L y \quad (27)$$

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

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

# 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

$$L_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}. \quad (26)$$

The bilinear form

$$\langle x, y \rangle_{L^T L} := x^T L^T L y \quad (27)$$

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

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

# 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

$$L_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}. \quad (26)$$

The bilinear form

$$\langle x, y \rangle_{L^T L} := x^T L^T L y \quad (27)$$

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

Singular vectors corresponding to  $\sigma_j = 2 - 2 \cos(j\pi/n)$ ,  $j = 0, \dots, n-1$  are  $u_j = (\cos((2i-1)j\pi/(2n)))_{i=1, \dots, n}$ , 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  $\varepsilon > 0$

$$\tilde{L}_{1D} = \begin{bmatrix} -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & -1 & 1 & \\ & & & & \varepsilon \end{bmatrix} \quad \text{or} \quad \tilde{L}_{1D} = \begin{bmatrix} \varepsilon & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & & -1 & 1 \end{bmatrix} \quad (28)$$

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  $\varepsilon$  forces either the first or last element to have small magnitude.

# 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  $\varepsilon > 0$

$$\tilde{L}_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{or} \quad \tilde{L}_{1D} = \begin{bmatrix} \varepsilon & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \quad (28)$$

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  $\varepsilon$  forces either the first or last element to have small magnitude.

# 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  $\varepsilon > 0$

$$\tilde{L}_{1D} = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & \varepsilon \end{bmatrix} \quad \text{or} \quad \tilde{L}_{1D} = \begin{bmatrix} \varepsilon & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \quad (28)$$

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  $\varepsilon$  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

$$L_{1D}^{2nd} = \begin{bmatrix} -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(n-2) \times n} \quad (29)$$

which does not affect constant and linear vectors.

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

$$\tilde{L}_{1D}^{2nd} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n} \quad (30)$$

which is obtained by adding one row at the top and 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.

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

$$L_{1D}^{2nd} = \begin{bmatrix} -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(n-2) \times n} \quad (29)$$

which does not affect constant and linear vectors.

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

$$\tilde{L}_{1D}^{2nd} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n} \quad (30)$$

which is obtained by adding one row at the top and 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  $\lambda$  are available, and most likely will never exist”

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

Let

$$x_\lambda = \sum_{i=1}^n \hat{f}_i \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^n \hat{f}_i \frac{u_i^T s}{\sigma_i} v_i \quad (31)$$

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

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

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

Let

$$x_\lambda = \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^n f_i \frac{u_i^T s}{\sigma_i} v_i \quad (31)$$

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

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

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

Let

$$x_\lambda = \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i + \varepsilon \sum_{i=1}^n f_i \frac{u_i^T s}{\sigma_i} v_i \quad (31)$$

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 (31) to  $x_{true}$ , i.e.

$$\left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - x_{true} \right\| = \left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^n \frac{u_i^T b}{\sigma_i} v_i \right\| \quad (32)$$

and the **perturbation error** is defined as the norm of the second term in (31), i.e.

$$\varepsilon \left\| \sum_{i=1}^n f_i \frac{u_i^T s}{\sigma_i} v_i \right\|. \quad (33)$$

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  $\lambda$  reduces the regularization error and increases the perturbation error. Methods are needed to balance these two quantities.

# Choice of regularization parameter cnt.

The **regularization error** is defined as the distance of the first term in (31) to  $x_{true}$ , i.e.

$$\left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - x_{true} \right\| = \left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^n \frac{u_i^T b}{\sigma_i} v_i \right\| \quad (32)$$

and the **perturbation error** is defined as the norm of the second term in (31), i.e.

$$\varepsilon \left\| \sum_{i=1}^n f_i \frac{u_i^T s}{\sigma_i} v_i \right\|. \quad (33)$$

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  $\lambda$  reduces the regularization error and increases the perturbation error. Methods are needed to balance these two quantities.



# Choice of regularization parameter cnt.

The **regularization error** is defined as the distance of the first term in (31) to  $x_{true}$ , i.e.

$$\left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - x_{true} \right\| = \left\| \sum_{i=1}^n f_i \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^n \frac{u_i^T b}{\sigma_i} v_i \right\| \quad (32)$$

and the **perturbation error** is defined as the norm of the second term in (31), i.e.

$$\varepsilon \left\| \sum_{i=1}^n f_i \frac{u_i^T s}{\sigma_i} v_i \right\|. \quad (33)$$

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  $\lambda$  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:

$$\|\mathbf{e}\| = \varepsilon \|\mathbf{s}\| \approx \delta_e.$$

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

$$\|d_\lambda\| = \|\mathbf{e}\|.$$

If the perturbation  $\mathbf{e}$  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  $\delta_e$  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  $\delta_e$ .

# Discrepancy principle

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

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

The solution  $x_\lambda$  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  $\delta_e$  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  $\delta_e$ .

# Discrepancy principle

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

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

The solution  $x_\lambda$  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  $\delta_e$  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  $\delta_e$ .

# 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 separating 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  $Lx_\lambda$  where  $x_\lambda$  is dominated by perturbation errors, i.e.  $\lambda$  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 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 separating 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  $Lx_\lambda$  where  $x_\lambda$  is dominated by perturbation errors, i.e.  $\lambda$  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 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 separating 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  $Lx_\lambda$  where  $x_\lambda$  is dominated by perturbation errors, i.e.  $\lambda$  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 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 separating 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  $Lx_\lambda$  where  $x_\lambda$  is dominated by perturbation errors, i.e.  $\lambda$  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 criterion

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

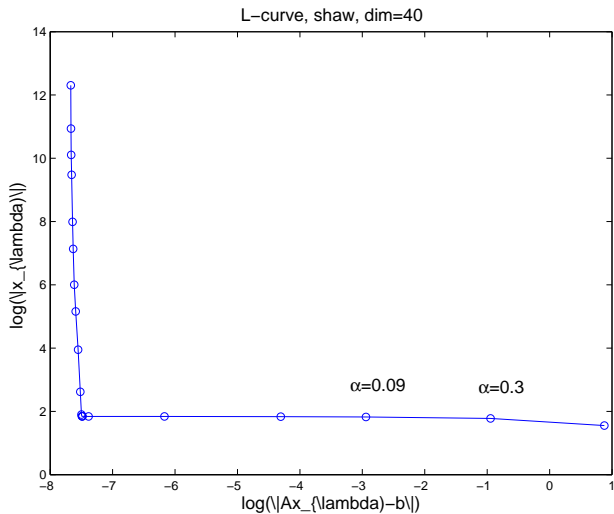
It is based on a graph of the penalty term  $\|Lx_\lambda\|$  versus the discrepancy norm  $\|\tilde{b} - Ax_\lambda\|$ . It is observed that when plotted in log-log scale this curve often has a steep part, a flat part, and a distinct corner separating these two parts. This explains the name **L-curve**.

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  $Lx_\lambda$  where  $x_\lambda$  is dominated by perturbation errors, i.e.  $\lambda$  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  $x = \text{ones}(n, 1)$  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

# 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  $x = \text{ones}(n, 1)$  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

# 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  $x = \text{ones}(n, 1)$  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

# 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  $x = \text{ones}(n, 1)$  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