Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

First lecture, January 19, 2011.

0 Practical issues

Information and materials

- The main information channel of the course is the homepage: `https://noppa.tkk.fi/noppa/kurssi/mat-1.3626/` .
- The text book is "J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005" (mainly Chapters 2 and 3).
- Lecture notes and exercise papers will be posted on the course homepage.

Exercises

- The first exercise session will held on Friday, January 21, i.e., the day after tomorrow.
- Each week there is one home assignment: The solution to the assignment in the exercise paper of the week m is to be *returned* to the course assistant Stratos Staboulis/Matti Leinonen before the exercise session of the week $m + 1$. (For example, the solution to the home assignment of the first exercise paper should be returned before the exercise session on Friday, January 28.)
- The course assistant will demonstrate 'model' solutions to the exercises.

Evaluation

The course grades will be based on the weekly *home assignments* and a *home exam*.

- The home assignments constitute 25% of the grade. Each returned solution is given $0-3$ points; at the end of the course, the obtained points will be summed and scaled appropriately.
- The home exam constitutes 75% of the grade. It will be held after the lectures have ended — the exact timing will be agreed upon later on. There will be a few, more extensive assignments that must be solved within a given time period (e.g., within one week).

Timetable

The course extends over nine or ten weeks (plus lecture breaks).

- The first half will concentrate on traditional regularization techniques (Staboulis as the course assistant).
- The second half will examine inverse problems from a statistical view point (Leinonen as the course assistant).

1 What is an ill-posed problem?

Well-posed problems

Jacques Salomon Hadamard (1865-1963):

1. A solution exists.
2. The solution is unique.
3. The solution depends continuously on the data, in some *reasonable* topology.

Ill-posed problems

Nuutti Hyvönen: The ill-posed problems are the complement of the well-posed problems in the space of all problems.

Examples:

- Interpolation.
- Finding the cause of a known consequence $\Longrightarrow$ inverse problems.
- Almost all problems encountered in everyday life.

When solving an ill-posed or inverse problem, it is essential to use all possible prior and expert knowledge about the possible solutions.

An example: Heat distribution in an insulated rod

Let us consider the problem

$$\begin{aligned} u_t &= u_{xx} && \text{in } (0,\pi)\times\mathbb{R}_+, \\ u_x(0,\cdot) &= u_x(\pi,\cdot) = 0 && \text{on } \mathbb{R}_+, \\ u(\cdot,0) &= f && \text{on } (0,\pi), \end{aligned}$$

where $u(\cdot,t)$ is the heat distribution at the time $t>0$, f is the initial heat distribution, and the boundary conditions indicate that the heat cannot flow out of the 'rod' $[0,\pi]$.

Forward problem: Determine the 'final' distribution $u(\cdot,T)\in L^2(0,\pi)$, $T>0$, if the initial distribution $f\in L^2(0,\pi)$ is known.

Inverse problem: Determine the initial distribution $f\in L^2(0,\pi)$, if the (noisy) 'final' distribution $u(\cdot,T)=:w\in L^2(0,\pi)$ is known.

Forward problem

The solution to the forward problem can be given explicitly:

$$u(x,T) = \sum_{n=0}^{\infty} \hat{f}_n e^{-n^2 T} \cos(nx),$$

where $\{\hat{f}_n\}_{n=0}^{\infty} \subset \mathbb{R}$ are Fourier cosine coefficients of the initial heat distribution f, i.e., $f = \sum_{n=0}^{\infty} \hat{f}_n \cos(nx)$ in the sense of $L^2(0,\pi)$.

It is relatively easy to see that the solution operator

$$E_T : f \mapsto u(\cdot, T), \quad L^2(0,\pi) \to L^2(0,\pi)$$

satisfies the following conditions:

- E_T is linear, bounded and *compact*.
- E_T is injective, i.e., $\mathrm{Ker}(E_T) = \{0\}$.
- $\mathrm{Ran}(E_T)$ is dense in $L^2(0,\pi)$.

Inverse problem

Solving the inverse problem for a general final heat distribution $w \in L^2(0,\pi)$ corresponds to inverting the compact operator $E_T : L^2(0,\pi) \to L^2(0,\pi)$, which is obviously impossible.

The *unbounded* solution operator

$$E_T^{-1} : \mathrm{Ran}(E_T) \to L^2(0,\pi)$$

is, however, well-defined. In other words, the inverse problem has a unique solution if $w = E_T f$ for some $f \in L^2(0,\pi)$, i.e., the measurement contains no noise.

Summary:

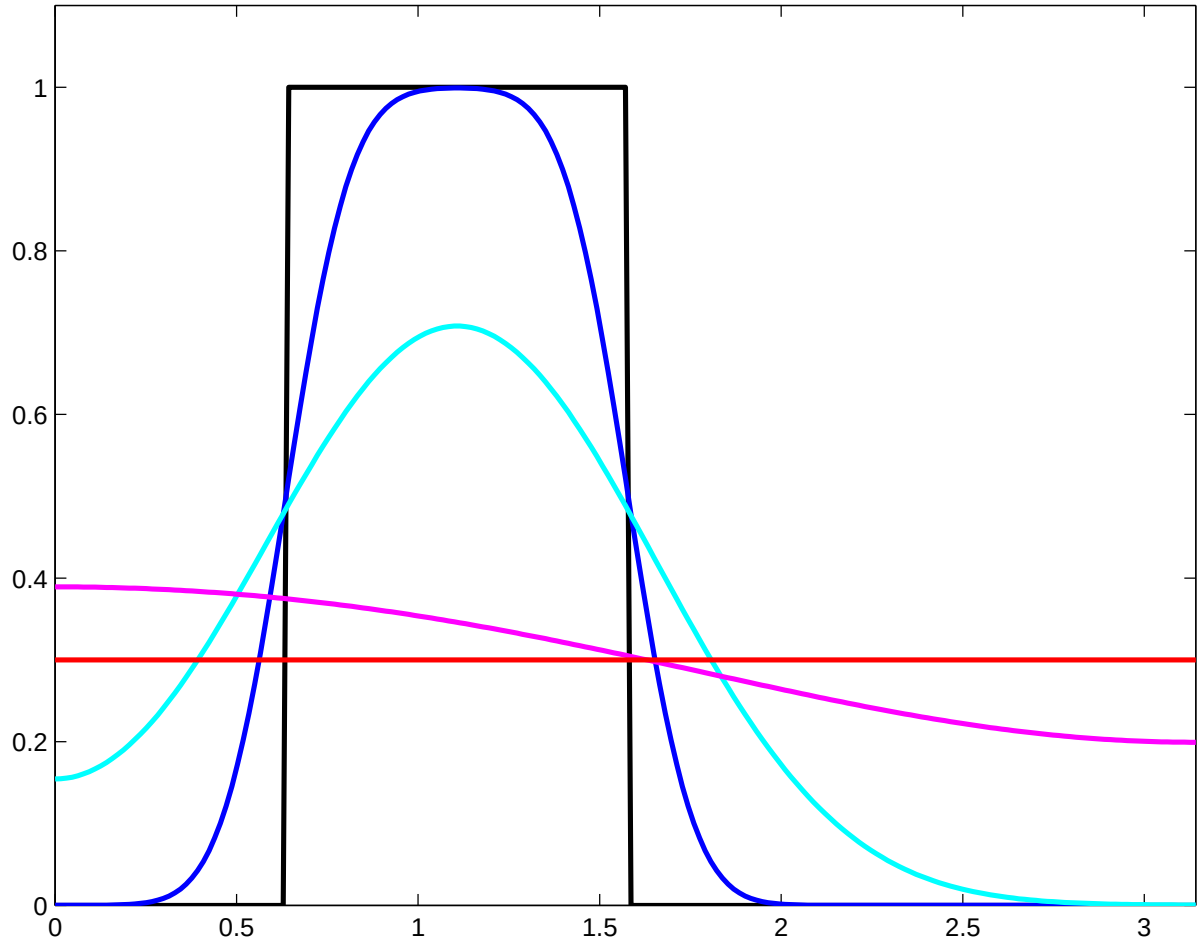
- If $w \in \mathrm{Ran}(E_T)$, the third Hadamard condition is not satisfied.
- If $w \notin \mathrm{Ran}(E_T)$, none of the Hadamard conditions is satisfied.

(Due to noise etc., the latter case is usually the valid one in practice.)

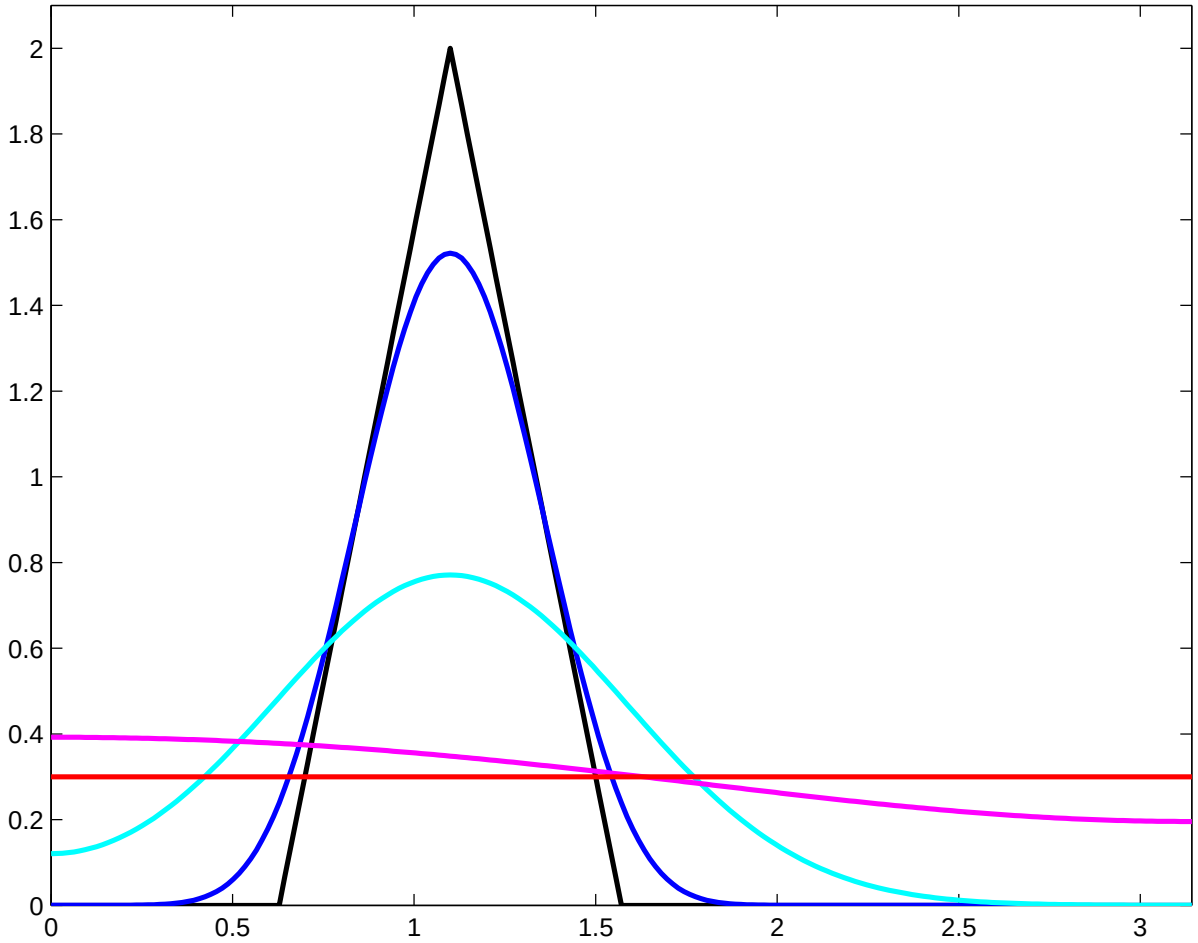
Question: Should one then ignore the ill-posed inverse problem?

Answer: No. The available measurement *always* contains *some* information about the initial heat distribution.

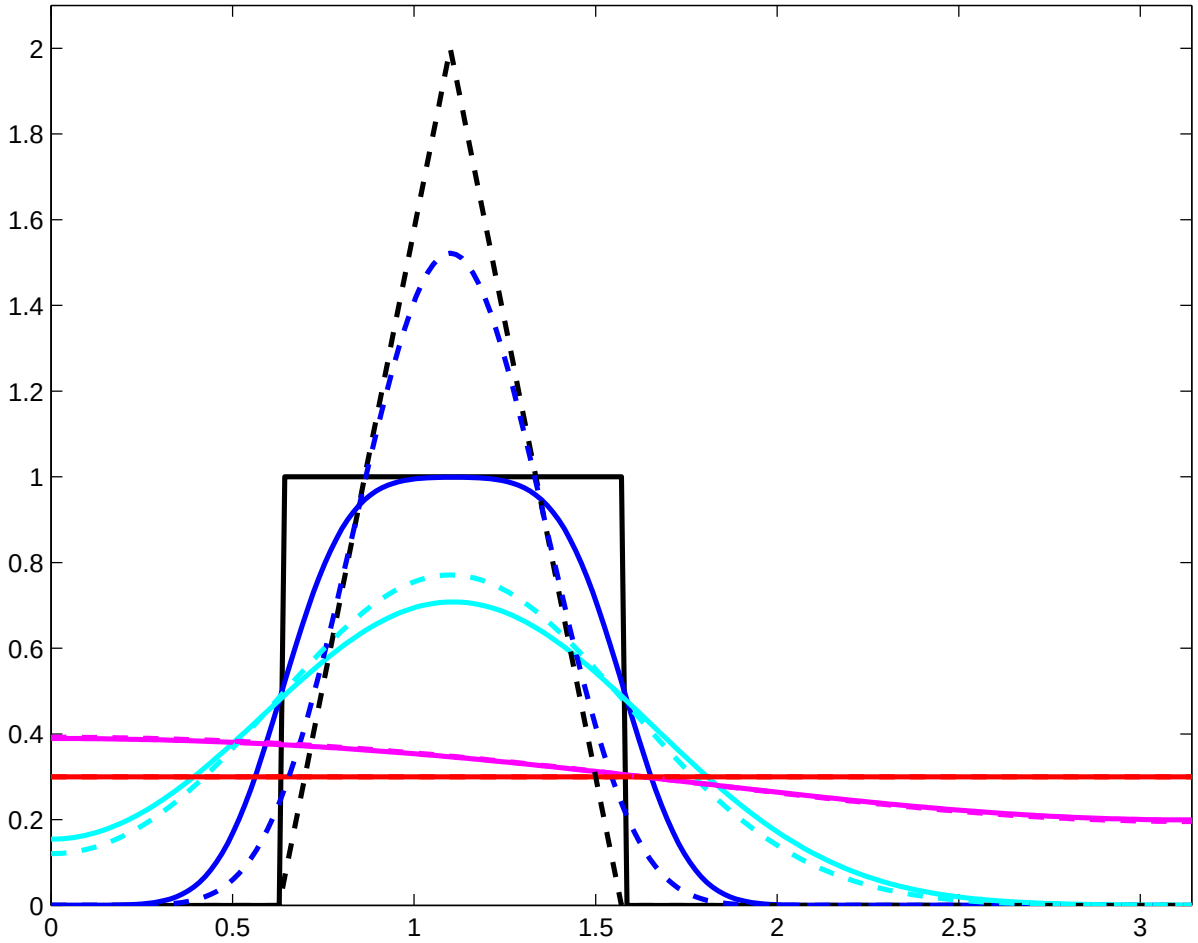
Heat distribution at $t = 0, 0.01, 0.1, 1$ and 10.



Another heat distribution at $t = 0, 0.01, 0.1, 1$ and 10.



Comparison of the two at $t = 0, 0.01, 0.1, 1$ and 10.



2 Classical regularization methods

2.1 Fredholm equation

Separable Hilbert space

A vector space H is a *real inner product space* if there exists a mapping $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ satisfying

1. $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in H$.
2. $\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ for all $x_1, x_2, y \in H$, $a, b \in \mathbb{R}$.
3. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

Furthermore, H is a *separable real Hilbert space* if, in addition,

1. H is *complete* with respect to the norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$.
2. There exists a *countable orthonormal basis* $\{\varphi_n\}$ of H with respect to the inner product $\langle \cdot, \cdot \rangle$. This means that

$$\langle \varphi_j, \varphi_k \rangle = \delta_{jk} \quad \text{and} \quad x = \sum_n \langle x, \varphi_n \rangle \varphi_n \quad \text{for all } x \in H.$$

Fredholm equation

Let $A : H_1 \to H_2$ be a *compact* linear operator between the real separable Hilbert spaces H_1 and H_2. In the first half of this course, we mainly concentrate on the problem of finding $x \in H_1$ satisfying the equation

$$Ax = y, \tag{1}$$

where $y \in H_2$ is given. (In this setting, compact operators are the closure of the finite-dimensional operators, i.e., loosely speaking matrices, in the operator topology.)

Examples:

- In the example of Section 1, we have $A = E_T$ and $H_1 = H_2 = L^2(0, \pi)$.
- The most important case on this course is $H_1 = \mathbb{R}^n$, $H_2 = \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ is a matrix.

2.2 Truncated singular value decomposition

Orthogonal decompositions

Let $A^* : H_2 \to H_1$ be the adjoint operator of $A : H_1 \to H_2$, i.e.,

$$\langle Ax, y\rangle = \langle x, A^* y\rangle \quad \text{for all } x \in H_1, y \in H_2.$$

We have the orthogonal decompositions

$$\begin{aligned} H_1 &= \mathrm{Ker}(A) \oplus (\mathrm{Ker}(A))^\perp = \mathrm{Ker}(A) \oplus \overline{\mathrm{Ran}(A^*)}, \\ H_2 &= \overline{\mathrm{Ran}(A)} \oplus (\mathrm{Ran}(A))^\perp = \overline{\mathrm{Ran}(A)} \oplus \mathrm{Ker}(A^*), \end{aligned}$$

where the "bar" denotes the closure of a set and

$$\begin{aligned} \mathrm{Ker}(A) &= \{x \in H_1 \mid Ax = 0\}, \\ \mathrm{Ran}(A) &= \{y \in H_2 \mid y = Ax \text{ for some } x \in H_1\}, \\ (\mathrm{Ker}(A))^\perp &= \{x \in H_1 \mid \langle x, z\rangle = 0 \text{ for all } z \in \mathrm{Ker}(A)\}, \quad \text{etc.} \end{aligned}$$

Characterization of compact operators

There exist (possible countably infinite) orthonormal sets of vectors $\{v_n\} \subset H_1$ and $\{u_n\} \subset H_2$, and a sequence of *positive* numbers $\{\lambda_n\}$, $\lambda_k \geq \lambda_{k+1}$ and $\lim_{n\to\infty} \lambda_n = 0$ in the countably infinite case, such that

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n \qquad \text{for all } x \in H_1 \tag{2}$$

and, in particular,

$$\overline{\mathrm{Ran}(A)} = \overline{\mathrm{span}\{u_n\}} \qquad \text{and} \qquad (\mathrm{Ker}(A))^{\perp} = \overline{\mathrm{span}\{v_n\}}.$$

(Conversely, if $A : H_1 \to H_2$ has this kind of decomposition, it is compact.)

The system $\{v_n, u_n, \lambda_n\}$ is called a *singular system* of A, and (2) is a *singular value decomposition* (SVD) of A. (Note that $1 \leq n \leq \infty$ or $1 \leq n \leq N < \infty$ depending on $\mathrm{rank}(A) := \dim(\mathrm{Ran}(A))$.)

Solvability of $Ax = y$

It follows from the orthonormality of $\{u_n\}$ that

$$P : H_2 \to \overline{\mathrm{Ran}(A)}, \quad y \mapsto \sum_n \langle y, u_n \rangle u_n,$$

is an orthogonal projection, i.e., $P^2 = P$ and $\mathrm{Ran}(P) \perp \mathrm{Ran}(I - P)$.

The equation $Ax = y$ has a solution *if and only if*

$$y = Py \qquad \text{and} \qquad \sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 < \infty. \tag{3}$$

In case that (3) is satisfied, all solutions of $Ax = y$ are of the form

$$x = x_0 + \sum_n \frac{1}{\lambda_n} \langle y, u_n \rangle v_n$$

for some $x_0 \in \mathrm{Ker}(A)$.

Intuitive interpretation of the solvability conditions:

- The first condition, $y = Py$, states that y cannot have components in the orthogonal complement of $\overline{\mathrm{Ran}(A)}$ if $y = Ax$.
- The second condition, i.e., the convergence of the series

$$\sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2,$$

  is redundant if $\mathrm{rank}(A) < \infty$, in which case $\overline{\mathrm{Ran}(A)} = \mathrm{Ran}(A)$. On the other hand, if $\mathrm{rank}(A) = \infty$, this condition is equivalent to asking that the norm of

$$x = x_0 + \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \langle y, u_n \rangle v_n, \qquad x_0 \in \mathrm{Ker}(A),$$

  is finite, i.e., the 'potential solutions' belong to H_1.

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Second lecture, January 21, 2011.

2.2 Truncated singular value decomposition (cont.)

Summary of the previous lecture

The problem: Find $x \in H_1$ that satisfies the equation

$$Ax = y,$$

where $y \in H_2$ is given and $A : H_1 \to H_2$ is a compact linear operator.

Singular value decomposition (SVD):

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n \qquad \text{for all } x \in H_1.$$

The solutions: If solutions exist, they are of the form

$$x = x_0 + \sum_n \frac{1}{\lambda_n} \langle y, u_n \rangle v_n,$$

where $x_0 \in \mathrm{Ker}(A)$.

Solvability conditions: There exists a solution *if and only if*

$$y = Py \qquad \text{and} \qquad \sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 < \infty,$$

where P is a projection onto $\overline{\mathrm{Ran}(A)} = \overline{\mathrm{span}\{u_n\}}$.

The natural way to circumvent problems with the first solvability condition is to consider the projected equation

$$Ax = PAx = Py$$

instead of $Ax = y$. However, this does not help with the second condition since there is no guarantee that

$$\sum_n \frac{1}{\lambda_n^2} |\langle Py, u_n \rangle|^2 < \infty$$

for a general $y \in H_2$, if $\mathrm{rank}(A) = \infty$, i.e., if $\mathrm{Ran}(A)$ is infinite-dimensional.

Truncated singular value decomposition (TSVD)

Let us define a family of finite-dimensional orthogonal projections by

$$P_k : H_2 \to \mathrm{span}\{u_1, \ldots, u_k\}, \quad y \mapsto \sum_{n=1}^{k} \langle y, u_n \rangle u_n.$$

Due to the orthogonality of $\{u_n\}$,

$$P(P_k y) = \sum_n \langle P_k y, u_n \rangle u_n = \sum_{n=1}^{k} \langle y, u_n \rangle u_n = P_k y,$$

and, moreover,

$$\sum_n \frac{1}{\lambda_n^2} |\langle P_k y, u_n \rangle|^2 = \sum_{n=1}^{k} \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 < \infty.$$

(Note that one must choose $k \leq \mathrm{rank}(A)$ if the latter is finite.)

In consequence, the problem

$$Ax \;=\; P_k y. \tag{4}$$

satisfies the solvability conditions (3). The corresponding solutions are given by

$$x \;=\; x_0 + \sum_n \frac{1}{\lambda_n}\langle P_k y, u_n\rangle v_n \;=\; x_0 + \sum_{n=1}^{k} \frac{1}{\lambda_n}\langle y, u_n\rangle v_n \in H_1.$$

By the *truncated SVD solution* of $Ax = y$ for a given $k \geq 1$, we mean the $x_k \in H_1$ that satisfies (4) and is orthogonal to the subspace $\mathrm{Ker}(A)$. Since $\{v_n\}$ span $(\mathrm{Ker}(A))^{\perp}$, it easily follows that such x_k is *unique*, has the *smallest norm* of the solutions to (4), and is given by

$$x_k \;=\; \sum_{n=1}^{k} \frac{1}{\lambda_n}\langle y, u_n\rangle v_n.$$

An example: Heat distribution in a rod (revisited)

Recall the heat equation

$$\begin{aligned} u_t &= u_{xx} && \text{in } (0,\pi)\times\mathbb{R}_+, \\ u_x(0,\cdot) &= u_x(\pi,\cdot) = 0 && \text{on } \mathbb{R}_+, \\ u(\cdot,0) &= f && \text{on } (0,\pi). \end{aligned}$$

The forward solution operator

$$E_T : f \mapsto u(\cdot,T), \quad H_1 = L^2(0,\pi) \to L^2(0,\pi) = H_2$$

is characterized by

$$E_T : v_n \mapsto \lambda_n v_n,$$

where $\{v_n\}_{n=0}^{\infty} = \{\sqrt{\frac{1}{\pi}}\} \cup \{\sqrt{\frac{2}{\pi}}\cos(n\,\cdot)\}_{n=1}^{\infty}$ form an orthonormal basis of $L^2(0,\pi)$, and $\lambda_n = \lambda_n(T) = e^{-n^2T} > 0$ converges to zero as $n \to \infty$.

In consequence, we have

$$E_T f = \sum_{n=0}^{\infty} \lambda_n \langle f, v_n \rangle v_n,$$

where the inner product of $L^2(0,\pi)$ is defined in the usual way:

$$\langle f, g \rangle = \int_0^{\pi} f g \, dx, \qquad f, g \in L^2(0,\pi).$$

In this case $u_n = v_n$ (because E_T is self-adjoint). Since $\{v_n\}_{n=0}^{\infty}$ are an orthonormal basis for $L^2(0,\pi)$, we have

$$(\mathrm{Ker}(E_T))^{\perp} = \overline{\mathrm{Ran}(E_T)} = L^2(0,\pi),$$

i.e., E_T is injective and has a dense range, as mentioned already earlier. In particular, the projection onto the closure of the range of E_T is the identity operator, i.e., $P = I$.

We thus deduce that there exists $f \in L^2(0,\pi)$ such that

$$E_T f = w,$$

for a given $w \in L^2(0,\pi)$, if and only if

$$\sum_{n=0}^{\infty} \frac{1}{\lambda_n^2} |\langle w, v_n \rangle|^2 = \sum_{n=0}^{\infty} e^{n^4 T^2} |\langle w, v_n \rangle|^2 < \infty,$$

which is a very restrictive condition and demonstrates why this inverse problem is extremely ill-posed.

Finally, note that the truncated SVD solution to this inverse problem is

$$f_k = \sum_{n=0}^{k} \frac{1}{\lambda_n} \langle w, v_n \rangle v_n = \sum_{n=0}^{k} e^{n^2 T} \langle w, v_n \rangle v_n, \qquad k \geq 0.$$

The special case: $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$

Let $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$, which means that

$$Ax = y$$

is a matrix equation or, in other words, a system of linear equations. In particular, $A \in \mathbb{R}^{m \times n}$.

Since all operators of finite rank, i.e., with finite-dimensional range, are compact, we have the representation

$$Ax = \sum_{j=1}^{p} \lambda_j (x^{\mathrm{T}} v_j) u_j = \sum_{j=1}^{p} \lambda_j u_j (v_j^{\mathrm{T}} x), \qquad p \leq \min\{n, m\},$$

where $\{v_j\}_{j=1}^p \subset \mathbb{R}^n$ and $\{u_j\}_{j=1}^p \subset \mathbb{R}^m$ are sets of orthonormal vectors and $\{\lambda_j\}_{j=1}^p$ are positive numbers such that $\lambda_j \geq \lambda_{j+1}$. (Note that $p = \mathrm{rank}(A)$.)

How can one write this decomposition in a neat matrix form?

Let us introduce, e.g., by Gram–Schmidt process, complementary sets of orthonormal vectors $\{v_j\}_{j=p+1}^{n}$ and $\{u_j\}_{j=p+1}^{m}$, such that the completed systems $\{v_j\}_{j=1}^{n}$ and $\{u_j\}_{j=1}^{m}$ are orthonormal basis for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Moreover, we set $\lambda_j = 0$ for $j = p+1, \ldots, \min\{n, m\}$.

Next, we define three auxiliary matrices:

$$\begin{aligned} V &= [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}, \\ U &= [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}, \\ \Lambda &= \mathrm{diag}(\lambda_1, \ldots, \lambda_{\min\{n,m\}}) \in \mathbb{R}^{m\times n}. \end{aligned}$$

Here, $\Lambda \in \mathbb{R}^{m\times n}$ is a diagonal matrix with the elements $\lambda_1, \ldots, \lambda_{\min\{n,m\}}$ on its diagonal; if $m > n$ (resp. $n > m$), there are $m - n$ empty rows (resp. $n - m$ empty columns) at the bottom of Λ (resp. at the right end of Λ). Note that due to the orthonormality of $\{v_j\}$ and $\{u_j\}$, the matrices V and U are orthogonal:

$$V^{\mathrm{T}}V = VV^{\mathrm{T}} = I \qquad \text{and} \qquad U^{\mathrm{T}}U = UU^{\mathrm{T}} = I.$$

A simple computation shows that

$$U\Lambda V^{\mathrm{T}}x \;=\; \sum_{j=1}^{p} \lambda_j u_j (v_j^{\mathrm{T}} x) \;=\; Ax$$

for all $x \in \mathbb{R}^n$. Hence, we have the decomposition

$$A \;=\; U\Lambda V^{\mathrm{T}}.$$

This is what we call the SVD in the case of matrices in $\mathbb{R}^{m\times n}$. In particular, this is how Matlab understands the SVD.

Note, in particular, that the singular values $\{\lambda_j\}_{j=1}^{\min\{n,m\}}$ are just *non-negative* — earlier they were assumed to be positive —, and

$$\begin{aligned}
\mathrm{Ran}(A) &= \mathrm{span}\{u_j \mid 1 \leq j \leq p\}, \\
\mathrm{Ker}(A) &= \mathrm{span}\{v_j \mid p+1 \leq j \leq n\}, \\
(\mathrm{Ran}(A))^{\perp} &= \mathrm{span}\{u_j \mid p+1 \leq j \leq m\}, \\
(\mathrm{Ker}(A))^{\perp} &= \mathrm{span}\{v_j \mid 1 \leq j \leq p\}.
\end{aligned}$$

Truncated SVD for a matrix $A \in \mathbb{R}^{m\times n}$

The truncated SVD solution, i.e., the solution of

$$Ax \;=\; P_k y \quad \text{and} \quad x \perp \mathrm{Ker}(A), \qquad\qquad 1 \le k \le p,$$

where $P_k \to \mathrm{span}\{u_1, \ldots, u_k\}$ is an orthogonal projection, is given in the matrix framework by

$$x_k \;=\; \sum_{j=1}^{k} \frac{1}{\lambda_j} \langle y, u_j \rangle v_j \;=\; \sum_{j=1}^{k} \frac{1}{\lambda_j} v_j (u_j^{\mathrm{T}} y) \;=\; V \Lambda_k^{\dagger} U^{\mathrm{T}} y.$$

Here, $\Lambda_k^{\dagger} \in \mathbb{R}^{n\times m}$ is a diagonal matrix, with $\min\{m, n\}$ number of non-negative elements $1/\lambda_1, \ldots, 1/\lambda_k, 0, \ldots, 0$ on its diagonal.

For the largest possible cut-off $k = p$, the matrix

$$A^\dagger := A_p^\dagger = V\Lambda_p^\dagger U^{\mathrm{T}} =: V\Lambda^\dagger U^{\mathrm{T}}$$

is called the *Moore–Penrose pseudoinverse*. It follows from the above material that $x^\dagger = A^\dagger y$ is the solution of the projected equation

$$Ax = P_p y = Py,$$

where $P : \mathbb{R}^m \to \mathbb{R}^m$ is, once again, the orthogonal projection onto $\mathrm{Ran}(A)$. However, since the smallest non-zero singular value λ_p is typically extremely small in inverse problems, the use of pseudoinverse is often very sensitive to inaccuracies in the data y.

An example: Heat distribution in a rod (revisited)

Recall once again the heat equation

$$\begin{aligned} u_t &= u_{xx} && \text{in } (0,\pi)\times\mathbb{R}_+, \\ u_x(0,\cdot) &= u_x(\pi,\cdot) = 0 && \text{on } \mathbb{R}_+, \\ u(\cdot,0) &= f && \text{on } (0,\pi). \end{aligned}$$

Our plan is to discretize the dependence on the spatial variable x, and then investigate the properties of the corresponding inverse problem numerically.

To begin with, we introduce the step size $h = \pi/100$ and the grid points $x_j = jh$, $j = 0,\ldots,100$. Furthermore, we denote $U_j(t) = u(x_j,t)$.

We approximate the second derivative of u with respect to x at the point (x_j, t) by the difference rule:

$$u_{xx}(x_j, t) \approx \frac{1}{h^2} \left(U_{j-1}(t) - 2\, U_j(t) + U_{j+1}(t) \right), \qquad 1 \leq j \leq 99.$$

Furthermore, we discretize the boundary conditions by requiring that

$$u_x(0, t) \approx \frac{1}{h} \left(U_1(t) - U_0(t) \right) = 0 = \frac{1}{h} \left(U_{100}(t) - U_{99}(t) \right) \approx u_x(\pi, t).$$

By solving this for $U_0(t)$ and $U_{100}(t)$ and substituting into the preceding difference rule, we obtain altogether that

$$\begin{aligned}
u_{xx}(x_1, t) &\approx \frac{1}{h^2} \left(-U_1(t) + U_2(t) \right), \\
u_{xx}(x_j, t) &\approx \frac{1}{h^2} \left(U_{j-1}(t) - 2\, U_j(t) + U_{j+1}(t) \right), \qquad 2 \leq j \leq 98, \\
u_{xx}(x_{99}, t) &\approx \frac{1}{h^2} \left(U_{98}(t) - U_{99}(t) \right).
\end{aligned}$$

Denoting $U(t) = (U_1(t), \ldots, U_{99}(t))^{\mathrm{T}}$ and $F = (f(x_1), \ldots, f(x_{99}))^{\mathrm{T}}$ and plugging the above approximations into the heat equation, we end up with a set of ordinary differential equations:

$$\begin{aligned} U'(t) &= B\,U(t), \qquad t \in \mathbb{R}_+, \\ U(0) &= F, \end{aligned}$$

where $B \in \mathbb{R}^{99\times 99}$ is a certain tridiagonal matrix (see next slide).

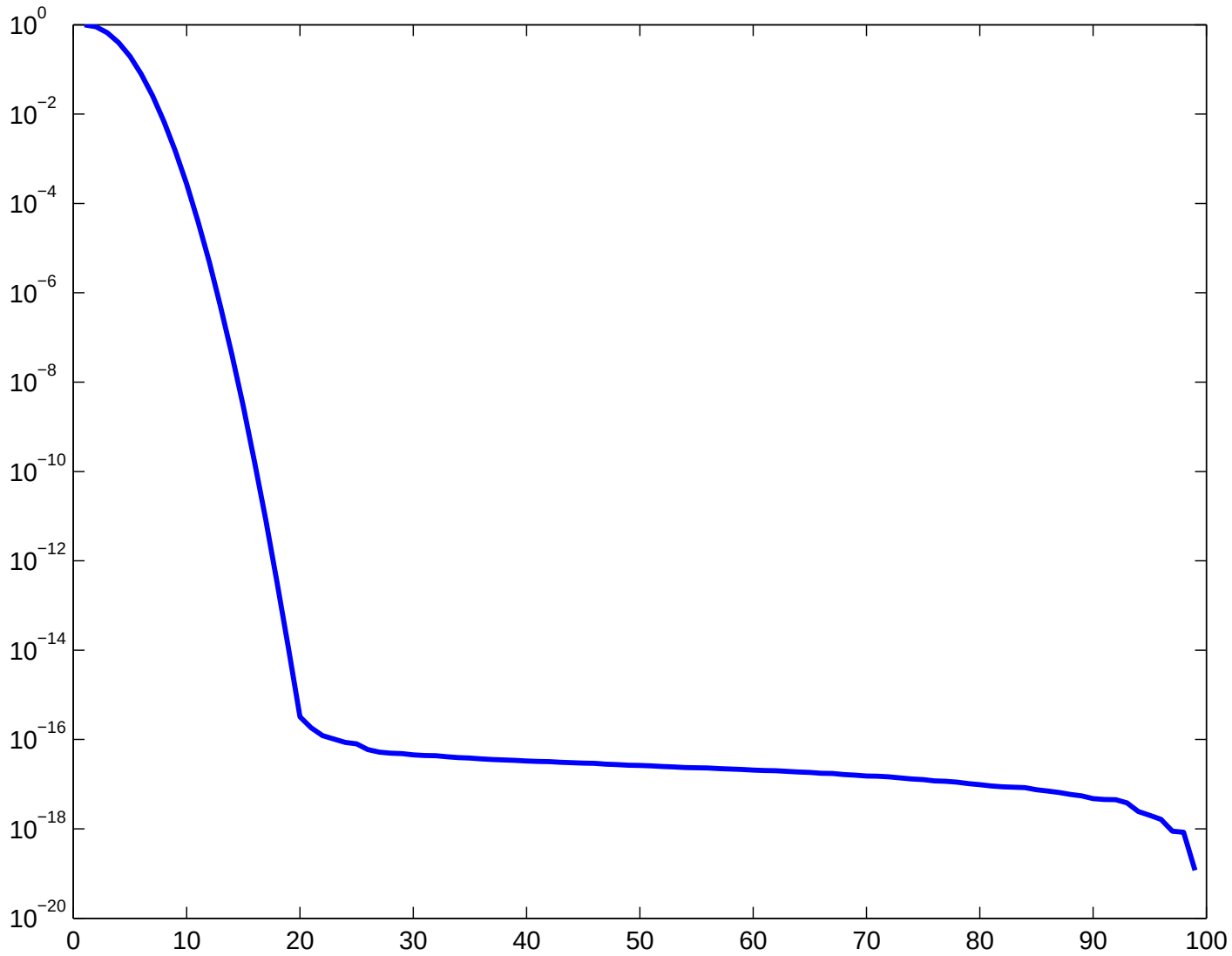
The forward solution corresponding to this space-discretized problem can be given with the help of the matrix exponent function as

$$U(T) = AF,$$

where $A = A(T) = e^{TB}$ and $T > 0$.

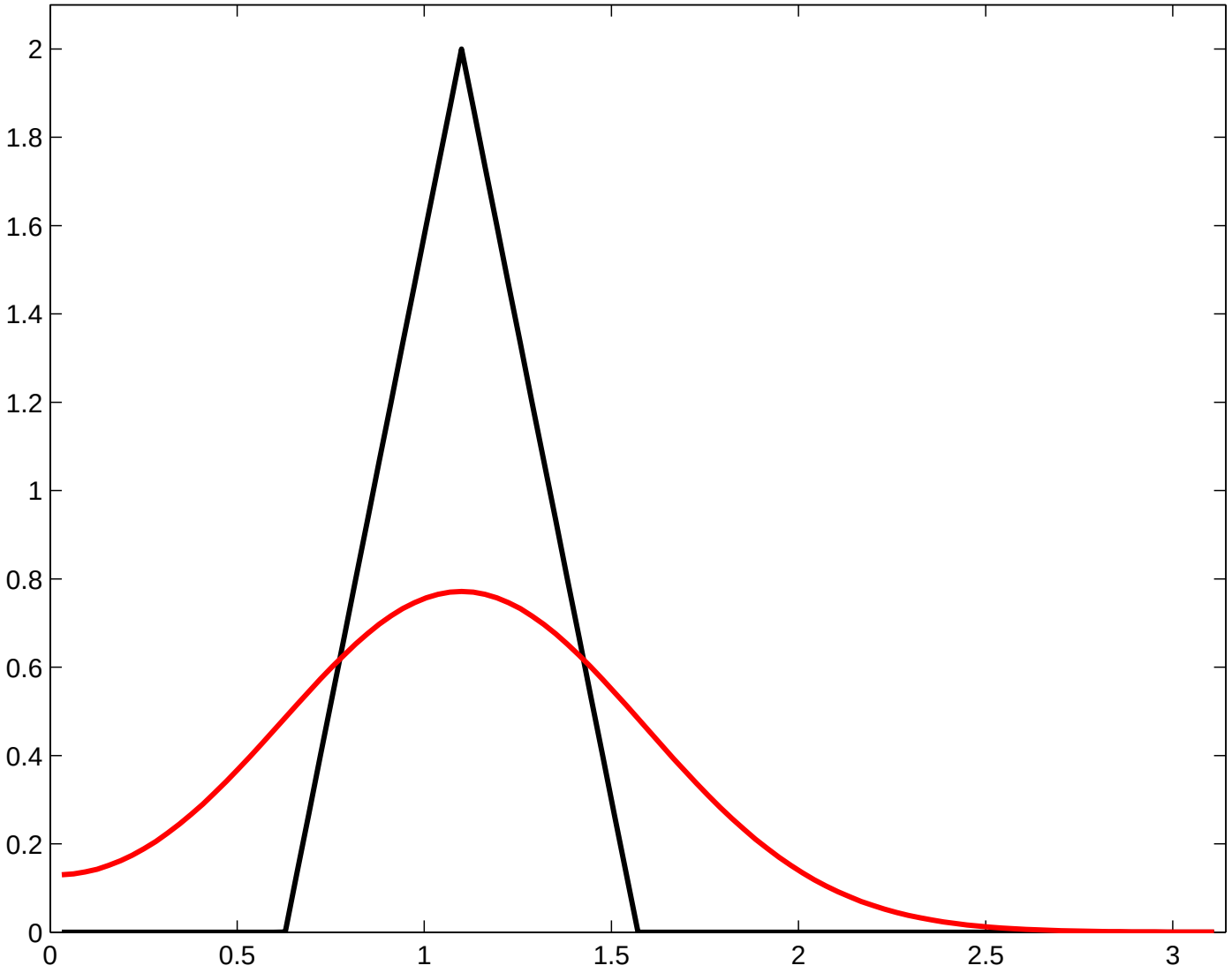
In Matlab, the matrices B and $A = e^{TB}$ can be formed by the following script, which also forms the SVD and plots the singular values for A:

```
T = 0.1; % say
N = 100;
h = pi/N;
B = diag(ones(N-2,1),-1) - 2*eye(N-1) + diag(ones(N-2,1),1);
B(1,1) = -1; % the left boundary condition
B(N-1,N-1) = -1; % the right boundary condition
B = B/h^2;
A = expm(T*B);
[U S V] = svd(A); % SVD
semilogy(diag(S), 'LineWidth', 2);
```


10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{-18}
10^{-20}
0
10
20
30
40
50
60
70
80
90
100

Let us next form a 'wedge function', which serves as the initial heat distribution, and compute the corresponding final distribution at $T = 0.1$:

```
x = linspace(h,pi-h,N-1); % the grid points
a = 40/3/pi; b1 = -8/3; b2 = 20/3; % coefficients
f = [a*x(1:35) + b1, -a*x(36:end) + b2]';
ind = f > 0;
f = f.*ind;
w = A*f; % final distribution
plot(x, f, 'k', 'LineWidth', 2);
hold on
plot(x, w, 'r', 'LineWidth', 2);
axis([0, pi, 0, 2.1])
hold off
```


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3

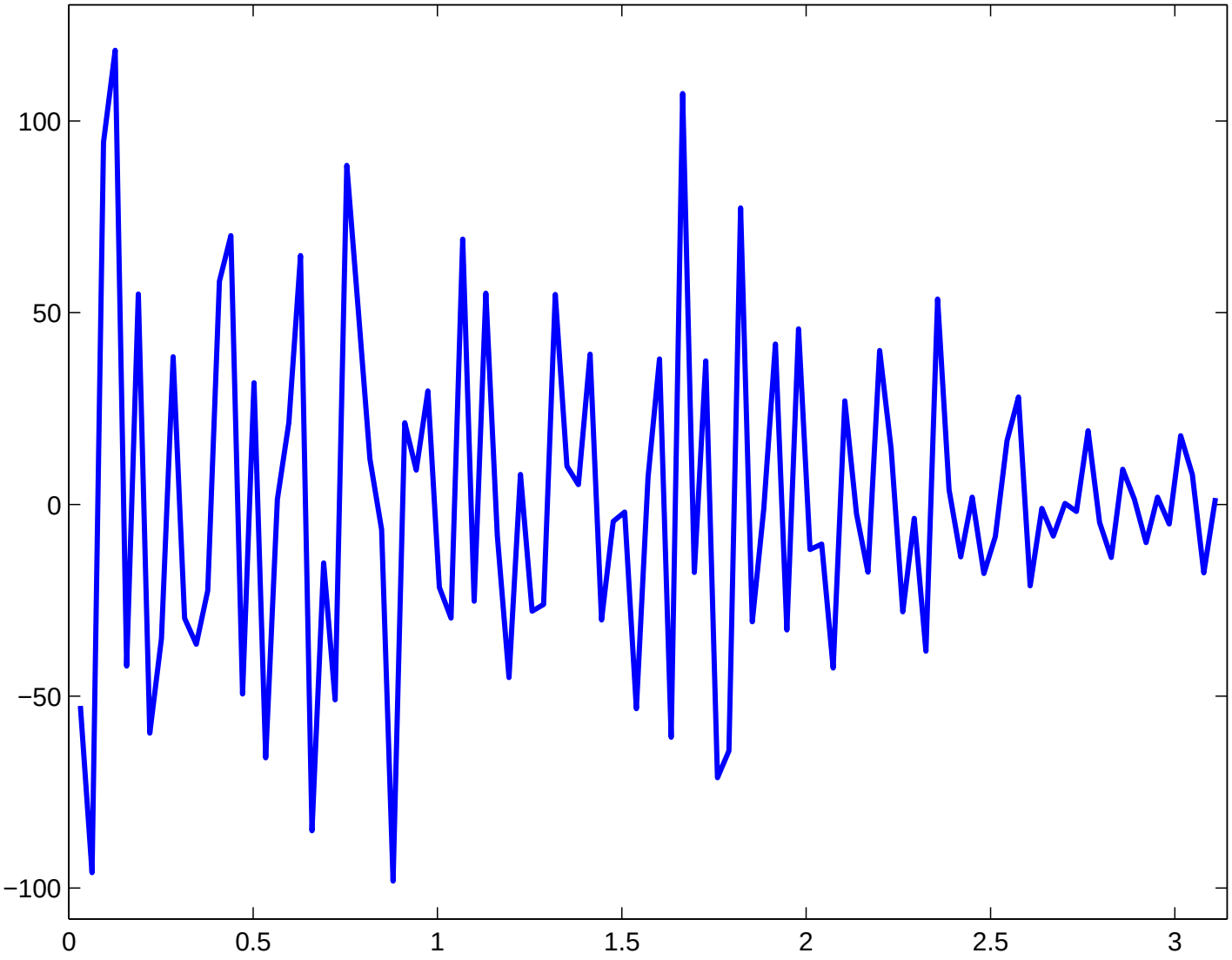
Let us be a bit silly and try to recover the initial heat distribution by inverting A:

```
f_stupid = A\w;
plot(x, f_stupid, 'LineWidth', 2);
```

which results in a catastrophe as demonstrated on the next slide. This is not surprising since writing

```
rank(A)
```

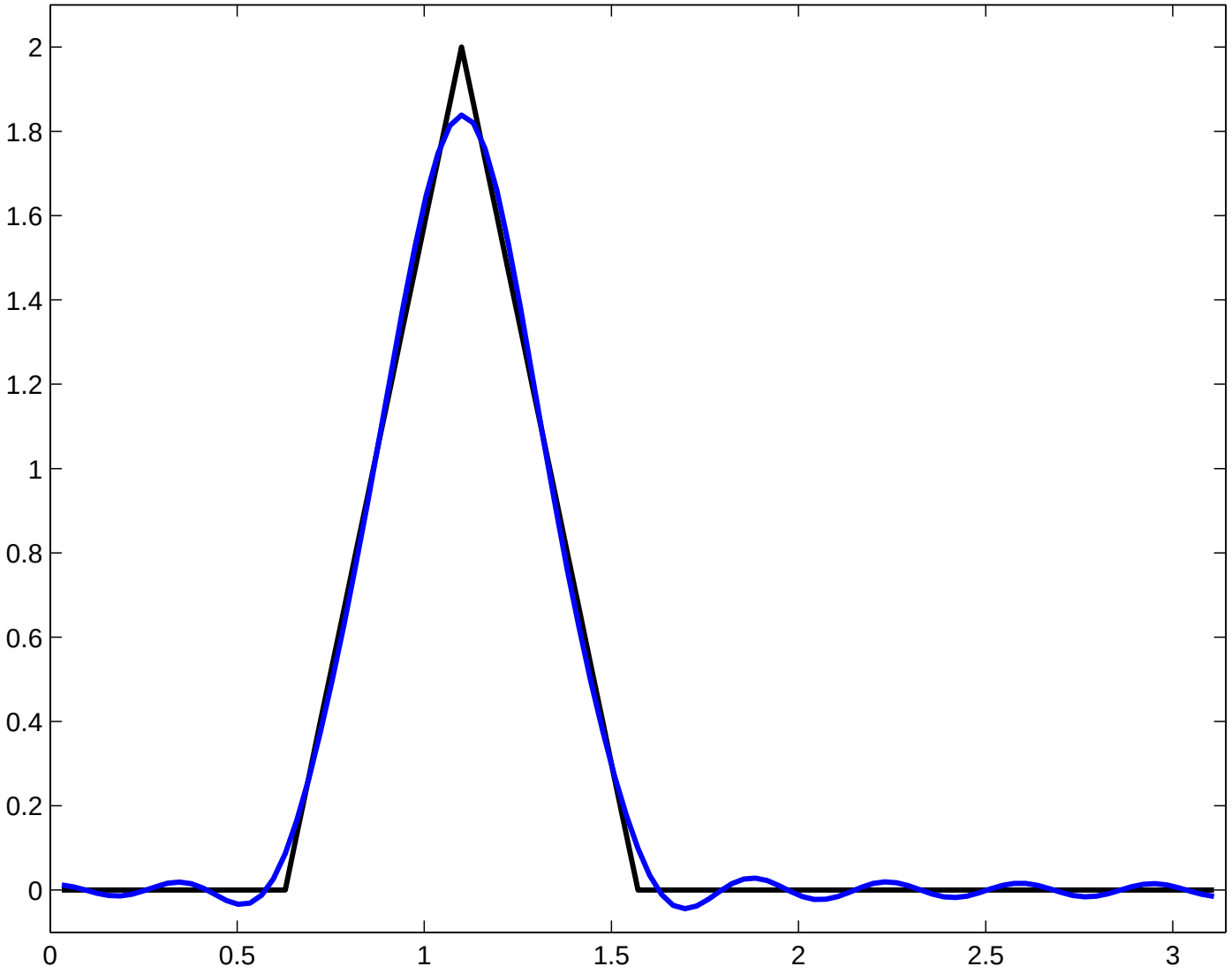
in Matlab, gives the value 18. In other words, from Matlab's numerical point of view, A has only 18 linearly independent columns — in particular, A is not (numerically) invertible.


100
50
0
−50
−100
0
0.5
1
1.5
2
2.5
3

Let us be more clever and compute the truncated SVD solution for $k = 18$:

```
k = 18; % the (numerical) rank of A
d = diag(S); % the singular values
idk = [1./d(1:k); zeros((N-1)-k,1)]; % invert only 18
iBk = V*diag(idk)*U'; % the corresponding 'inverse'
fk = iBk*w; % the 'solution'
plot(x, f, 'k','LineWidth', 2); hold on
plot(x, fk, 'LineWidth', 2); hold off
```

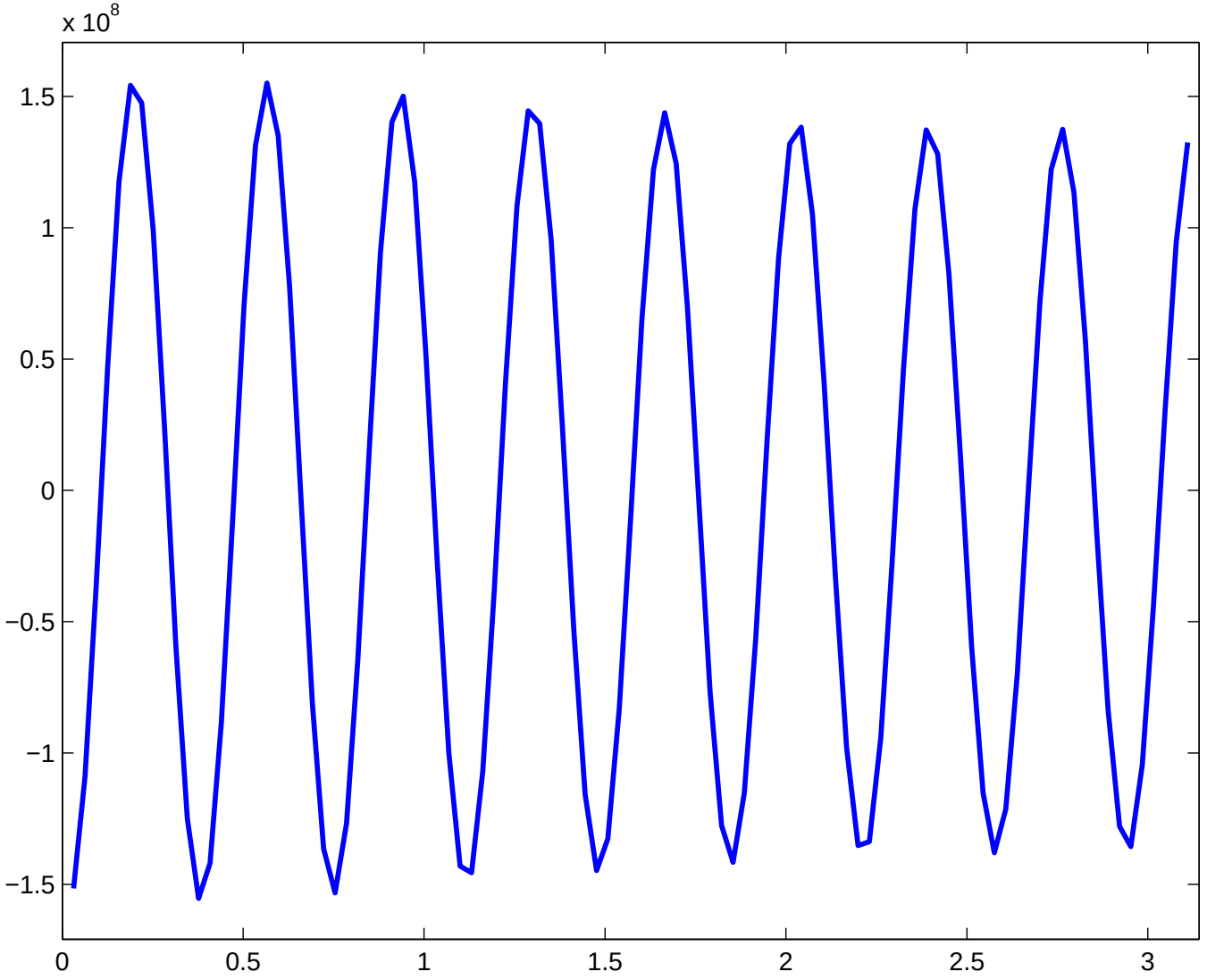
We have, actually, committed a severe *inverse crime*: If an inverse problem is solved using the same discretization with which the data was generated, the results are typically overly optimistic. This problem could be circumvented, e.g., by interpolating onto a sparser grid before the inversion. The 'inverse crime effect' can also be reduced by the addition of artificial noise.


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3

In practice, the measurement is always inaccurate. Let us thus add just a tiny bit of noise in the measurement — so tiny that one could barely recognize it with naked eye. (In fact, this noise level corresponds approximately to the discrepancy between data sets simulated with the above introduced difference scheme and with an alternative method based on FFT and the SVD of the original solution operator E_T.)

```
wn = w + 0.001*randn(N-1,1); % noisy data
fkn_stupid = iBk*wn;
plot(x, fkn_stupid, 'LineWidth', 2);
```

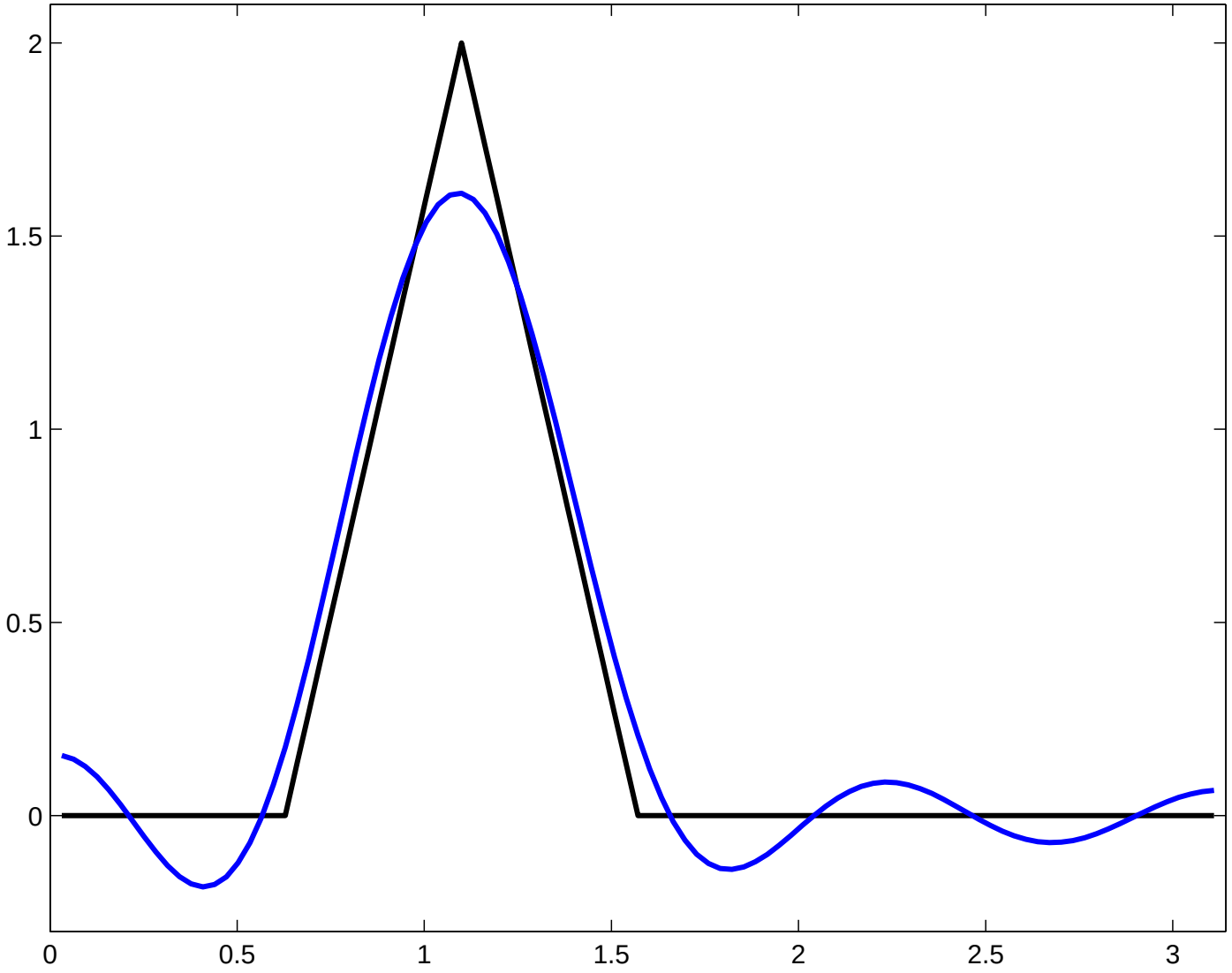
As demonstrated on the next slide, this approach does not work anymore. The reason is the following: The inverse of the 18th singular value is approximately $3.15 \cdot 10^{12}$, which means that the (ever so tiny) component of the noise vector in the direction v_{18} is heavily magnified.


x 10^8
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.5
1
1.5
2
2.5
3

By trial and error, we decide to take the largest $k = 8$ singular values into account when computing the truncated SVD solution:

```
k = 8;
idk = [1./d(1:k); zeros((N-1)-k,1)];
iBk = V*diag(idk)*U';
fkn = iBk*wn;
plot(x, f, 'k','LineWidth', 2);
hold on
plot(x, fkn, 'LineWidth', 2);
hold off
```

This is pretty much the best one can do without additional information about the shape of the initial heat distribution. (For example, if we knew beforehand that f is piecewise linear, such information could be incorporated in the inversion algorithm, which would surely result in better reconstructions.)


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Third lecture, January 26, 2011.

Summary of the previous lecture

The truncated SVD solution: For $\mathbb{N} \ni k \leq \mathrm{rank}(A)$, there exist unique $x_k \in H_1$ such that

$$Ax_k = P_k y \qquad \text{and} \qquad x_k \perp \mathrm{Ker}(A).$$

where $P_k : H_2 \to \mathrm{span}\{u_1, \ldots, u_k\}$ is an orthogonal projection. This solution can be given as

$$x_k = \sum_{n=1}^{k} \frac{1}{\lambda_n} \langle y, u_n \rangle v_n.$$

SVD notations for matrices : For a matrix $A \in \mathbb{R}^{m\times n}$, the SVD is usually written as

$$A = U\Lambda V^{\mathrm{T}},$$

where $\Lambda \in \mathbb{R}^{m\times n}$ has the (non-negative!) singular values on its diagonal, and the columns of $V \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ are composed of the (extended!) orthonormal basis $\{v_j\}_{j=1}^n$ and $\{u_j\}_{j=1}^m$, respectively.

The truncated SVD solution for $1 \le k \le p := \mathrm{rank}(A)$ is given by

$$x_k = V\Lambda_k^\dagger U^{\mathrm{T}} y$$

where $\Lambda \in \mathbb{R}^{n\times m}$ has the elements $1/\lambda_1, \ldots, 1/\lambda_k, 0, \ldots, 0$ on its diagonal. The matrix $A^\dagger = V\Lambda_p^\dagger U^{\mathrm{T}}$ is called the Moore–Penrose pseudoinverse of A.

Morozov discrepancy principle

(Let us return to the case where H_1 and H_2 are general separable real Hilbert spaces, and $A : H_1 \to H_2$ is a compact linear operator.)

To make the truncated SVD a more useful tool, one should come up with some rule for choosing the spectral cut-off index $k \geq 1$ appearing in the truncated SVD problem

$$Ax = P_k y \qquad \text{and} \qquad x \perp \mathrm{Ker}(A).$$

Unfortunately, it is difficult (if not impossible) to invent a reliable general scheme of doing this.

However, there exists a widely used rule of thumb called the *Morozov discrepancy principle*.

Assume that the measurement $y \in H_2$ is a noisy version of some underlying 'exact' data vector $y_0 \in H_2$. Furthermore, suppose that we have some estimate on the discrepancy between y and y_0, i.e.,

$$\|y - y_0\| \approx \epsilon > 0.$$

For example, it may be known that

$$y = y_0 + n,$$

where the vector $n \in H_2$ is a realization of some random variable with known probability distribution. Knowledge about the statistics of n could be due to, e.g., calibrations of the measurement device.

The idea of the Morozov discrepancy principle is to choose the smallest $k \geq 1$ such that the *residual* satisfies

$$\|y - Ax_k\| \leq \epsilon.$$

Intuitively this means that we cannot expect the approximate solution to yield a smaller residual than the measurement error — otherwise we would be fitting the solution to noise.

Does such k exist?

Yes, it does if $\epsilon > \|Py - y\|$, as explained below.

If $\mathrm{rank}(A) = \infty$, it follows from $\overline{\mathrm{Ran}(A)} = \mathrm{Ran}(P) \perp \mathrm{Ran}(I - P)$ that

$$\begin{aligned} \|Ax_k - y\|^2 &= \|(Ax_k - Py) + (Py - y)\|^2 \\ &= \|Ax_k - Py\|^2 + \|(P - I)y\|^2 \\ &= \sum_{n=k+1}^{\infty} |\langle y, u_n \rangle|^2 + \|(P - I)y\|^2 \\ &\to \|Py - y\|^2 \quad \text{as } k \to \infty, \end{aligned}$$

which is the best one can do since $\inf_{z \in \mathrm{Ran}(A)} \|z - y\| = \|Py - y\|$ by virtue of the *projection theorem*. (However, there is no guarantee that $\|x_k\|$ would not explode as $k \to \infty$.)

On the other hand, if $p = \mathrm{rank}(A) < \infty$,

$$\|Ax_p - y\| = \|P_p y - y\| = \|Py - y\|.$$

(Usually, one should not choose this large *spectral cut-off* in practice.)

2.3 Tikhonov regularization

Motivation of Tikhonov regularization

As pointed out on the previous slide, the norm of the residual

$$\|Ax - y\|$$

is minimized by the sequence of truncated SVD solutions $\{x_k\}$ as k tends to $\mathrm{rank}(A)$. Unfortunately, when inverse/ill-posed problems are considered, we typically also have

$$\|x_k\| \to \infty \qquad \text{as } k \to \mathrm{rank}(A).$$

(If $\mathrm{rank}(A) = \infty$, this can be understood literally; if $\mathrm{rank}(A) = p < \infty$, this should be understood in the sense that the x_p is usually rubbish — especially, if the measurement y is noisy.)

As a consequence, it seems well-motivated to try minimizing the residual and the norm of the solution *simultaneously*.

Tikhonov regularized solution

A *Tikhonov regularized solution* $x_\delta \in H_1$ is a minimizer of the Tikhonov functional

$$F_\delta(x) := \|Ax - y\|^2 + \delta\|x\|^2,$$

where $\delta > 0$ is called the *regularization parameter*.

Theorem. *A Tikhonov regularized solution exists, is unique, and is given by*

$$x_\delta = (A^*A + \delta I)^{-1}A^*y = \sum_{n=1}^{p} \frac{\lambda_n}{\lambda_n^2 + \delta}\langle y, u_n\rangle v_n,$$

where $p = \operatorname{rank}(A) \leq \infty$.

Proof: Let us prove this claim only in the case that $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$; the general result follows from the same ideas, but requires some more sophisticated functional analysis.

To begin with, note that

$$x^{\mathrm{T}}(A^{\mathrm{T}}A + \delta I)x \;=\; \|Ax\|^2 + \delta\|x\|^2 \;\geq\; \delta\|x\|^2 > 0$$

if $x \neq 0$. In particular, $A^{\mathrm{T}}A + \delta I \in \mathbb{R}^{n\times n}$ is injective, which means that it is invertible due to the *fundamental theorem of linear algebra*.

Hence,

$$x_\delta \;:=\; (A^{\mathrm{T}}A + \delta I)^{-1}A^{\mathrm{T}}y \in H_1$$

is well-defined.

Let $\{\lambda_j\}_{j=1}^p$ be the positive singular values of A, and $\{v_j\}_{j=1}^p$ and $\{u_j\}_{j=1}^p$ the corresponding sets of singular vectors that span $\mathrm{Ker}(A)^\perp$ and $\mathrm{Ran}(A)$, respectively.

We expand $x_\delta = \sum (v_j^{\mathrm{T}} x_\delta) v_j + Q x_\delta$, where $Q : \mathbb{R}^n \to \mathrm{Ker}(A)$ is an orthogonal projection. According to the first exercise of the first exercise session,

$$(A^{\mathrm{T}} A + \delta I) x_\delta = \sum_{j=1}^{p} (\lambda_j^2 + \delta)(v_j^{\mathrm{T}} x_\delta) v_j + \delta\, Q x_\delta.$$

Similarly,

$$A^{\mathrm{T}} y = \sum_{j=1}^{p} \lambda_j (u_j^{\mathrm{T}} y) v_j.$$

Equating these two expressions results in

$$(v_j^{\mathrm{T}} x_\delta) \;=\; \frac{\lambda_j}{\lambda_j^2 + \delta}(u_j^{\mathrm{T}} y), \qquad 1 \le j \le p,$$

and $Qx_\delta = 0$, which altogether means that

$$x_\delta \;=\; \sum_{n=1}^{p} \frac{\lambda_n}{\lambda_n^2 + \delta}(u_j^{\mathrm{T}} y) v_n \;=\; \sum_{n=1}^{p} \frac{\lambda_n}{\lambda_n^2 + \delta}\langle y, u_n \rangle v_n .$$

Finally, consider $x = x_\delta + z$, where $z \in \mathbb{R}^n$ is arbitrary. We have

$$\begin{aligned} F_\delta(x) &= \|(Ax_\delta - y) + Az\|^2 + \delta\|x_\delta + z\|^2 \\ &= \|Ax_\delta - y\|^2 + 2\,(Az)^{\mathrm{T}}(Ax_\delta - y) + \|Az\|^2 \\ &\quad + \delta\left(\|x_\delta\|^2 + 2z^{\mathrm{T}}x_\delta + \|z\|^2\right) \\ &= F_\delta(x_\delta) + \|Az\|^2 + \delta\|z\|^2 \\ &\quad + 2z^{\mathrm{T}}\left(\left(A^{\mathrm{T}}A + \delta I\right)x_\delta - A^{\mathrm{T}}y\right) \\ &= F_\delta(x_\delta) + \|Az\|^2 + \delta\|z\|^2 \;\geq\; F_\delta(x_\delta), \end{aligned}$$

where the equality holds if and only if $z = 0$. This shows that $x_\delta = (A^{\mathrm{T}}A + \delta I)^{-1}A^{\mathrm{T}}y$ is the unique minimizer of the Tikhonov functional. □

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Fourth lecture, January 28, 2011.

Summary of the previous lecture

Morozov discrepancy principle: According to the Morozov discrepancy principle, for the truncated SVD solution $x_k \in H_1$ one should choose the smallest spectral cut-off index $\mathbb{N} \ni k \leq \mathrm{rank}(A)$ such that

$$\|Ax_k - y\| \leq \epsilon,$$

where $\epsilon > 0$ corresponds to the anticipated inaccuracy in the data vector $y \in H_2$. How to estimate such ϵ is not trivial — one can even argue that it is not unambiguous. Be that as it may, k is uniquely determined by the Morozov discrepancy principle if

$$\|y - Py\| < \epsilon.$$

Tikhonov regularization: The *Tikhonov regularized solution* $x_\delta \in H_1$ is the unique minimizer of the Tikhonov functional

$$F_\delta(x) := \|Ax - y\|^2 + \delta\|x\|^2, \qquad \delta > 0.$$

It is given explicitly by the formula

$$x_\delta = (A^*A + \delta I)^{-1}A^*y = \sum_n \frac{\lambda_n}{\lambda_n^2 + \delta}\langle y, u_n\rangle v_n$$

Note that the family of Tikhonov regularized solutions $\{x_\delta\}_{\delta\in\mathbb{R}_+}$ is parameterized by the positive real parameter $\delta > 0$. (In the case of truncated SVD, the regularized solutions are parameterized discretely as $\{x_k\}_{k=1}^p$, where $p = \operatorname{rank}(A)$.)

Properties of the Tikhonov regularized solution

The Tikhonov regularized solution has the following intuitive properties. The proof of this theorem is omitted.

Theorem. *Let $P : H_2 \to \overline{\mathrm{Ran}(A)}$ be an orthogonal projection. The residual $\|Ax_\delta - y\|$ is strictly increasing as a function of δ and it satisfies*

$$\lim_{\delta \to 0} \|Ax_\delta - y\| = \|Py - y\| \qquad \text{and} \qquad \lim_{\delta \to \infty} \|Ax_\delta - y\| = \|y\|.$$

Moreover, if $Py \in \mathrm{Ran}(A)$, then x_δ converges to the solution of the problem

$$Ax = Py \qquad \text{and} \qquad x \perp \mathrm{Ker}(A)$$

as $\delta \to 0$. On the other hand, if $Py \notin \mathrm{Ran}(A)$, then the norm $\|x_\delta\|$ tends to infinity as δ goes to zero.

The Morozov principle for Tikhonov regularization

Assume once again that the measurement $y \in H_2$ is a noisy version of some underlying 'exact' data vector $y_0 \in H_2$, and that

$$\|y - y_0\| \approx \epsilon > 0.$$

In the framework of the Tikhonov regularization, the Morozov discrepancy principle advises to choose the regularization parameter $\delta > 0$ so that the residual satisfies

$$\|y - Ax_\delta\| = \epsilon.$$

Such a regularization parameter exists if

$$\|y - Py\| < \epsilon < \|y\|.$$

This follows from the above theorem because the residual $\|y - Ax_\delta\|$ is continuous with respect to δ.

Tikhonov regularized solution for matrices

Assume once again that $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$. In this case, the Tikhonov functional can be given as

$$F_\delta(x) = \left\| \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2, \qquad I \in \mathbb{R}^{n \times n},\ 0 \in \mathbb{R}^n. \tag{5}$$

It is interesting to notice that the normal equation corresponding to this *least squares problem* is (see 3. exercise of 1. exercise session)

$$\begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} x = \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} y \\ 0 \end{bmatrix},$$

or equivalently

$$(A^{\mathrm{T}} A + \delta I) x = A^{\mathrm{T}} y.$$

Bear in mind that one does not, actually, need to form this normal equation in Matlab when using Tikhonov regularization: After defining

$$K = \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} \in \mathbb{R}^{(n+m)\times n} \qquad \text{and} \qquad z = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n+m},$$

the command

```
xdelta = K\z
```

computes the Tikhonov regularized solution.

Explanation: For non-square matrices the `mldivide` command of Matlab tries to solve the corresponding least squares problem. As unique minimizer is known to exist, this corresponds to multiplying z from the left by the Moore–Penrose pseudoinverse of K (see 3. exercise of 1. session). As all n singular values of K are larger than $\sqrt{\delta}$ (see 1. exercise of 2. session) this pseudoinverse is well-behaved.

An example: Heat distribution in a rod (revisited)

Recall the discretized inverse heat conduction problem that was discussed during the second and third lectures. Let `w` be the simulated heat distribution at `T=0.1` with the 'wedge function' as the initial data, and `A` the corresponding propagation matrix `A=expm(TB)`. We add the same small amount of noise as previously and compute the Tikhonov regularized solution:

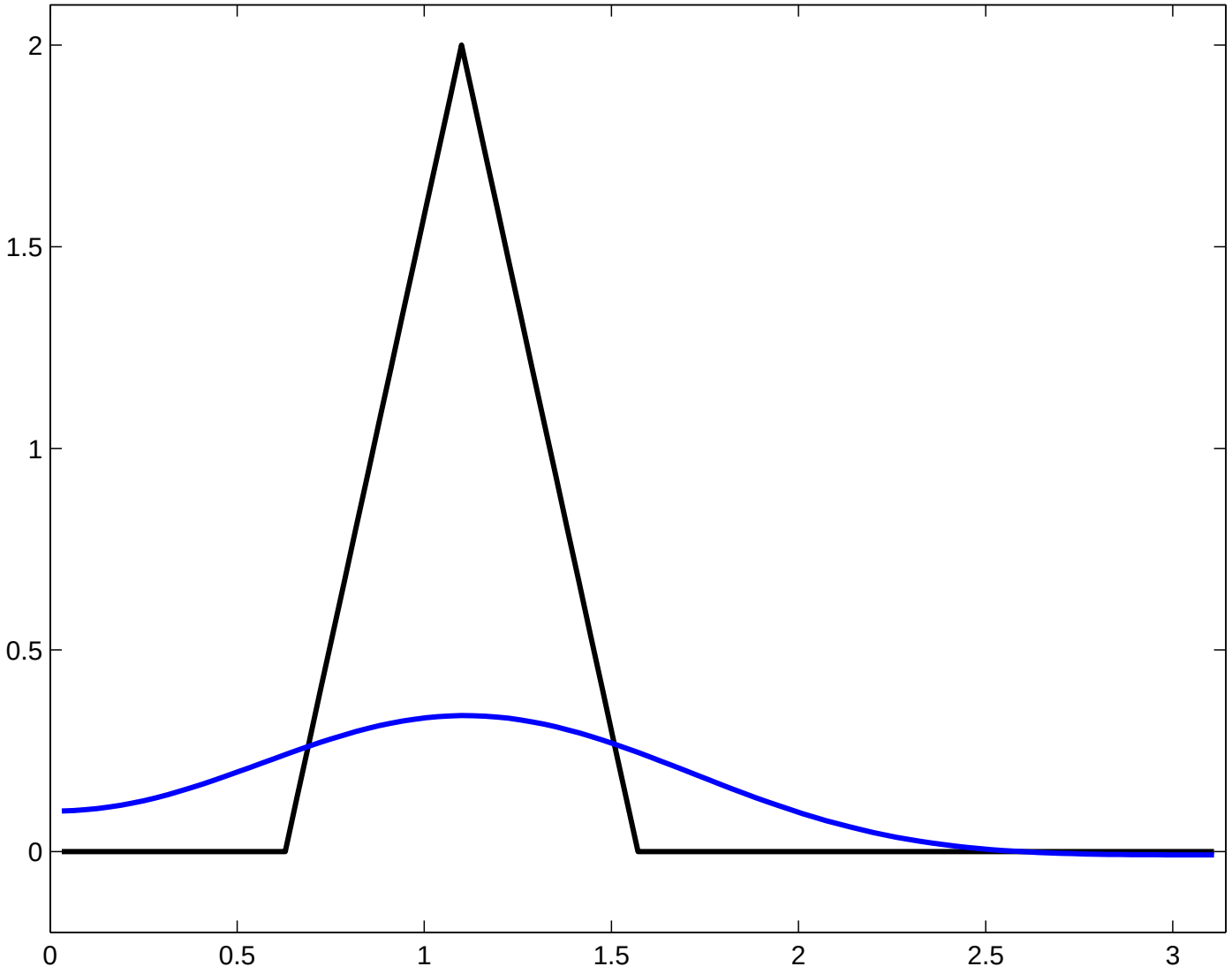
```
wn = w + 0.001*randn(N-1,1);
zn = [wn; zeros(N-1,1)]; %  augmented data vector
K = [A; sqrt(delta)*eye(N-1)]; % augmented system matrix
fdelta = K\zn; % Tikhonov regularized solution
```

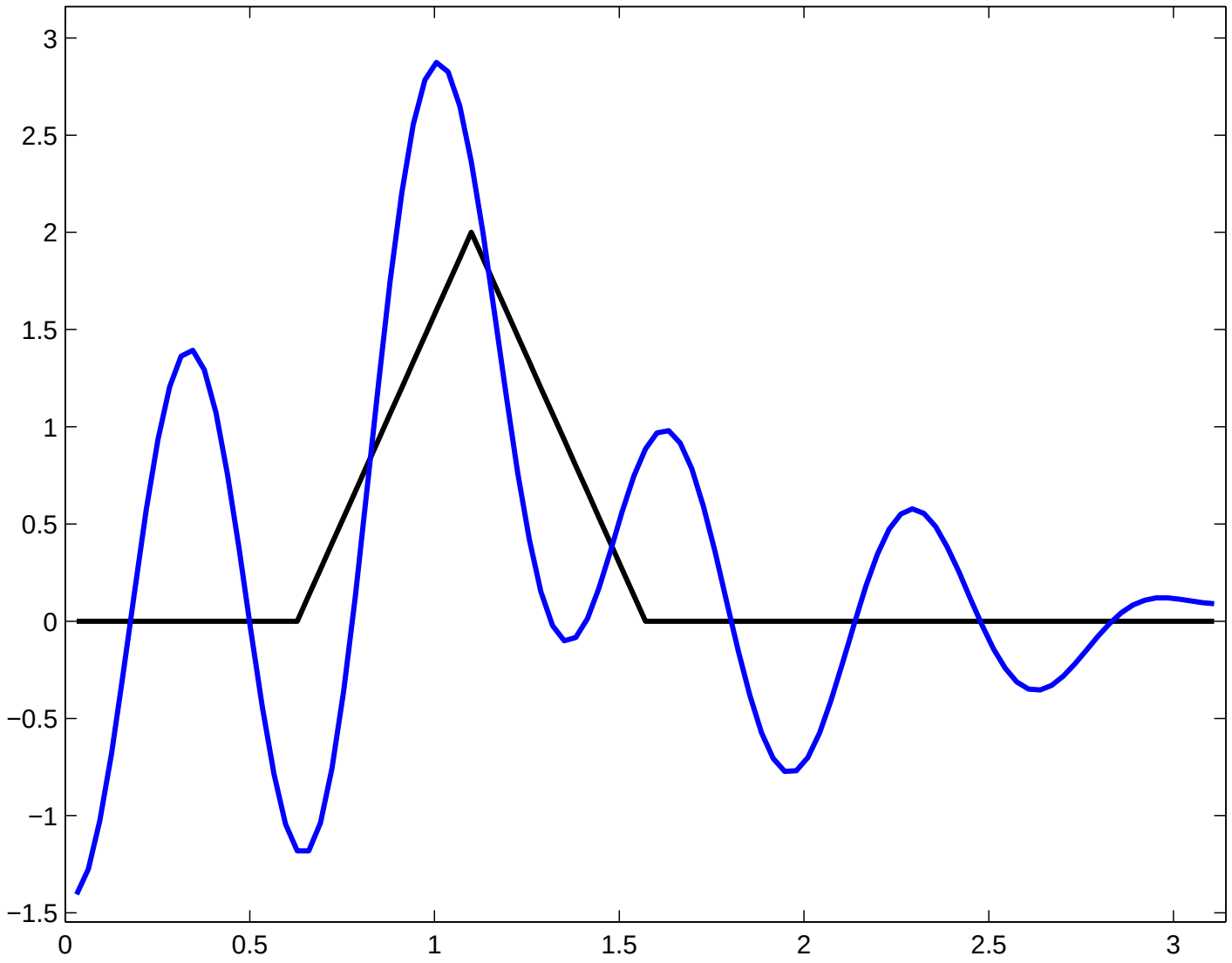
We do this for three different values of the regularization parameter $\delta = 1$ (too large), $\delta = 10^{-8}$ (too small), and $\delta = 5.95 \cdot 10^{-5}$, which corresponds to the Morozov discrepancy principle: We assume here that the discrepancy between the measured data and the underlying 'exact' data equals the square root of the expectation value of the squared norm of the noise vector, i.e.,

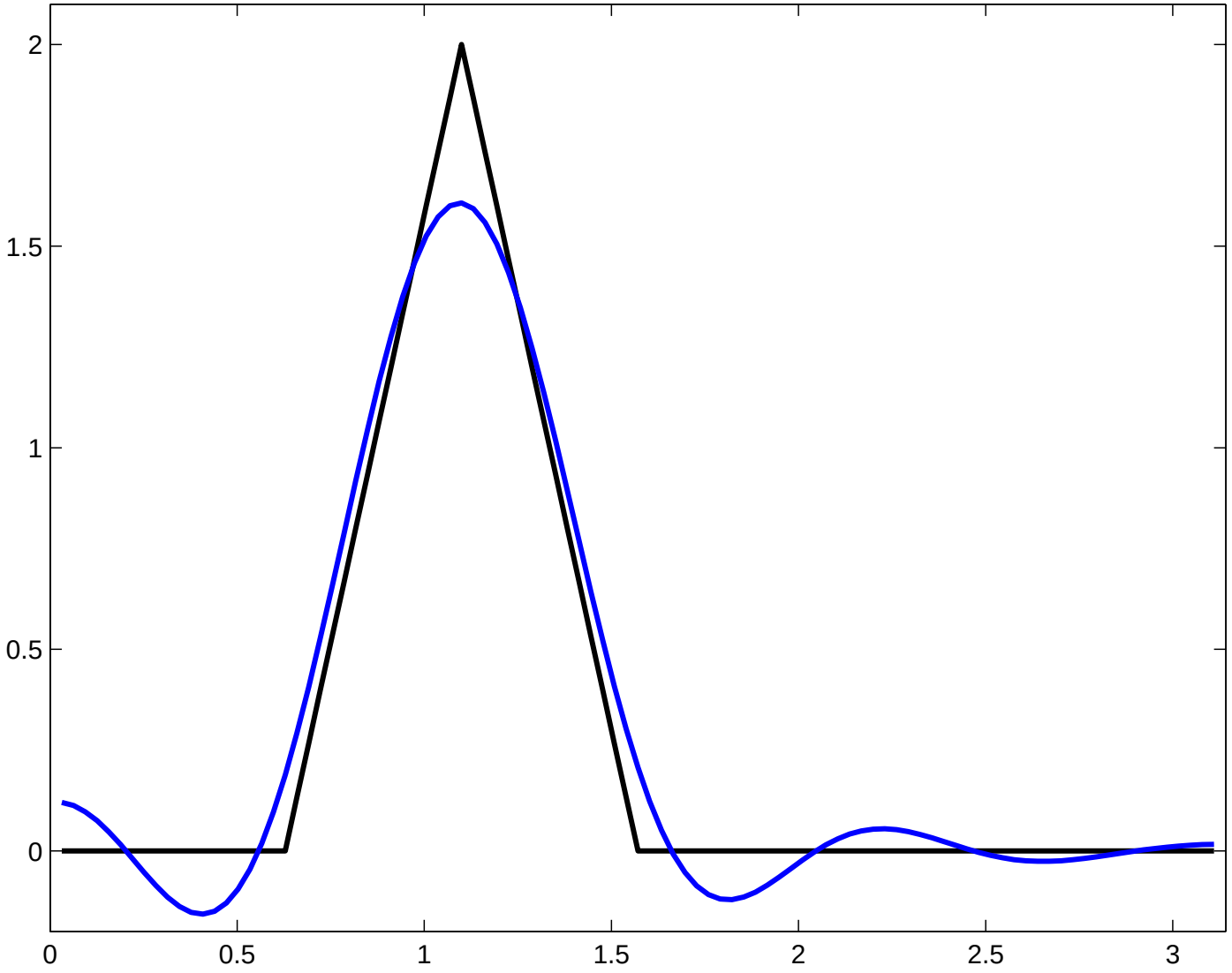
$$\epsilon = \sqrt{99 \cdot 0.001^2} \approx 9.95 \cdot 10^{-3}.$$

Note that the value of δ given by the discrepancy principle depends on the particular realization of the noise vector even though ϵ does not.

The expectation value of the norm of the noise vector would be as — if not more — logical choice for ϵ, but it is more difficult to write down explicitly. (Luckily, these two choices do not differ that much in the considered case: numerical tests suggest that the latter gives $\epsilon \approx 9.92 \cdot 10^{-3}$.)


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3


3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.5
1
1.5
2
2.5
3


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Generalizations of Tikhonov regularization

Tikhonov regularization for nonlinear problems

Let us briefly consider the nonlinear case, where $A : H_1 \to H_2$ is a nonlinear operator and the examined equation is of the form

$$A(x) = y.$$

A standard way of solving such a problem is via sequential linearizations, which leads to solving a set of linear problems involving the derivative operator of A.

As an example, in Newton's method one would first pick an initial guess $x_0 \in H_1$ and then try to produce the $(j+1)$th iterate by solving the linearized problem

$$A(x_j) + A'(x_j)(x_{j+1} - x_j) = y, \qquad j = 0, 1, \ldots,$$

recursively for x_{j+1}. (In the general setting A' is the *Fréchet derivative* of A, but for finite-dimensional operators it is just the Jacobian matrix.)

Unfortunately, if large alterations of x produce only small changes in $A(x)$, i.e., if the original equation is ill-posed, there is no guarantee that the corresponding linearized problems can be solved as such — not even in the least squares sense. Hence, regularization is needed.

Unlike the truncated SVD method, Tikhonov regularization generalizes easily to this nonlinear framework. Now, it amounts to searching for $x_\delta \in H_1$ that minimizes the functional

$$F_\delta(x) = \|A(x) - y\|^2 + \delta\|x\|^2, \qquad \delta > 0.$$

Since F_δ is no longer quadratic in x, it is not clear that a unique minimizer exists. Furthermore, even if a Tikhonov regularized solution exists, it cannot usually be given by an explicit formula.

Be that as it may, one can try to minimize $F_\delta(x)$ by using some nonlinear optimization technique. One — but probably not the best — way of doing this, is to pick an initial guess $x_{\delta,0} \in H_1$ and then recursively define the $(j+1)$th iterate $x_{\delta,j+1} \in H_1$ to be the unique minimizer of the $x_{\delta,j}$-dependent Tikhonov functional

$$\begin{aligned}\tilde{F}_{\delta,j}(x) &= \|A(x_{\delta,j}) + A'(x_{\delta,j})(x - x_{\delta,j}) - y\|^2 + \delta\|x\|^2 \\ &= \|A'(x_{\delta,j})x - [y - A(x_{\delta,j}) + A'(x_{\delta,j})x_{\delta,j}]\|^2 + \delta\|x\|^2,\end{aligned}$$

where the dependence of A on x has been linearized with $x_{\delta,j}$ as the base point. Since this Tikhonov functional is of the 'standard form', $x_{\delta,j+1}$ can be given explicitly with the help of $A'(x_{\delta,j})$, $A(x_{\delta,j})$, $x_{\delta,j}$, y and δ. (In practice, evaluating $A'(x_{\delta,j})$ is often the most difficult part.)

Combining this with some reasonable stopping criterion does indeed give reasonable solutions for many nonlinear inverse problems.

More general penalty terms

A more general way of defining the Tikhonov functional is

$$F_\delta(x) = \|Ax - y\|^2 + \delta G(x),$$

where the penalty function $G : H_1 \to \mathbb{R}$ takes non-negative values. The existence of a unique minimizer for this kind of functional depends on the properties of G, as does the workload needed for finding the minimizer.

One typical way of defining G is

$$G(x) = \|L(x - x_0)\|^2, \tag{6}$$

where $x_0 \in H_1$ is a given reference vector and L is some linear operator. The choice of x_0 and L reflects our prior knowledge about the 'feasible' solutions: Lx is some property that is known to be relatively close to the reference value Lx_0 for all reasonable solutions. (In standard case $x_0 = 0$ and $L = I$, the solutions are 'known' to lie relatively close to the origin.)

The numerical implementation of Tikhonov regularization with G of (6) is approximately as easy as for the standard penalty term:

In the case that $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$, the operator L is just some matrix in $\mathbb{R}^{l \times n}$ and the Tikhonov functional can be given as

$$F_\delta(x) \;=\; \|Kx - z\|^2 \tag{7}$$

where

$$K \;=\; \begin{bmatrix} A \\ \sqrt{\delta} L \end{bmatrix} \qquad \text{and} \qquad z = \begin{bmatrix} y \\ \sqrt{\delta} L x_0 \end{bmatrix}.$$

Assuming that the matrix L is chosen so cleverly that all n singular values of K are (well) larger than zero, the Tikhonov regularized solution can be computed in Matlab by applying the pseudoinverse of K on z by the command

```
xdelta = K\z
```

Explanation: As shown in 3. exercise of 1. session, all minimizers of (7) satisfy the normal equation

$$K^{\mathrm{T}} K x = K^{\mathrm{T}} z.$$

On the other hand, it was proved in 1. exercise of 1. session that the symmetric matrix $K^{\mathrm{T}} K \in \mathbb{R}^{n \times n}$ has n positive eigenvalues that are the squares of the singular values of K. In particular, this means that $K^{\mathrm{T}} K$ is invertible, and thus there is exactly one minimizer for (7). This is given by $K^{\dagger} z$ due to 3. exercise of 1. session.

(The fact that a symmetric matrix with nonzero eigenvalues is invertible follows, e.g., from the eigenvalue decomposition.)

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Fifth lecture, February 2, 2011.

2.4 Regularization by truncated iterative methods

For simplicity, in the rest of Chapter 2 we will only consider the case when

$$Ax = y$$

is a system of linear equations, i.e., $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

In the literature there are lots of iterative methods for solving this kind of matrix equations. By "iterative" we mean a method that attempts to solve the problem by finding successive approximations for the solution, starting from some initial guess. Typically, computation of such iterations involves multiplications by A and its adjoint, but not explicit computation of inverse operators. (The *Gaussian elimination* is an example of the opposite: it is a direct, i.e., non-iterative, method that tries to come up with a solution in a finite number of steps.)

Iterative methods are sometimes the only feasible choice if the problem involves a large number of variables (sometimes of the order of millions), making direct methods prohibitively expensive. Iterations are especially practical if multiplications by A are cheap. This is the case, e.g., when A is a multi-diagonal matrix originating from a difference or element approximation for some boundary value problem for an elliptic partial differential operator. (There exist lots of other examples, as well.)

Although iterative solvers have not usually been designed for ill-posed equations, they often posses regularizing properties: If the iterations are terminated before "the solution starts to fit to noise", one often obtains reasonable solutions for inverse problems.

2.4.1 Landweber–Fridman iteration

Banach fixed point iteration

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a vector-valued function. We say that $S \subset \mathbb{R}^n$ is an invariant set for T if

$$T(S) \subset S, \quad \text{i.e.,} \quad T(x) \in S \quad \text{for all } x \in S.$$

Moreover, T is a contraction on an invariant set S if there exists $0 \leq \kappa < 1$ such that

$$\|T(x) - T(y)\| < \kappa \|x - y\| \qquad \text{for all } x, y \in S.$$

Finally, a vector $x \in \mathbb{R}^n$ is called a fixed point of T if

$$T(x) = x.$$

.

Theorem. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a contraction on the closed invariant set S. Then there exists a unique fixed point $x \in S$ of T. Furthermore, this fixed point can be found by the following fixed point iteration:*

$$x = \lim_{k \to \infty} x_k, \qquad \text{where } x_{k+1} = T(x_k),$$

for any $x_0 \in S$.

Proof. The proof — although not very complicated — is omitted.

A simple example: Consider the function $T : x \mapsto x^2$ from $\mathbb{R}$ to itself.

(i) Let $S = [0, 1/3]$. Clearly, $T(S) = [0, 1/9] \subset S$ and

$$|T(x) - T(y)| = |x^2 - y^2| = |x + y||x - y| \leq 2/3|x - y|.$$

Hence, there is a unique fixed point, which is given by $\lim x_0^{2^k} = 0$ for every $x_0 \in S$.

(ii) If $S = (0, 1/3]$, the fixed point does not anymore lie in S.

(iii) If $S = [0, 1]$, $T(S) = S$, but T is no longer a contraction:

$$|T(3/4) - T(1/2)| = 5/16 > 1/4 = |3/4 - 1/2|.$$

In this case there are two fixed points: $T(0) = 0$ and $T(1) = 1$.

(iv) If, e.g., $S = [0, 5/6]$, there is a unique fixed point $0 \in S$, but its existence is not predicted by the fixed point theorem since T is not a contraction on S.

Landweber–Fridman scheme

Instead of the original equation

$$Ax = y,$$

we will consider the normal equation

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}y.$$

According to 3. exercise of 1. session, $x \in \mathbb{R}^n$ satisfies the normal equation of and only if it minimizes the residual

$$\|Ax - y\|.$$

Moreover, there exist a unique element of $\mathbb{R}^n$, given by $x^\dagger := A^\dagger y \in \mathbb{R}^n$, that solves the normal equation and is orthogonal to $\mathrm{Ker}(A)$.

(Bear in mind, however, that the use of the pseudoinverse $A^\dagger$ is suspect if the matrix is ill-conditioned, i.e., if $\lambda_1/\lambda_p \gg 1$, where $p = \mathrm{rank}(A)$.)

We define an affine mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T(x) = x + \beta(A^{\mathrm{T}}y - A^{\mathrm{T}}Ax), \qquad \beta \in \mathbb{R}.$$

Notice that any solution of the normal equation is a fixed point of T. We will show that if β is small enough there is only one fixed point of T in $\mathrm{Ker}(A)^{\perp}$, namely $x^{\dagger}$, and it can be reached by the fixed point iteration if $x_0 = 0$.

Theorem. *Let $0 < \beta < 2/\lambda_1^2$ be fixed. Then, the fixed point iteration*

$$x_{k+1} = T(x_k), \qquad x_0 = 0,$$

converges towards $x^{\dagger}$ as $k \to \infty$.

Proof. Set $S = \mathrm{Ker}(A)^{\perp} = \mathrm{Ran}(A^{\mathrm{T}})$. Clearly, $T(S) \subset S$ since

$$T(x) \;=\; x + A^{\mathrm{T}}\left(\beta y - \beta A x\right) \in \mathrm{Ran}(A^{\mathrm{T}})$$

for all $x \in \mathrm{Ran}(A^{\mathrm{T}})$. Thus, S is invariant under T.

Recall that A and its transpose can be represented with the help of A's singular system as

$$Ax \;=\; \sum_{j=1}^{p} \lambda_j (v_j^{\mathrm{T}} x) u_j \qquad \text{and} \qquad A^{\mathrm{T}} y \;=\; \sum_{j=1}^{p} \lambda_j (u_j^{\mathrm{T}} y) v_j,$$

where $p = \mathrm{rank}(A)$ and λ_j are the positive singular values of A. The orthonormal sets of vectors $\{v_j\}_{j=1}^{p}$ and $\{u_j\}_{j=1}^{p}$ span $S = \mathrm{Ker}(A)^{\perp}$ and $\mathrm{Ran}(A)$, respectively. In particular,

$$x \;=\; \sum_{j=1}^{p} (v_j^{\mathrm{T}} x) v_j \qquad \text{for all } x \in S.$$

Let $x, z \in S$ and note that also $x - z \in S$. We have

$$\begin{aligned} T(x) - T(z) &= (x - z) - \beta A^{\mathrm{T}} A(x - z) \\ &= \sum_{j=1}^{p} (v_j^{\mathrm{T}}(x - z)) v_j - \beta \sum_{j=1}^{p} \lambda_j^2 (v_j^{\mathrm{T}}(x - z)) v_j \\ &= \sum_{j=1}^{p} (1 - \beta\lambda_j^2)(v_j^{\mathrm{T}}(x - z)) v_j. \end{aligned}$$

As λ_1 is the largest of the singular values, it holds by assumption that

$$-1 < \beta\lambda_j^2 - 1 \leq \beta\lambda_1^2 - 1 < 2 - 1 = 1, \qquad \text{for all } j = 1, \ldots, p.$$

Hence, we see that

$$\kappa := \max_{j=1,\ldots,p} |\beta\lambda_j^2 - 1| < 1.$$

In consequence,

$$\begin{aligned}\|T(x) - T(z)\|^2 &\leq \sum_{j=1}^{p}(1-\beta\lambda_j^2)^2(v_j^{\mathrm{T}}(x-z))^2 \\ &\leq \kappa^2 \sum_{j=1}^{p}(v_j^{\mathrm{T}}(x-z))^2 \;=\; \kappa^2\|x-z\|^2,\end{aligned}$$

which shows that T is a contraction on S. As S is also a closed invariant set for T, we know that there exists a unique fixed point of T in S.

To complete the proof, we recall that $x^\dagger = A^\dagger y$ belongs to $S = \mathrm{Ker}(A)^\perp$ and satisfies the normal equation (see exercise 3. of session 1.). Furthermore, since $x_0 = 0$ is in S — it is orthogonal to all vectors —, the fixed point iteration starting from x_0 converges to $x^\dagger$. $\square$

Regularization properties of Landweber–Fridman

From now on we will assume that $0 < \beta < 2/\lambda_1^2$.

In the third exercise session, it will be shown that the kth iterate of the Landweber–Fridman iteration can be written explicitly:

$$x_k = \sum_{j=1}^{p} \frac{1}{\lambda_j} \big(1 - (1 - \beta\lambda_j^2)^k\big)(u_j^{\mathrm{T}} y)v_j, \qquad k = 0, 1, \ldots . \tag{8}$$

Since $|1 - \beta\lambda_j^2| < 1$ by assumption,

$$(1 - \beta\lambda_j^2)^k \to 0 \qquad \text{as } k \to \infty,$$

which is what one would expect since

$$x^\dagger = \sum_{j=1}^{p} \frac{1}{\lambda_j} (u_j^{\mathrm{T}} y)v_j.$$

However, while $k \in \mathbb{N}$ is finite, the coefficients of the terms $(u_j^{\mathrm{T}} y)v_j$ appearing in the series representation (8) satisfy

$$\begin{aligned}
\frac{1}{\lambda_j}\left(1-(1-\beta\lambda_j^2)^k\right) &= \frac{1}{\lambda_j}\left(1-\sum_{l=0}^{k}\binom{k}{l}(-1)^l\beta^l\lambda_j^{2l}\right) \\
&= \frac{1}{\lambda_j}\sum_{l=1}^{k}\binom{k}{l}(-1)^{l+1}\beta^l\lambda_j^{2l} \\
&= \sum_{l=1}^{k}\binom{k}{l}(-1)^{l+1}\beta^l\lambda_j^{2l-1},
\end{aligned}$$

which converges to zero as $\lambda_j \to 0$ (for a fixed k).

As a consequence, while k is 'small enough', no coefficient of $(u_j^{\mathrm{T}} y)v_j$ in (8) is so large that the component of the measurement noise in the direction u_j is amplified in an uncontrolled manner. (Recall that the corresponding coefficients for Tikhonov regularization are $\lambda_j/(\lambda_j^2+\delta)$.)

Discrepancy principle for Landweber–Fridman

Let the measurement $y \in \mathbb{R}^m$ be a noisy version of some underlying 'exact' data vector $y_0 \in \mathbb{R}^m$, and assume that

$$\|y - y_0\| \approx \epsilon > 0.$$

The Morozov discrepancy principle works for the Landweber–Fridman iteration in approximately the same way as for the truncated SVD and the Tikhonov regularization: Choose the smallest $k \geq 0$ such that the residual satisfies

$$\|y - Ax_k\| \leq \epsilon.$$

Such a stopping rule exists if

$$\epsilon > \|y - Py\| = \|y - A(A^\dagger y)\|,$$

where $P = AA^\dagger$ (see 1. ses., 2. ex.) is the orthogonal projection onto the range of A. Indeed, since the sequence $\{x_k\}_{k=0}^\infty$ converges to $x^\dagger = A^\dagger y$, for any $\epsilon > \|y - Ax^\dagger\|$ there exists $k = k_\epsilon \in \mathbb{N}$ such that

$$\|x_k - x^\dagger\| \leq \frac{1}{\|A\|}(\epsilon - \|y - Ax^\dagger\|),$$

and thus by the reverse triangle inequality,

$$\begin{aligned} \|y - Ax_k\| - \|y - Ax^\dagger\| &\leq \|(y - Ax_k) - (y - Ax^\dagger)\| \\ &\leq \|A\|\|x_k - x^\dagger\| \\ &\leq \epsilon - \|y - Ax^\dagger\|, \end{aligned}$$

which just means that $\|y - Ax_k\| \leq \epsilon$.

An example: Heat distribution in a rod (revisited)

Recall again the discretized inverse heat conduction problem that was discussed during the second and third lectures. Let `w` be the simulated heat distribution at `T=0.1` with the 'wedge function' as the initial data, and `A` the corresponding propagation matrix `A=expm(TB)`. We add again the same small amount of noise to the measurement:

```
wn = w + 0.001*randn(N-1,1);
```

and use the Morozov discrepancy principle with

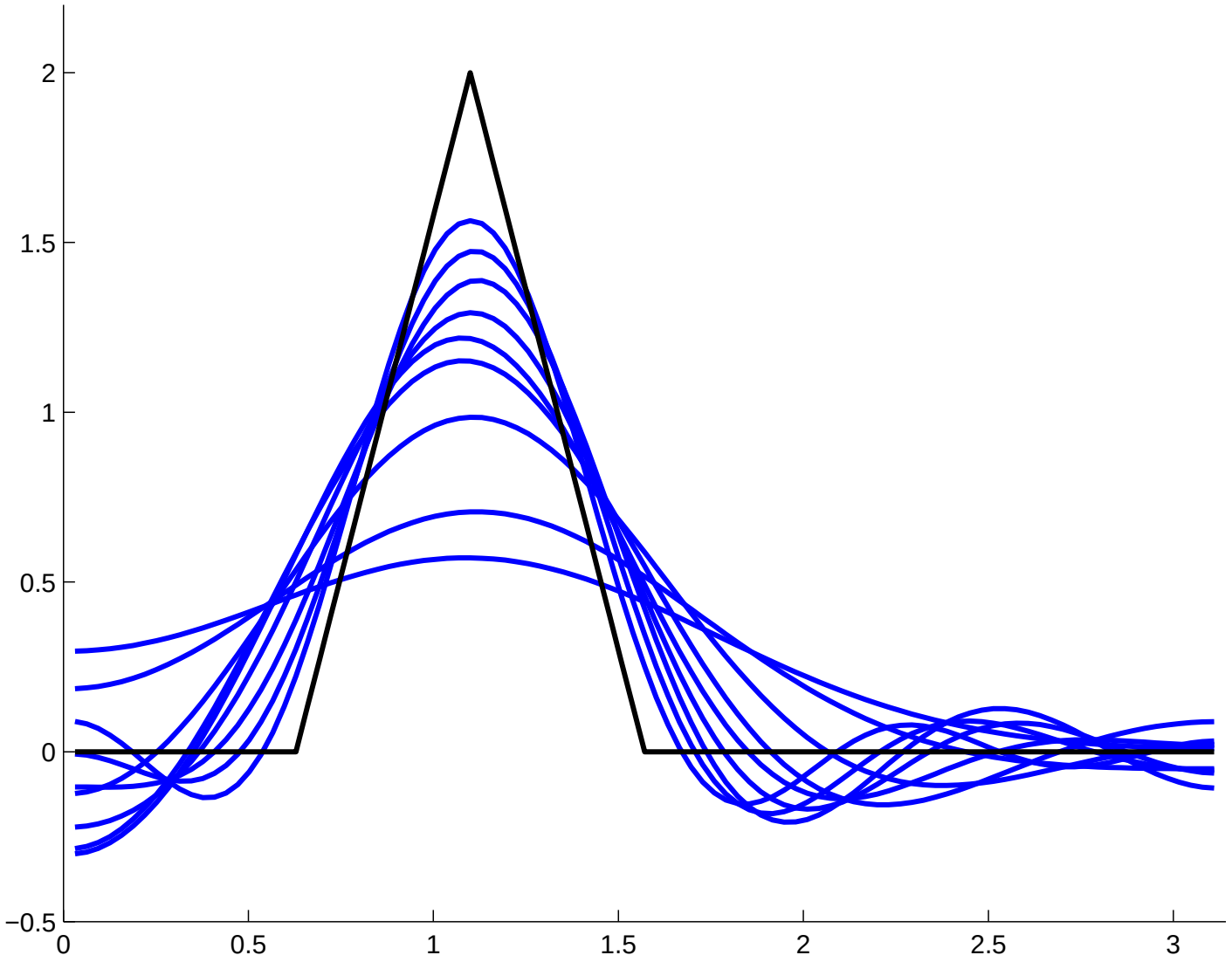
$$\epsilon = \sqrt{99 \cdot 0.001^2} \approx 9.95 \cdot 10^{-3}.$$

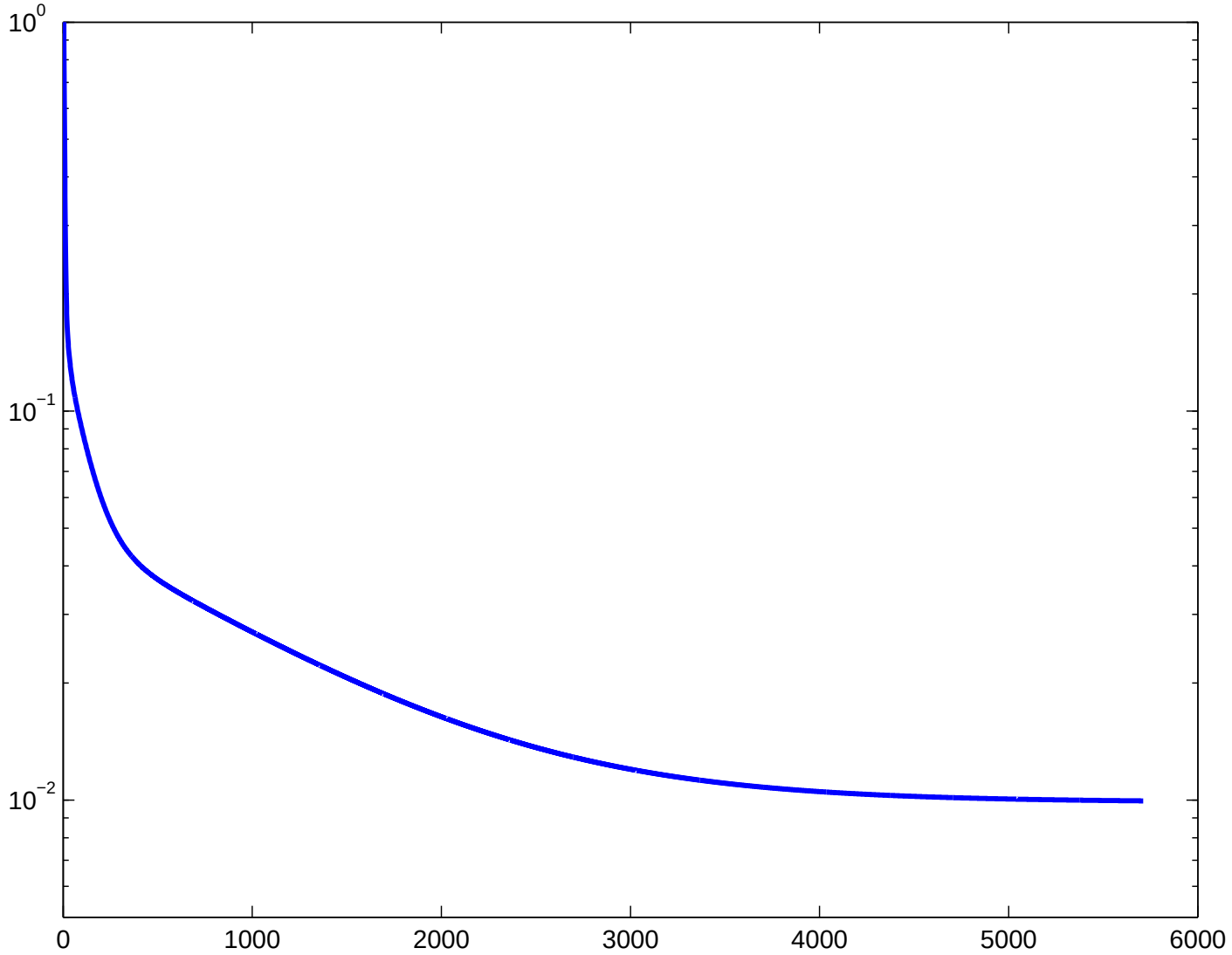
Because the largest singular value of the solution operator $E_T : L^2(0,\pi) \to L^2(0,\pi)$ in the corresponding infinite-dimensional case is 1, it is reasonable to anticipate that the same is also approximately true for A. Thus, we choose $\beta = 1 < 2/1 \approx 2/\lambda_1^2$.

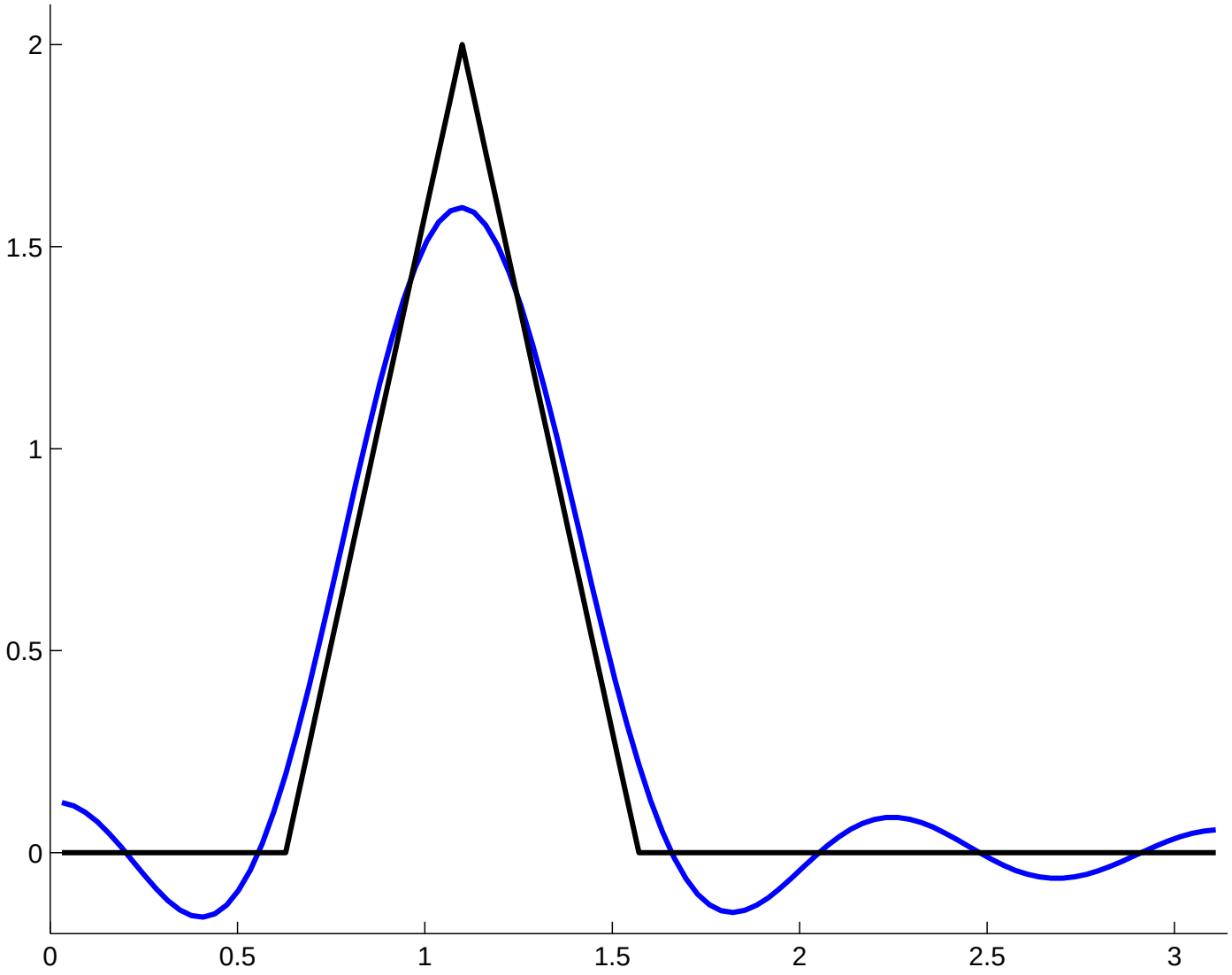
The implementation of the Landweber–Fridman iteration with the Morozov discrepancy principle in Matlab is straightforward. Bear in mind, however, that matrix-matrix products are far more expensive to compute than matrix-vector products. Hence, you should either compute and store the product $A^{\mathrm{T}}A$ before you start iterating or use parentheses to avoid computing this product during the iteration:

```
flw = flw + beta*(A'*wn - A'*(A*flw));
```

With the particular realization of the measurement noise, the Morozov discrepancy principle was satisfied by the iterate corresponding to $k = 5712$. In the following, we visualize the evolution of the Landweber–Fridman iteration for $k = 1, 2, 7, 20, 54, 148, 403, 1096, 2980,$ show the residual as a function of k, and plot the solution corresponding to the discrepancy principle.


2
1.5
1
0.5
0
−0.5
0
0.5
1
1.5
2
2.5
3


10^0
10^{-1}
10^{-2}
0
1000
2000
3000
4000
5000
6000


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Sixth lecture, February 4, 2011.

2.4.1 Kaczmarz iteration and ART

Partition of the original problem

Let us continue to consider the matrix equation

$$Ax = y,$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

Suppose that we can write the system matrix A in the form

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_l \end{bmatrix}, \qquad A_j \in \mathbb{R}^{k_j \times n},\ j = 1, \ldots l,$$

where $k_1 + \cdots + k_l = m$ and each submatrice A_j is assumed to have k_j linearly independent row vectors, i.e., $\mathrm{rank}(A_j) = k_j \leq n$. In particular, A_j defines a surjective mapping from $\mathbb{R}^n$ to $\mathbb{R}^{k_j}$. (Recall that the rank of a matrix equals the number of linearly independent columns/rows.)

Similarly, we decompose $y \in \mathbb{R}^m$ into l subvectors:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_l \end{bmatrix}, \qquad y_j \in \mathbb{R}^{k_j},\ j = 1, \ldots, l.$$

Now, the original equation can be given as the system

$$A_j x = y_j, \qquad j = 1, \ldots, l.$$

The jth of these matrix problems is composed of $k_j \leq n$ linearly independent linear equations, and thus the corresponding 'solution space'

$$X_j = \{x \in \mathbb{R}^n \mid A_j x = y_j\}$$

is a $n - k_j$ dimensional hyperplane in $\mathbb{R}^n$. (Notice that this hyperplane is a subspace, i.e., it passes through the origin, if and only if $y_j = 0$.)

The Kaczmarz sequence

Although X_j is not in general a subspace, we can define an orthogonal projection $\mathcal{P}_j : \mathbb{R}^n \to X_j$ by requiring that

$$\mathcal{P}_j z \in X_j \quad \text{and} \quad (I - \mathcal{P}_j)z \perp (w_1 - w_2)$$

for all $z \in \mathbb{R}^n$ and $w_1, w_2 \in X_j$. In other words, $\mathcal{P}_j z$ is the point closest to z in X_j. Furthermore, we define the sequential 'projection' $\mathcal{P} : \mathbb{R}^n \to \mathbb{R}^n$ via

$$\mathcal{P} = \mathcal{P}_l \mathcal{P}_{l-1} \dots \mathcal{P}_2 \mathcal{P}_1.$$

The Kaczmarz sequence $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ is defined recursively as

$$x_{k+1} = \mathcal{P} x_k, \qquad x_0 = 0.$$

Theorem. *Assume that* $X = \bigcap_{j=1}^{l} X_j \neq \emptyset$*, i.e., the original equation has at least one solution. Then the Kaczmarz sequence* $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ *converges to the minimum norm solution as* k *goes to infinity. In other words,*

$$\lim_{k \to \infty} x_k = x^{\dagger},$$

where $x^{\dagger} = A^{\dagger} y$ *satisfies* $Ax^{\dagger} = y$ *and* $x^{\dagger} \perp \mathrm{Ker}(A)$*.*

Proof. The text book presents the (relatively complicated) proof in the more general case where A operates between separable Hilbert spaces. Here, we omit the proof.

Algebraic reconstruction technique (ART)

Let us consider the special case where the original problem $Ax = y$, $A \in \mathbb{R}^{m \times n}$, is partitioned into m subproblems, i.e., linear equations:

$$A_j x = a_j^{\mathrm{T}} x = y_j, \qquad j = 1, \ldots, m,$$

where a_j^{T} is the jth row of A — with $a_j \in \mathbb{R}^n$ treated as a column vector — and $y_j \in \mathbb{R}$ is just the jth component of the vector $y \in \mathbb{R}^m$. Notice that in this case the condition that $A_j : \mathbb{R}^n \to \mathbb{R}$ is a surjection for every $1 \leq j \leq m$ is equivalent to requiring that A does not have any empty rows.

The Kaczmarz iteration corresponding to this setting is called the algebraic reconstruction technique (ART) — at least, this is what we call ART on this course. ART is used extensively in X-ray tomography.

Examples of ART iterations

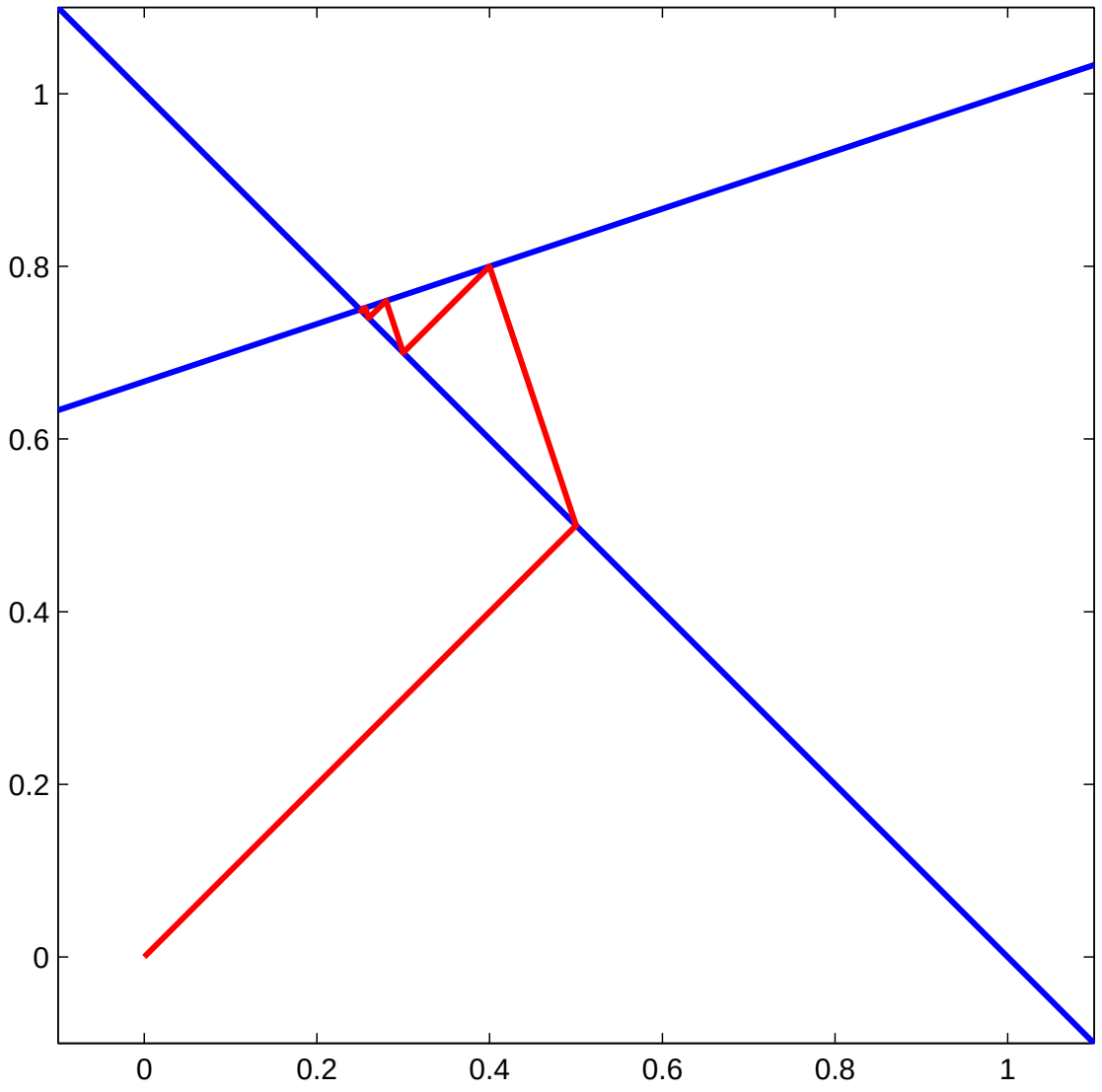
Let us first consider the case where

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

In particular, A is invertible and the corresponding hyperplanes, i.e., lines in $\mathbb{R}^2$, are given by

$$\begin{aligned} X_1 &= \{x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2 \mid x_1 + x_2 = 1\}, \\ X_2 &= \{x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2 \mid -x_1 + 3x_2 = 2\}. \end{aligned}$$

In this case, the ART algorithm should converge towards the unique solution $x = (1/4, 3/4)^{\mathrm{T}}$. In the following, we visualize each projection by $\mathcal{P}_j$, $j = 1, 2$, not just the sequential projections by $\mathcal{P} = \mathcal{P}_2\mathcal{P}_1$.


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Let us then add one row to A and one component to y:

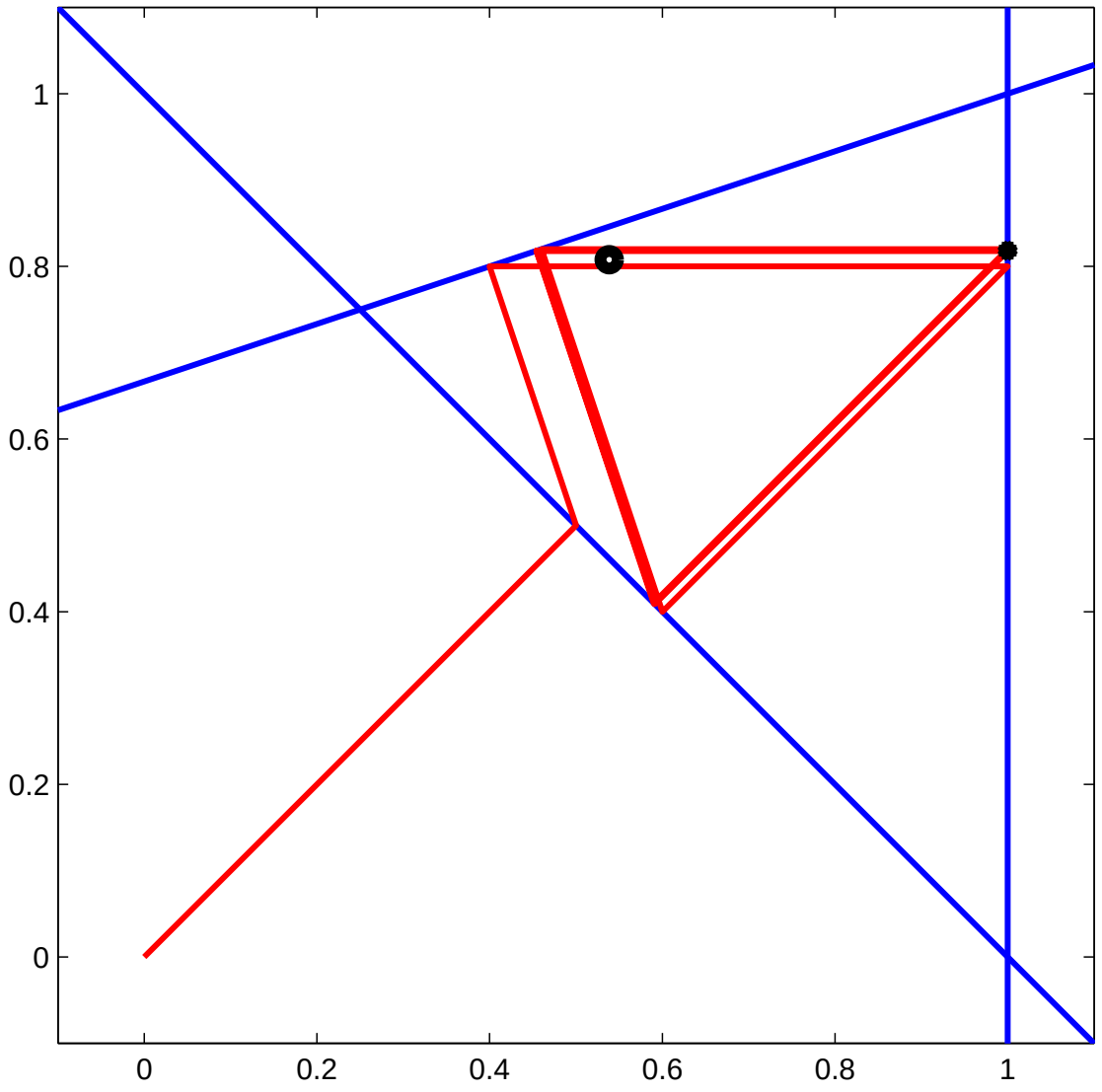
$$A = \begin{bmatrix} 1 & 1 \\ -1 & 3 \\ 1 & 0 \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix},$$

which adds the third hyperplane

$$X_3 = \{x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2 \mid x_1 = 1\}.$$

into play.

In this case, the equation $Ax = y$ does not have a solution. The ART iteration seems to converge to a point on X_3 depicted by an asterisk in the following figure — note that this does not mean that nothing happens *within* each iteration step. For comparison, the 'ring' marks the least squares solution $x^\dagger = A^\dagger y$.


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

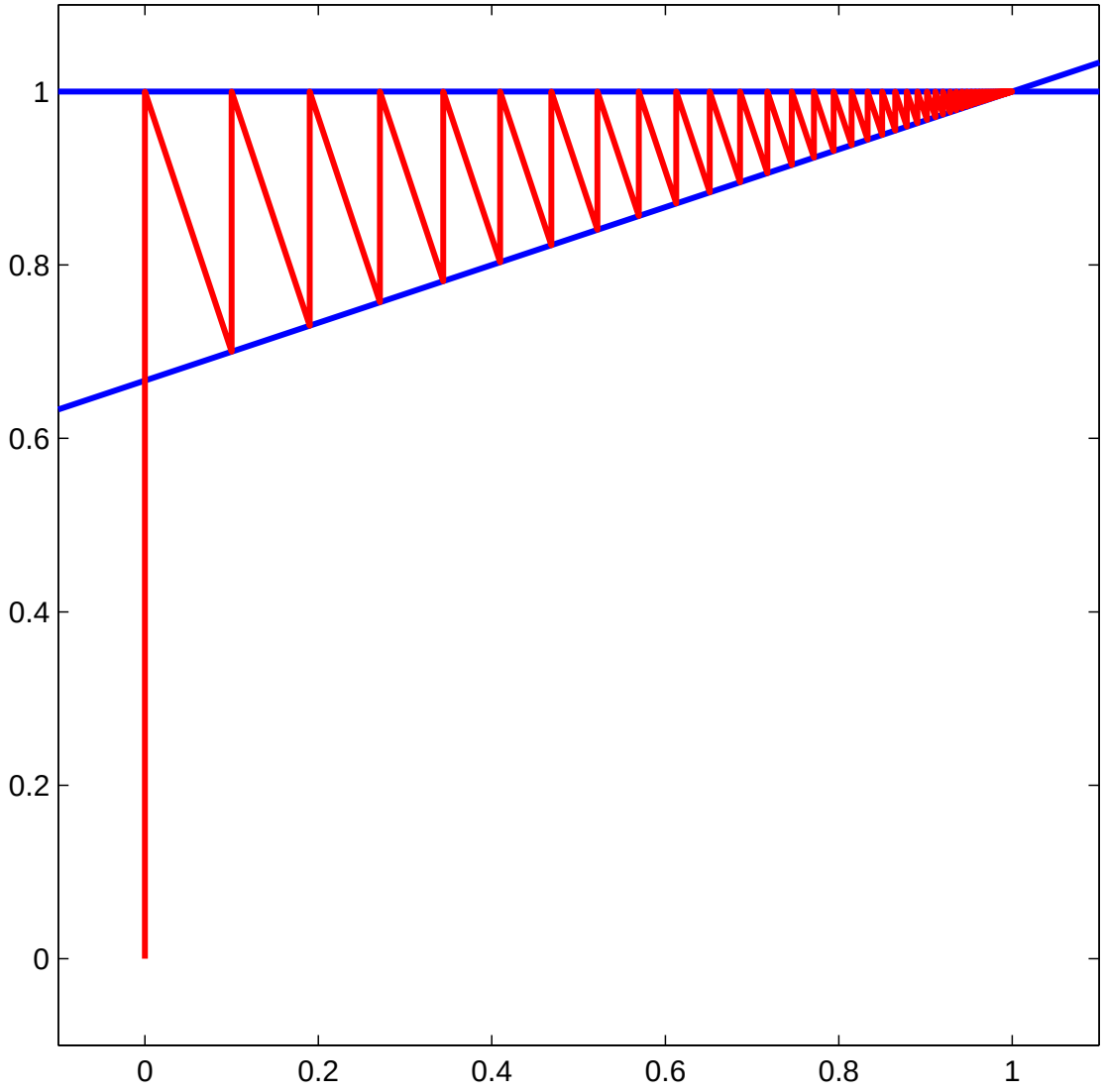
Finally, we return to the case of square matrices, but choose A so that its rows are somewhat 'closer' to being linearly dependent:

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 3 \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Once again, A is invertible and the corresponding 'solution hyperplanes' are given by

$$\begin{aligned} X_1 &= \{x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2 \mid x_2 = 1\}, \\ X_2 &= \{x = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2 \mid -x_1 + 3x_2 = 2\}. \end{aligned}$$

The ART algorithm converges towards the unique solution $x = (1, 1)^{\mathrm{T}}$, but extremely slowly.


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Computation of the projections $\mathcal{P}_j$

Consider still the equation $Ax = y$, $A \in \mathbb{R}^{n \times m}$, and assume still that there exists a partition

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_l \end{bmatrix}, \quad A_j \in \mathbb{R}^{k_j \times n}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_l \end{bmatrix}, \quad y_j \in \mathbb{R}^{k_j},$$

such that each A_j is surjective, i.e., $\operatorname{rank}(A_j) = k_j \leq n$. As before, let X_j denote the (non-empty) hyperplane composed of the solutions to $A_j x = y_j$, and $\mathcal{P}_j : \mathbb{R}^n \to X_j$ the orthogonal projection onto such hyperplane. Furthermore, we define

$$Q_j : \mathbb{R}^n \to \operatorname{Ker}(A_j), \qquad j = 1, \ldots, l,$$

to be the orthogonal projection onto the kernel of A_j.

In the fourth exercise session, it will be shown that

$$\mathcal{P}_j x \;=\; z + Q_j(x - z)$$

for all $x \in \mathbb{R}^n$ and any $z \in X_j$. In particular, this formula is independent of the particular choice of z.

Lemma. The projection $\mathcal{P}_j$ can be written explicitly as

$$\mathcal{P}_j x \;=\; x + A_j^{\mathrm{T}}(A_j A_j^{\mathrm{T}})^{-1}(y_j - A_j x)$$

for all $x \in \mathbb{R}^n$ and $j = 1, \ldots, l$.

Proof. We start by proving that $A_j A_j^{\mathrm{T}} \in \mathbb{R}^{k_j \times k_j}$ is invertible. Since $A_j : \mathbb{R}^n \to \mathbb{R}^{k_j}$ is surjective, it follows that

$$\mathrm{Ker}(A_j^{\mathrm{T}})^{\perp} \,=\, \mathrm{Ran}(A_j) \,=\, \mathbb{R}^{k_j}.$$

Hence, $\mathrm{Ker}(A_j^{\mathrm{T}}) = \{0\}$, i.e., A_j^{T} is injective. This means, in fact, that also $A_j A_j^{\mathrm{T}}$ is injective:

$$A_j A_j^{\mathrm{T}} z = 0 \;\Rightarrow\; z^{\mathrm{T}} A_j A_j^{\mathrm{T}} z = 0 \;\Rightarrow\; \|A_j^{\mathrm{T}} z\|^2 = 0 \;\Rightarrow\; z = 0.$$

Due to the fundamental theorem of linear algebra, the injective square matrix $A_j A_j^{\mathrm{T}}$ is invertible.

Fix an arbitrary $x \in \mathbb{R}^n$ and let us write

$$\mathcal{P}_j x \,=\, z + Q_j(x - z)$$

with some $z \in X_j$, as suggested before the lemma.

Since $Q_j : \mathbb{R}^n \to \mathrm{Ker}(A_j)$ is an orthogonal projection, $I - Q_j$ maps $\mathbb{R}^n$ onto $\mathrm{Ker}(A_j)^\perp$ (and is, in fact, also an orthogonal projection). Hence, we have

$$x - \mathcal{P}_j x \;=\; (I - Q_j)(x - z) \in \mathrm{Ker}(A_j)^\perp = \mathrm{Ran}(A_j^{\mathrm{T}}).$$

This means that there exist $w \in \mathbb{R}^{k_j}$ such that

$$A_j^{\mathrm{T}} w \;=\; x - \mathcal{P}_j x, \tag{9}$$

and, consequently,

$$A_j A_j^{\mathrm{T}} w \;=\; A_j x - A_j \mathcal{P}_j x \;=\; A_j x - y_j$$

because $\mathcal{P}_j x \in X_j$. Solving this equation for w and substituting into (9) results in

$$A_j^{\mathrm{T}} (A_j A_j^{\mathrm{T}})^{-1} (A_j x - y_j) \;=\; x - \mathcal{P}_j x,$$

which completes the proof. □

Algorithmic implementation of ART

In the case of ART, i.e., when the submatrices $A_j = a_j^{\mathrm{T}}$, $j = 1, \ldots, m$, are the rows of the original system matrix A, and y_j, $j = 1, \ldots, m$, are the components of y, the inverse needed above

$$(A_j A_j^{\mathrm{T}})^{-1} = (a_j^{\mathrm{T}} a_j)^{-1} = 1/\|a_j\|^2$$

is just a real number. Thus, the ART algorithm reads as

Set $k = 0$ and $x_0 = 0$;

Repeat until the chosen stopping rule is satisfied:

$z_0 = x_k$;

for $j = 1, \ldots, m$

$z_j = z_{j-1} + (1/\|a_j\|^2)(y_j - a_j^{\mathrm{T}} z_{j-1}) a_j$;

end

$x_{k+1} = z_m$; $\quad k \leftarrow k + 1$;

end

Discrepancy principle for the Kaczmarz iteration

As you probably guess, we let the measurement $y \in \mathbb{R}^m$ be a noisy version of some underlying 'exact' data vector $y_0 \in \mathbb{R}^m$, and assume that

$$\|y - y_0\| \approx \epsilon > 0.$$

The Morozov discrepancy principle works for the Kaczmarz iteration as follows: Choose the smallest $k \geq 0$ such that the residual satisfies

$$\|y - Ax_k\| \leq \epsilon,$$

if such k exists.

Unlike for the truncated SVD and the Landweber–Fridman iteration, the condition

$$\epsilon > \|y - Py\|,$$

where P is the projection onto the range of A, is not sufficient to guarantee the existence of such a stopping index k without further assumptions. As an example, in the second example of this lecture

$$\|y - Ax_k\| \rightsquigarrow 0.98 \qquad \text{as } k \to \infty,$$

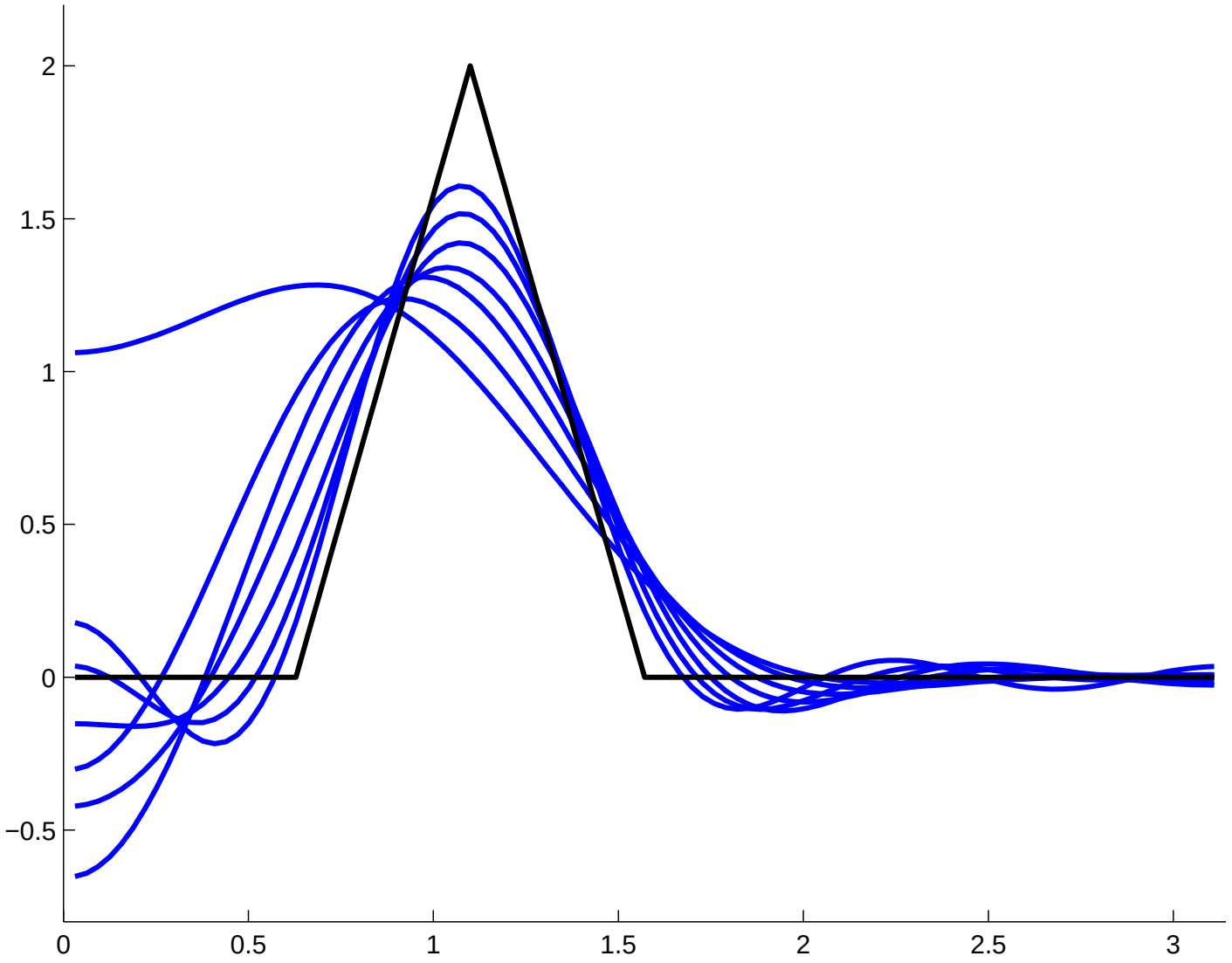
while $\|y - Py\| \approx 0.59$.

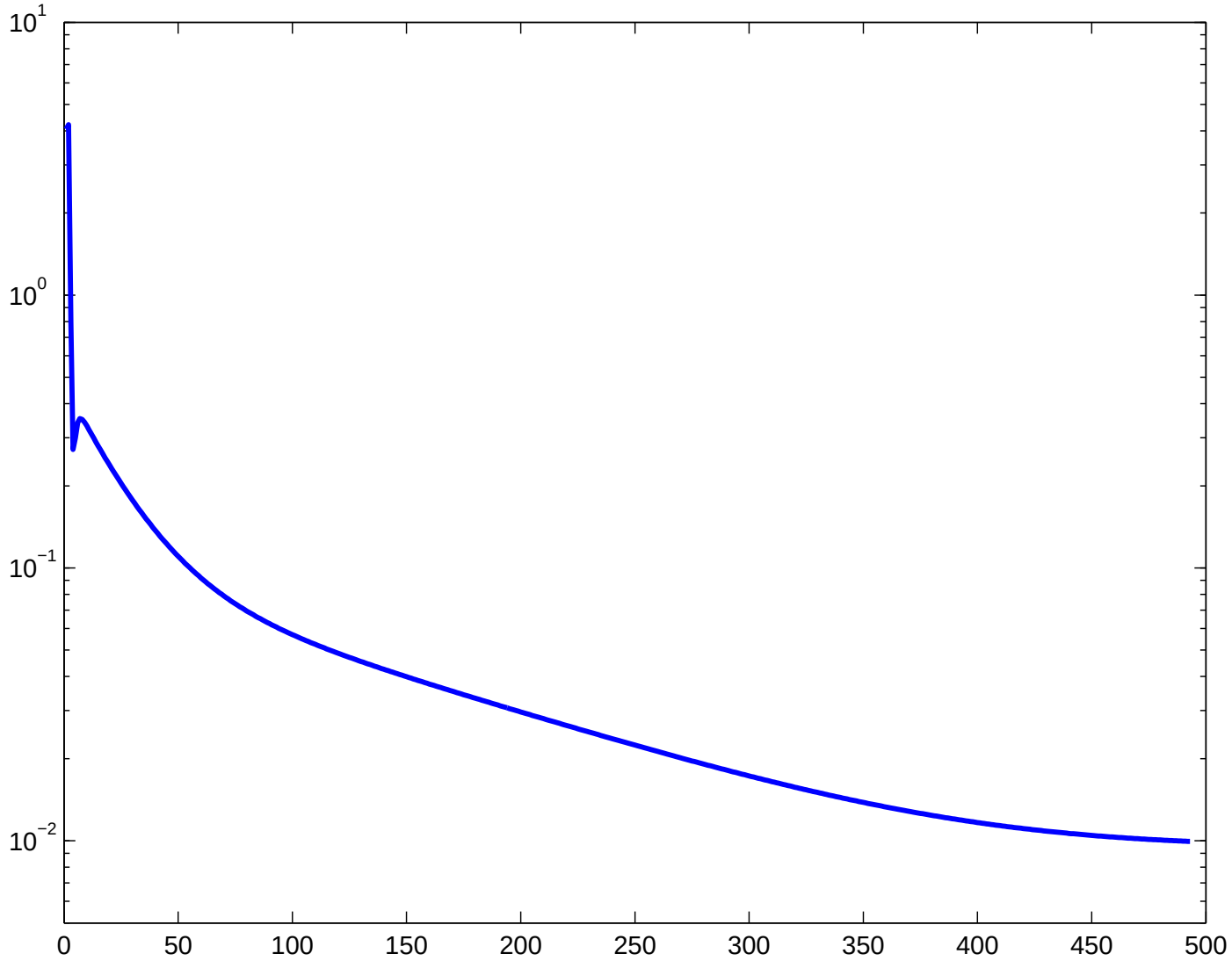
However, one can always try to apply the Morozov discrepancy principle and hope for the best.

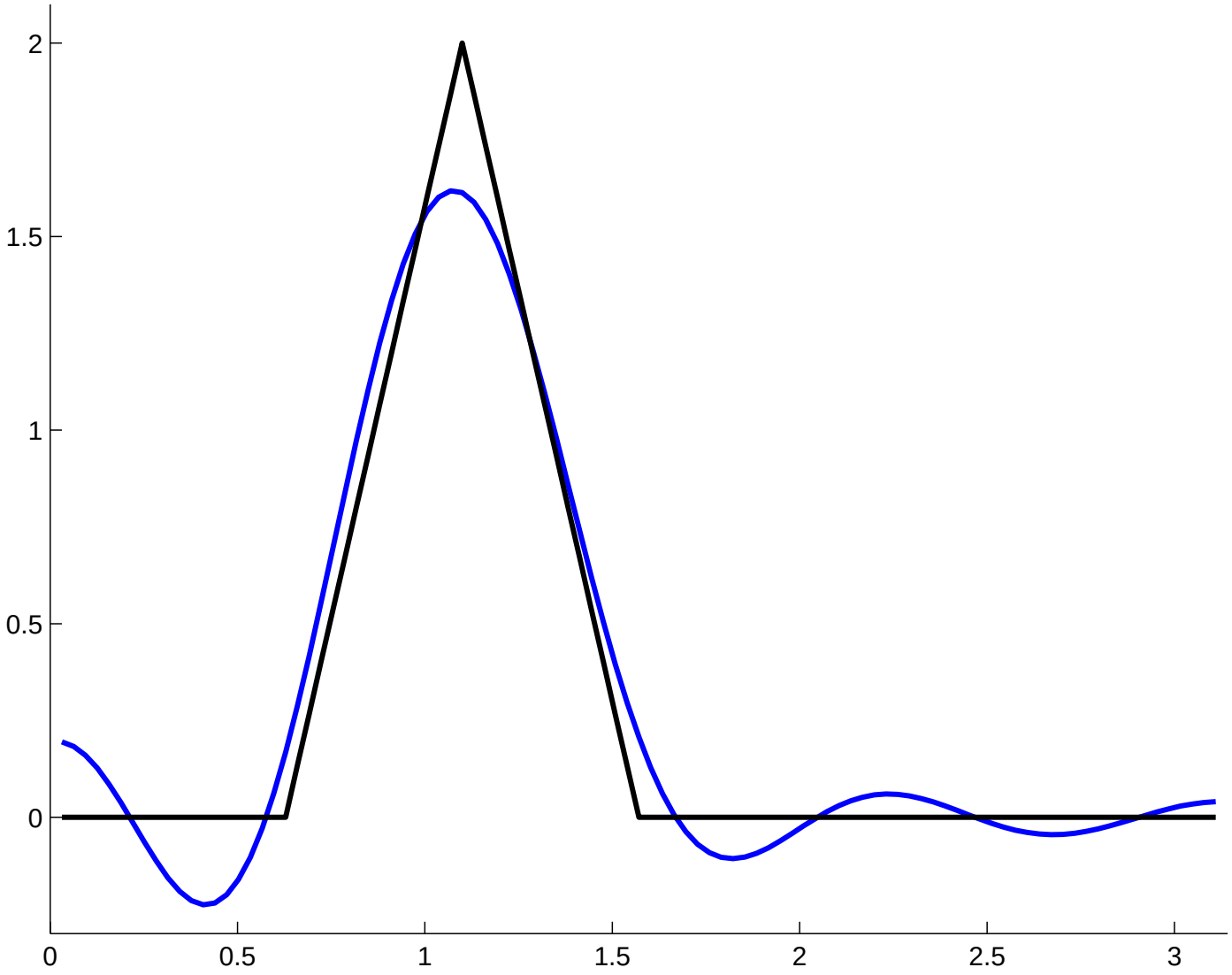
An example: Heat distribution in a rod (revisited)

Let us once again consider the discretized inverse heat conduction problem in an insulated rod. We simulate the data in the exactly same way as above, add the same amount of noise and use the same value of ϵ for the Morozov discrepancy principle. The implementation of ART with the discrepancy principle in Matlab is straightforward.

With the particular realization of the measurement noise, the Morozov discrepancy principle was satisfied by the iterate corresponding to $k = 493$. In the following, we visualize the evolution of the ART iteration for $k = 1, 2, 7, 20, 54, 148, 403$, show the residual as a function of k, and plot the solution corresponding to the discrepancy principle.


2
1.5
1
0.5
0
−0.5
0
0.5
1
1.5
2
2.5
3


10^{1}
10^{0}
10^{-1}
10^{-2}
0
50
100
150
200
250
300
350
400
450
500


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Seventh lecture, February 9, 2011.

2.4.3 Krylov subspace methods

Krylov subspace methods

The Krylov subspace methods are iterative solvers for (large) matrix equations of the form $Ax = y$, $A \in \mathbb{R}^{n \times n}$. Loosely speaking, such methods try to approximate the solution vector $x \in \mathbb{R}^n$ as a linear combination of vectors of the type u, Au, A^2u etc., with some given $u \in \mathbb{R}^n$. If multiplication by A is cheap — e.g., if A is sparse —, the Krylov subspace methods are especially efficient.

On this course, we only consider the most well-known Krylov subspace method, the conjugate gradient method. Other methods of this class include, e.g., the generalized minimal residual method (GMRES), and the biconjugate gradient method (BiCG).

The regularizing properties of the conjugate gradient method can be analyzed explicitly; see, e.g., the monograph

M. Hanke, *Conjugate gradient type methods for ill-posed problems*, Pitman Research Notes in Mathematics Series, 327.

However, here we content ourselves with introducing the basic ideas behind the conjugate gradient scheme and demonstrating numerically how application of an 'early stopping rule' provides reasonable solutions for inverse problems.

Assumptions on A and a related inner product

We assume that the system matrix $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, i.e.,

$$A^{\mathrm{T}} = A \qquad \text{and} \qquad u^{\mathrm{T}} A u > 0 \quad \text{for } u \neq 0.$$

In particular, this means that the square matrix A is injective, and consequently invertible due to the fundamental theorem of linear algebra. It is easy to see that the inverse $A^{-1} \in \mathbb{R}^{n \times n}$ is also symmetric and positive definite.

We define an A-dependent inner product and the corresponding norm via

$$\langle u, v \rangle_A = u^{\mathrm{T}} A v \qquad \text{and} \qquad \|u\|_A = \langle u, u \rangle_A^{1/2}.$$

It follows from the assumptions on A that $\langle \cdot, \cdot \rangle_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ really is an inner product on $\mathbb{R}^n$, and consequently $\| \cdot \|_A : \mathbb{R}^n \to \mathbb{R}$ is a norm.

The error, the residual and a minimization problem

Let $x_* = A^{-1}y \in \mathbb{R}^n$ be the unique solution of the equation

$$Ax = y$$

for a given $y \in \mathbb{R}^n$. We define the error and the residual corresponding to some approximative solution $x \in \mathbb{R}^n$ by

$$e = x_* - x \qquad \text{and} \qquad r = y - Ax = Ae.$$

Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be the A-dependent quadratic functional

$$\phi(x) = \|e\|_A^2 = e^{\mathrm{T}}Ae = r^{\mathrm{T}}A^{-1}r = \|r\|_{A^{-1}}^2.$$

Since $\|\cdot\|_A$ is a norm, $\phi(x)$ is non-negative and equals zero if and only if

$$e = 0 \qquad \Longleftrightarrow \qquad x = x_*.$$

Hence, minimizing ϕ is equivalent to solving the original equation.

Minimizing ϕ in a given direction

Evaluating ϕ would require the knowledge of x_* or, equivalently, that of A^{-1}; since our ultimate goal is to approximate the solution x_* iteratively, assuming it known is not a feasible option.

Fortunately, if we have some initial guess $x_0 \in \mathbb{R}^n$ and some search direction $0 \neq s_0 \in \mathbb{R}^n$, we can find the minimizer of ϕ over the line

$$\mathcal{S}_0 = \{x \in \mathbb{R}^n \mid x = x_0 + \alpha s_0,\ \alpha \in \mathbb{R}\}$$

without knowing x_*.

Lemma. *The function*

$$\alpha \mapsto \phi(x_0 + \alpha s_0), \quad \mathbb{R} \to \mathbb{R},$$

attains its minimum at

$$\alpha = \alpha_0 := \frac{s_0^{\mathrm{T}} r_0}{\|s_0\|_A^2} = \frac{s_0^{\mathrm{T}} r_0}{s_0^{\mathrm{T}} A s_0},$$

where r_0 is the residual corresponding to the initial guess:

$$r_0 = y - A x_0.$$

Proof. The residual corresponding to $x = x_0 + \alpha s_0$ is

$$r \;=\; y - Ax \;=\; y - Ax_0 - \alpha A s_0 \;=\; r_0 - \alpha A s_0.$$

In consequence,

$$\begin{aligned} \phi(x) &= r^{\mathrm{T}} A^{-1} r \\ &= (r_0 - \alpha A s_0)^{\mathrm{T}} A^{-1} (r_0 - \alpha A s_0) \\ &= \alpha^2 s_0^{\mathrm{T}} A s_0 - 2\alpha s_0^{\mathrm{T}} r_0 + r_0^{\mathrm{T}} A^{-1} r_0, \end{aligned}$$

which, as a function of α, is a parabola that opens upwards, because $s_0^{\mathrm{T}} A s_0 > 0$. Hence, its minimum is at the unique zero of the derivative with respect to α, i.e., at $\alpha = \alpha_0$. □

About the choice of the search directions

Given a sequence of (non-zero) search directions $\{s_k\} \subset \mathbb{R}^n$, we can thus produce a sequence of approximate solutions by first choosing x_0 and then finding iteratively the minimizer of ϕ on the line passing through x_k in the direction s_k as follows:

$$x_{k+1} = x_k + \alpha_k s_k, \qquad \text{with } \alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k}, \qquad k = 0, 1, \ldots,$$

where r_k is the residual corresponding to the kth iterate, i.e.,

$$r_k = y - A x_k.$$

Notice that $\{\phi(x_k)\}$ is a decreasing sequence of real numbers because $\phi(x_{k+1})$ is always smaller than — or as small as — $\phi(x_k)$.

However, an efficient choice of the search directions $\{s_k\}$ is a subtle issue.

Probably, one of the first ideas that comes to mind is to choose

$$s_k = -\nabla\phi(x_k) = 2(y - Ax_k), \qquad k = 0, 1, \ldots,$$

because it gives the direction of the *steepest descent*. However, this does not in general provide a sequence $\{x_k\}$ that converges fast towards the global minimizer $x_* = A^{-1}y$, as demonstrated by the following example:

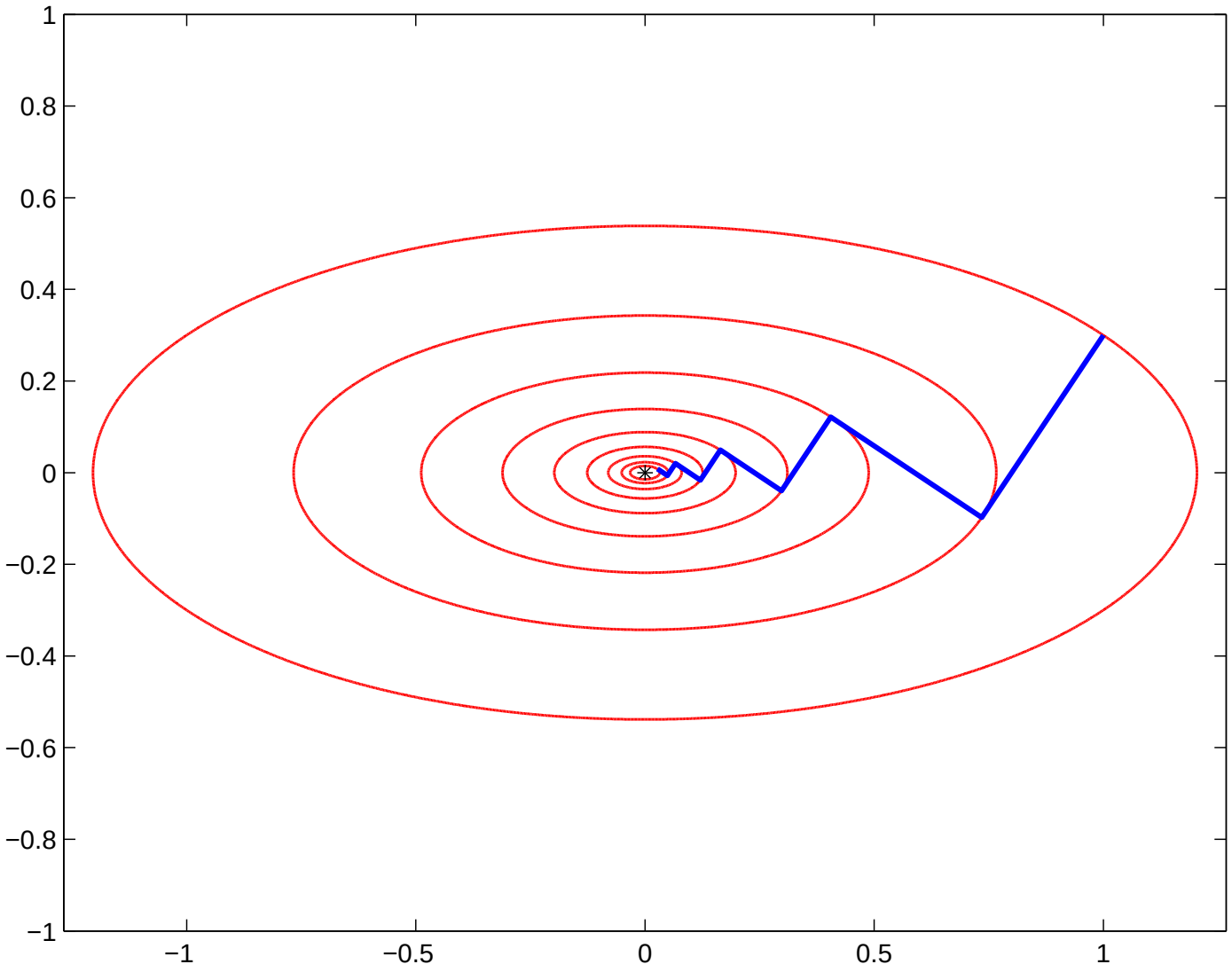
Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \qquad \text{and} \qquad y = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which means, in particular, that

$$\phi(x) = \phi(x^{(1)}, x^{(2)}) = (x^{(1)})^2 + 5(x^{(2)})^2.$$

The following image shows level contours of ϕ and the sequence $\{x_k\}_{k=0}^{9}$ starting from $x_0 = (1, 0.3)^{\mathrm{T}}$. The actual solution $x_* = (0, 0)^{\mathrm{T}}$ is marked with an asterisk.


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1

Minimizing ϕ over a hyperplane

Let $\{s_0, \ldots, s_k\}$ be a set of linearly independent search direction. Next, we consider finding the minimizer of ϕ on the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\},$$

where $x_0 \in \mathbb{R}^n$ is the initial guess and $S_k = [s_0, \ldots, s_k] \in \mathbb{R}^{n\times(k+1)}$.

Lemma. *The function*

$$h \mapsto \phi(x_0 + S_k h), \quad \mathbb{R}^{k+1} \to \mathbb{R},$$

attains its minimum at

$$h = h_* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0,$$

where $r_0 = y - Ax_0$ is the residual corresponding to the initial guess.

Proof. Let us first prove that $S_k^{\mathrm{T}} A S_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is invertible: Due to the positive definiteness of A, we have

$$S_k^{\mathrm{T}} A S_k z = 0 \quad \Longrightarrow \quad z^{\mathrm{T}} S_k^{\mathrm{T}} A S_k z = 0 \quad \Longrightarrow \quad S_k z = 0,$$

which means that $z = 0$ since the columns of S_k are linearly independent. Hence, $\mathrm{Ker}(S_k^{\mathrm{T}} A S_k) = \{0\}$, i.e., $S_k^{\mathrm{T}} A S_k$ is injective, and thus $(S_k^{\mathrm{T}} A S_k)^{-1}$ exists by the fundamental theorem of linear algebra.

The residual corresponding to $x = x_0 + S_k h$ satisfies

$$r \;=\; y - A(x_0 + S_k h) \;=\; r_0 - A S_k h,$$

and thus

$$\begin{aligned} \phi(x_0 + S_k h) &= (r_0 - A S_k h)^{\mathrm{T}} A^{-1} (r_0 - A S_k h) \\ &= h^{\mathrm{T}} S_k^{\mathrm{T}} A S_k h - 2 r_0^{\mathrm{T}} S_k h + r_0^{\mathrm{T}} A^{-1} r_0. \end{aligned}$$

In particular, the coefficient matrix $S_k^{\mathrm{T}} A S_k$ of the quadratic term of $\phi(x_0 + S_k h)$ in h is positive definite:

$$u^{\mathrm{T}}(S_k^{\mathrm{T}} A S_k)u = (S_k u)^{\mathrm{T}} A (S_k u) \geq 0, \qquad u \in \mathbb{R}^{k+1},$$

where the equality holds if and only if $S_k u = 0$, i.e., $u = 0$. Thus, the basics of quadratic programming tell us that the unique zero of the gradient of $\phi(x_0 + S_k h)$ with respect to h, i.e.,

$$h_* \; = \; (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0,$$

is the unique minimizer of $\phi(x_0 + S_k h)$ over $h \in \mathbb{R}^{k+1}$. $\square$

A-conjugate search directions

Since finding the minimizer of ϕ over the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\}$$

involves inverting a $(k+1) \times (k+1)$ matrix, such an approach is not necessarily very attractive.

On the other hand, as demonstrated by the numerical example above, minimizing ϕ sequentially in the directions $s_0, \ldots, s_k$ does not, in general, result in as good approximate solution as doing the minimization over the whole hyperplane $\mathcal{S}_k$ at once. (Clearly, the first two search directions of the numerical example were linearly independent, and thus minimization over the hyperplane $\mathcal{S}_2$, i.e., the whole $\mathbb{R}^2$, would have given the global minimizer $x_* = (0, 0)^{\mathrm{T}}$.)

However, the sequential minimization *does* produce the minimizer over $\mathcal{S}_k$ if the search directions $\{s_0, \ldots, s_k\}$ are chosen in a clever way.

We say that non-zero vectors $\{s_0, \ldots, s_k\} \subset \mathbb{R}^n$ are A-conjugate if

$$\langle s_i, s_j \rangle_A \;=\; s_i^{\mathrm{T}} A s_j \;=\; 0$$

for $i \neq j$. In other words, the vectors $\{s_0, \ldots, s_k\}$ are A-conjugate if they are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_A$.

The A-conjugacy condition can be expressed neatly with the help of the matrix $S_k = [s_0, \ldots, s_k] \in \mathbb{R}^{n \times (k+1)}$:

$$S_k^{\mathrm{T}} A S_k \;=\; \begin{bmatrix} s_0^{\mathrm{T}} \\ \vdots \\ s_k^{\mathrm{T}} \end{bmatrix} [As_0, \ldots, As_k] \;=\; \mathrm{diag}(d_0, d_1 \ldots, d_k) \in \mathbb{R}^{(k+1)\times(k+1)},$$

where $d_j = s_j^{\mathrm{T}} A s_j > 0$, $j = 0, \ldots, k$, due to the positive definiteness of the matrix A.

The following theorem demonstrates that it is useful to choose the search directions to be A-conjugate.

Theorem. *Let $x_0 \in \mathbb{R}^n$ be an initial guess and assume that the vectors $\{s_0, \ldots, s_k\} \subset \mathbb{R}^n$ are non-zero and A-conjugate. Then, the sequential minimizer of ϕ over these directions, i.e., $x_{k+1} \in \mathbb{R}^n$ obtained by the iteration*

$$x_{j+1} = x_j + \alpha_j s_j, \qquad \text{with } \alpha_j = \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j}, \qquad j = 0, \ldots, k,$$

is the minimizer of ϕ on the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\}.$$

To put it short,

$$x_{k+1} = x_0 + S_k h_* = x_0 + S_k (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0.$$

Proof. Let $a_j = (\alpha_0, \ldots, \alpha_j)^{\mathrm{T}} \in \mathbb{R}^{j+1}$. With this notation we have

$$x_j \;=\; x_0 + \sum_{i=0}^{j-1} \alpha_i s_i \;=\; x_0 + S_{j-1} a_{j-1}, \qquad j = 1, \ldots, k+1.$$

Moreover the residual corresponding to x_j is

$$r_j \;=\; y - Ax_j \;=\; (y - Ax_0) - AS_{j-1}a_{j-1} \;=\; r_0 - AS_{j-1}a_{j-1}.$$

In particular,

$$s_j^{\mathrm{T}} r_j \;=\; s_j^{\mathrm{T}} r_0 - s_j^{\mathrm{T}} AS_{j-1}a_{j-1} \;=\; s_j^{\mathrm{T}} r_0 + s_j^{\mathrm{T}} [As_0, \ldots, As_{j-1}]a_{j-1},$$

where the last term vanishes since s_j is A-conjugate to $\{s_0, \ldots, s_{j-1}\}$.

Hence,

$$\alpha_j \;=\; \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j} \;=\; \frac{s_j^{\mathrm{T}} r_0}{s_j^{\mathrm{T}} A s_j}, \qquad j = 0, \ldots, k.$$

On the other hand, since $\{s_0, \ldots, s_k\}$ are A-conjugate, we have

$$\begin{aligned} (S_k^{\mathrm{T}} A S_k)^{-1} &= \big(\mathrm{diag}(s_0^{\mathrm{T}} A s_0, \ldots, s_k^{\mathrm{T}} A s_k)\big)^{-1} \\ &= \mathrm{diag}(1/(s_0^{\mathrm{T}} A s_0), \ldots, 1/(s_k^{\mathrm{T}} A s_k)), \end{aligned}$$

which means that

$$h_* \;=\; (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0 \;=\; (S_k^{\mathrm{T}} A S_k)^{-1} \begin{bmatrix} s_0^{\mathrm{T}} r_0 \\ \vdots \\ s_k^{\mathrm{T}} r_0 \end{bmatrix} \;=\; \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{bmatrix}.$$

Consequently, $a_k = h_*$ and

$$x_{k+1} \;=\; x_0 + S_k a_k \;=\; x_0 + S_k h_*. \qquad \square$$

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Eighth lecture, February 11, 2011.

Summary of the previous lecture

A Minimization problem: Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Instead of solving the original equation $Ax = y$ directly, we consider minimizing the functional

$$\phi(x) = (x_*-x)^{\mathrm{T}}A(x_*-x) = e^{\mathrm{T}}Ae = (y-Ax)^{\mathrm{T}}A^{-1}(y-Ax) = r^{\mathrm{T}}A^{-1}r,$$

where $x_* = A^{-1}y$ is the actual solution, and e and r are called the error and the residual corresponding to the approximate solution x. The unique minimizer of this functional is the solution of the original problem, i.e., x_*.

A sequence of minimizers: Given an initial guess x_0 and a set of non-zero search directions $\{s_j\}_{j=0}^{k} \subset \mathbb{R}^n$, we define the approximate solution x_{j+1}, $j = 1, \ldots, k$, recursively as the minimizer of the functional ϕ on the line

$$\mathcal{S}_j \;=\; \{x \in \mathbb{R}^n \mid x = x_j + \alpha s_j,\ \alpha \in \mathbb{R}\}.$$

This can be done through the iteration

$$x_{j+1} \;=\; x_j + \alpha_j s_j, \qquad \text{with} \;\; \alpha_j \;=\; \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j}, \qquad j = 0, \ldots, k,$$

where $r_j = y - Ax_j$ is the residual corresponding to x_j.

A-conjugate search directions: The non-zero vectors $\{s_j\}_{j=0}^k$ are called A-conjugate if

$$\langle s_i, s_j \rangle_A \;=\; s_i^{\mathrm{T}} A s_j \;=\; 0 \qquad \text{for } i \neq j.$$

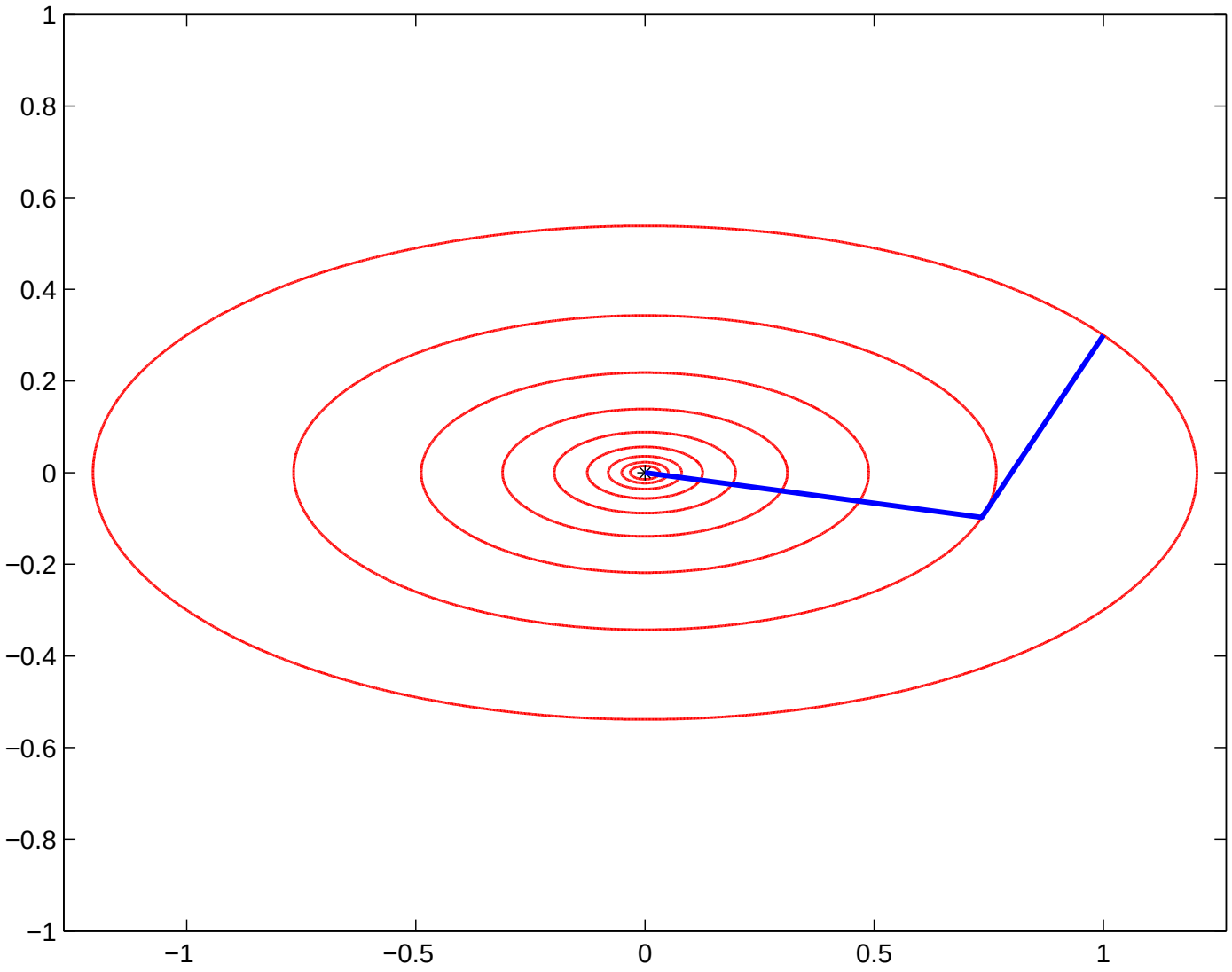
If the search directions are chosen this cleverly, the iterate x_{k+1} is the minimizer of ϕ over the whole hyperplane

$$\mathcal{S}_k \;=\; \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\; h \in \mathbb{R}^{k+1}\},$$

i.e., over all vectors of the form $x = x_0 + \sum_{j=0}^k h_j s_j$, where $h_0, \ldots, h_k$ are real numbers. This minimizer can be given explicitly as

$$x_{k+1} \;=\; x_0 + S_k h_*, \qquad h_* \;=\; (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0,$$

where $S_k = [s_0, \ldots, s_k] \in \mathbb{R}^{n \times (k+1)}$. In particular, x_n is the global minimizer, i.e., $x_n = x_*$.


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1

A useful corollary about the residuals

If the search directions are chosen to be A-conjugate, we have also extra information about the residuals:

Corollary. *If the non-zero search directions* $\{s_j\}_{j=0}^{k} \subset \mathbb{R}^n$ *are* A*-conjugate, then the residual* $r_{k+1} = y - Ax_{k+1}$ *satisfies*

$$r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\},$$

where the orthogonality is in the sense of the standard inner product.

Proof. Since $x_{k+1} = x_0 + S_k h_*$, it holds that

$$r_{k+1} \;=\; (y - Ax_0) - AS_k h_* \;=\; r_0 - AS_k h_*.$$

In consequence,

$$[r_{k+1}^{\mathrm{T}} s_0, \ldots, r_{k+1}^{\mathrm{T}} s_k] \;=\; r_{k+1}^{\mathrm{T}} S_k \;=\; r_0^{\mathrm{T}} S_k - h_*^{\mathrm{T}} S_k^{\mathrm{T}} A S_k \;=\; 0$$

because $h_*^{\mathrm{T}} = ((S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0)^{\mathrm{T}} = r_0^{\mathrm{T}} S_k (S_k^{\mathrm{T}} A S_k)^{-1}$.

How to construct A-conjugate search directions?

There are many ways to construct a set of A-conjugate search directions. If one chooses to use Krylov subspaces the result is the conjugate gradient algorithm:

Definition: *The kth Krylov subspace of A with the initial vector $r_0 = y - Ax_0$ is defined as*

$$\mathcal{K}_k = \mathcal{K}(A, r_0) = \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}, \qquad k = 1, 2, \ldots.$$

Note, in particular, that $A(\mathcal{K}_k) \subset \mathcal{K}_{k+1}$.

Take also note that $\mathcal{K}_{k-1} \subset \mathcal{K}_k$, where the dimension of the latter is *at most* k, and it is *at most* one higher than that of the former. (For example, if r_0 is an eigenvector of A, then the vectors spanning $\mathcal{K}_k$ are scalar multiples of each other, which means that $\dim(\mathcal{K}_k) = 1$ for all $k \geq 1$. Fortunately, it turns out that this is not a hindrance.)

The logic of the conjugate gradient algorithm

Let us construct a sequence of A-conjugate search directions inductively. The leading idea is that, given a set of A-conjugate search direction, we can either find a new A-conjugate direction or the previous iterate is already the global minimizer x_*, i.e., the unique solution of $Ax = y$.

1. Choose an initial guess $x_0 \in \mathbb{R}^n$.

2. If $r_0 = y - Ax_0 = 0$, we have found the solution $x_* = x_0$. Otherwise, set $s_0 = r_0$ (, which is, by the way, the steepest descent direction). Note, in particular, that the set of a single search direction $\{s_0\}$ is trivially A-conjugate and

$$\mathcal{K}_1 = \mathrm{span}\{s_0\} = \mathrm{span}\{r_0\}.$$

3. Suppose that we have non-zero and A-conjugate search directions $\{s_j\}_{j=0}^{k-1}$, $k \geq 1$, such that

$$\mathcal{K}_m = \text{span}\{s_0, \ldots, s_{m-1}\} = \text{span}\{r_0, \ldots, r_{m-1}\}, \qquad m = 1, \ldots, k, \tag{10}$$

where $r_j = y - Ax_j$, $j = 0, \ldots, k-1$, are the residuals corresponding to the iterates $\{x_j\}_{j=0}^{k-1}$ of the sequential minimization algorithm.

If $r_k = 0$, the algorithm has converged to $x_* = x_k$. Otherwise, we try to choose another A-conjugate and non-zero search direction $s_k \in \mathbb{R}^n$ so that (10) remains valid if k is replaced by $k+1$.

Assume thus that $r_k \neq 0$. Since

$$r_k = y - Ax_k = y - A(x_{k-1} + \alpha_{k-1}s_{k-1}) = r_{k-1} - \alpha_{k-1}As_{k-1}$$

and r_{k-1} and s_{k-1} belong by assumption to $\mathcal{K}_k$, the new residual r_k belongs to $\mathcal{K}_{k+1}$. Since r_k is orthogonal to $\{s_0, \ldots, s_{k-1}\}$, which span $\mathcal{K}_k$ and belong to $\mathcal{K}_{k+1}$, we must have

$$\mathcal{K}_{k+1} = \mathrm{span}\{s_0, \ldots, s_{k-1}, r_k\} = \mathrm{span}\{r_0, \ldots, r_{k-1}, r_k\}.$$

Let us try to find the new search direction s_k in the form

$$s_k = r_k + \beta_{k-1}s_{k-1}, \qquad \beta_{k-1} \in \mathbb{R}.$$

Note that this kind of vector belongs to $\mathcal{K}_{k+1}$ and, furthermore,

$$\mathcal{K}_{k+1} = \mathrm{span}\{s_0, \ldots, s_{k-1}, r_k\} = \mathrm{span}\{s_0, \ldots, s_{k-1}, s_k\}.$$

Consequently, all we have to worry about is the A-conjugacy condition:

We want to choose $\beta_{k-1} \in \mathbb{R}^k$ so that

$$\begin{aligned} s_j^{\mathrm{T}} A s_k &= s_j^{\mathrm{T}} A r_k + \beta_{k-1} s_j^{\mathrm{T}} A s_{k-1} \\ &= (A s_j)^{\mathrm{T}} r_k + \beta_{k-1} s_j^{\mathrm{T}} A s_{k-1} = 0 \end{aligned} \tag{11}$$

for $j = 0, \ldots, k-1$. Because $\{s_0, \ldots, s_{k-2}\} \subset \mathcal{K}_{k-1}$, we have

$$\{A s_0, \ldots, A s_{k-2}\} \subset \mathcal{K}_k = \operatorname{span}\{s_0, \ldots, s_{k-1}\},$$

and thus the vectors $\{A s_0, \ldots, A s_{k-2}\}$ are orthogonal to r_k. Hence, the A-conjugacy of $\{s_0, \ldots, s_{k-1}\}$ yields that only the last of the equations (11) is non-trivial.

Solving this equation for β_{k-1} results in the needed update rule

$$s_k = r_k + \beta_{k-1} s_{k-1}, \qquad \beta_{k-1} = -\frac{s_{k-1}^{\mathrm{T}} A r_k}{s_{k-1}^{\mathrm{T}} A s_{k-1}}.$$

Conjugate gradient method

To sum up, we have arrived at the following algorithm

Choose x_0.

Set $k = 0$, $r_0 = y - Ax_0$, $s_0 = r_0$;

Repeat until the chosen stopping rule is satisfied:

$\alpha_k = (s_k^{\mathrm{T}} r_k)/(s_k^{\mathrm{T}} A s_k)$;

$x_{k+1} = x_k + \alpha_k s_k$;

$r_{k+1} = r_k - \alpha_k A s_k$; % Note: $r_{k+1} = y - Ax_k - \alpha_k A s_k$

$\beta_k = -(s_k^{\mathrm{T}} A r_{k+1})/(s_k^{\mathrm{T}} A s_k)$;

$s_{k+1} = r_{k+1} + \beta_k s_k$;

$k \leftarrow k + 1$;

end

However, the algorithm is usually presented in a slightly different form. Assuming that the iteration has not yet converged at the iterate x_k, we deduce the following formulae:

Since $r_k \perp s_{k-1}$,

$$s_k^{\mathrm{T}} r_k \;=\; (r_k + \beta_{k-1} s_{k-1})^{\mathrm{T}} r_k \;=\; \|r_k\|^2,$$

resulting in

$$\alpha_k = \frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k}.$$

In particular, since $r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\} = \mathcal{K}_{k+1} \ni r_k$, this means that

$$\|r_{k+1}\|^2 \;=\; r_{k+1}^{\mathrm{T}}(r_k - \alpha_k A s_k) \;=\; -\frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k}\, r_{k+1}^{\mathrm{T}} A s_k \;=\; \beta_k \|r_k\|^2.$$

Solving for β_k and plugging the obtained formulae for α_k and β_k into the preliminary conjugate gradient algorithm leads to the standard form of the method:

Choose x_0.

Set $k = 0$, $r_0 = y - Ax_0$, $s_0 = r_0$;

Repeat until the chosen stopping rule is satisfied:

$$\alpha_k = \|r_k\|^2 / (s_k^{\mathrm{T}} A s_k);$$
$$x_{k+1} = x_k + \alpha_k s_k;$$
$$r_{k+1} = r_k - \alpha_k A s_k;$$
$$\beta_k = \|r_{k+1}\|^2 / \|r_k\|^2;$$
$$s_{k+1} = r_{k+1} + \beta_k s_k;$$
$$k \leftarrow k + 1;$$

end

NB: *There is an error in the update formula for x_{k+1} in the textbook.*

Conjugate gradient method for inverse problems

According to the above construction, if you apply the conjugate gradient method to the equation

$$Ax = y,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, you obtain the exact solution — up to rounding errors — in at most n iteration steps, i.e., $x_n = x_* = A^{-1}y$. However, such extensive iterating is not usually necessary: The algorithm typically converges satisfactorily much quicker; see, e.g., 2. exercise of the 4. session, where a (pessimistic) convergence rate is provided.

When dealing with ill-posed problems, one should be even more careful and terminate the iterations well before convergence, in order to avoid fitting the solution to noise. One should, actually, be extremely cautious because the conjugate gradient method often converges very fast.

Let us be a bit more precise and consider a general ill-posed matrix equation

$$Ax = y,$$

where $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ are given.

In some cases, one may have $m = n$ and, in addition, some prior information stating that A is — at least in theory — positive (semi-)definite. In such situation, one can apply the conjugate gradient algorithm directly on this original equation.

In the general case, one may still consider the normal equation

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}y,$$

which corresponds, in essence, to solving the original equation in the least squares sense.

Now, the system matrix $A^{\mathrm{T}}A = (A^{\mathrm{T}}A)^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite:

$$u^{\mathrm{T}}A^{\mathrm{T}}Au \;=\; \|Au\|^2 \;>\; 0 \qquad \text{for all } \; u \in (\mathbb{R}^n \setminus \mathrm{Ker}(A)).$$

Hence, the conditions of the conjugate gradient algorithm are almost satisfied, and one may look for the solution of the inverse problem by using the conjugate gradient algorithm with A replaced by $A^{\mathrm{T}}A$ and y by $A^{\mathrm{T}}y$. (When implementing the algorithm in Matlab, bear in mind that matrix-matrix products are typically far more expensive than matrix-vector products.)

As a stopping condition, one may try, e.g., the Morozov principle for the original equation: Terminate the iteration when

$$\|y - Ax_k\| \;\leq\; \epsilon$$

for some $\epsilon > 0$, which measures the amount of noise in y in some sense.

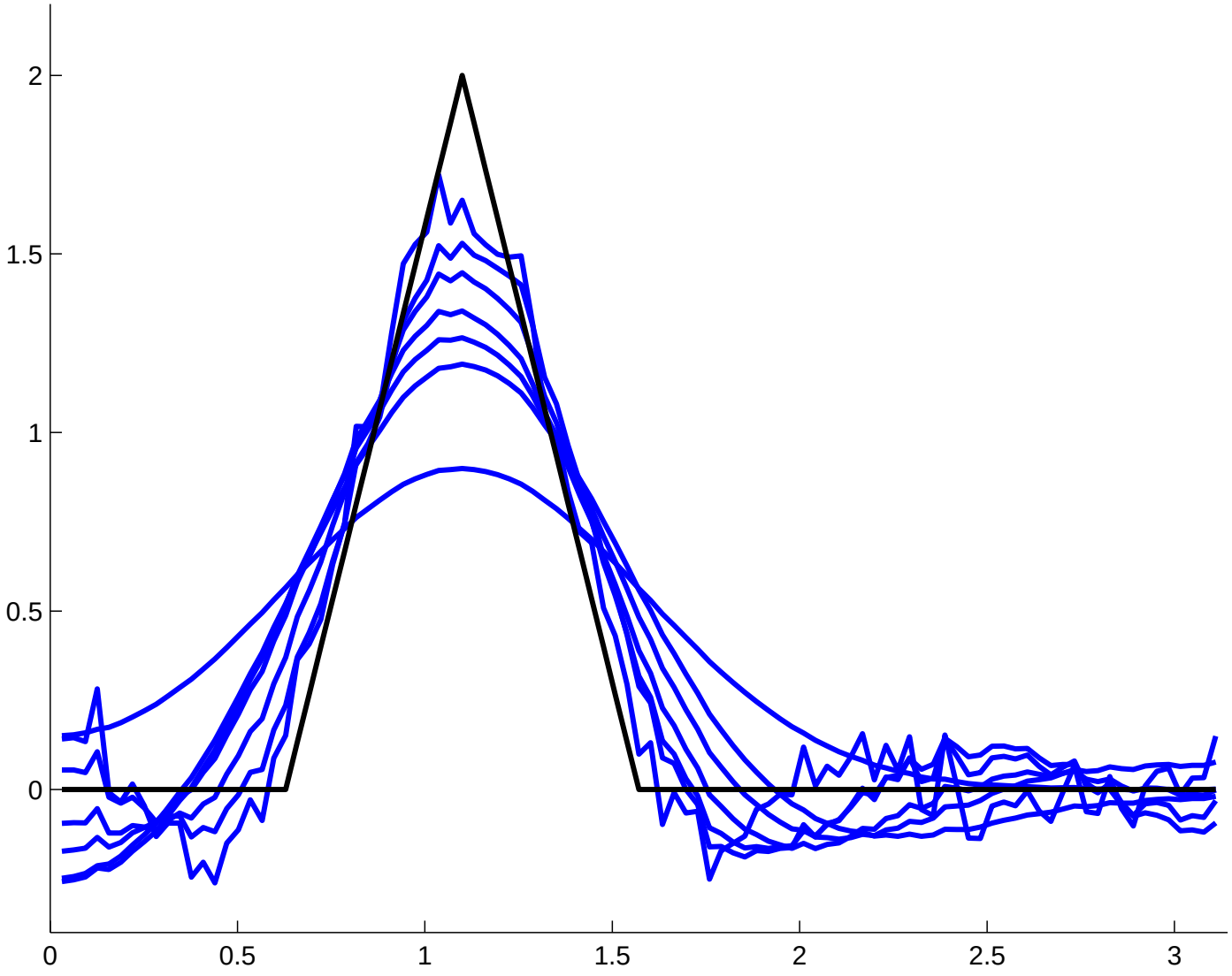
An example: Heat distribution in a rod (revisited)

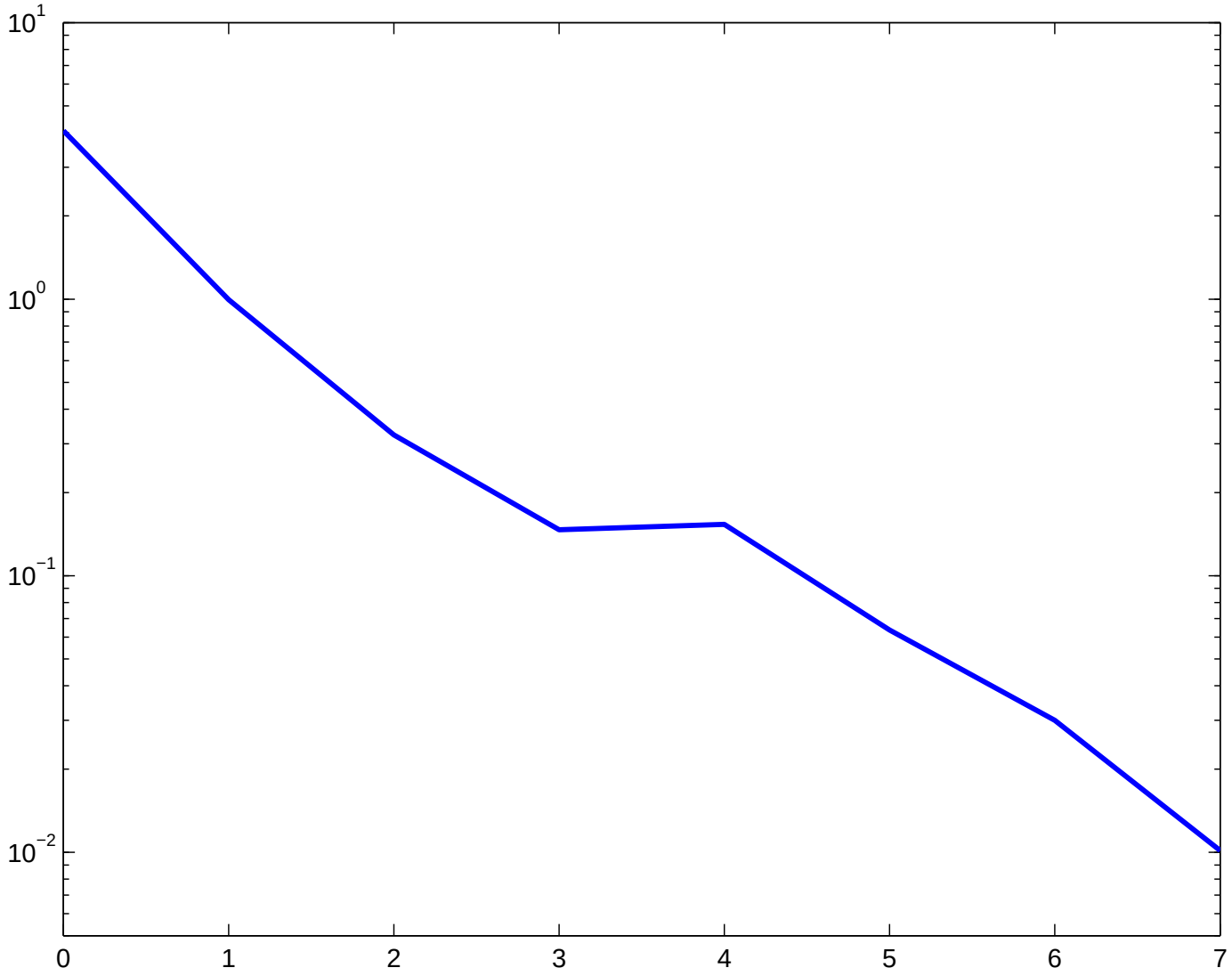
Let us once again consider the discretized inverse heat conduction problem in an insulated rod. We simulate the data in the exactly same way as above and add the same amount of noise.

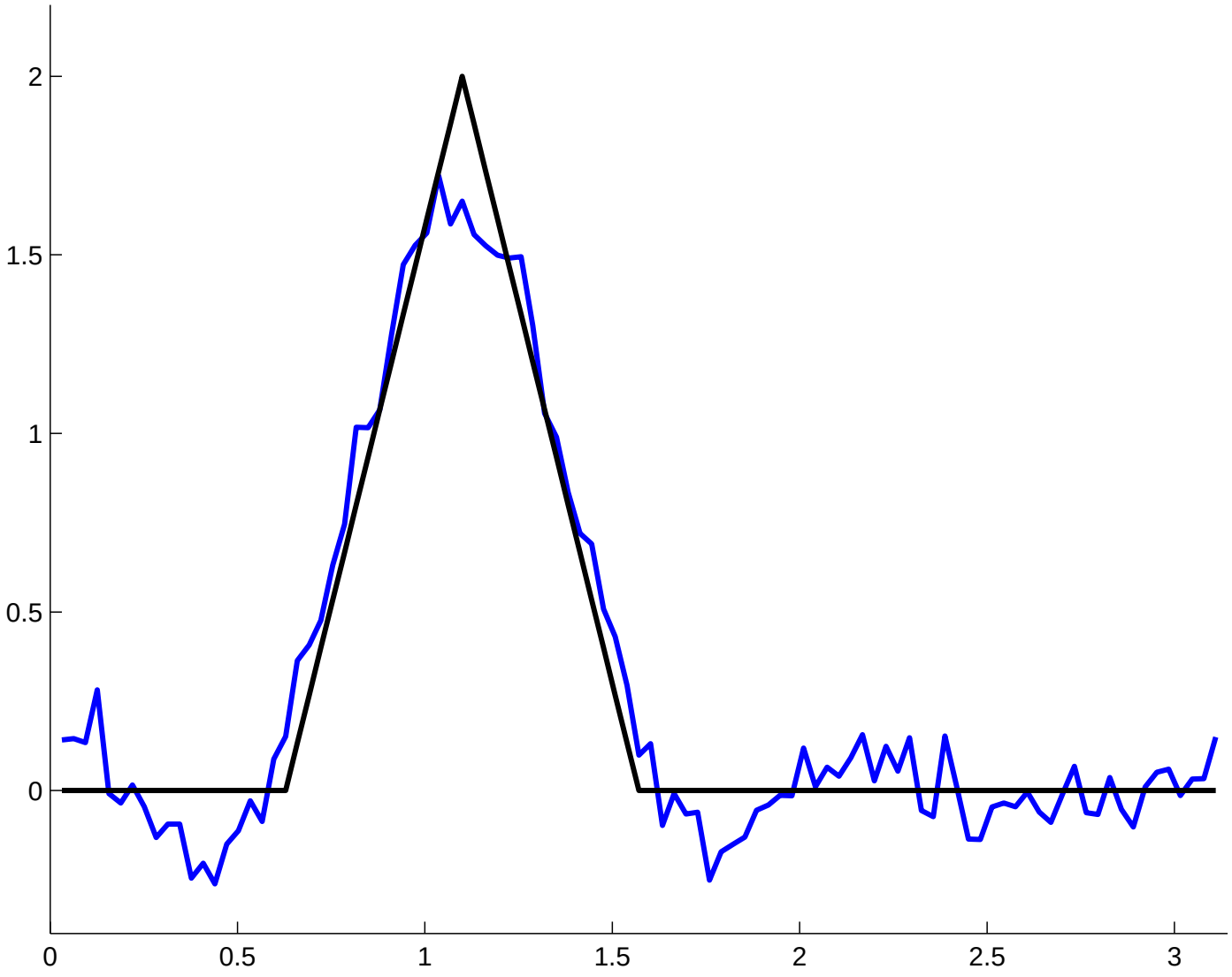
The system matrix $A = e^{TB}$, $T = 0.1$, is symmetric since B is symmetric. Moreover, the infinite-dimensional version of A, i.e., E_T, is positive definite, and thus it is not far-fetched to assume that A is, at least, close to being positive semi-definite. (A symmetric matrix is positive definite if and only if all of its eigenvalues are positive; according to Matlab the eigenvalues of A are either positive or extremely close to zero.) Hence, it seems reasonable to try applying the conjugate gradient method directly to the original equation.

If we use the same value $\epsilon = \sqrt{99 \cdot 0.001^2} = 9.95 \cdot 10^{-3}$ for the Morozov discrepancy principle as in the previous examples, the conjugate gradient method becomes unstable before the stopping rule is satisfied. However, for the value $1.2 \cdot \sqrt{99 \cdot 0.001^2}$ the stopping rule is satisfied after seven iterations.

In the following, we visualize the evolution of the conjugate gradient iteration, show the norm of the residual $\|y - Ax_k\|$ as a function of k, and plot the solution corresponding to the (fine-tuned) discrepancy principle.


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3


10^{1}
10^{0}
10^{-1}
10^{-2}
0
1
2
3
4
5
6
7

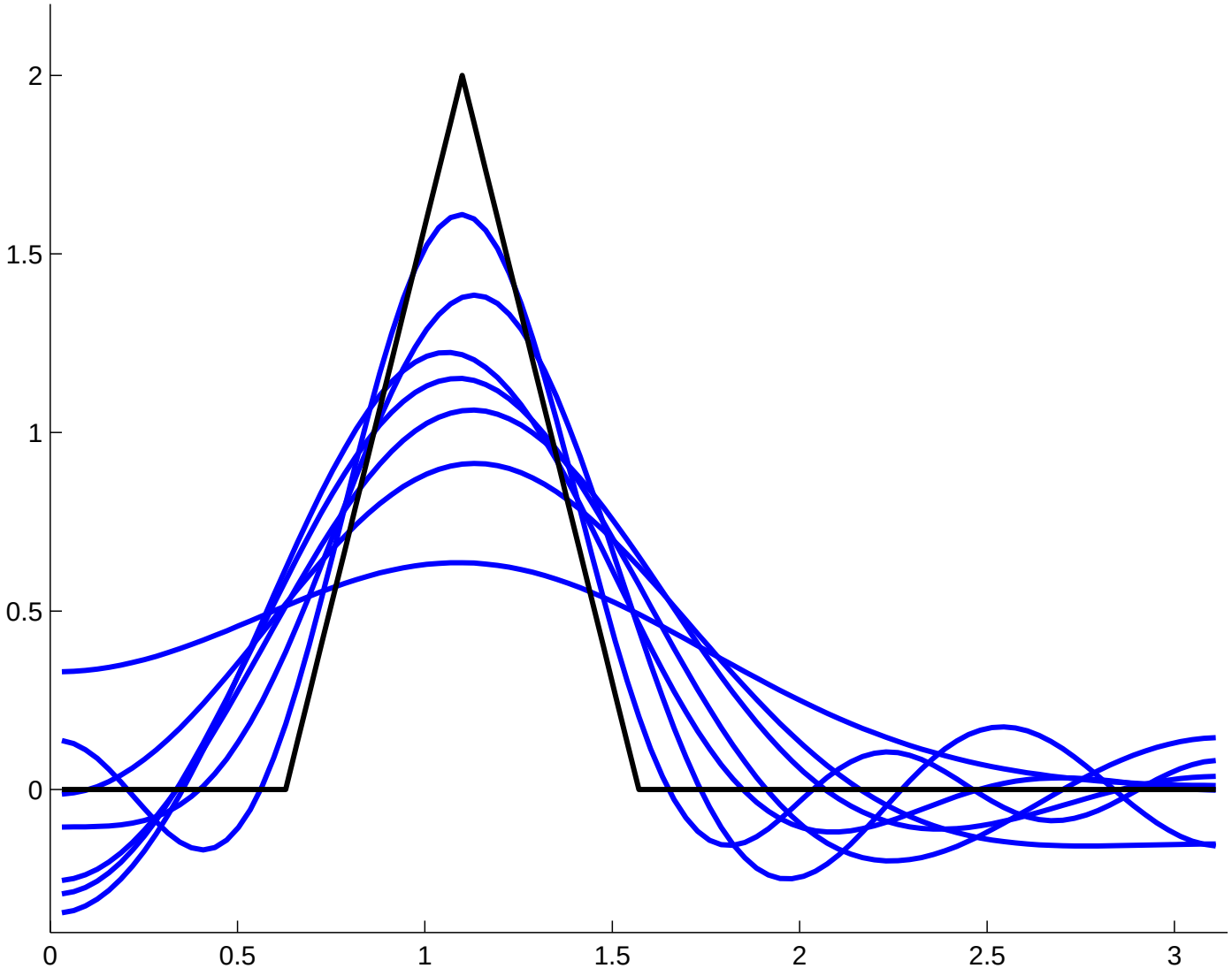

2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

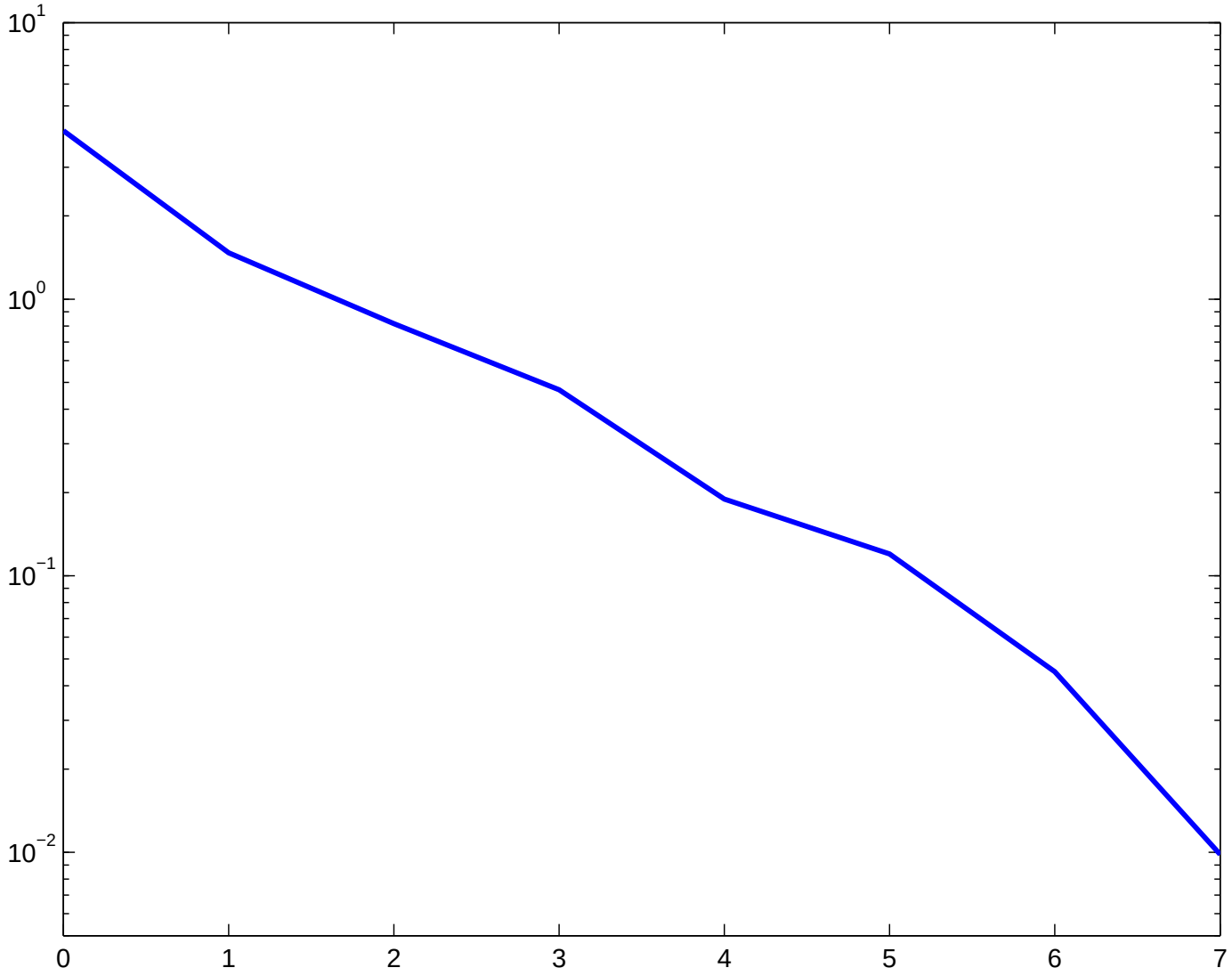
Next, we consider the exactly same problem, but this time apply the conjugate gradient method to the normal equation. As a stopping rule we use the Morozov discrepancy principle for the original equation, i.e., we stop the iteration when

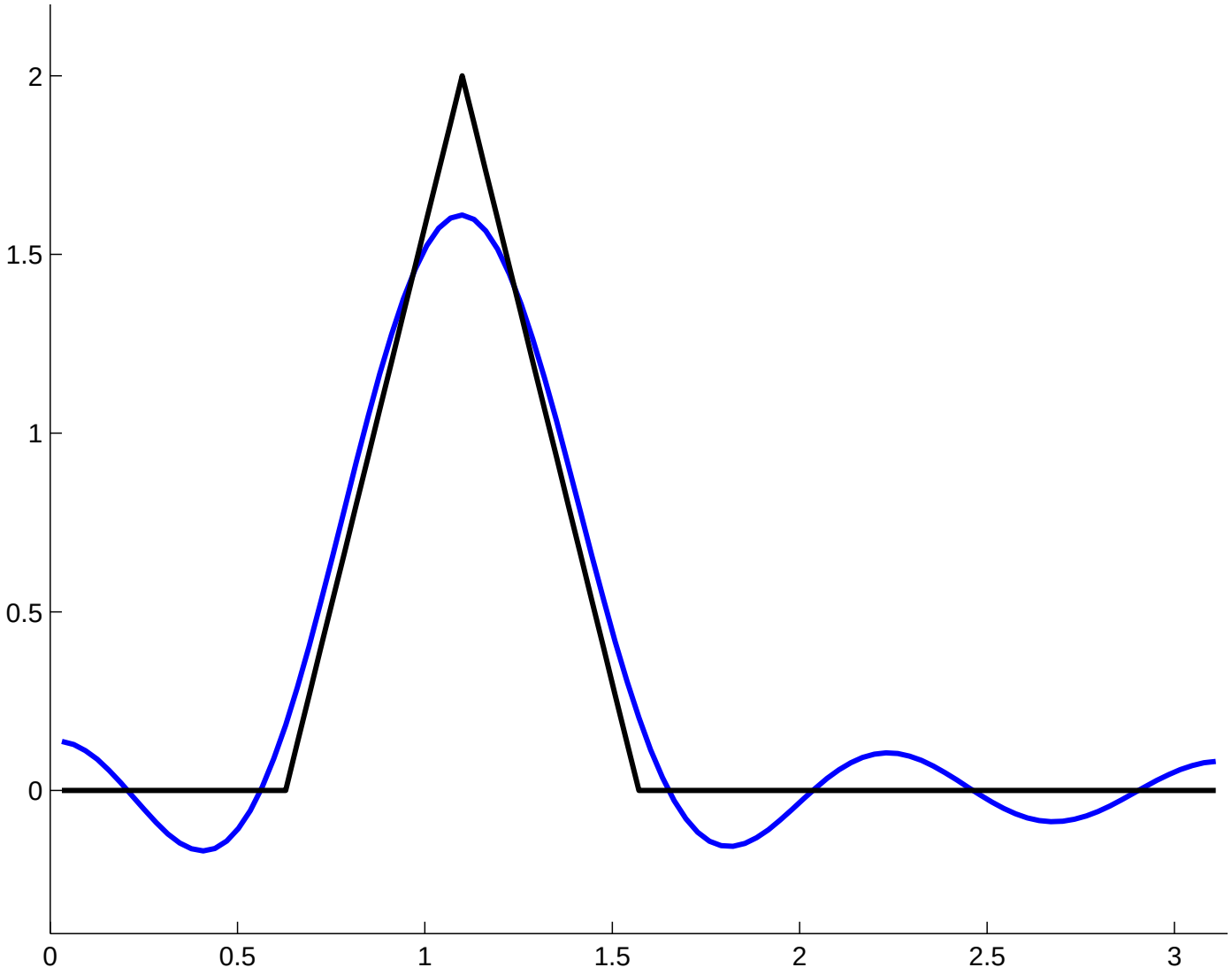
$$\|y - Ax_k\| \leq \epsilon,$$

where we use the 'standard' $\epsilon = \sqrt{99 \cdot 0.001^2} = 9.95 \cdot 10^{-3}$.

For some reason, the use of the normal equation makes the algorithm more stable: the discrepancy principle for this 'original' ϵ is satisfied after seven iterations and the solution looks nicer than when applying the algorithm directly to the original equation. (Bear in mind, however, that considering the normal equation makes the algorithm slower since more matrix-matrix or matrix-vector products need to be computed.)


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3


10^{1}
10^{0}
10^{-1}
10^{-2}
0
1
2
3
4
5
6
7


2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3

Computational methods in inverse problems

Nuutti Hyvönen, Matti Leinonen and Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Ninth lecture, February 16, 2011.

Recap

An example: Laplace transform

Let $f : [0, \infty) \to \mathbb{R}$ be some unknown function and assume that we have access to noisy samples of its Laplace transform

$$\mathcal{L}f(s) = \int_0^\infty e^{-st} f(t)\, dt, \qquad s \geq 0,$$

at some measurement points s_j, $j = 1, \ldots, m$. The task is to approximate f using the noisy values $\{\mathcal{L}f(s_j)\}_{j=1}^m$ as data.

Observe that for large t the kernel e^{-st} is typically very small, and hence the 'tail' of f does not affect the Laplace transform as much as its values close to the origin. In consequence, reconstructing f is an ill-posed inverse problem.

Discretization

In order to come up with a computational model, we approximate the integral of the Laplace transform as

$$\mathcal{L}f(s_j) \approx \int_0^T e^{-s_j t} f(t)\, dt \approx \sum_{k=1}^{n} w_k e^{-s_j t_k} f(t_k), \qquad j = 1, \ldots, m,$$

where $t_1, \ldots, t_n \in [0, T]$ are the nodes and $w = (w_1, \ldots, w_n)^{\mathrm{T}} \in \mathbb{R}^n$ the corresponding weights of the chosen quadrature rule. Notice that it is implicitly assumed that $e^{-st} f(t)$ is 'small' for all t that are larger than the threshold $T > 0$.

For example, if we decided to use the trapezoid rule on an equidistant mesh in the interval $[0, T]$, we would choose $h = T/(n-1)$ and

$$w = (h/2, h, h, \ldots, h, h, h/2)^{\mathrm{T}} \qquad \text{and} \qquad t_k = (k-1)h$$

for $k = 1, \ldots, n$.

The above quadrature rule can be written in the matrix form

$$y = Ax,$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are given by

$$\begin{aligned} x &= (f(t_1), \ldots, f(t_n))^{\mathrm{T}} \\ y &= (\mathcal{L}f(s_1), \ldots, \mathcal{L}f(s_m))^{\mathrm{T}}, \end{aligned}$$

and the elements of the matrix $A \in \mathbb{R}^{m \times n}$ are defined as

$$(A)_{jk} = w_k e^{-s_j t_k}, \qquad j = 1, \ldots, m, \quad k = 1, \ldots, n.$$

In the following numerical examples, we choose $m = 91$ sampling points on a logarithmic grid:

$$\log s_j \;=\; -\log 10 + 2\frac{(j-1)}{m-1}\log 10, \qquad j = 1, \ldots, m,$$

where $\log$ denotes the natural logarithm. Now, the points $\{\log s_j\}_{j=1}^m$ form a uniform grid in the interval $[-\log(10), \log(10)]$, and thus $\{s_j\}_{j=1}^m$ lie in the interval $[0.1, 10]$, with half of the points between 0.1 and 1. This reflects our knowledge that the information in the Laplace transform is — *very loosely speaking* — concentrated close to the origin.

We set $n = 101$ and choose the nodes $\{t_k\}_{k=1}^n$ and the weights $w \in \mathbb{R}^n$ according to the Gauss–Legendre quadrature rule in the interval $[0, 5]$. (One could use something less sophisticated, such as trapezoid rule in this same interval, as well.)

Simulation of data

We choose

$$f(t) = \begin{cases} t^3 - 4t^2 + 4t, & 0 \le t < 2, \\ 0, & t \ge 2. \end{cases}$$

In this simple case, the Laplace transform can be calculated explicitly with the help of partial integration:

$$\mathcal{L}f(s) = \frac{4}{s^2} - \frac{4}{s^3}(2 + e^{-2s}) + \frac{6}{s^4}(1 - e^{-2s}), \qquad s > 0.$$

Consequently, we just compute the value of $\mathcal{L}f(s)$ at the chosen sampling points $\{s_j\}_{j=1}^m$ using this formula, add realizations of a normally distributed random variable with zero mean and standard deviation 10^{-3} to each sample, plug the resulting data into the vector y, and we are ready to go.

On inverse crimes

The most obvious form of *inverse crime* is to use the exactly same numerical model to simulate the data and to carry out the inversion. Such a procedure results typically in overly optimistic reconstructions.

Here, this form of inverse crime is avoided because the data is simulated using an analytic formula and the reconstruction process is based on a quadrature rule. However, if the explicit form of $\mathcal{L}f$ was not known, we could operate as follows:

1. Choose two sets of node and sampling points $\{\tilde{s}_j\}_{j=1}^{m_0}$ and $\{\tilde{t}_k\}_{k=1}^{n_0}$, and $\{s_j\}_{j=1}^{m}$ and $\{t_k\}_{k=1}^{n}$.
2. Use the first sets of points, $\{\tilde{s}_j\}_{j=1}^{m_0}$ and $\{\tilde{t}_k\}_{k=1}^{n_0}$, and the corresponding 'quadrature matrix' $A = A_0$ to compute $\mathcal{L}f$ at the points $\{\tilde{s}_j\}_{j=1}^{m_0}$.
3. Use interpolation to approximate the value of $\mathcal{L}f$ at the (typically sparser) set of sampling points $\{s_j\}_{j=1}^{m}$, and add noise. (See `interp1` and `interp2` in Matlab.)
4. Test your inversion method by using the hereby obtained noisy versions of $\{\mathcal{L}f(s_j)\}_{j=1}^{m}$ as data and the 'quadrature matrix' corresponding to the sets of points $\{s_j\}_{j=1}^{m}$ and $\{t_k\}_{k=1}^{n}$ as the system matrix A.

Notice that in 'real life' these kinds of problems do not occur because you do not simulate the data yourself.

Numerical experiments

In the following, we will apply the considered inversion methods to the above introduced discretized "inverse Laplace transform problem":

$$Ax = y.$$

If not stated otherwise, we utilize the Morozov discrepancy principle with

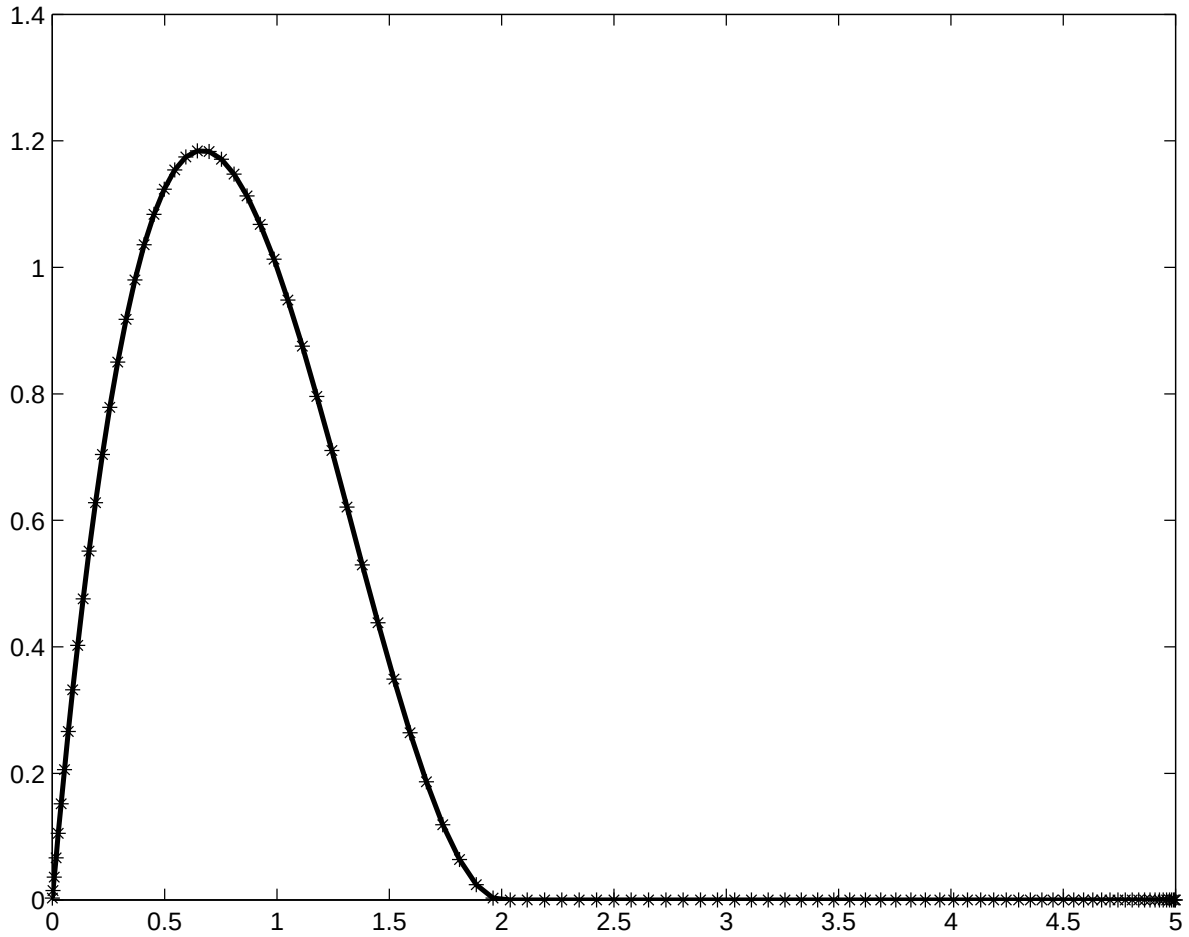
$$\epsilon = 10^{-3}\sqrt{m} \approx 9.5 \cdot 10^{-3}$$

as the stopping rule, i.e., we terminate the iterations, or pick a spectral cut-off index, or choose a regularization parameter so that the approximate solution $\tilde{x}$ satisfies

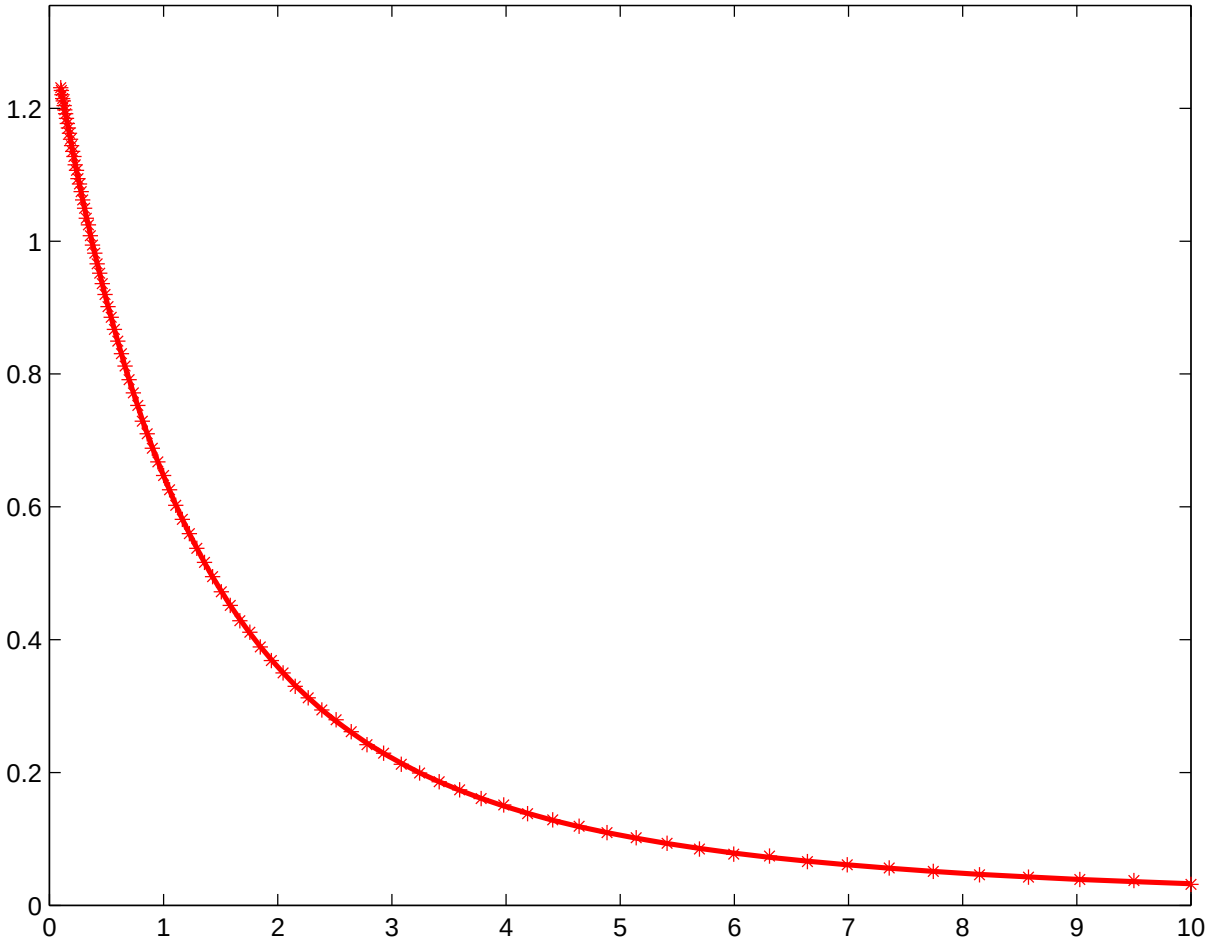
$$\|y - A\tilde{x}\| \simeq \epsilon.$$

For the exact implementation of the Morozov stopping criterion for different algorithms, see the material of the previous lectures.

Target function and the nodes



Laplace transform and the noisy measurements



Truncated singular value decomposition

The singular value decomposition of A is

$$A = U\Lambda V^{\mathrm{T}},$$

where $\Lambda \in \mathbb{R}^{m\times n}$ has the (non-negative) singular values on its diagonal, and the columns of $V \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ are composed of the (extended) orthonormal basis $\{v_j\}_{j=1}^{n}$ and $\{u_j\}_{j=1}^{m}$, respectively.

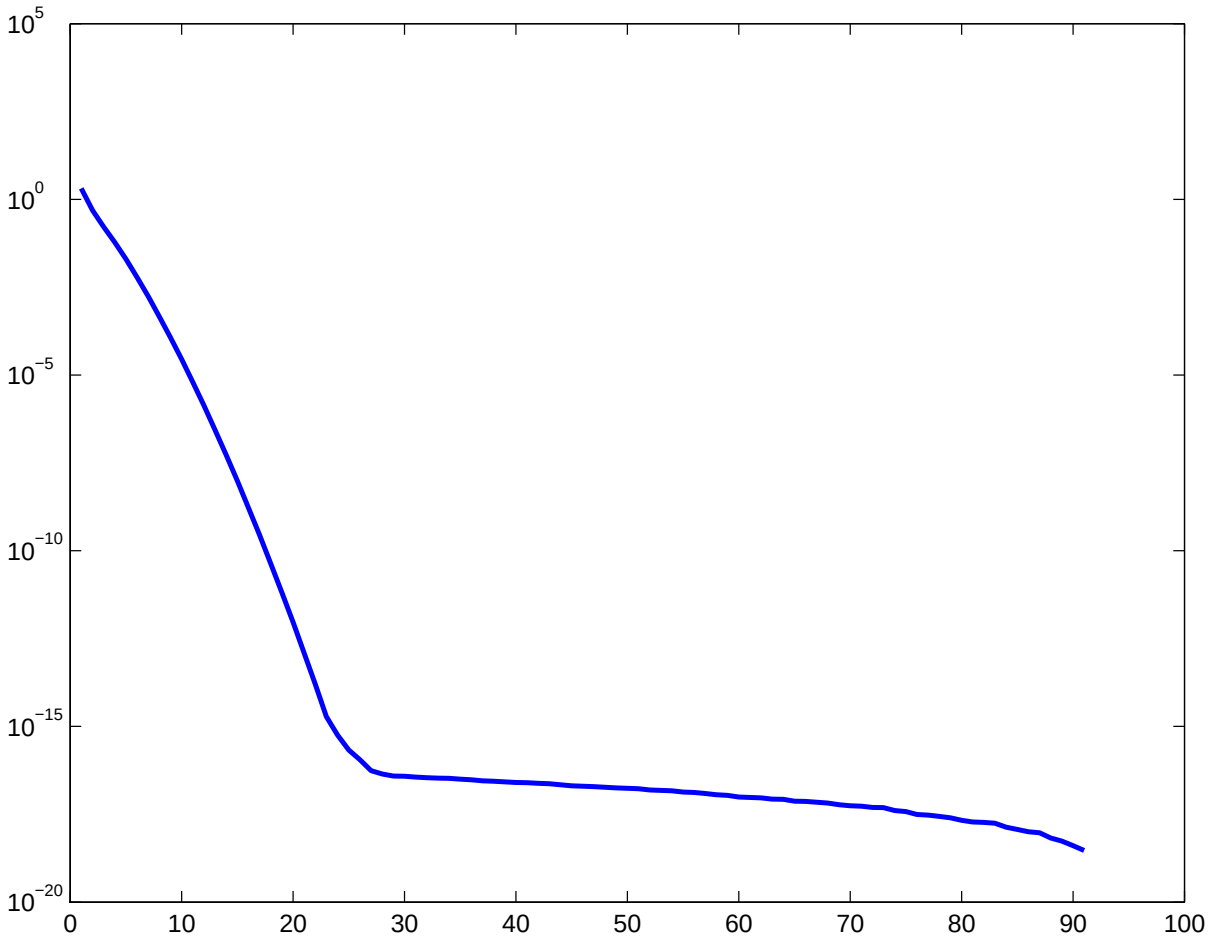
The truncated SVD solution for $1 \leq k \leq \mathrm{rank}(A)$ is given by

$$x_k = V\Lambda_k^{\dagger}U^{\mathrm{T}}y$$

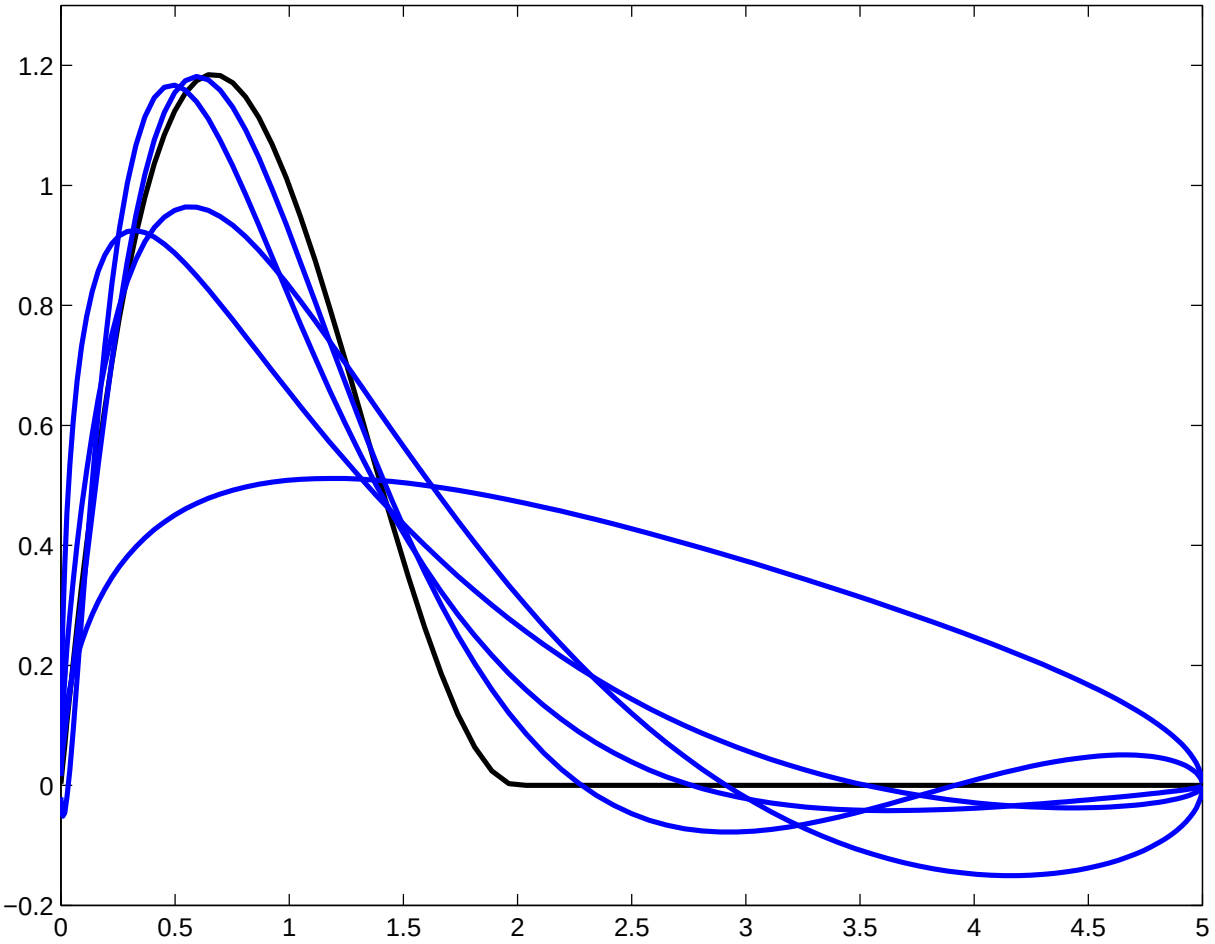
where $\Lambda_k^{\dagger} \in \mathbb{R}^{n\times m}$ has the elements $1/\lambda_1, \ldots, 1/\lambda_k, 0, \ldots, 0$ on its diagonal. (The singular values of our A are plotted on the next slide.)

In the following, we show the evolution of x_k as a function of k, present the Morozov discrepancy principle solution and, for comparison, present the truncated SVD solution for no noise and $k = 21 = \mathtt{rank(A)} - 1$.

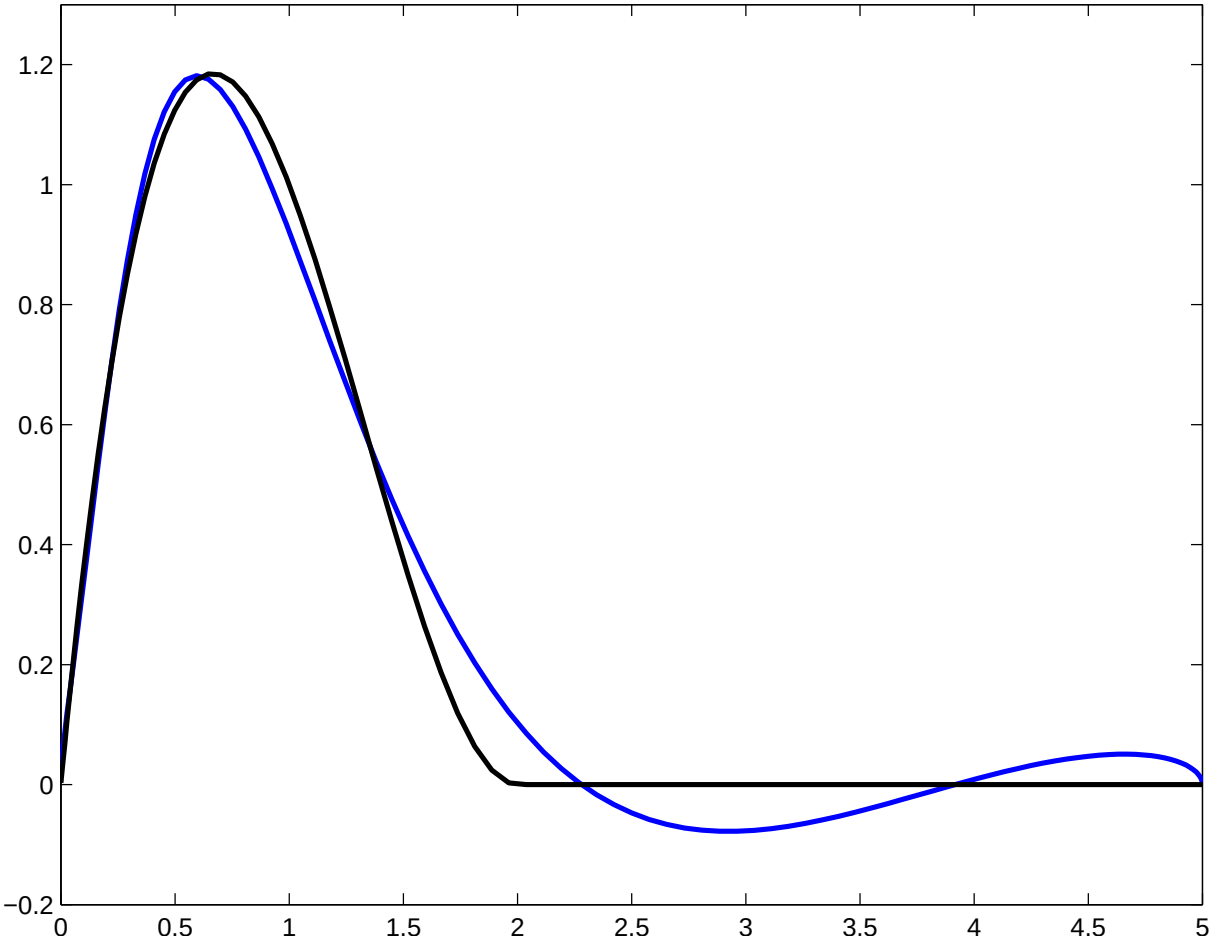
Singular values of A



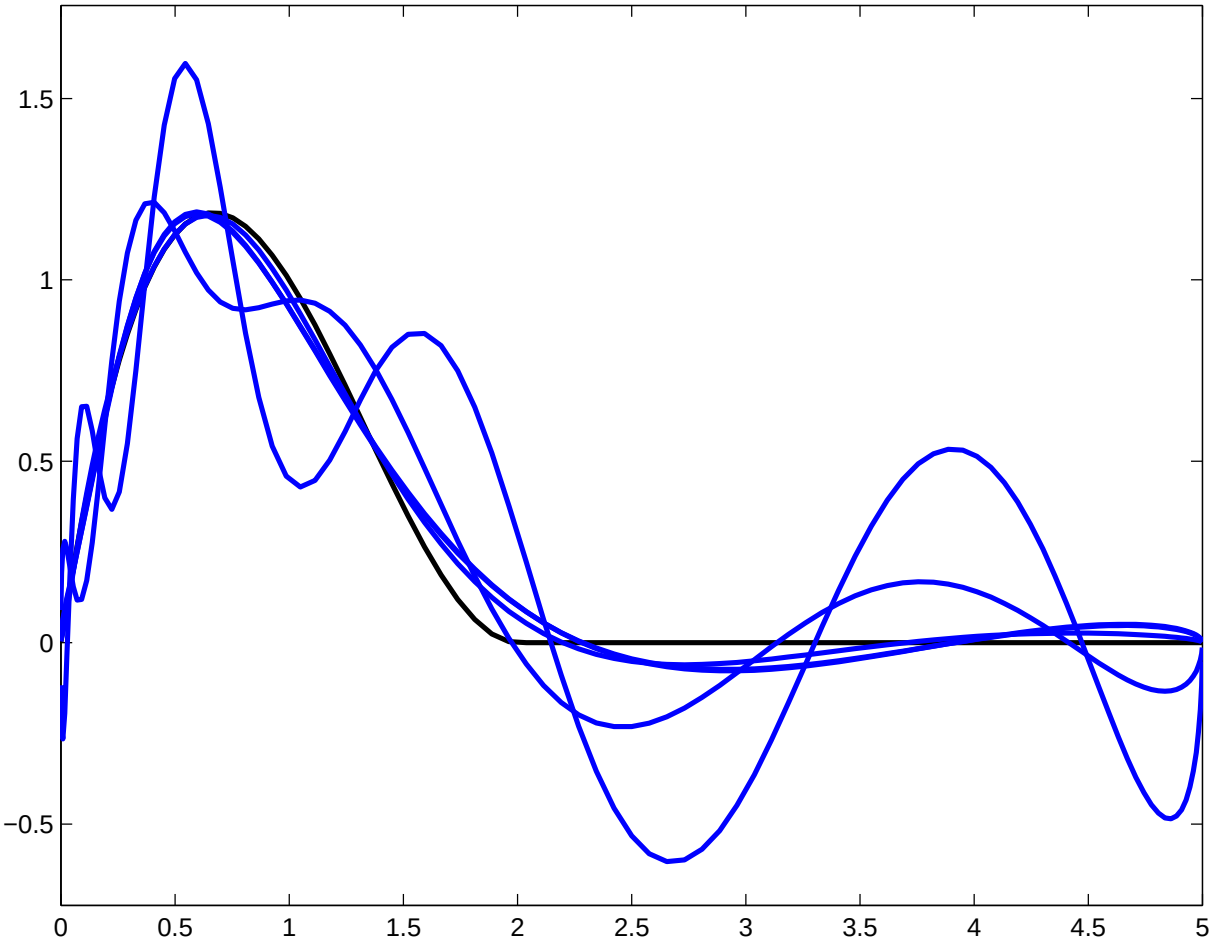
Truncated SVD solutions for $k = 1, \ldots, 5$



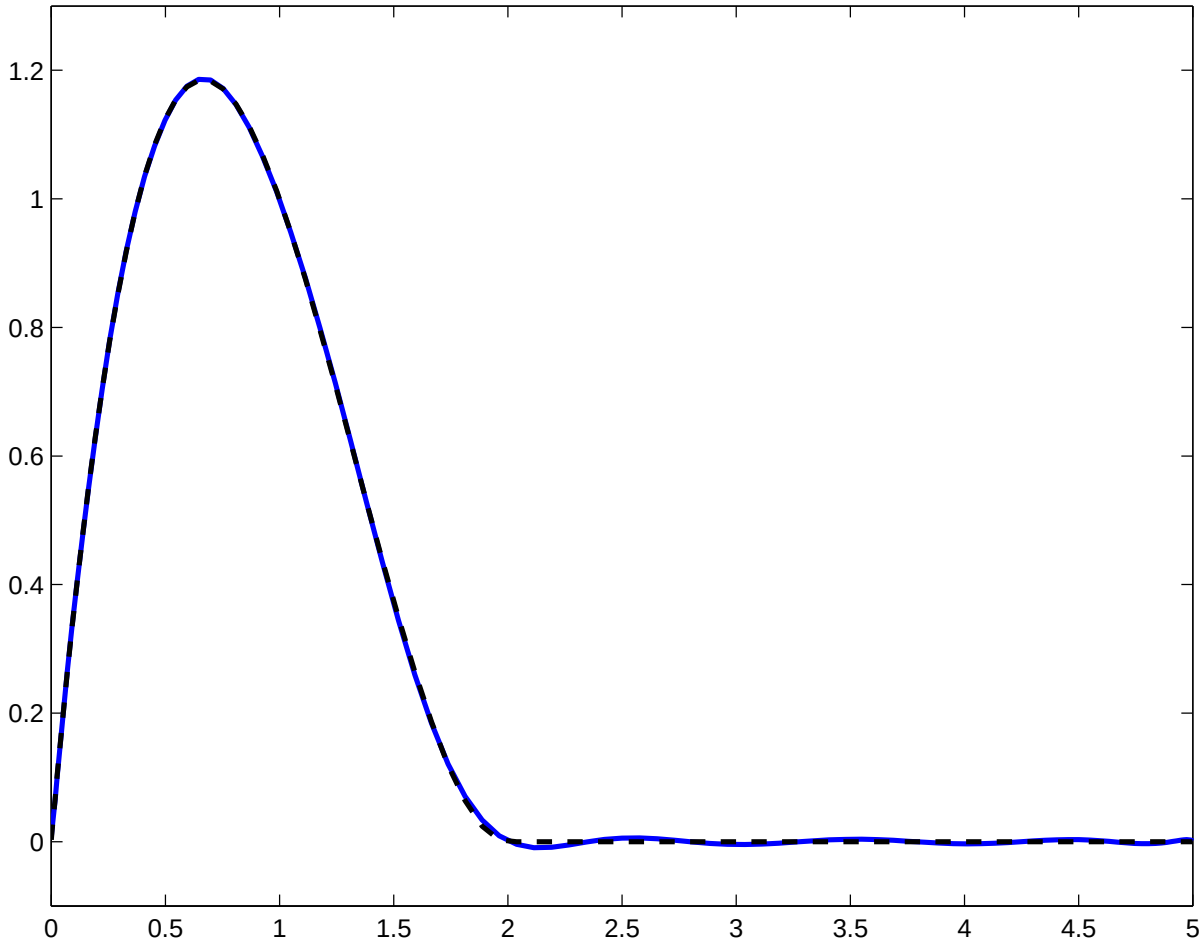
Morozov discrepancy solution ($k = 5$)



Truncated SVD solutions for $k = 5, \ldots, 8$



Truncated SVD solutions for $k = 21$ and no noise



Tikhonov regularized solution

The Tikhonov regularized solution $x_\delta \in \mathbb{R}^n$ is the unique minimizer of the Tikhonov functional

$$\|Ax - y\|^2 + \delta\|x\|^2, \qquad \delta > 0.$$

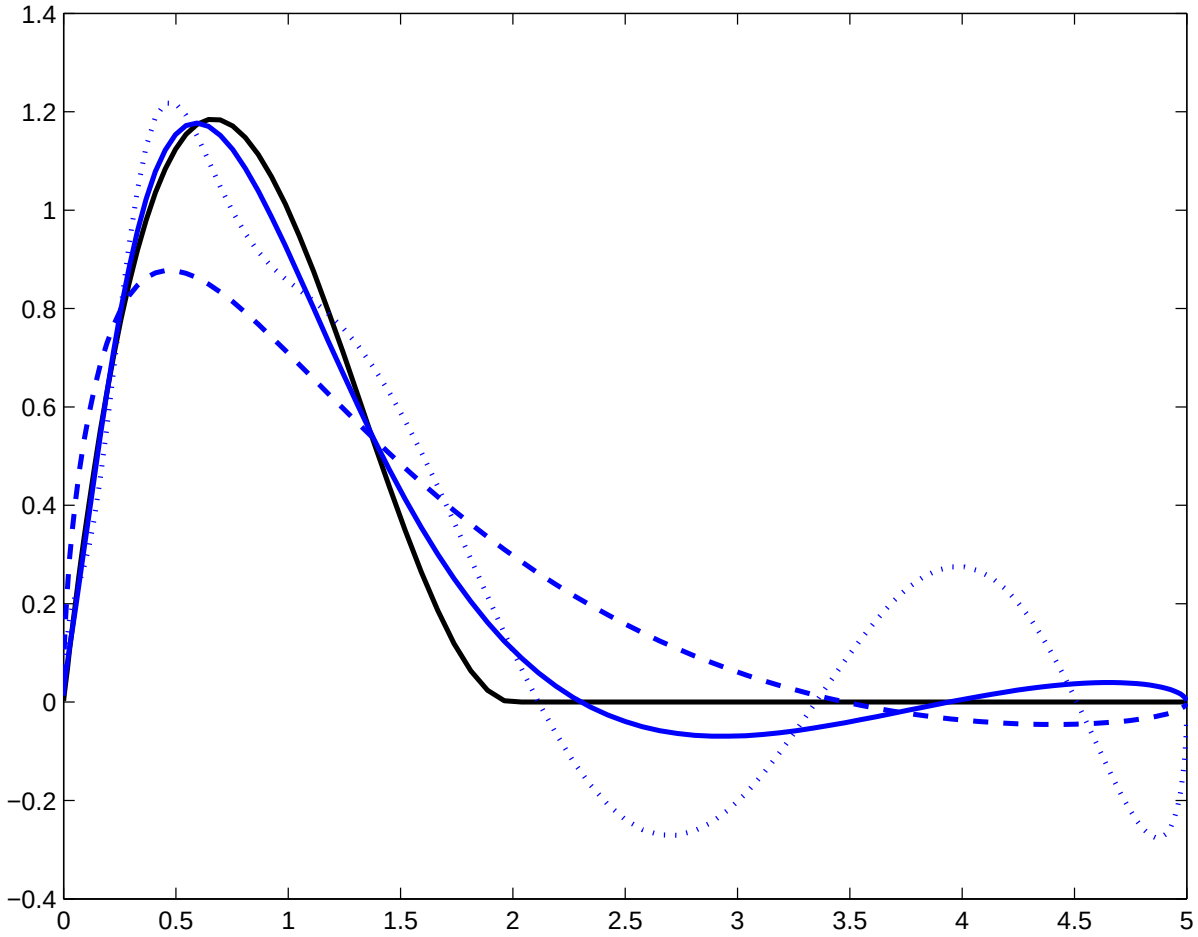
It is given explicitly by the formula

$$x_\delta = (A^{\mathrm{T}}A + \delta I)^{-1}A^{\mathrm{T}}y. \tag{12}$$

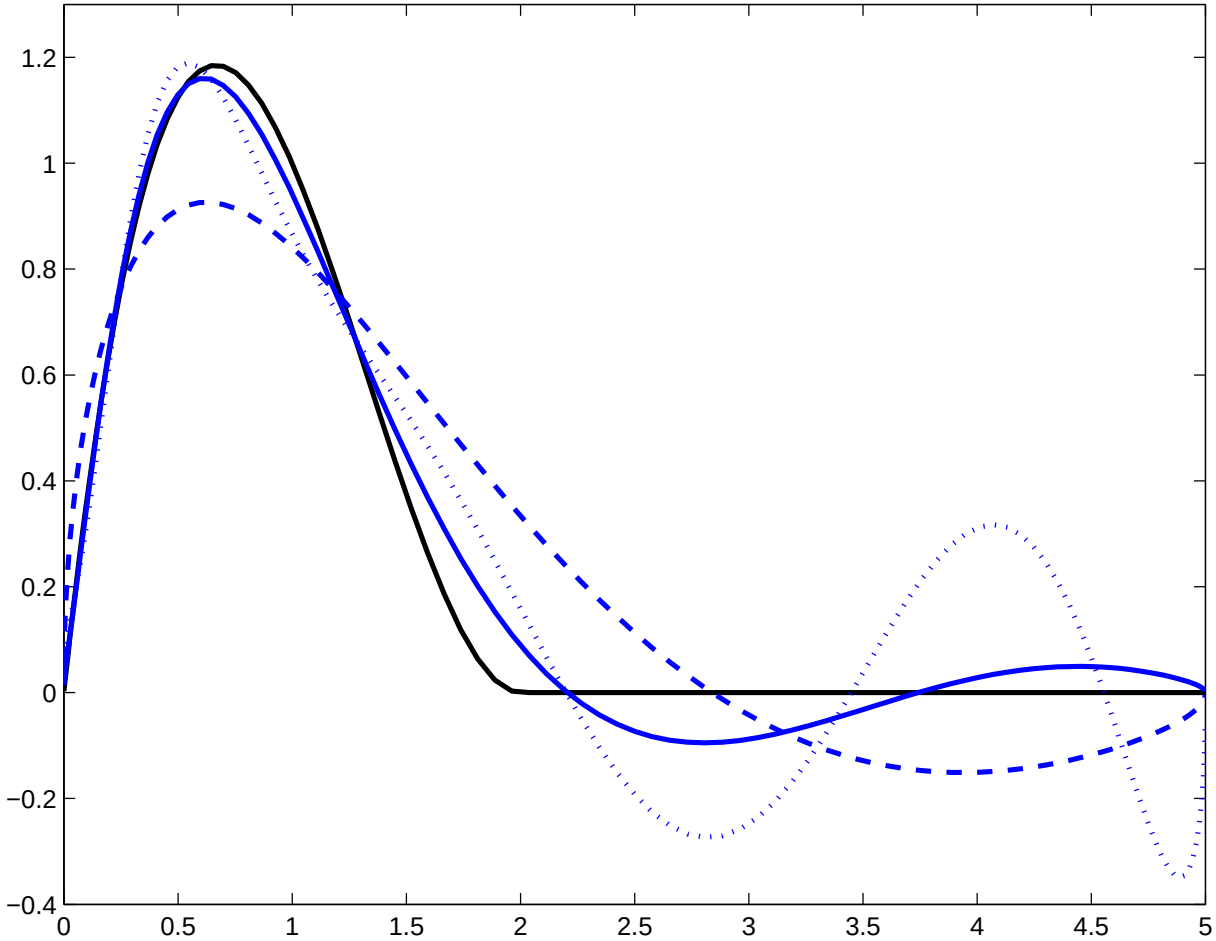
If one replaces x in the penalty term of the Tikhonov functional by Lx, for some $L \in \mathbb{R}^{l\times n}$, then the identity operator in (335) is replaced by $L^{\mathrm{T}}L$ — at least formally.

In the following, we first use traditional Tikhonov regularization, and then plug Lx in the penalty term, with $L \in \mathbb{R}^{n\times n}$ being a difference matrix that approximates the second spatial derivative on the interval $[0, 5]$.

Traditional Tikhonov with $\delta = \delta_{\mathrm{Mor}} \approx 3.6 \cdot 10^{-5}$ (solid), $\delta = 10^{3}\delta_{\mathrm{Mor}}$ (slashed) and $\delta = 10^{-3}\delta_{\mathrm{Mor}}$ (dotted)



Smoothness Tikhonov with $\delta_{\mathrm{Mor}} \approx 3.8 \cdot 10^{-10}$ (solid), $\delta = 10^{3}\delta_{\mathrm{Mor}}$ (slashed) and $\delta = 10^{-3}\delta_{\mathrm{Mor}}$ (dotted)



Landweber–Fridman iteration

The Landweber–Fridman iteration produces a sequence of approximate solutions $\{x_k\}_{k=0}^{\infty}$ according to the recursion rule

$$x_{k+1} = T(x_k), \qquad x_0 = 0,$$

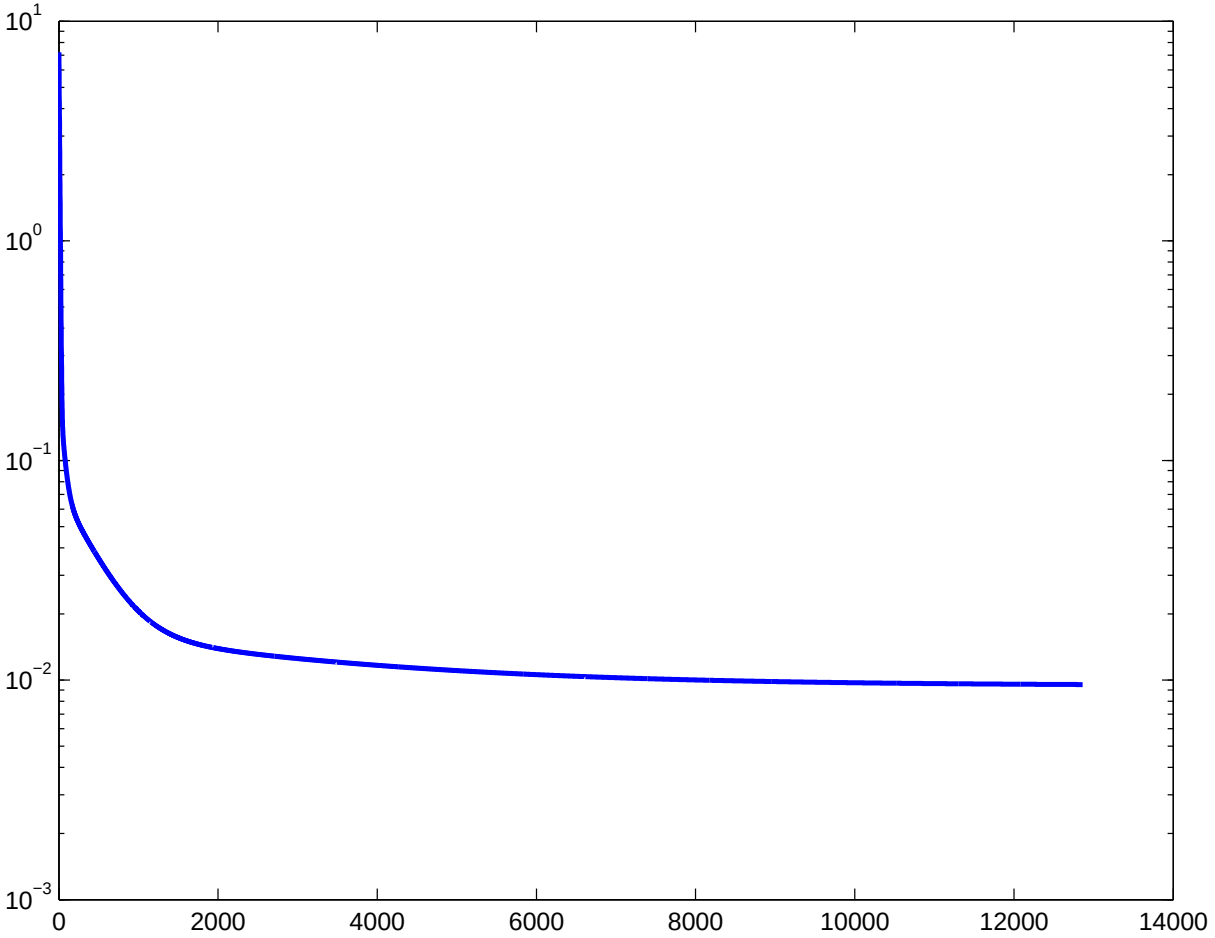
where

$$T(x) = x + \beta(A^{\mathrm{T}}y - A^{\mathrm{T}}Ax), \qquad \beta \in \mathbb{R}.$$

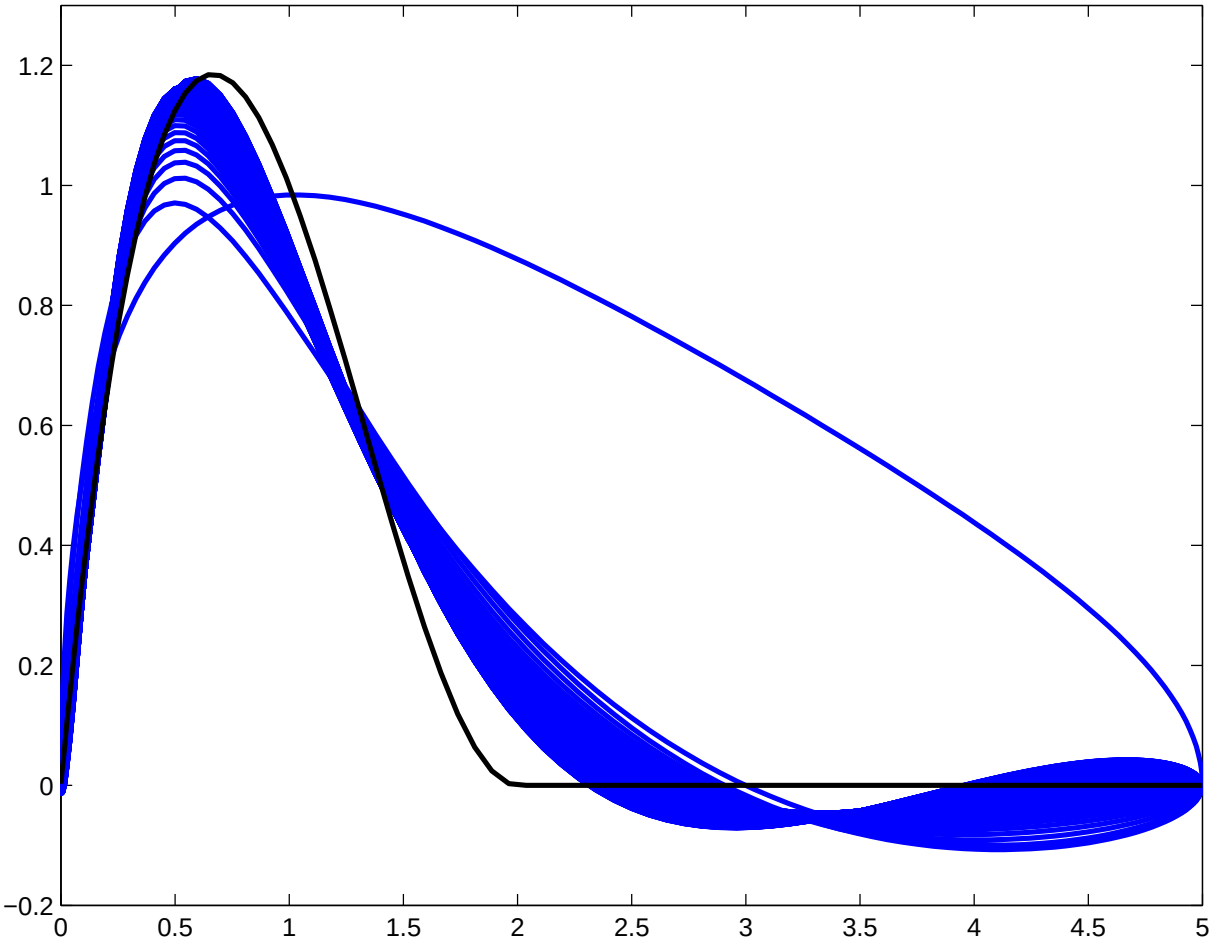
In order to achieve convergence, the free parameter β should be chosen from the interval $(0, 2/\lambda_1^2)$, where λ_1 is the largest singular value of A, i.e., the matrix norm of A. The larger the value of β in this interval, the faster the convergence. Here, $\|A\| \approx 2.05$ and we choose $\beta = 0.45$.

In the following, we visualize the evolution of the Landweber–Fridman sequence and show the solution corresponding to the Morozov discrepancy principle. (Note that the convergence is really slow; there is no real possibility for fitting the solution to noise.)

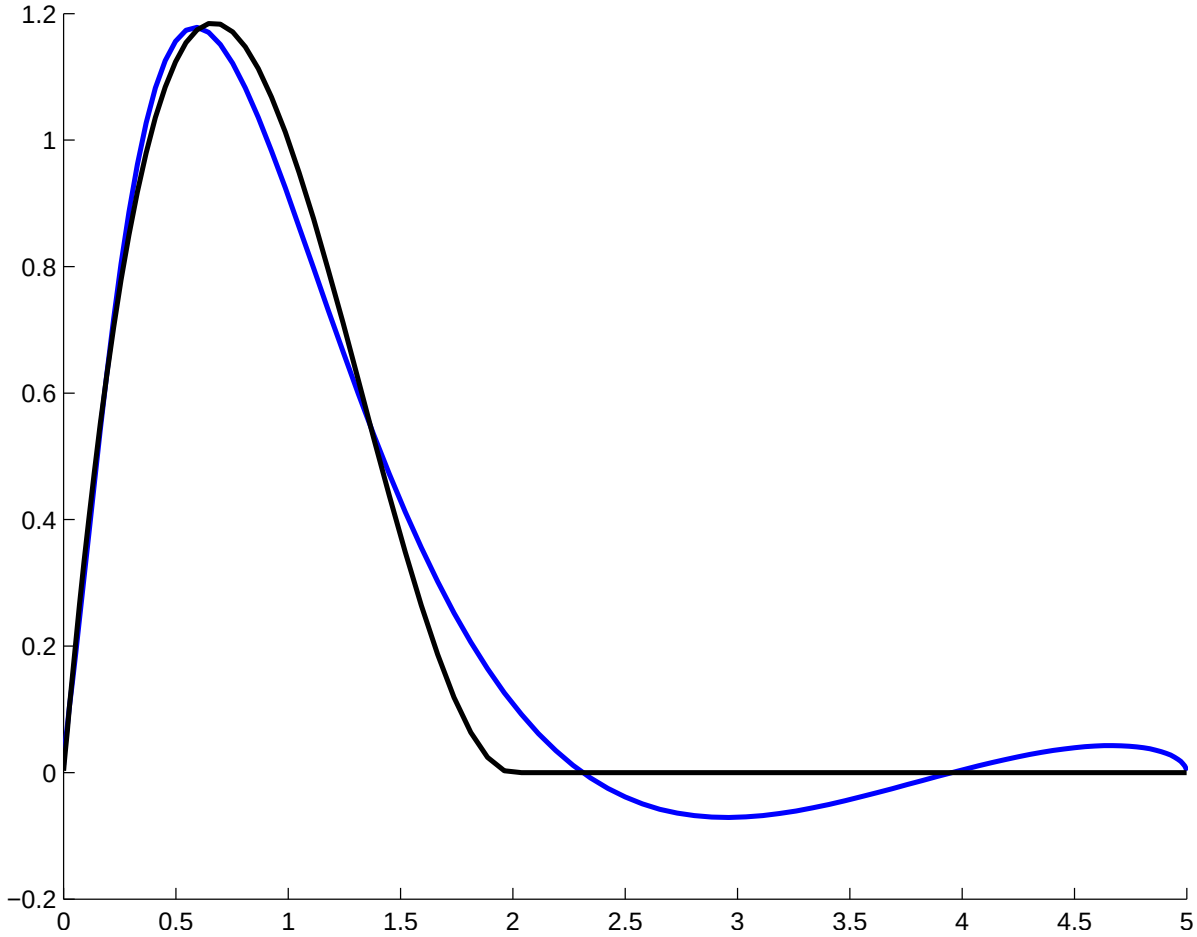
Residual $\|y - Ax_k\|$ as a function of k



Approximate solutions x_k, $k = 1, 101, 201, \ldots$



Morozov discrepancy solution ($k = 12\,861$)



Kaczmarz iteration (ART)

The most basic form of Kaczmarz iteration is to take zero as the initial guess and then iterate by projecting recursively onto the hyperplanes defined by the rows of the considered matrix equation. If $a_j^{\mathrm{T}} \in \mathbb{R}^{1\times n}$ denotes the jth row of the matrix A, then this algorithm is as follows:

Set $k = 0$ and $x_0 = 0$;

Repeat until the chosen stopping rule is satisfied:

$z_0 = x_k$;

for $j = 1, \ldots, m$

$z_j = z_{j-1} + (1/\|a_j\|^2)(y_j - a_j^{\mathrm{T}} z_{j-1})a_j$;

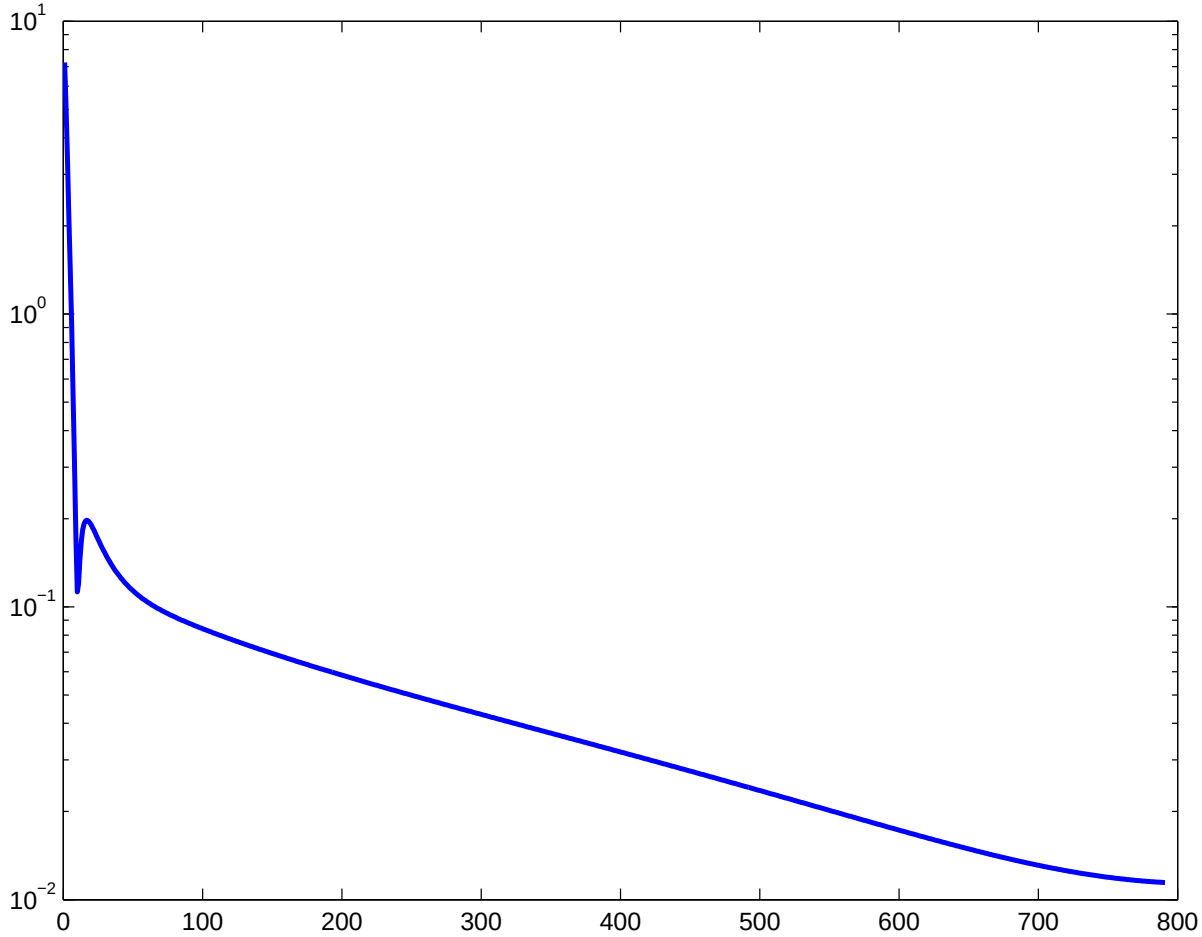
end

$x_{k+1} = z_m$; $\quad k \leftarrow k + 1$;

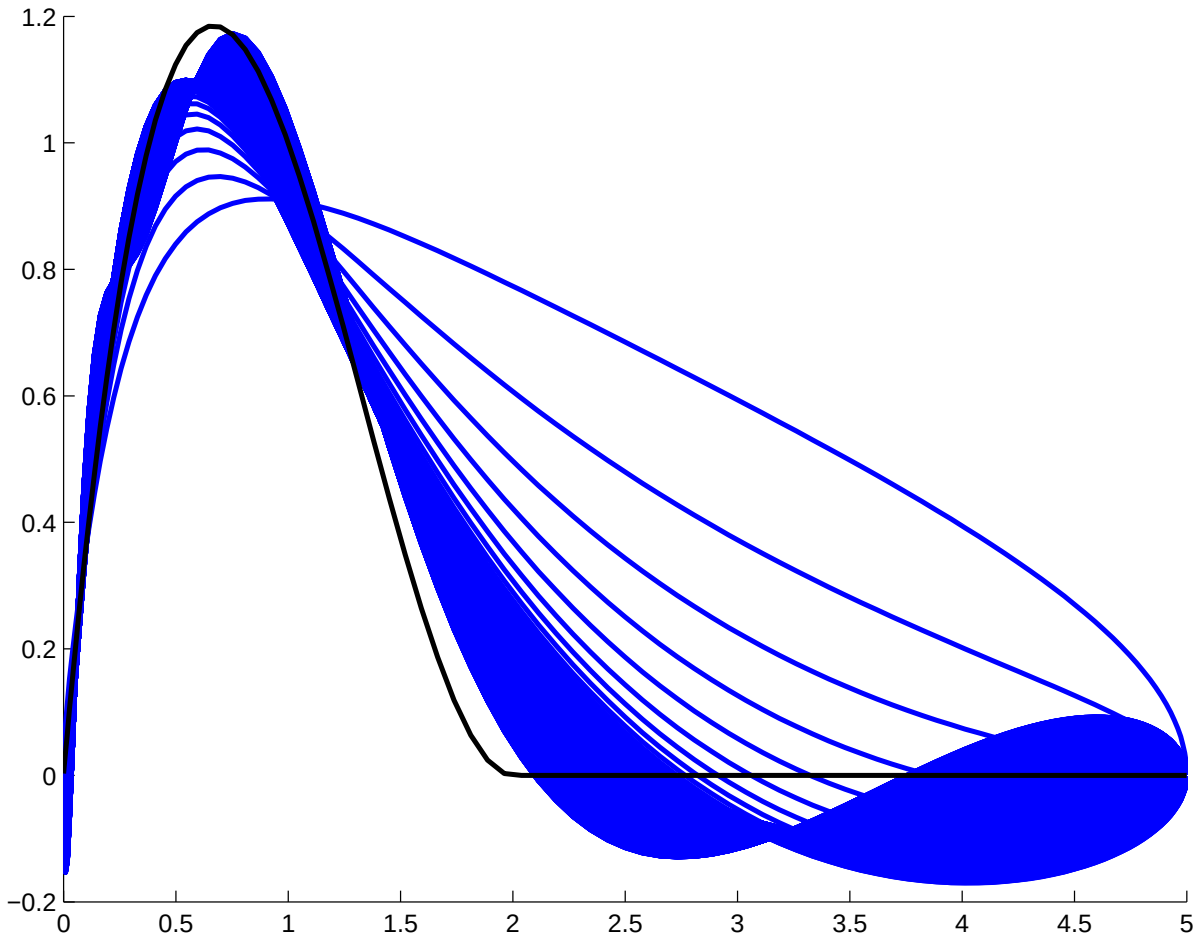
end

In the considered case, ART does not seem to converge for the original ϵ, and thus we use the discrepancy principle with 1.2ϵ here.

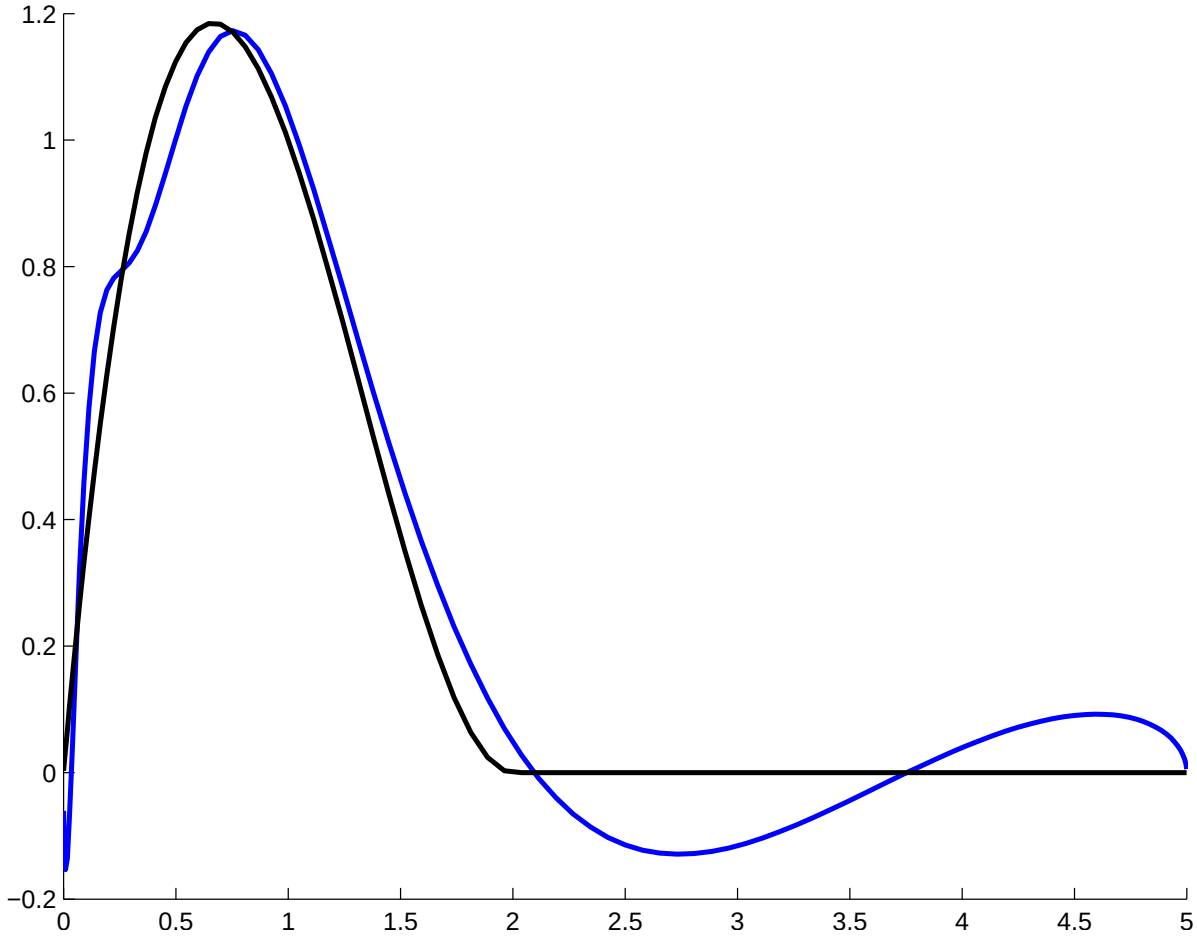
Residual $\|y - Ax_k\|$ as a function of k



Approximate solutions x_k, $k = 1, \ldots, 790$



Morozov discrepancy solution ($k = 790$)



Conjugate gradient method

With conjugate gradient method one is forced to consider the normal equation

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}y.$$

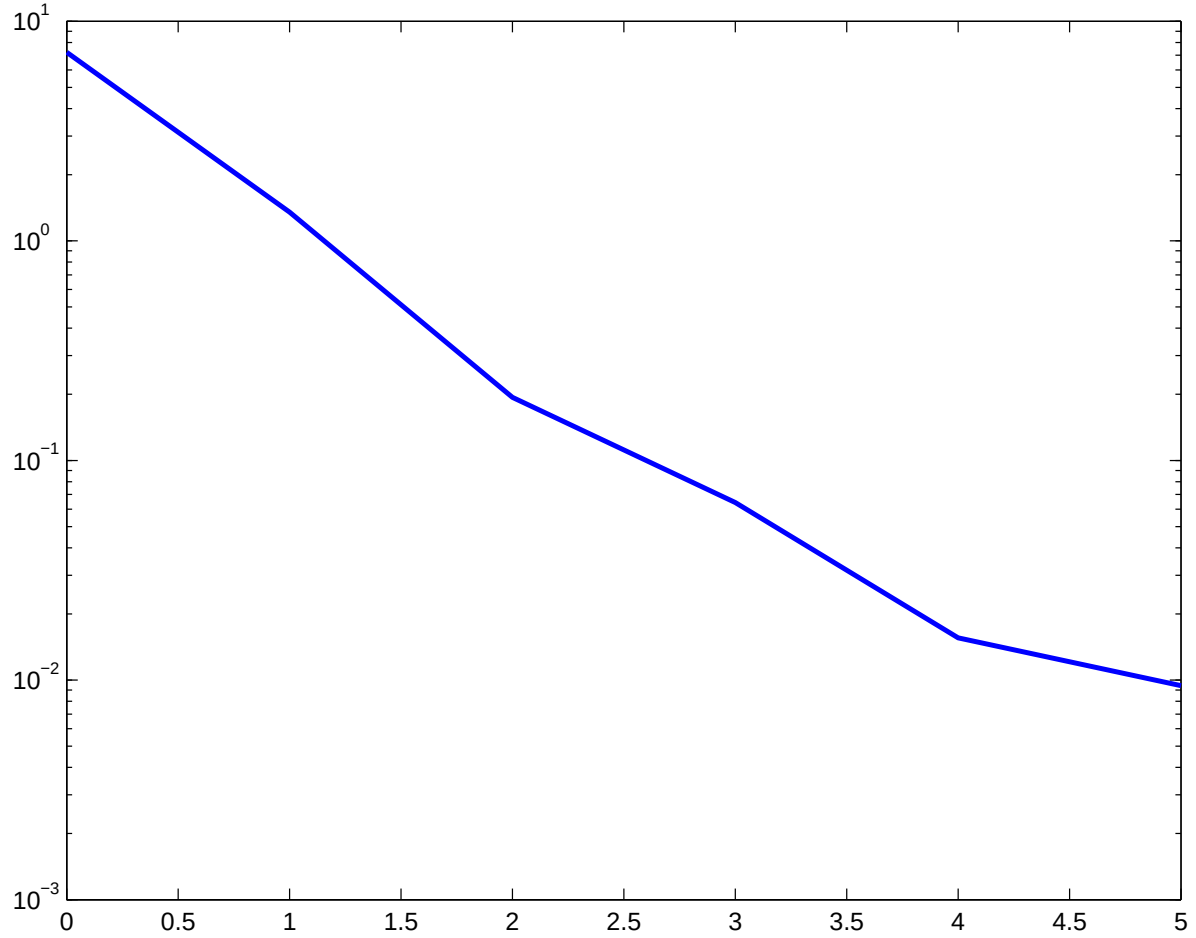
In this case, the algorithm can be written, e.g., as follows (here $x_0 = 0$):

Choose x_0. Set $k = 0$, $r_0 = A^{\mathrm{T}}y - A^{\mathrm{T}}(Ax_0)$, $s_0 = r_0$;
Repeat until the chosen stopping rule is satisfied:

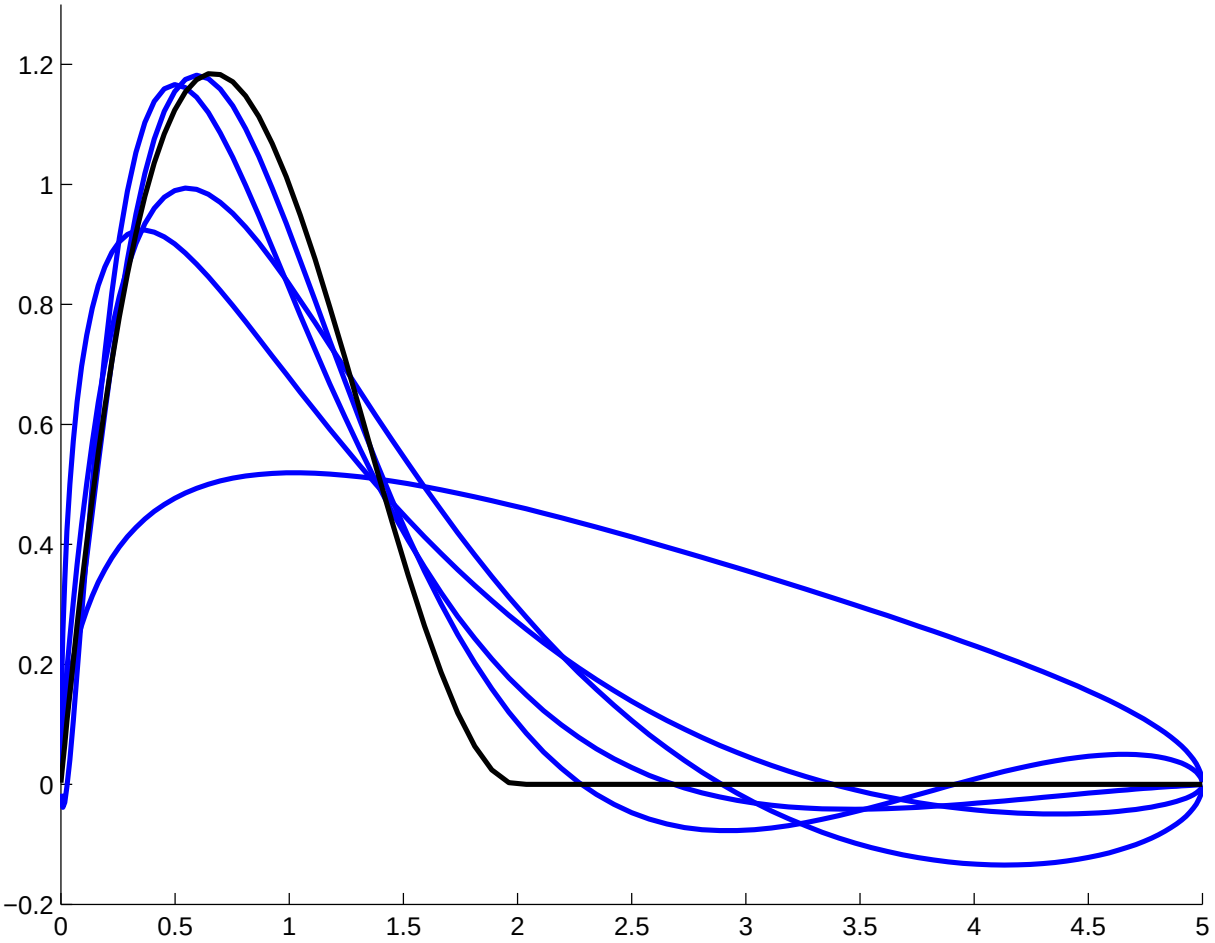
$z_k = A^{\mathrm{T}}(As_k)$;
$\alpha_k = \|r_k\|^2/(s_k^{\mathrm{T}}z_k)$;
$x_{k+1} = x_k + \alpha_k s_k$;
$r_{k+1} = r_k - \alpha_k z_k$;
$\beta_k = \|r_{k+1}\|^2/\|r_k\|^2$;
$s_{k+1} = r_{k+1} + \beta_k s_k$;
$k \leftarrow k + 1$;

end

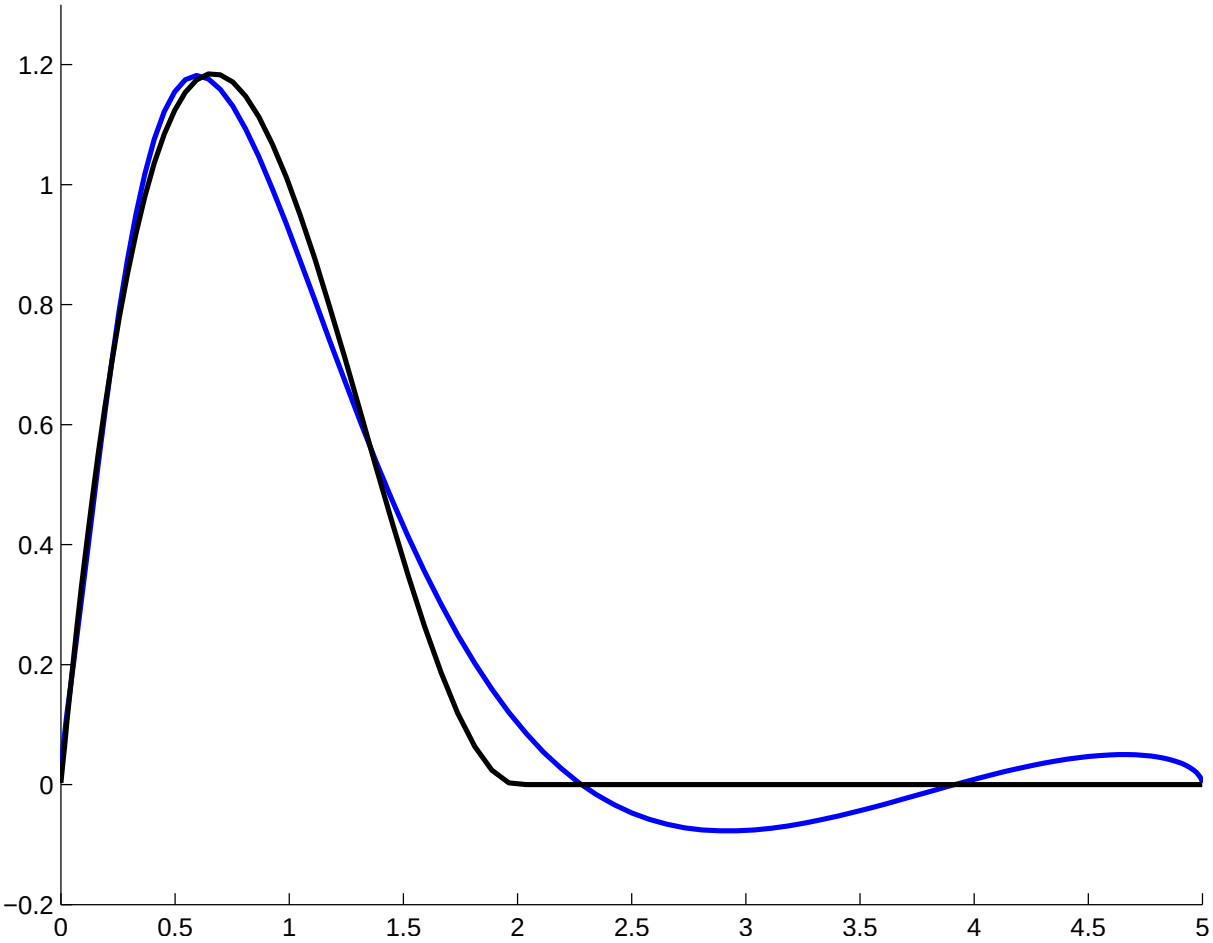
Residual $\|y - Ax_k\|$ as a function of k



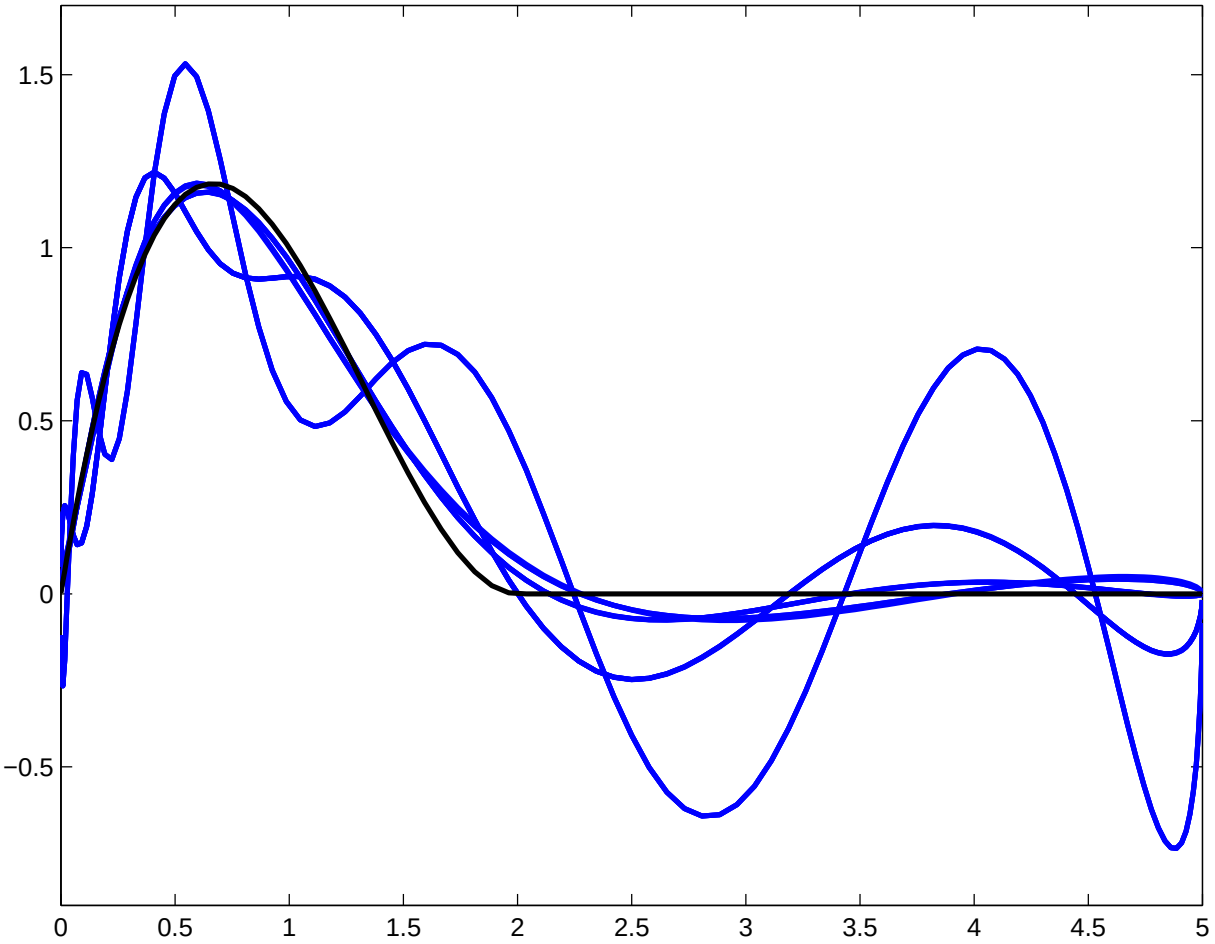
Approximate solutions x_k, $k = 1, \ldots, 5$



Morozov discrepancy solution ($k = 5$)



Approximate solutions x_k, $k = 5, \ldots, 17$



Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Tenth lecture, February 18, 2011.

Computational methods in inverse problems, part II

The second part of the course concentrates on the Bayesian approach to inverse problems.

The lectures are mainly based on the books:

- "J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2005" (parts of Chapter 3),
- "D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing. Ten Lectures on Subjective Computing*, Springer, 2007".

Statistical inversion

In the statistical approach to inverse problems, the leading idea is to recast the inverse problem in the form of statistical *quest for information*.

- Quantities are either directly observable or unobservable.
- Some of the unobservable quantities are of primary interest, others may be considered to be of secondary interest.
- Quantities depend on each other through models.
- The objective of statistical inversion is to extract information on the unknown quantities of interest based on all available knowledge about the measurements, models between the parameters, and information that is available prior to the measurement.

The statistical approach is based on the following principles:

1. All variables are modelled as random variables.
2. The randomness describes our degree of (or lack of) information on their realizations.
3. The information concerning the values of the random variables is coded in probability distributions.
4. The solution of the inverse problem is the *posterior* probability distribution of the quantities of interest (given the measurement).

A classical regularization method typically produces a single estimate, using often a more or less ad hoc removal of the ill-posedness of the problem.

In the statistical framework, the solution is a probability distribution that contains all information on the possible values of the variable of interest. This distribution can be used to obtain different estimates and to evaluate their reliability, e.g., single estimates and credibility intervals. The statistical approach removes the ill-posedness by considering a well-posed extension of the inverse problem in the space of probability densities. When constructing the well-posed extension, the prior beliefs are more explicitly stated than in traditional regularization.

Subjective probability

Example: Tossing a coin.

Assume that the odds of getting heads or tails are equal, i.e.,

$$P(\text{heads}) = P(\text{tails}) = \frac{1}{2}.$$

Such an assumption is generally accepted and can be verified empirically (empirical probability). This example reflects the *frequentist* view, where probability can be seen as the relative frequency of occurrence in a set of repeated experiments.

In connection to Bayesian approach, one sometimes talks about *subjective* probabilities. The inference process commonly incorporates subjective components that reflect the beliefs of, e.g., the person doing the inference (e.g., in the form of prior beliefs about the behaviour of the unknown).

Examples:

What is the probability of rain tomorrow?

What is the probability that Finland will win a gold medal in the next Olympic games?

On random variables and probability densities

Probabilities and events (very informal)

Let Ω contain all possible events, and consider a subset $E \subset \Omega$. For the probability $P(E)$ of an event E, we require

$$0 \leq P(E) \leq 1.$$

Furthermore, it is assumed that

$$P(\Omega) = 1 \qquad \text{and} \qquad P(\emptyset) = 0.$$

Additivity: If $A \cap B = \emptyset$ for $A, B \subset \Omega$, then

$$P(A \cup B) = P(A) + P(B).$$

Two events A and B are called *independent*, if

$$P(A \cap B) = P(A)P(B).$$

The *conditional probability* of A on B is the probability that A happens provided that B happens,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

If A and B are mutually independent,

$$P(A \mid B) = P(A), \quad P(B \mid A) = P(B).$$

Real valued random variables (still informal)

We denote random variables by capital letters and their realizations with lower case letters. Let $X : \Omega \to \mathbb{R}$ be a real valued random variable and denote its *probability density* by $\pi(x) = \pi_X(x) \geq 0$.

The probability of the event $x \in B$, $B \subset \mathbb{R}$ is obtained through integration

$$P\{X(\omega) \in B\} = P(X^{-1}(B)) = \int_B \pi(x)dx.$$

In particular,

$$P\{X(\omega) \in \mathbb{R}\} = P(\Omega) = \int_{-\infty}^{\infty} \pi(x)dx = 1.$$

The *expectation* is the center of mass of the probability density

$$\mathrm{E}(X) = \int_{\mathbb{R}} x\pi(x)dx =: \bar{x}.$$

The *variance* is the expectation of the squared deviation from the expectation

$$\mathrm{var}(X) = \sigma_X^2 = \mathrm{E}\{(X - \bar{x})^2\} = \int_{\mathbb{R}} (x - \bar{x})^2 \pi(x)dx.$$

The *joint probability density* $\pi(x, y) = \pi_{X,Y}(x, y)$ of two random variables X and Y is

$$P\{X \in A,\, Y \in B\} = \iint_{A\times B} \pi(x, y)dxdy.$$

The random variables X and Y are *independent* if

$$\pi(x, y) = \pi(x)\pi(y).$$

The *covariance* of X and Y is

$$\mathrm{cov}(X,Y) = \mathrm{E}\{(X-\bar{x})(Y-\bar{y})\}.$$

Note that

$$\mathrm{cov}(X,Y) = \mathrm{E}\{XY\} - \mathrm{E}\{X\}\mathrm{E}\{Y\}.$$

The *correlation coefficient* of X and Y is

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}, \quad \sigma_X = \sqrt{\mathrm{var}(X)},\ \sigma_Y = \sqrt{\mathrm{var}(Y)},$$

or, equivalently, with the help of normalized random variables,

$$\mathrm{corr}(X,Y) = \mathrm{E}\{\tilde{X}\tilde{Y}\}, \quad \tilde{X} = \frac{X-\bar{x}}{\sigma_X}, \quad \tilde{Y} = \frac{Y-\bar{y}}{\sigma_Y}.$$

Random variables are *uncorrelated* if their covariance (or correlation coefficient) vanishes,

$$\mathrm{cov}(X,Y) = 0.$$

If X and Y are independent, they are uncorrelated, since

$$\mathrm{E}\{(X-\bar{x})(Y-\bar{y})\} = \mathrm{E}\{X-\bar{x}\}\mathrm{E}\{Y-\bar{y}\} = 0.$$

On the other hand, uncorrelated random variables are not necessarily independent.

Given two random variables X and Y with joint probability density $\pi(x,y)$, the *marginal density* of X when Y may take any value, is

$$\pi(x) = \int_{\mathbb{R}} \pi(x,y)dy.$$

Analogously, the marginal density of Y is

$$\pi(y) = \int_{\mathbb{R}} \pi(x,y)dx.$$

The *conditional probability density* of X given Y is the probability density of X assuming that $Y = y$:

$$\pi(x \,|\, y) = \frac{\pi(x, y)}{\pi(y)} \qquad \text{if } \pi(y) \neq 0.$$

Note that by the symmetry of the roles of X and Y, we have

$$\pi(x, y) = \pi(x \,|\, y)\pi(y) = \pi(y \,|\, x)\pi(x),$$

which leads to an important identity

$$\pi(x \,|\, y) = \frac{\pi(y \,|\, x)\pi(x)}{\pi(y)},$$

known as the *Bayes* formula.

The *conditional expectation* or the *conditional mean* is the expectation of X given that $Y = y$:

$$\mathrm{E}\{X \mid y\} = \int_{\mathbb{R}} x\pi(x \mid y)dx.$$

The expectation of X can be computed also via its conditional expectation:

$$\begin{aligned}
\mathrm{E}\{X\} &= \int x\pi(x)dx = \int x \left(\int \pi(x, y)dy \right) dx \\
&= \int x \left(\int \pi(x \mid y)\pi(y)dy \right) dx \\
&= \int \left(\int x\pi(x \mid y)dx \right) \pi(y)dy \\
&= \int \mathrm{E}\{X \mid y\}\pi(y)dy.
\end{aligned}$$

Multivariate random variables

A multivariate random variable is a random variable

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},$$

where each component X_i is a real scalar valued random variable.

The probability density of X is the joint probability density $\pi_X(x) = \pi(x) = \pi(x_1, \ldots, x_n)$ of its components.

The corresponding expectation is

$$\bar{x} = \int_{\mathbb{R}^n} x\pi(x)dx \ \in \mathbb{R}^n,$$

or, componentwise,

$$\bar{x}_i = \int_{\mathbb{R}^n} x_i \pi(x) dx \ \in \mathbb{R}, \quad 1 \leq i \leq n.$$

The *covariance matrix* is defined as

$$\mathrm{cov}(X) = \int_{\mathbb{R}^n} (x - \bar{x})(x - \bar{x})^{\mathrm{T}} \pi(x) dx \ \in \mathbb{R}^{n \times n},$$

or, componentwise,

$$\mathrm{cov}(X)_{ij} = \int_{\mathbb{R}^n} (x_i - \bar{x}_i)(x_j - \bar{x}_j)^{\mathrm{T}} \pi(x) dx \ \in \mathbb{R}, \quad 1 \leq i, j \leq n.$$

The covariance matrix is symmetric and positive semi-definite.

The symmetry is implicit in the definition of the covariance matrix, whereas the positive semi-definiteness follows by writing for $v \in \mathbb{R}^n$ that

$$\begin{aligned} v^{\mathrm{T}}\mathrm{cov}(X)v &= \int_{\mathbb{R}^n} [v^{\mathrm{T}}(x-\bar{x})][(x-\bar{x})^{\mathrm{T}}v]\pi(x)dx \\ &= \int_{\mathbb{R}^n} (v^{\mathrm{T}}(x-\bar{x}))^2\pi(x)dx \geq 0. \end{aligned}$$

Note that the above expression measures the variance of X in the direction v.

The diagonal entries of the covariance matrix are the variances of the individual components of X. Indeed, let us denote by $x_i' \in \mathbb{R}^{n-1}$ the vector x with the ith component deleted, i.e.,

$$x_i' = [x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]^{\mathrm{T}}.$$

Then, we have

$$\begin{aligned}
\mathrm{cov}(X)_{ii} &= \int_{\mathbb{R}^n} (x_i - \bar{x}_i)^2 \pi(x) dx \\
&= \int_{\mathbb{R}} (x_i - \bar{x}_i)^2 \left(\int_{\mathbb{R}^{n-1}} \pi(x_i, x_i') dx_i' \right) dx_i \\
&= \int_{\mathbb{R}} (x_i - \bar{x}_i)^2 \pi(x_i) dx_i \\
&= \mathrm{var}(X_i).
\end{aligned}$$

The marginal and conditional probabilities for multivariate random variables are defined by the same formulas as for the univariate random variables.

Example: Random variables waiting for the train

Assume that every day, except on Sundays, a train for your destination leaves every S minutes from the station. On Sundays, the interval between trains is $2S$ minutes. You arrive at the station with no information about the timetable of the trains (or of the day!!). What is your expected waiting time?

Define a random variable, $T =$ waiting time, whose distribution on working days is

$$T \sim \pi(t \,|\, \text{working day}) = \frac{1}{S}\chi_S(t), \qquad \chi_S(t) = \begin{cases} 1, & 0 \le t < S, \\ 0, & \text{otherwise.} \end{cases}$$

On Sundays, the distribution of T is

$$T \sim \pi(t \,|\, \text{Sunday}) = \frac{1}{2S}\chi_{2S}(t).$$

On a working day, the expected waiting time is

$$E\{T \,|\, \text{working day}\} = \int t\pi(t \,|\, \text{working day})dt = \frac{1}{S}\int_0^S tdt = \frac{S}{2}.$$

On Sundays, the expected waiting time is two times as long.

If you have no idea which day of the week it is, you can give equal probability to each day. Thus,

$$\pi(\text{working day}) = \frac{6}{7}, \quad \pi(\text{Sunday}) = \frac{1}{7}.$$

To get the expected waiting time regardless of the day of the week, marginalize over the days of the week:

$$\begin{aligned} E\{T\} &= E\{T \,|\, \text{working day}\}\pi(\text{working day}) + E\{T \,|\, \text{Sunday}\}\pi(\text{Sunday}) \\ &= \frac{3S}{7} + \frac{S}{7} = \frac{4S}{7}. \end{aligned}$$

Example: Poisson distribution

A weak light source emits photons that are counted with a CCD (*Charged Coupled Device*). The counting process $N(t)$,

$$N(t) = \text{number of particles observed in } [0, t] \in \mathbb{N}$$

is an integer-valued random variable.

Under some assumptions, it can be shown that N is a *Poisson process*:

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad \lambda > 0.$$

We now fix $t = T =$ the recording time, define a random variable $N = N(T)$, and let $\theta = \lambda T$. We write

$$N \sim \text{Poisson}(\theta).$$

We want to calculate the expectation and variance of this Poisson random variable. Since the discrete probability density is

$$\pi(n) = P\{N = n\} = \frac{\theta^n}{n!} e^{-\theta}, \quad \theta > 0,$$

and our random variable takes on discrete values, in the definition of the expectation we have an infinite sum instead of an integral (a countable number of probability masses), that is

$$
\begin{aligned}
\mathrm{E}\{N\} &= \sum_{n=0}^{\infty} n\pi(n) = e^{-\theta} \sum_{n=0}^{\infty} n \frac{\theta^n}{n!} \\
&= e^{-\theta} \sum_{n=1}^{\infty} \frac{\theta^n}{(n-1)!} = e^{-\theta} \sum_{n=0}^{\infty} \frac{\theta^{n+1}}{n!} \\
&= \theta e^{-\theta} \underbrace{\sum_{n=0}^{\infty} \frac{\theta^n}{n!}}_{e^{\theta}} = \theta.
\end{aligned}
$$

We calculate the variance of a Poisson random variable in a similar way, writing first

$$\begin{aligned}\mathrm{var}(N) = \mathrm{E}\{(N-\theta)^2\} &= \mathrm{E}\{N^2\} - 2\theta \underbrace{\mathrm{E}\{N\}}_{=\theta} + \theta^2 \\ &= \mathrm{E}\{N^2\} - \theta^2 \\ &= \sum_{n=0}^{\infty} n^2 \pi(n) - \theta^2.\end{aligned}$$

Substituting the expression of $\pi(n)$, we thus get

$$
\begin{aligned}
\operatorname{var}(N) &= e^{-\theta} \sum_{n=0}^{\infty} n^2 \frac{\theta^n}{n!} - \theta^2 = e^{-\theta} \sum_{n=1}^{\infty} n \frac{\theta^n}{(n-1)!} - \theta^2 \\
&= e^{-\theta} \sum_{n=0}^{\infty} (n+1) \frac{\theta^{n+1}}{n!} - \theta^2 \\
&= \theta e^{-\theta} \sum_{n=0}^{\infty} n \frac{\theta^n}{n!} + \theta e^{-\theta} \sum_{n=0}^{\infty} \frac{\theta^n}{n!} - \theta^2 \\
&= \theta e^{-\theta} \left((\theta + 1) e^{\theta} \right) - \theta^2 \\
&= \theta,
\end{aligned}
$$

that is, the mean and the variance coincide.

Normal distributions

A random variable $X \in \mathbb{R}$ is normally distributed, or Gaussian, i.e.,

$$X \sim \mathcal{N}(x_0, \sigma^2),$$

if

$$P\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right) dx.$$

For $X \sim \mathcal{N}(x_0, \sigma^2)$, it holds that

$$\mathrm{E}\{X\} = x_0, \quad \mathrm{var}(X) = \sigma^2.$$

As a generalization, $X \in \mathbb{R}^n$ is Gaussian if its probability density is

$$\pi(x) = \left(\frac{1}{(2\pi)^n \det(\Gamma)}\right)^{1/2} \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}} \Gamma^{-1} (x - x_0)\right),$$

where $x_0 \in \mathbb{R}^n$, and $\Gamma \in \mathbb{R}^{n \times n}$ is symmetric and positive definite.

Gaussian random variables are widely used in statistics. They appear naturally when *macroscopic* measurements are averages of individual *microscopic* random effects.

Examples: pressure and temperature.

The Central Limit Theorem sheds light on this:

Central Limit Theorem: Assume that real valued random variables $X_1, X_2, \ldots$ are independent and identically distributed, each with expectation μ and variance σ^2. Then the distribution of

$$Z_n = \frac{1}{\sigma\sqrt{n}}(X_1 + X_2 + \ldots + X_n - n\mu)$$

converges to the distribution of a standard normal random variable

$$\lim_{n\to\infty} P\{Z_n \leq x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{t^2/2} dt.$$

Another interpretation of the Central Limit Theorem: If

$$Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

then for large n a good approximation for the probability distribution of Y is

$$Y \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

Example: Poisson distribution (revisited)

One implication of the Central Limit Theorem is that the Poisson distribution can be approximated with a Gaussian distribution if the expectation θ is large.

Intuitive reasoning based on the CCD camera: Assume for simplicity that the expectation θ is a positive integer. The total photon count can then be viewed as a sum of sub-counts on $\theta \in \mathbb{N}$ smaller counter units of equal size. These sub-counts can in turn be viewed as mutually independent Poisson distributed random variables with expectation (and variance) 1. Now, it follows from the Central Limit Theorem that as θ increases, the sum of the sub-counts approaches a normally distributed variable with mean and variance θ.

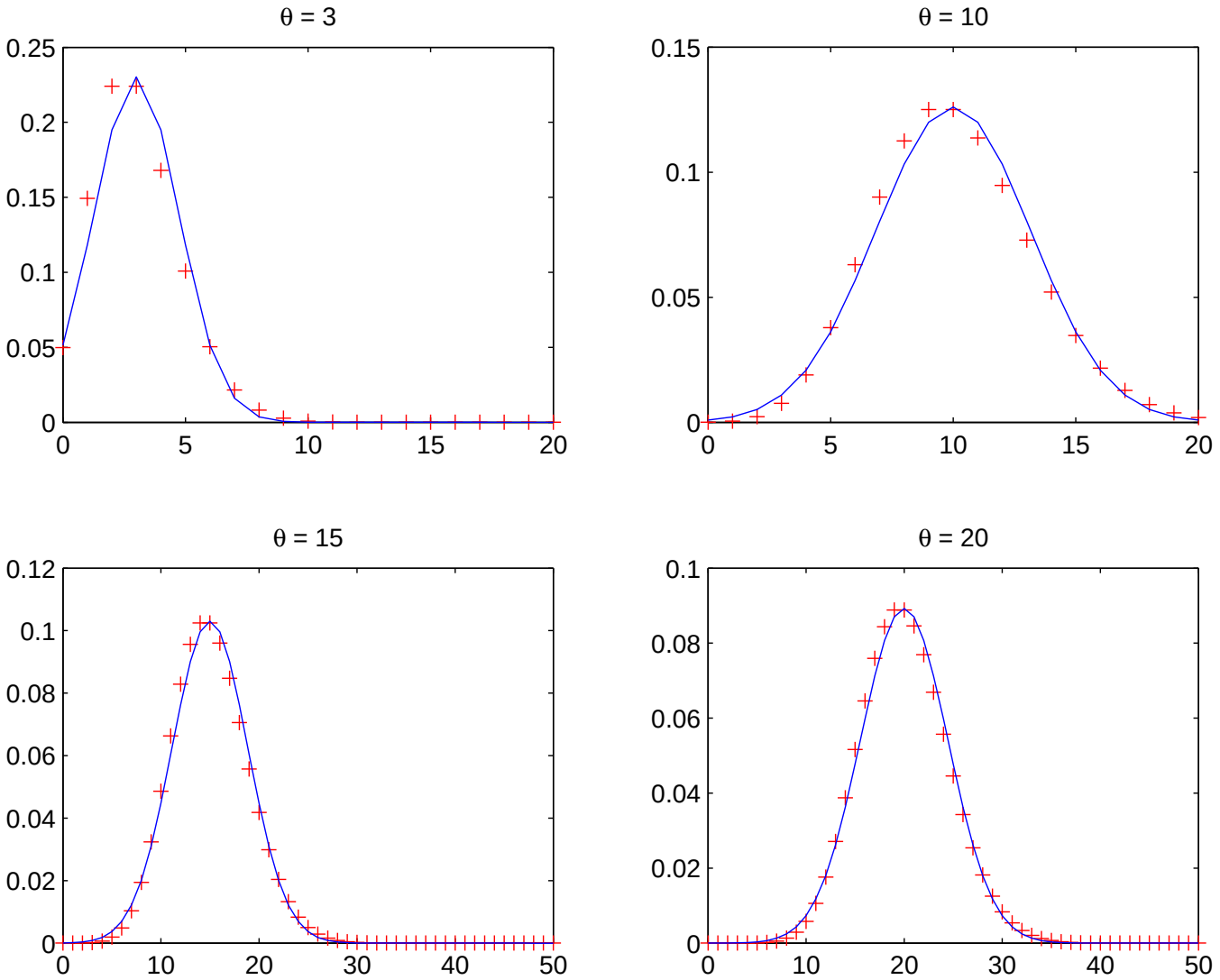
Let us test this hypothesis numerically. We plot the Poisson probability distribution

$$\pi_{\mathrm{Poisson}}(n \,|\, \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

as a function of $n \in \mathbb{N}$, and compare it to the Gaussian approximation

$$\pi_{\mathrm{Gaussian}}(x \,|\, \theta, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(x-\theta)^2\right)$$

as a function of $x \in \mathbb{R}_+$, for increasing values of $\theta > 0$.


θ = 3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
θ = 10
0.15
0.1
0.05
0
0
5
10
15
20
θ = 15
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
θ = 20
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Eleventh lecture, February 23, 2011.

Inverse problems and Bayes' formula

Classical setup for inverse problems:

$$y = f(x, e),$$

where

- $y \in \mathbb{R}^m$ is the measured quantity,
- $x \in \mathbb{R}^n$ is the quantity we seek to get information about,
- $e \in \mathbb{R}^k$ contains the poorly known parameters and noise, and
- $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is the model.

In the statistical setup, all parameters are viewed as random variables, and the classical model is replaced by

$$Y = f(X, E).$$

Notice that the probability distributions of the three random variables X, Y and E depend on each other.

Nomenclature:
Y is called the *measurement*, and its realization y_{obs} the *data*.
X is the unobservable variable of primary interest and called the *unknown*.
The other variables E that are neither observable nor of primary interest are called *parameters* or *noise*.

Prior density

Even before performing the measurement, we typically have some knowledge about the variable X. This information is coded in a probability density $x \mapsto \pi_{\mathrm{pr}}(x)$ called the *prior density*.

Likelihood function

The conditional probability density of Y in case we know the value of the unknown, i.e., $X = x$, is called the *likelihood function*:

$$\pi(y \,|\, x) = \frac{\pi(x, y)}{\pi_{\mathrm{pr}}(x)}, \qquad \text{if } \pi_{\mathrm{pr}}(x) \neq 0.$$

Posterior density

Given the measurement data $Y = y_{\text{obs}}$, the conditional probability density

$$\pi(x \,|\, y_{\text{obs}}) = \frac{\pi(x, y_{\text{obs}})}{\pi(y_{\text{obs}})}, \qquad \text{if } \pi(y_{\text{obs}}) = \int_{\mathbb{R}^n} \pi(x, y_{\text{obs}}) dx \neq 0,$$

is called the *posterior density* of X.

The posterior density expresses what we know about X after realizing the observation $Y = y_{\text{obs}}$.

Inverse problem in the Bayesian framework

Given the data $Y = y_{\mathrm{obs}}$, *find the conditional probability density* $\pi(x \,|\, y_{\mathrm{obs}})$ *of the variable* X.

Bayes theorem of inverse problems

Assume that the random variable $X \in \mathbb{R}^n$ has a known prior probability density $\pi_{\mathrm{pr}}(x)$ and the data consist of the observed value y_{obs} of an observable random variable $Y \in \mathbb{R}^m$ such that $\pi(y_{\mathrm{obs}}) > 0$. Then, the posterior probability density of X, given the data y_{obs}, is

$$\pi_{\mathrm{post}}(x) = \pi(x \,|\, y_{\mathrm{obs}}) = \frac{\pi_{\mathrm{pr}}(x)\pi(y_{\mathrm{obs}} \,|\, x)}{\pi(y_{\mathrm{obs}})}.$$

In practice, the marginal density $\pi(y_{\mathrm{obs}})$ plays a role of a norming constant and is often not important.

Solving an inverse problem in the Bayesian framework

1. Based on all available prior information on the unknown X, find a prior probability density π_{pr} that reflects this information as well as possible.
2. Find the likelihood function $\pi(y \mid x)$ that describes the interrelation between the observation and the unknown.
3. Develop methods to explore the posterior probability density.

Estimators

Maximum a posteriori estimate (MAP)

$$x_{\mathrm{MAP}} = \arg\max_{x \in \mathbb{R}^n} \pi(x \,|\, y)$$

Existence or uniqueness is not guaranteed.

Finding the MAP estimate requires solution of an optimization problem, using, e.g, iterative gradient-based methods.

Conditional mean (CM) estimate is defined as

$$x_{\mathrm{CM}} = E\{x \mid y\} = \int_{\mathbb{R}^n} x\pi(x \,|\, y)dx$$

provided that the integral converges.

Requires solving an integration problem. In high-dimensional spaces, this may require special techniques (sampling).

Maximum likelihood (ML) estimate

$$x_{\mathrm{ML}} = \arg \max_{x \in \mathbb{R}^n} \pi(y \,|\, x)$$

Answers the question: *Which value of the unknown is most likely to produce the measured data?*

The ML estimate is a non-Bayesian estimate, and in the case of ill-posed inverse problems, often not useful. Loosely speaking, it corresponds to solving a classical inverse problem without regularization.

Conditional covariance is a 'spread estimator':

$$\operatorname{cov}(x \,|\, y) = \int_{\mathbb{R}^n} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}} \pi(x \,|\, y) dx \in \mathbb{R}^{n \times n}$$

Requires solving an integration problem.

Bayesian credibility set

Given p, $0 < p < 100$, the credibility set D_p of $p\%$ is defined through the conditions

$$\int_{D_p} \pi(x \,|\, y) dx = \frac{p}{100}, \qquad \pi(x \,|\, y)\big|_{x \in \partial D_p} = \text{constant},$$

and $\pi(x \,|\, y) \geq \pi(z \,|\, y)$ for all $x \in D_p$ and $z \notin D_p$. The boundary of D_p is an equiprobability hypersurface enclosing $p\%$ of the mass of the posterior distribution. (Notice that D_p is not necessarily well defined.)

For a single component, one can look at the symmetric interval of a given credibility: The conditional marginal density of the kth component X_k of X is obtained as

$$\pi(x_k \,|\, y) = \int_{\mathbb{R}^{n-1}} \pi(x_1, \ldots, x_n \,|\, y) dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_n.$$

The end points a and b, $a < b$, of the credibility interval $I_k(p) \subset \mathbb{R}$ with a given p, $0 < p < 100$, are determined from the conditions

$$\int_{-\infty}^{a} \pi(x_k \,|\, y) dx_k = \int_{b}^{\infty} \pi(x_k \,|\, y) dx_k = \frac{1}{2} - \frac{p}{200}.$$

(Unfortunately, these conditions do not always define $I_k(p)$ uniquely.)

An Example: x_{MAP} and x_{CM} estimates

In this example, we compare the x_{MAP} and x_{CM} estimates in a simple one-dimensional case. Let $X \in \mathbb{R}$ and assume that the posterior density $\pi_{\mathrm{post}}(x)$ of X is given by

$$\pi_{\mathsf{post}}(x) = \frac{\alpha}{\sigma_0}\phi\left(\frac{x}{\sigma_0}\right) + \frac{1-\alpha}{\sigma_1}\phi\left(\frac{x-1}{\sigma_1}\right),$$

where $0 < \alpha < 1$, $\sigma_0, \sigma_1 > 0$, and ψ is the standard Gaussian density,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

In this case, we have

$$x_{\mathrm{CM}} = 1 - \alpha,$$

and for small σ_0 and σ_1 it is a good estimate that

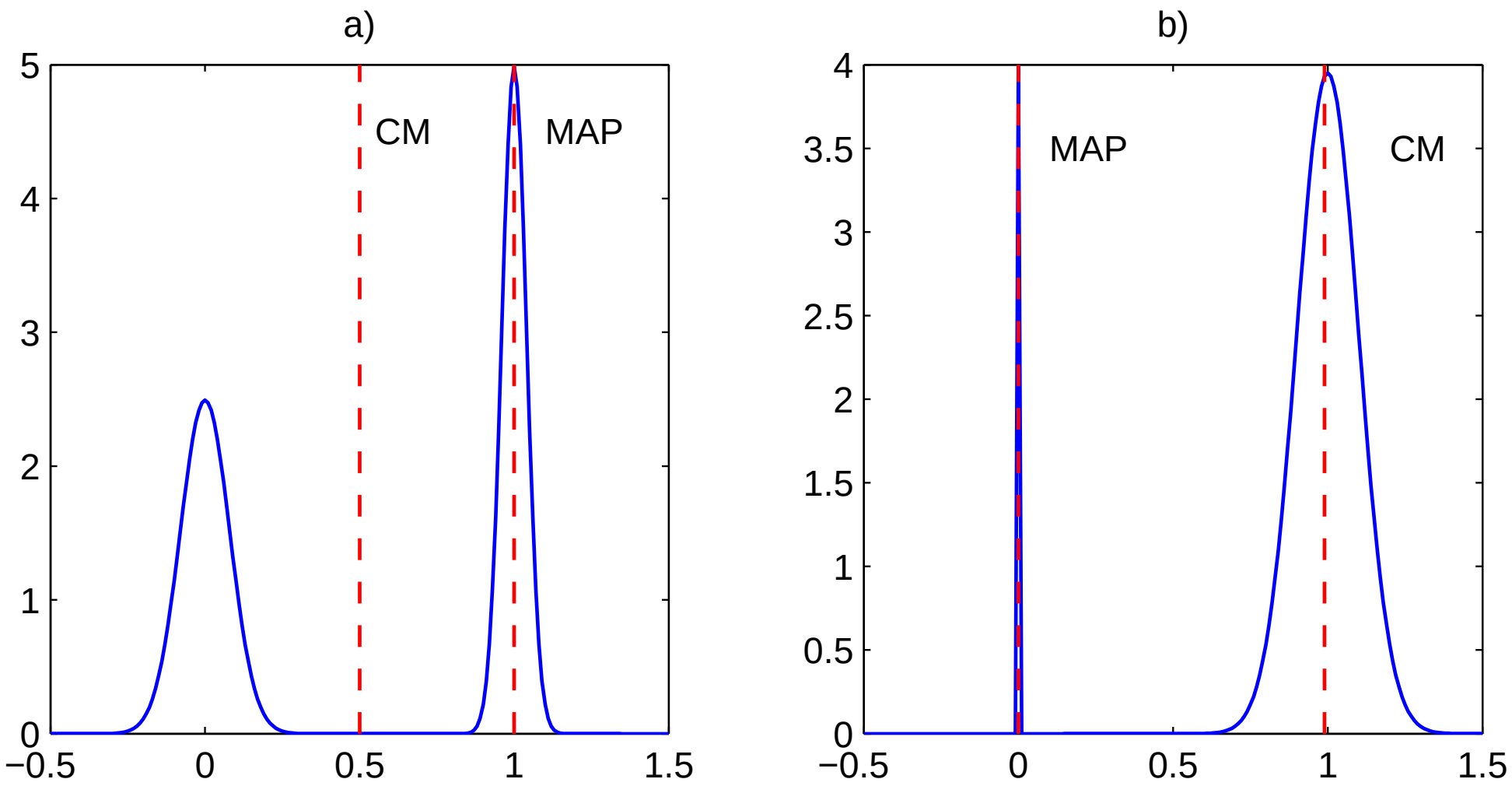
$$x_{\mathrm{MAP}} \approx \begin{cases} 0 & \text{if } \alpha/\sigma_0 \gtrsim (1-\alpha)/\sigma_1, \\ 1 & \text{if } \alpha/\sigma_0 \lesssim (1-\alpha)/\sigma_1. \end{cases}$$

We investigate two different choices of the parameters α, σ_0, σ_1, namely

a) $\alpha = 0.5$, $\sigma_0 = 0.08$ and $\sigma_1 = 0.04$,

b) $\alpha = 0.01$, $\sigma_0 = 0.001$ and $\sigma_1 = 0.1$.

Note that in case b), $\alpha = \sigma_0/\sigma_1$, which means that $\alpha/\sigma_0 > (1-\alpha)/\sigma_1$, and thus $x_{\mathrm{MAP}} \approx 0$ should be the valid case. (You can easily verify this fact numerically.)


a)
CM
MAP
5
4
3
2
1
0
−0.5
0
0.5
1
1.5
b)
MAP
CM
4
3.5
3
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
1
1.5

Let us also consider the posterior variance

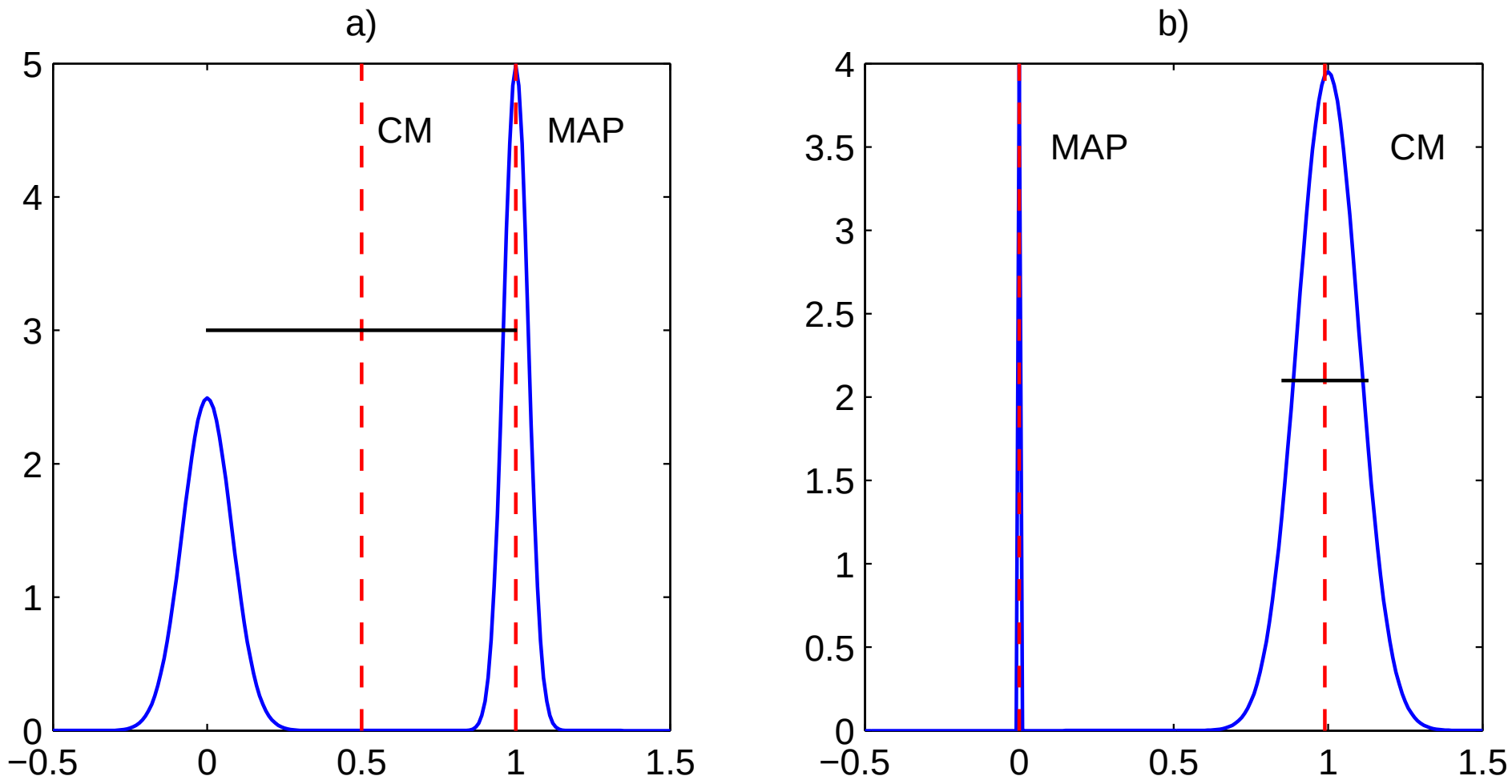
$$\sigma^2 = \int_{-\infty}^{\infty} (x - x_{\mathrm{CM}})^2 \pi_{\mathrm{post}}(x) dx = \int_{-\infty}^{\infty} x^2 \pi_{\mathrm{post}}(x) dx - x_{\mathrm{CM}}^2,$$

which can be calculated analytically in our simple setting:

$$\sigma^2 = \alpha\sigma_0^2 + (1-\alpha)(\sigma_1^2 + 1) - (1-\alpha)^2.$$

In the following images, we have visualized the intervals of length 2σ, i.e., of length two times the standard deviation, centered at x_{CM} for both sets of parameters.

Notice that when the conditional mean gives a poor estimate, this is reflected as a larger variance.


a)
CM
MAP
5
4
3
2
1
0
−0.5
0
0.5
1
1.5
b)
MAP
CM
4
3.5
3
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
1
1.5

Construction of the likelihood function

The likelihood function answers the question: *If we knew the unknown x, how would the measurements be distributed?*

What makes the data deviate from the predicted value given by our observation model?

Some common sources:

1. measurement noise in the data,
2. incompleteness of the observation model (e.g., discretization errors, the reduced nature of the model as compared to the "reality").

Commonly used techniques in construction of the likelihood function (and priors) include conditioning (inspect one variable at the time) and marginalization (eliminate variables of secondary interest).

Additive noise

Very often, the noise is modelled as additive and independent of X. This means that the stochastic model is

$$Y = f(X) + E.$$

Let us assume that the probability distribution of the noise is known:

$$P\{E \in B\} = \int_B \pi_{\text{noise}}(e)de, \qquad B \in \mathbb{R}^m.$$

Because X and E are mutually independent, fixing $X = x$ does not alter the probability distribution of E. Hence, Y conditioned on $X = x$ is distributed as E shifted by the constant $f(x)$:

$$\pi(y \,|\, x) = \pi_{\text{noise}}(y - f(x)).$$

If the prior probability density of X is π_{pr}, we thus obtain from the Bayes formula that

$$\pi(x \,|\, y) \propto \pi_{\mathrm{pr}}(x)\pi_{\mathrm{noise}}(y - f(x)).$$

If the unknown X and the noise E are *not* mutually independent, we need to know the conditional density of the noise

$$P\{E \in B \,|\, X = x\} = \int_B \pi_{\mathrm{noise}}(e \,|\, x)de.$$

Then, we may write

$$\pi(y \,|\, x) = \int_{\mathbb{R}^m} \pi(y, e \,|\, x)de = \int_{\mathbb{R}^m} \pi(y \,|\, x, e)\pi_{\mathrm{noise}}(e \,|\, x)de.$$

If both $X = x$ and $E = e$ are fixed, $Y = f(x) + e$, and hence

$$\pi(y \,|\, x, e) = \delta(y - f(x) - e).$$

Substituting $\pi(y \,|\, x, e)$ into the last formula of the preceding slide thus yields

$$\pi(y \,|\, x) = \pi_{\text{noise}}(y - f(x) \,|\, x),$$

and once again from the Bayes formula we get that

$$\pi(x \,|\, y) \propto \pi_{\text{pr}}(x)\pi_{\text{noise}}(y - f(x) \,|\, x).$$

Example: Additive independent noise

A simple low-dimensional example: a linear model

$$Y = AX + E,$$

where $X \in \mathbb{R}^2$ and $Y, E \in \mathbb{R}^3$ are random variables, and

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -2 \\ 2 & 1 \end{bmatrix}$$

is deterministic. Assume that E has mutually independent normally distributed components with zero mean and variance $\sigma^2 = 0.09$, i.e.,

$$\pi_{\text{noise}}(e) \propto \exp\left(-\frac{1}{2\sigma^2}\|e\|^2\right).$$

Our only prior information is that

$$P\{|X_j| > 2\} = 0, \qquad j = 1, 2,$$

which we write in the form of a prior density via

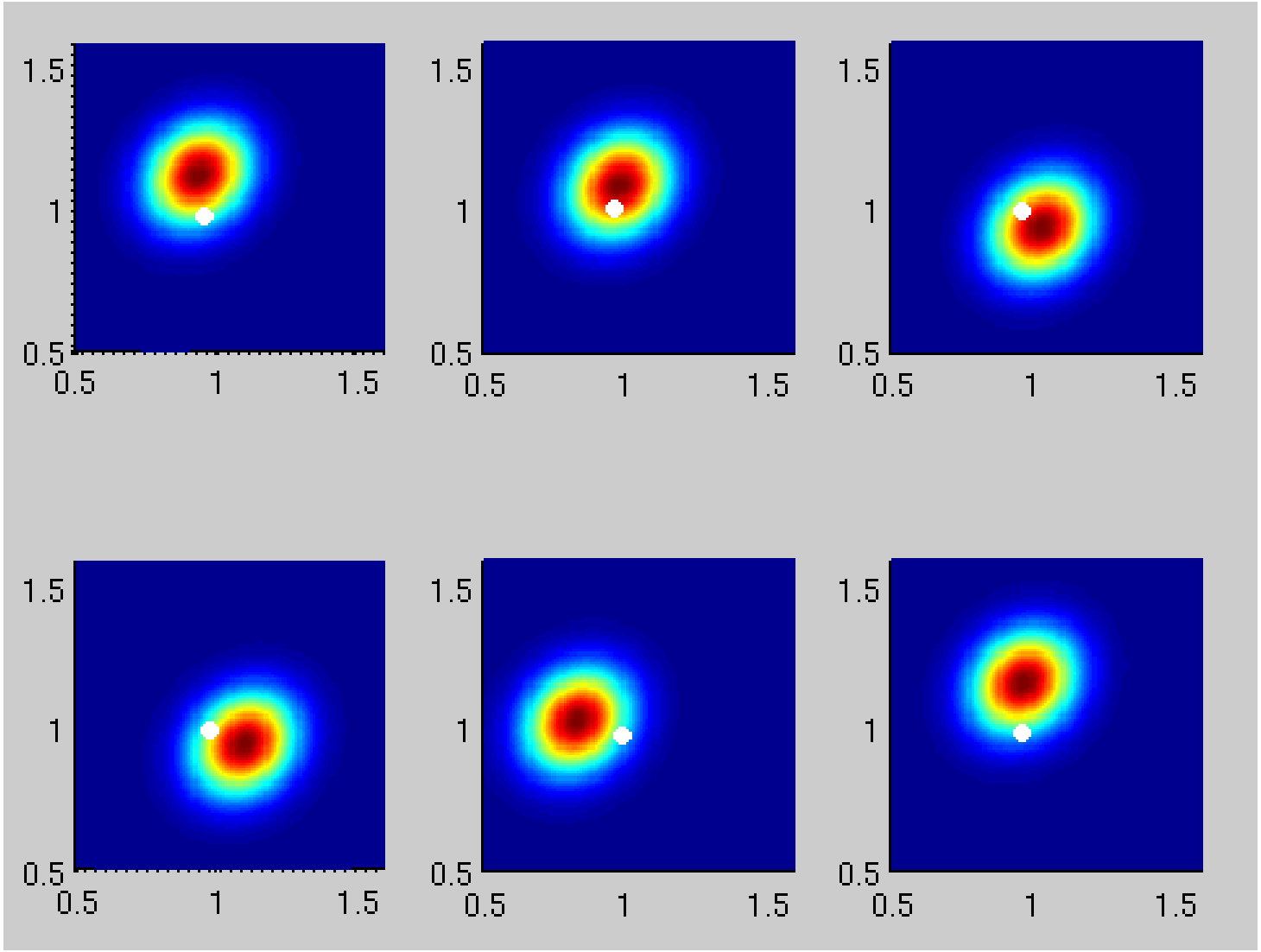
$$\pi_{\mathrm{pr}}(x) = \frac{\chi_Q(x)}{16},$$

where χ_Q is the characteristic function of the square $[-2, 2] \times [-2, 2]$. The posterior density is then

$$\pi(x \,|\, y) \propto \chi_Q(x) \exp\left(-\frac{1}{2\sigma^2} \|y - Ax\|^2\right).$$

Suppose that the true value of X is $x_0 = [1, 1]^{\mathrm{T}}$. We simulate the data through $y = Ax_0 + e$, where e is drawn from π_{noise}.

The following figure illustrates the posterior density with six different realizations of E. Note that in this case the prior hardly plays any role.


1.5
1
0.5
0.5
1
1.5

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Twelfth lecture, February 25, 2011.

Construction of the likelihood function (continued)

General noise model

Assume that we have an observation model of the type $Y = f(X, E)$, where $X \in \mathbb{R}^n$ is the unknown, $Y \in \mathbb{R}^m$ is the measurement and $E \in \mathbb{R}^k$ is the noise/parameter vector. Since fixing X and E determines the value of Y, we may write

$$\pi(y \,|\, x, e) = \delta(y - f(x, e)).$$

In consequence,

$$\pi(y \,|\, x) = \int_{\mathbb{R}^k} \pi(y, e \,|\, x) de = \int_{\mathbb{R}^k} \delta(y - f(x, e)) \pi_{\text{noise}}(e \,|\, x) de.$$

Change of variables

Consider two random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$ that are related via the formula

$$Y = f(X),$$

where f is continuously differentiable and injective (these conditions can be relaxed). Suppose we know the probability density of Y, namely π_Y.

Then, for a Borel set $B \subset \mathbb{R}^n$, it holds that

$$\begin{aligned} P\{X \in B\} &= P\{Y \in f(B)\} = \int_{f(B)} \pi_Y(y)dy \\ &= \int_B \pi_Y(f(x))|\det Df(x)|dx \end{aligned}$$

where $Df(x) \in \mathbb{R}^{n\times n}$ is the *differential* or the *Jacobian matrix* of f. As a consequence,

$$\pi_X(x) = \pi_Y(f(x))|\det Df(x)|.$$

Example: multiplicative noise

Consider an amplifier that takes in a signal $f(t) > 0$ and sends it out multiplied by a constant factor $\alpha > 1$. The ideal model for the output is thus

$$g(t) = \alpha f(t), \qquad 0 \le t \le T.$$

Suppose that the amplification factor is not a constant but fluctuates slightly around a mean value $\alpha_0 > 0$ as a function of time. In order to write a likelihood model for the output, we first discretize the signal:

$$x_j = f(t_j), \quad y_j = g(t_j), \qquad 0 = t_1 < t_2 < \cdots < t_n = T.$$

Let the amplification at $t = t_j$ be a_j, i.e.,

$$y_j = a_j x_j, \qquad 1 \le j \le n,$$

and introduce the stochastic extension:

$$Y_j = A_j X_j, \qquad 1 \le j \le n.$$

In vector notation, this reads

$$Y = A.X,$$

with the dot denoting componentwise multiplication of the vectors $A, X \in \mathbb{R}^n$; we also use a similar notation for componentwise division.

Assume that A is independent of X and has the probability density

$$A \sim \pi_{\text{noise}}(a).$$

To find the likelihood density of Y, conditioned on $X = x$ such that $x_j > 0$ for all $j = 1, \ldots n$, we write

$$A_j = \frac{Y_j}{x_j}, \qquad 1 \leq j \leq n.$$

Thus, we obtain by the change of variables formula that

$$\pi(y \,|\, x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\text{noise}} \left(\frac{y.}{x} \right).$$

As an example, assume that the components of $A \in \mathbb{R}^n$ are mutually independent and *log-normally* distributed:

$$W_i := \log A_i \sim \mathcal{N}(w_0, \sigma^2), \qquad w_0 = \log \alpha_0.$$

To find an explicit formula for the density of A, we note that if $w = \log a$, where the logarithm is applied componentwise, we have

$$dw = \frac{1}{a_1 a_2 \cdots a_n} da \qquad \text{for } a_1, \ldots, a_n > 0.$$

Thus, the probability density of A vanishes if any of the components of a is zero or negative, and otherwise it holds that

$$\pi_{\text{noise}}(a) = \frac{1}{a_1 a_2 \cdots a_n} \exp\left(-\frac{1}{2\sigma^2} \|\log(a/\alpha_0)\|^2 \right).$$

By substituting this formula in

$$\pi(y \,|\, x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\text{noise}} \left(\frac{y_\cdot}{x} \right),$$

we find that

$$\pi(y \,|\, x) \propto \frac{1}{y_1 y_2 \cdots y_n} \exp \left(-\frac{1}{2\sigma^2} \left\| \log \left(\frac{y_\cdot}{\alpha_0 x} \right) \right\|^2 \right).$$

for $y \in \mathbb{R}^n$ such that $y_j > 0$ for all $j = 1, \ldots, n$, and zero for other $y \in \mathbb{R}^n$. (Recall that it was assumed to begin with that the components of x are positive.)

Incompletely known forward model

Consider having a noisy measurement with an incompletely known forward model: The deterministic model with additive noise is $y = A(v)x + e$, $y, e \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ and $A(v) \in \mathbb{R}^{m \times n}$, where $A(v)$ depends on a parameter vector $v \in \mathbb{R}^k$.

The corresponding stochastic extension is

$$Y = A(V)X + E.$$

Assume that E, X and V are mutually independent. How to construct the likelihood model $\pi(y \,|\, x)$, assuming that the noise is distributed according to π_{noise} and the parameter according to π_{param}?

To begin with, fix $X = x$ and $V = v$ in order to get the conditional density of Y:

$$\pi(y \,|\, x, v) = \pi_{\text{noise}}(y - A(v)x).$$

Subsequently, we marginalize with respect to the parameter V which is of secondary interest:

$$\begin{aligned} \pi(y \,|\, x) &= \int_{\mathbb{R}^k} \pi(y, v \,|\, x) dv = \int_{\mathbb{R}^k} \pi(y \,|\, x, v) \pi_{\text{param}}(v) dv \\ &= \int_{\mathbb{R}^k} \pi_{\text{noise}}(y - A(v)x) \pi_{\text{param}}(v) dv. \end{aligned}$$

On sampling

Before moving on to construction of priors, we touch the subject of how to draw a sample of realizations from a given probability distribution.

Why is such consideration relevant?

- Visual inspection of priors, and
- estimation of integrals of the type

$$I = \int f(x)\pi(x)dx$$

  with the help of Markov chain Monte Carlo (MCMC) techniques.

In what follows, we assume to have random number generators for two elementary distributions at our disposal:

- Standard normal distribution

$$\pi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right);$$

  in Matlab the command `randn`.

- Uniform distribution over the interval $[0, 1]$,

$$\pi(x) = \chi_{[0,1]}(x);$$

  in Matlab the command `rand`.

Sampling from Gaussian distributions

Suppose that we want to create a sample of realizations for a multivariate Gaussian random variable $X \sim \mathcal{N}(x_0, \Gamma)$, with the probability density

$$\pi(x) = \left(\frac{1}{(2\pi)^n \det(\Gamma)} \right)^{1/2} \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}} \Gamma^{-1} (x - x_0) \right).$$

Since Γ^{-1} is (by assumption) symmetric and positive definite, it has a Cholesky decomposition

$$\Gamma^{-1} = R^{\mathrm{T}} R,$$

where R is an upper triangular matrix. Notice that the probability density of X can alternatively be written as

$$\pi(x) = \left(\frac{1}{(2\pi)^n \det(\Gamma)} \right)^{1/2} \exp\left(-\frac{1}{2} \| R(x - x_0) \|^2 \right).$$

Encouraged by this observation, we define a new random variable

$$W = R(X - x_0) \quad \Longleftrightarrow \quad X = R^{-1}W + x_0,$$

which, in particular, means that

$$\pi_W(w) = \pi_X(R^{-1}w + x_0)|\det(R^{-1})| = \pi_X(R^{-1}w + x_0)|\det(R)|^{-1}.$$

Using the identity

$$\det(\Gamma)^{-1} = \det(\Gamma^{-1}) = \det(R^{\mathrm{T}})\det(R) = \det(R)^2,$$

leads finally to the formula

$$\pi(w) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|w\|^2\right).$$

In consequence, W is *Gaussian white noise*, i.e.,

$$W \sim \mathcal{N}(0, I).$$

This transformation is called the *whitening* of X and the Cholesky factor R of the inverse of the covariance the *whitening matrix*.

If the whitening matrix is known, a random draw from a general Gaussian density can be generated as follows:

1. Draw $w \in \mathbb{R}^n$ from the Gaussian white noise density.
2. Compute the sought for realization $x \in \mathbb{R}^n$ by solving the linear system

$$w = R(x - x_0),$$

which is almost trivial since R is triangular.

Random draws from non-Gaussian densities using direct sampling

Let us next consider how to draw a random sample directly from the actual distribution in one dimension.

Let X be a real valued random variable with probability density $\pi(x)$ such that $\pi(x) = 0$ only at isolated points (this assumption can be relaxed). Define the cumulative distribution function via

$$\Phi(z) = \int_{-\infty}^{z} \pi(x)dx.$$

Due to the assumptions on π, it follows from the fundamental theorem of calculus that Φ is strictly increasing. In particular, $\Phi : \mathbb{R} \to (0, 1)$ has an inverse $\Phi^{-1} : (0, 1) \to \mathbb{R}$.

Define a new random variable,

$$T = \Phi(X).$$

Lemma. $T \sim \mathrm{Uniform}([0,1])$.

Proof. Observe first that,

$$P\{T < a\} = P\{\Phi(X) < a\} = P\{X < \Phi^{-1}(a)\}, \quad 0 < a < 1.$$

On the other hand, due to the definition of a probability density,

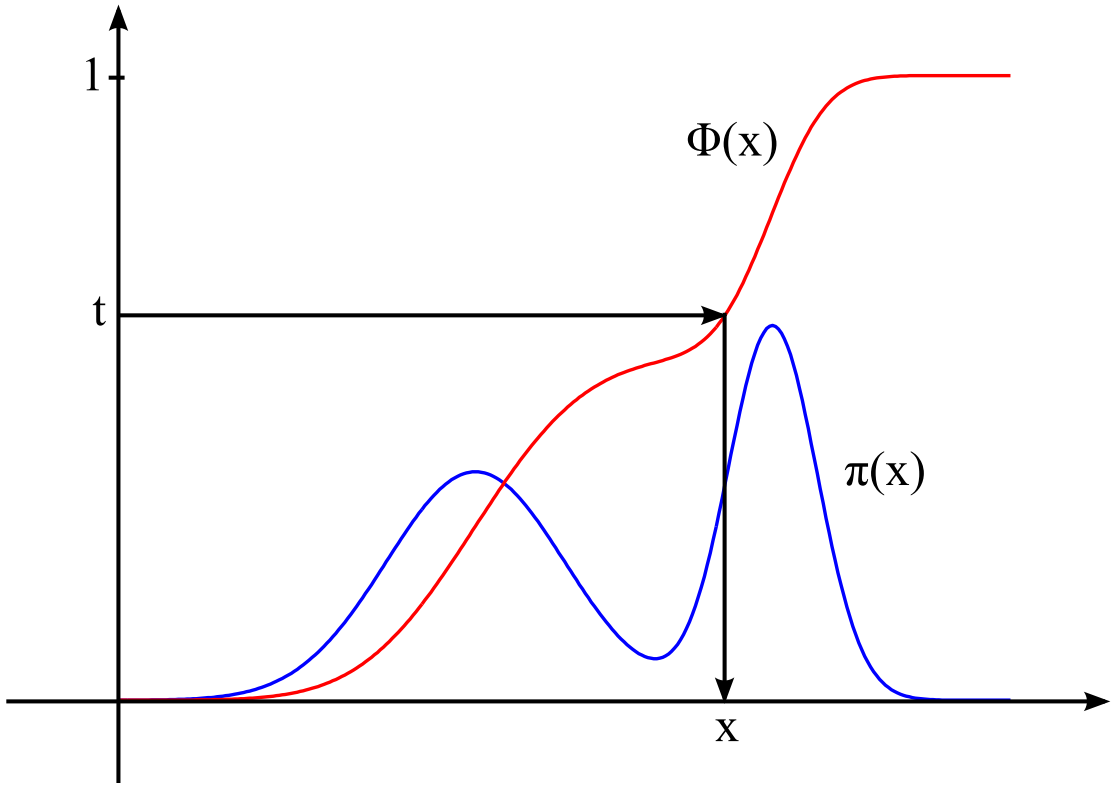
$$\begin{aligned} P\{X < \Phi^{-1}(a)\} &= \int_{-\infty}^{\Phi^{-1}(a)} \pi(x)dx = \int_{-\infty}^{\Phi^{-1}(a)} \Phi'(x)dx \\ &= \Phi(\Phi^{-1}(a)) - \lim_{x \to -\infty} \Phi(x) = a - 0 = a, \end{aligned}$$

which just means that T is distributed uniformly over the interval $[0,1]$.

An algorithm for drawing from the density π:

1. Draw $t \sim \text{Uniform}([0, 1])$,
2. Calculate $x = \Phi^{-1}(t)$.

This technique is sometimes referred to as the *Golden Rule*.



Example: Gaussian distribution with a bound constraint

Consider a one-dimensional normal distribution with a bound constraint,

$$\pi(x) \propto \pi_c(x) \exp\left(-\frac{1}{2}x^2\right),$$

where

$$\pi_c(x) = \begin{cases} 1 & \text{if } x > c, \\ 0 & \text{if } x \leq c \end{cases}$$

for some $c \in \mathbb{R}$. Our aim is to generate a sample from this distribution.

In this case, the cumulative distribution function is

$$\Phi(z) = C \int_c^z e^{-x^2/2} dx, \qquad C = \left(\int_c^\infty e^{-x^2/2} dx\right)^{-1},$$

where $C > 0$ is the normalizing constant of the corresponding probability density.

The function Φ has to be calculated numerically. Fortunately, there are routines available to do the needed integration: In Matlab, the built-in *error function*, `erf`, is defined as

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds.$$

We observe that

$$\begin{aligned}\Phi(z) &= C \left(\int_0^z - \int_0^c \right) e^{-x^2/2} dx = \sqrt{2} C \left(\int_0^{z/\sqrt{2}} - \int_0^{c/\sqrt{2}} \right) e^{-s^2} ds \\ &= \sqrt{\frac{\pi}{2}}\, C \left(\operatorname{erf}(z/\sqrt{2}) - \operatorname{erf}(c/\sqrt{2}) \right).\end{aligned}$$

Since $\operatorname{erf}(t) \to 1$ as $t \to \infty$, the same logic also shows that

$$C = \left(\sqrt{\frac{\pi}{2}} \left(1 - \operatorname{erf}(c/\sqrt{2}) \right) \right)^{-1}.$$

Altogether we have

$$\Phi(z) = \frac{\operatorname{erf}(z/\sqrt{2}) - \operatorname{erf}(c/\sqrt{2})}{1 - \operatorname{erf}(c/\sqrt{2})}.$$

How about the inverse then?

Setting

$$\Phi(z) = t \quad \Longleftrightarrow \quad z = \Phi^{-1}(t),$$

we find through a straightforward algebraic manipulation that

$$\operatorname{erf}(z/\sqrt{2}) = t\big(1 - \operatorname{erf}(c/\sqrt{2})\big) + \operatorname{erf}(c/\sqrt{2}),$$

or in other words (see `erfinv` in Matlab)

$$\Phi^{-1}(t) = \sqrt{2}\operatorname{erf}^{-1}\Big(t\big(1 - \operatorname{erf}(c/\sqrt{2})\big) + \operatorname{erf}(c/\sqrt{2})\Big).$$

The generation of random draws in Matlab is then very simple:

```
a = erf(c/sqrt(2));
t = rand;
z = sqrt(2)*erfinv(t*(1-a)+a);
```

Note: If the bound c is large, the above program does not work because the error function saturates quickly to unity. To be more precise, e.g. for `c=10`, Matlab interprets that `a` in the above code is exactly one, which means that the value of `z` is `Inf` independently of the random draw `t`. An alternative implementation in this case is to perform the numerical integration only at the region we are interested in. This approach is discussed at the exercises.

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Thirteenth lecture, March 2, 2011.

Prior models

The prior density should reflect our beliefs on the unknown variable of interest *before* taking the measurements into account.

Often, the prior knowledge is *qualitative* in nature, and transferring the information into *quantitative* form expressed through a prior density can be challenging.

A good prior should have the following property: Denote by x a possible realization of a random variable $X \sim \pi_{\mathrm{pr}}(x)$. If E is a collection of expected (i.e., something you would expect to see) vectors x and U is a collection of unexpected ones, then it should hold that

$$\pi_{\mathrm{pr}}(x) \gg \pi_{\mathrm{pr}}(x') \qquad \text{when } x \in E,\ x' \in U,$$

i.e., the prior assigns a clearly higher probability to the realization that we expect to see.

Example: Impulse prior densities

Consider, e.g., an imaging problem where the unknown is the discretized distribution of a physical parameter, i.e., a pixel image.

Assume that our prior information is that the image contains small and well localized objects in almost constant background. In such a case, one may try impulse prior densities, which have low average amplitude but allow outliers. (The 'tail' of an impulse prior density is long, although the expected value is small.)

Examples of impulse prior densities: Let $x \in \mathbb{R}^n$ represent a pixel image, where the component x_j is the intensity of the jth pixel. (In all of the following examples, X_j and X_k are assumed to be independent for $j \neq k$.)

The ℓ_1 prior:

$$\pi_{\mathrm{pr}}(x) = \left(\frac{\alpha}{2}\right)^n \exp(-\alpha\|x\|_1), \qquad \alpha > 0.$$

where the ℓ_1-norm is defined as

$$\|x\|_1 = \sum_{j=1}^{n} |x_j|.$$

More enhanced impulse noise effect can be obtained by taking even smaller power of the components of x:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\alpha \sum_{j=1}^{n} |x_j|^p\right), \qquad 0 < p < 1, \quad \alpha > 0.$$

Such priors are studied in the seventh exercise session.

Another choice is the *Cauchy density* that is defined via

$$\pi_{\mathrm{pr}}(x) = \Big(\frac{\alpha}{\pi}\Big)^n \prod_{j=1}^{n} \frac{1}{1+\alpha^2 x_j^2}, \qquad \alpha > 0.$$

The entropy of an image is defined as

$$\mathcal{E}(x) = -\sum_{j=1}^{n} x_j \log \frac{x_j}{x_0},$$

where it is assumed that $x_j > 0$, $j = 1, \ldots n$, and $x_0 > 0$ is a given constant. The *entropy density* is then of the form

$$\pi(x) \propto \exp\big(\alpha \mathcal{E}(x)\big), \qquad \alpha > 0.$$

Log-normal density: The logarithm of a single pixel $x \in \mathbb{R}$ is normally distributed, i.e.,

$$w = \log x, \qquad w \sim \mathcal{N}(w_0, \sigma^2).$$

The explicit density of x is then

$$\pi(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - w_0)^2\right), \qquad x > 0.$$

Do these priors represent our beliefs? How do these priors looks like?

To underline the interpretation as a pixel image, we add a positivity constraint to the above introduced priors, that is, we make the replacement

$$\pi_{\mathrm{pr}}(x) \to C\pi_{+}(x)\pi_{\mathrm{pr}}(x),$$

where $\pi_{+}(x)$ is one if all components of x are positive, and zero otherwise. Here, C is a normalizing constant: If $\pi_{\mathrm{pr}}(x)$ is a probability density, the same does not typically apply to $\pi_{+}(x)\pi_{\mathrm{pr}}(x)$ without appropriate scaling.

For visual inspection we make random draws of pixel images from the constrained densities. As all components are independent, drawing can be done componentwise.

To make the draws from one-dimensional densities, we calculate the cumulative distribution of the prior density and employ the Golden Rule, as presented at the previous lecture.

Example: Drawing from ℓ_1 prior

The one-dimensional cumulative distribution of the positively constrained ℓ_1 prior is

$$\Phi(t) = \alpha \int_0^t e^{-\alpha s} ds = 1 - e^{-\alpha t}.$$

The inverse cumulative distribution is thus

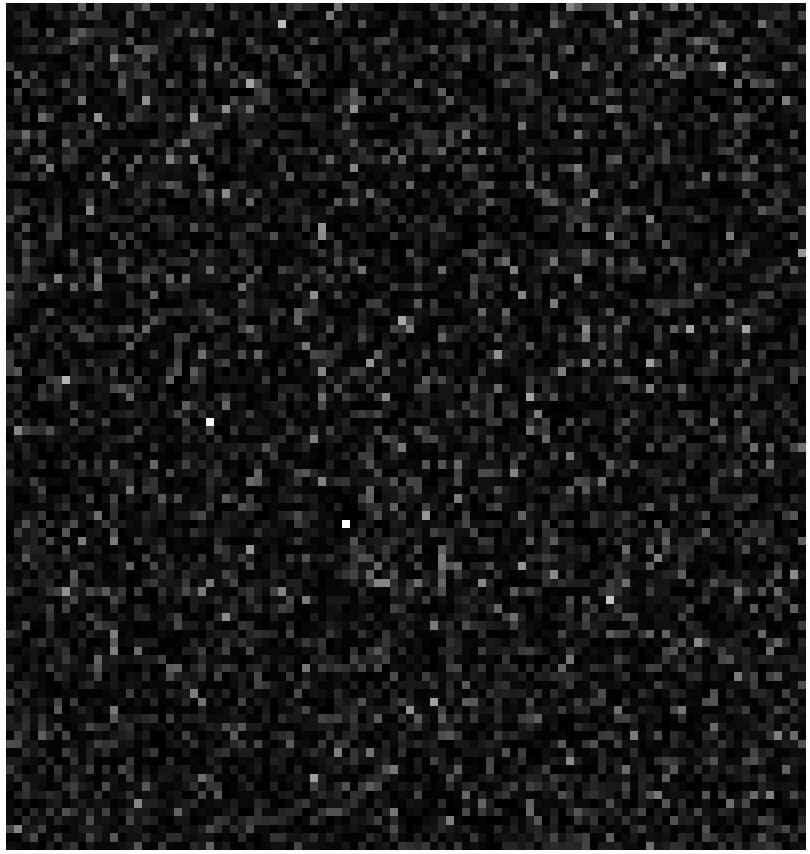
$$\Phi^{-1}(t) = -\frac{1}{\alpha} \log(1 - t).$$

For each pixel x_j, we draw t_j from the uniform distribution $\mathrm{Uniform}([0, 1])$ and calculate $x_j = -1/\alpha \log(1 - t_j)$.

The Matlab code for doing this is very simple:

```
A=rand(100,100);
alfa=1;
AL1inv=-1/alfa*log(1-A);
figure
imagesc(AL1inv)
colormap gray
axis square
```

Two random draws of pixel images from a ℓ_1-prior.



Example: Drawing from Cauchy prior

The one-dimensional cumulative distribution of the positively constrained Cauchy prior is

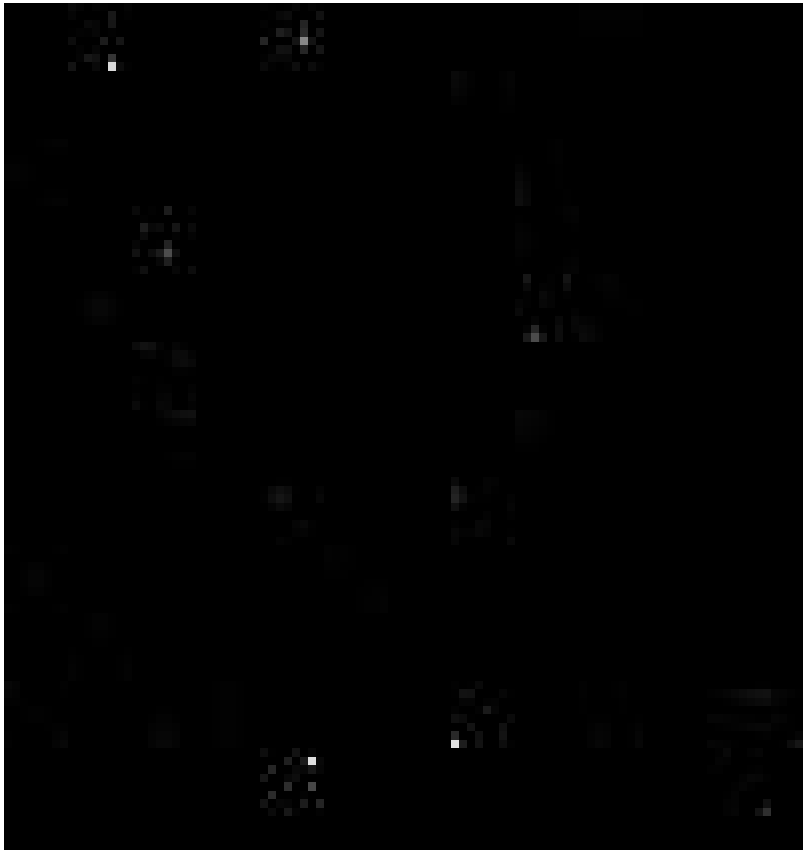
$$\Phi(t) = \frac{2\alpha}{\pi} \int_0^t \frac{1}{1+\alpha^2 s^2} ds = \frac{2}{\pi} \arctan(\alpha t),$$

meaning that the inverse cumulative distribution is

$$\Phi^{-1}(t) = \frac{1}{\alpha} \tan \frac{\pi t}{2}.$$

As in the case of the ℓ_1-prior, we draw t_j from the uniform distribution and then calculate $x_j = 1/\alpha \tan(\pi t/2)$.

Two random draws of pixel images from a Cauchy prior.



How do these priors compare to white noise?

Let us consider a Gaussian prior with a positivity constraint, i.e.,

$$\pi_{\mathrm{pr}}(x) \propto \pi_{+}(x) \exp\left(-\frac{1}{2\alpha^2}\|x\|^2\right), \qquad \alpha > 0.$$

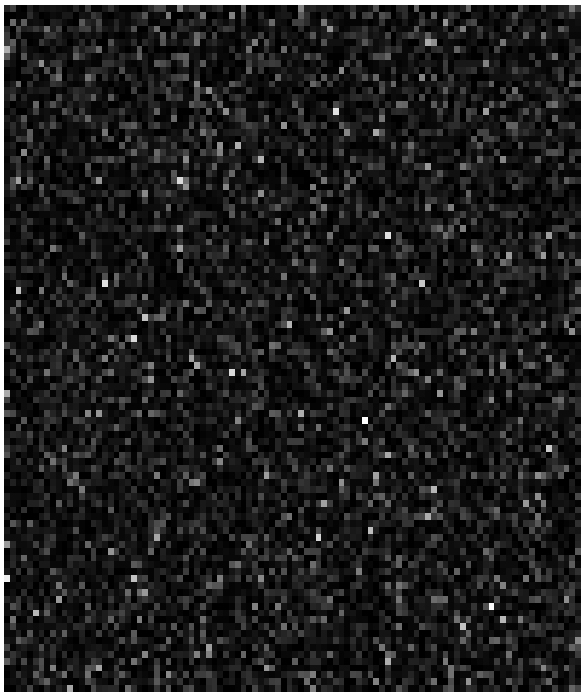
Recall that at the previous lecture we implemented drawing from a standard Gaussian distribution with a bound c. In particular, we were able to calculate the one-dimensional cumulative distribution function

$$\Phi^{-1}(t) = \sqrt{2}\,\mathrm{erf}^{-1}\Big(t\big(1-\mathrm{erf}(c/\sqrt{2})\big) + \mathrm{erf}(c/\sqrt{2})\Big).$$

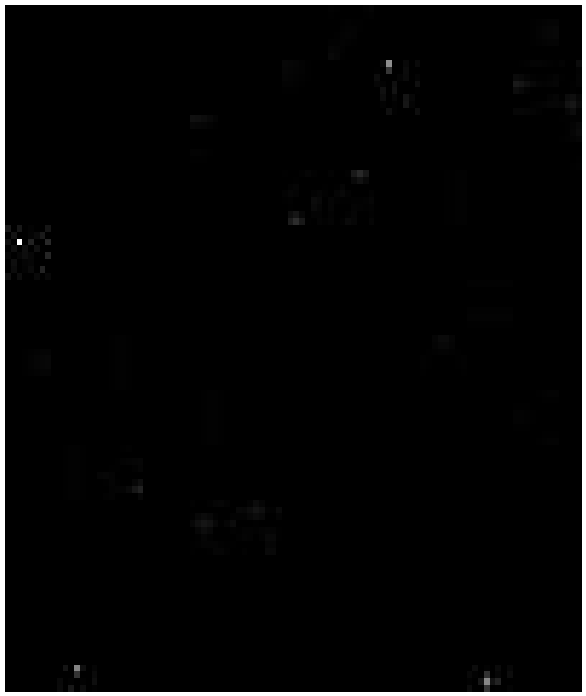
A similar derivation for $c = 0$ and the variance α^2 instead of 1 yields in the current case that

$$\Phi^{-1}(t) = \sqrt{2}\alpha\,\mathrm{erf}^{-1}(t).$$

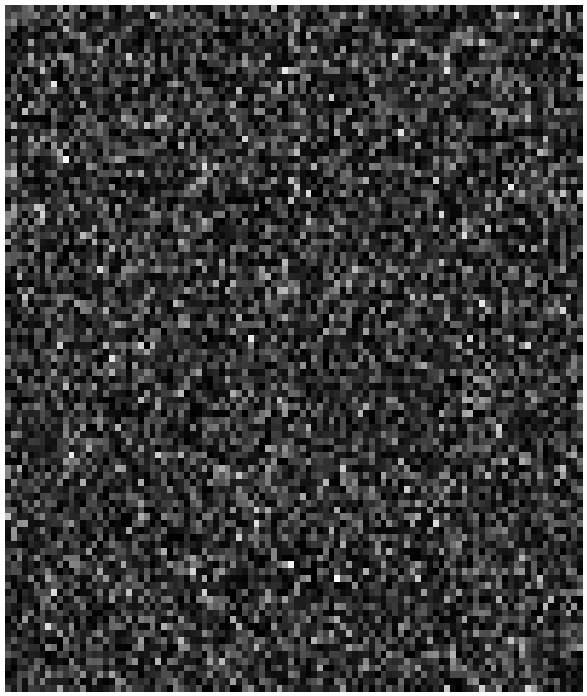
L_1 prior



Cauchy prior



White noise prior



Discontinuities

Prior information: The unknown is a function of, say, time. It is known to be relatively stable for long periods of time, but contains now and then discontinuities. We may also have information on the size of the jumps or the rate of occurrence of the discontinuities.

A more concrete example: Unknown is a function $f : [0, 1] \to \mathbb{R}$. We know that $f(0) = 0$ and that the function may have large jumps at a few locations.

After discretizing f, impulse priors can be used to construct a prior on the finite difference approximation of the *derivative* of f.

Discretization of the interval $[0,1]$: Choose grid points $t_j = j/N$, $j = 0, \ldots, N$, and set $x_j = f(t_j)$.

We write a Cauchy-type prior density

$$\pi_{\mathrm{pr}}(x) = \left(\frac{\alpha}{\pi}\right)^N \prod_{j=1}^{N} \frac{1}{1 + \alpha^2 (x_j - x_{j-1})^2}$$

that controls the jumps between the adjacent components of $x \in \mathbb{R}^{N+1}$. In particular, the components of X are not independent. (In addition to this prior, we know that $X_0 = x_0 = 0$.)

To make draws from the above density, we define new variables

$$\xi_j = x_j - x_{j-1}, \qquad 1 \le j \le N,$$

which are the changes in the function of interest between adjacent grid points.

Notice that $\tilde{x} = [x_1, \ldots, x_N]^{\mathrm{T}} \in \mathbb{R}^N$ satisfies

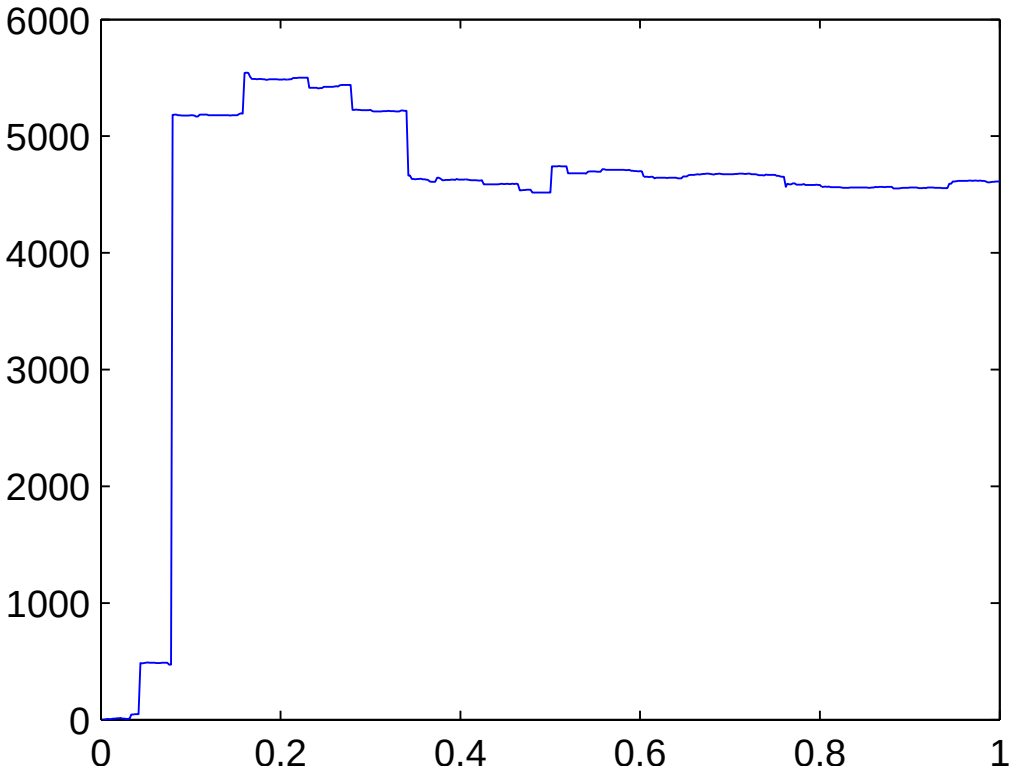
$$\tilde{x} = A\xi,$$

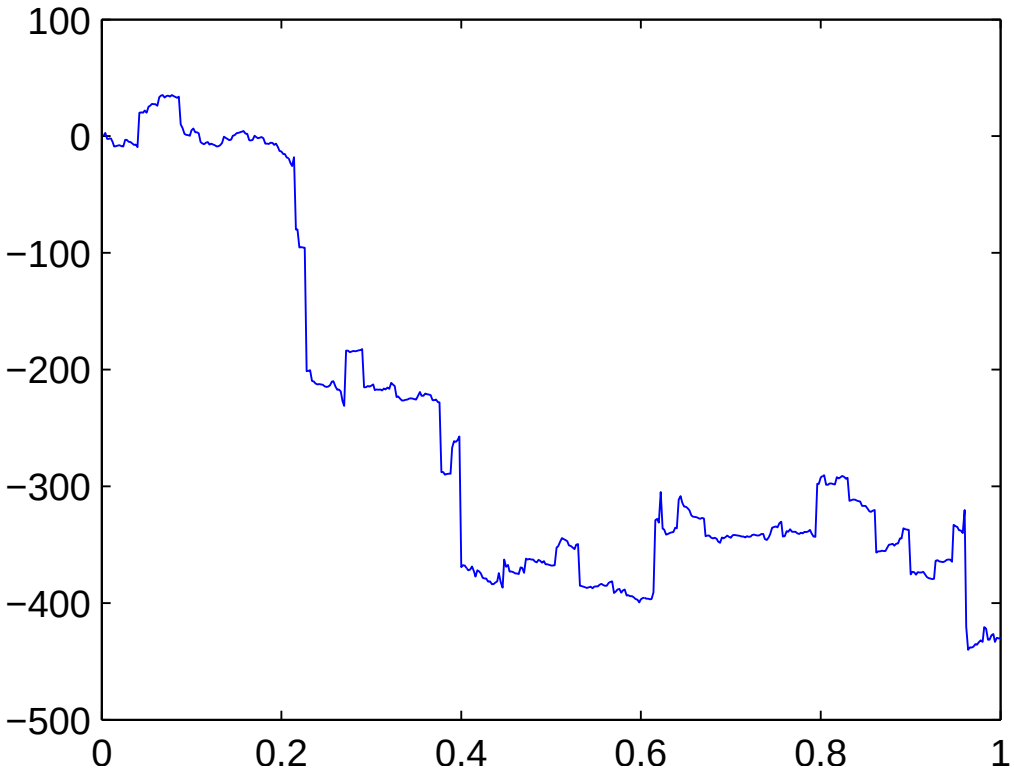
where $A \in \mathbb{R}^{N \times N}$ is a lower triangular matrix such that $A_{jk} = 1$ for $j \geq k$. Hence, it follows, e.g., from the change of variables rule for probability densities that

$$\pi_{\mathrm{pr}}(\xi) = \left(\frac{\alpha}{\pi}\right)^N \prod_{j=1}^{N} \frac{1}{1 + \alpha^2 \xi_j^2}.$$

In particular, due to the product form of $\pi_{\mathrm{pr}}(\xi)$, the components of Ξ are mutually independent, and can thus be drawn from a one-dimensional Cauchy density.

Subsequently, a random draw from the distribution of X can be constructed by recalling that $x_0 = 0$ and using the relation $\tilde{x} = A\xi$.


6000
5000
4000
3000
2000
1000
0
0
0.2
0.4
0.6
0.8
1


100
0
−100
−200
−300
−400
−500
0
0.2
0.4
0.6
0.8
1

Sample-based densities

Assume that we have a large sample of realizations of a random variable $X \in \mathbb{R}^n$:

$$S = \{x^1, x^2, \ldots, x^N\}.$$

One way to construct a prior density for X is to approximate $\pi(x)$ based on S.

Estimates of the mean and the covariance:

$$E\{X\} \approx \frac{1}{N} \sum_{j=1}^{N} x^j =: \bar{x},$$

$$\mathrm{cov}(X) = E\{XX^{\mathrm{T}}\} - E\{X\}E\{X\}^{\mathrm{T}} \approx \frac{1}{N} \sum_{j=1}^{N} x^j (x^j)^{\mathrm{T}} - \bar{x}\bar{x}^{\mathrm{T}} =: \Gamma.$$

(Notice that Γ is not the unbiased sample covariance estimator, but let us anyway follow the notation of the text book.)

The eigenvalue decomposition of Γ is

$$\Gamma = UDU^{\mathrm{T}},$$

where $U \in \mathbb{R}^{n \times n}$ is orthogonal and has the eigenvectors of Γ as its columns, and $D \in \mathbb{R}^{n \times n}$ is diagonal with the eigenvalues $d_1 \geq \ldots \geq d_n \geq 0$ as its diagonal entries. (Note that Γ is clearly symmetric and positive semi-definite, and thus it has a full set of eigenvectors with non-negative eigenvalues.)

The vectors x^j, $j = 1, \ldots, N$, are typically 'somewhat similar' and the matrix Γ can consequently be singular or almost singular: The eigenvalues often satisfy $d_j \approx 0$ for $j > r$, where $1 < r < n$ is some cut-off index. In other words, the difference $X - E\{X\}$ does not seem to vary much in the direction of the eigenvectors $u_{r+1}, \ldots, u_n$.

Assume this is the case. Then, one can postulate that the values of the random variable $X - E(X)$ lie 'with a high probability' in the subspace spanned by the first r eigenvectors of Γ. One way of trying to state this information quantitatively, is to introduce a *subspace prior*

$$\pi(x) \propto \exp\left(-\alpha \|(1-P)(x-\bar{x})\|^2\right),$$

where P is the orthogonal projector $\mathbb{R}^n \to \mathrm{span}\{u_1, \ldots, u_r\}$. The parameter $\alpha > 0$ controls how much $X - \bar{x}$ is allowed to vary from the subspace $\mathrm{span}\{u_1, \ldots, u_r\}$. (Take note that such a subspace prior is not a probability density in the traditional sense.)

If Γ is not almost singular, the inverse Γ^{-1} can be computed stably. In this case, the most straightforward way of approximating the (prior) probability density of X is to introduce the Gaussian approximation:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}(x-\bar{x})^{\mathrm{T}}\Gamma^{-1}(x-\bar{x})\right).$$

Depending on the higher order statistics of X, this may or may not provide a good approximation for the distribution of X.

Posterior density and a simple linear model

Consider a linear system of equations with noisy right hand side,

$$y = Ax + e, \qquad x \in \mathbb{R}^n,\ y, e \in \mathbb{R}^m,\ A \in \mathbb{R}^{m \times n}.$$

The corresponding stochastic extension reads

$$Y = AX + E,$$

where X, Y and E are random variables.

A very common assumption: X and E are independent and Gaussian,

$$X \sim \mathcal{N}(0, \gamma^2 \Gamma), \qquad E \sim \mathcal{N}(0, \sigma^2 I),$$

where we have assumed that both X and E have zero mean. (If this was not the case, the means could be subtracted from the respective random variables.)

The covariance of the noise indicates that the components of Y are contaminated by independent and identically distributed Gaussian random variables of variance σ^2. On the other hand, the prior distribution of X is assumed to have a bit more structure: Γ need not be diagonal and the parameter γ^2 is introduced for controlling the 'magnitude' of the (prior) covariance.

In other words, the prior density is of the form

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2\gamma^2} x^{\mathrm{T}} \Gamma^{-1} x\right),$$

and assuming that the noise level σ^2 is known, the likelihood function reads as

$$\pi(y \,|\, x) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - Ax\|^2\right).$$

It follows from the Bayes formula that the posterior density is

$$\begin{aligned}\pi(x \,|\, y) &\propto \pi_{\mathrm{pr}}(x)\pi(y \,|\, x) \\ &\propto \exp\left(-\frac{1}{2\gamma^2}x^{\mathrm{T}}\Gamma^{-1}x - \frac{1}{2\sigma^2}\|y - Ax\|^2\right) \\ &= \exp(-V(x \,|\, y)),\end{aligned}$$

where

$$V(x \,|\, y) = \frac{1}{2\gamma^2}x^{\mathrm{T}}\Gamma^{-1}x + \frac{1}{2\sigma^2}\|y - Ax\|^2.$$

If Γ is symmetric and positive definite, so is Γ^{-1}. Hence, we can introduce a Cholesky factorization:

$$\Gamma^{-1} = R^{\mathrm{T}} R.$$

With this notation,

$$x^{\mathrm{T}} \Gamma^{-1} x = x^{\mathrm{T}} R^{\mathrm{T}} R x = \|Rx\|^2,$$

and we define

$$T(x) = 2\sigma^2 V(x \,|\, y) = \|y - Ax\|^2 + \delta \|Rx\|^2, \qquad \delta := \frac{\sigma^2}{\gamma^2}.$$

The functional T is sometimes referred to as the *Tikhonov functional*.

Recall that the *maximum a posteriori (MAP)* estimator maximizes the posterior probability density of the unknowns:

$$x_{\mathrm{MAP}} = \arg\max_{x \in \mathbb{R}^n} \pi(x \,|\, y).$$

In our setting,

$$x_{\mathrm{MAP}} = \arg\min V(x \,|\, y) \quad \text{because} \quad V(x \,|\, y) = -\log \pi(x \,|\, y).$$

With the help of the Tikhonov functional, this reads

$$x_{\mathrm{MAP}} = \arg\min T(x) = \arg\min \left(\|y - Ax\|^2 + \delta \|Rx\|^2 \right).$$

Recall that the Tikhonov regularized solution of $y = Ax$ — with the penalty term $\|Rx\|$ — is the minimizer of $T(x)$. In consequence, the Tikhonov regularized solution and x_{MAP} coincide if the regularization parameter is chosen to be $\delta = \sigma^2/\gamma^2$.

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Fourteenth lecture, March 4, 2011.

Example: Laplace transform (revisited)

Recall the problem of finding a function f from noisy samples of its Laplace transform. This problem was discussed at the ninth lecture and solved using various classical regularization techniques.

We take another look at the problem, and interpret its Tikhonov regularized solution from the statistical viewpoint.

Laplace transform

Let $f : [0, \infty) \to \mathbb{R}$ be some unknown function and assume that we have access to noisy samples of its Laplace transform

$$\mathcal{L}f(s) = \int_0^\infty e^{-st} f(t)\, dt, \qquad s \geq 0,$$

at some measurement points s_j, $j = 1, \ldots, m$. The task is to approximate f using the noisy values $\{\mathcal{L}f(s_j)\}_{j=1}^m$ as data.

Observe that for large t the kernel e^{-st} is typically very small, and hence the 'tail' of f does not affect the Laplace transform as much as its values close to the origin. In consequence, reconstructing f is an ill-posed inverse problem.

Discretization

In order to come up with a computational model, we approximate the integral of the Laplace transform as

$$\mathcal{L}f(s_j) \approx \int_0^T e^{-s_j t} f(t)\, dt \approx \sum_{k=1}^{n} w_k e^{-s_j t_k} f(t_k), \qquad j = 1, \ldots, m,$$

where $t_1, \ldots, t_n \in [0, T]$ are the nodes and $w = (w_1, \ldots, w_n)^{\mathrm{T}} \in \mathbb{R}^n$ the corresponding weights of the chosen quadrature rule. Notice that it is implicitly assumed that $e^{-st} f(t)$ is 'small' for all t that are larger than the threshold $T > 0$.

For example, if we decided to use the trapezoid rule on an equidistant mesh in the interval $[0, T]$, we would choose $h = T/(n-1)$ and

$$w = (h/2, h, h, \ldots, h, h, h/2)^{\mathrm{T}} \qquad \text{and} \qquad t_k = (k-1)h$$

for $k = 1, \ldots, n$.

The above quadrature rule can be written in the matrix form

$$y = Ax,$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are given by

$$\begin{aligned} x &= (f(t_1), \ldots, f(t_n))^{\mathrm{T}} \\ y &= (\mathcal{L}f(s_1), \ldots, \mathcal{L}f(s_m))^{\mathrm{T}}, \end{aligned}$$

and the elements of the matrix $A \in \mathbb{R}^{m \times n}$ are defined as

$$(A)_{jk} = w_k e^{-s_j t_k}, \qquad j = 1, \ldots, m, \quad k = 1, \ldots, n.$$

In the following numerical examples, we choose $m = 91$ sampling points on a logarithmic grid:

$$\log s_j \;=\; -\log 10 + 2\frac{(j-1)}{m-1}\log 10, \qquad j = 1, \ldots, m,$$

where $\log$ denotes the natural logarithm. Now, the points $\{\log s_j\}_{j=1}^m$ form a uniform grid in the interval $[-\log(10), \log(10)]$, and thus $\{s_j\}_{j=1}^m$ lie in the interval $[0.1, 10]$, with half of the points between 0.1 and 1. This reflects our knowledge that the information in the Laplace transform is — *very loosely speaking* — concentrated close to the origin.

We set $n = 101$ and choose the nodes $\{t_k\}_{k=1}^n$ and the weights $w \in \mathbb{R}^n$ according to the Gauss–Legendre quadrature rule in the interval $[0, 5]$. (One could use something less sophisticated, such as trapezoid rule in this same interval, as well.)

Simulation of data

We choose

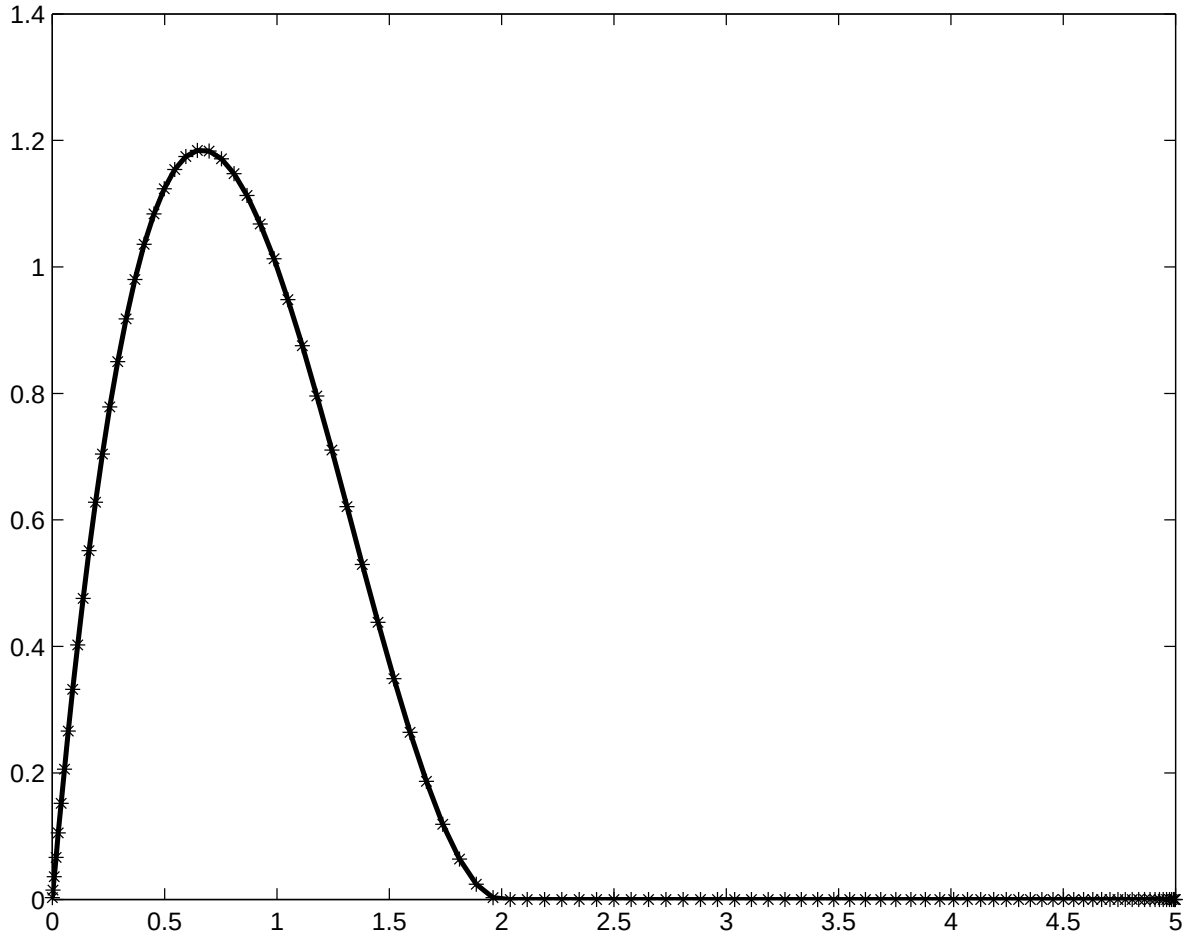
$$f(t) = \begin{cases} t^3 - 4t^2 + 4t, & 0 \le t < 2, \\ 0, & t \ge 2. \end{cases}$$

In this simple case, the Laplace transform can be calculated explicitly with the help of partial integration:

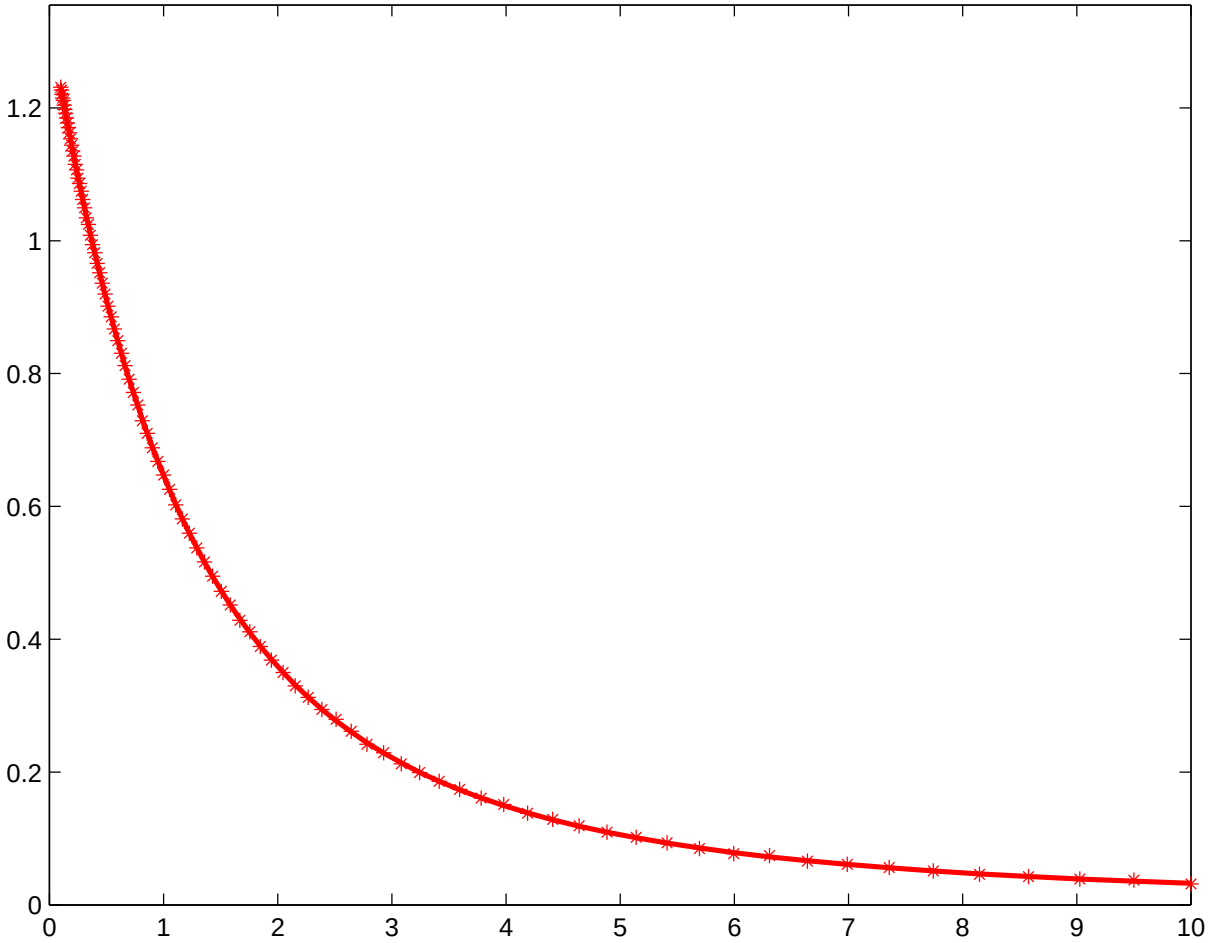
$$\mathcal{L}f(s) = \frac{4}{s^2} - \frac{4}{s^3}(2 + e^{-2s}) + \frac{6}{s^4}(1 - e^{-2s}), \qquad s > 0.$$

Consequently, we just compute the value of $\mathcal{L}f(s)$ at the chosen sampling points $\{s_j\}_{j=1}^m$ using this formula, add realizations of a normally distributed random variable with zero mean and standard deviation 10^{-3} to each sample, plug the resulting data into the vector y, and we are ready to go.

Target function and the nodes



Laplace transform and the noisy measurements



Tikhonov regularized solution

We consider the above introduced discretized "inverse Laplace transform problem"

$$Ax = y.$$

Recall that the Tikhonov regularized solution $x_\delta \in \mathbb{R}^n$ is the unique minimizer of the Tikhonov functional

$$\|Ax - y\|^2 + \delta\|x\|^2, \qquad \delta > 0.$$

It is given explicitly by the formula

$$x_\delta = (A^{\mathrm{T}}A + \delta I)^{-1}A^{\mathrm{T}}y.$$

According to the Morozov discrepancy principle a feasible choice for the regularization parameter is such $\delta = \delta_{\mathrm{Mor}}$ that the corresponding solution satisfies

$$\|y - Ax_{\delta_{\mathrm{Mor}}}\| \approx \epsilon = 10^{-3} \cdot \sqrt{m} \approx 9.5 \cdot 10^{-3}.$$

Statistical model

Let us introduce the stochastic extension

$$Y = AX + E,$$

where $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$ and $E \in \mathbb{R}^m$ are random variables. We assume that X and E are independent and Gaussian,

$$X \sim \mathcal{N}(0, \gamma^2 I), \qquad E \sim \mathcal{N}(0, \sigma^2 I).$$

Recall from the previous lecture that with these assumptions the maximum a posteriori estimate

$$x_{\mathrm{MAP}} = \arg\max \pi(x \,|\, y)$$

is given as

$$x_{\mathrm{MAP}} = \arg\min \left(\|y - Ax\|^2 + \delta \|x\|^2 \right), \qquad \delta = \frac{\sigma^2}{\gamma^2}.$$

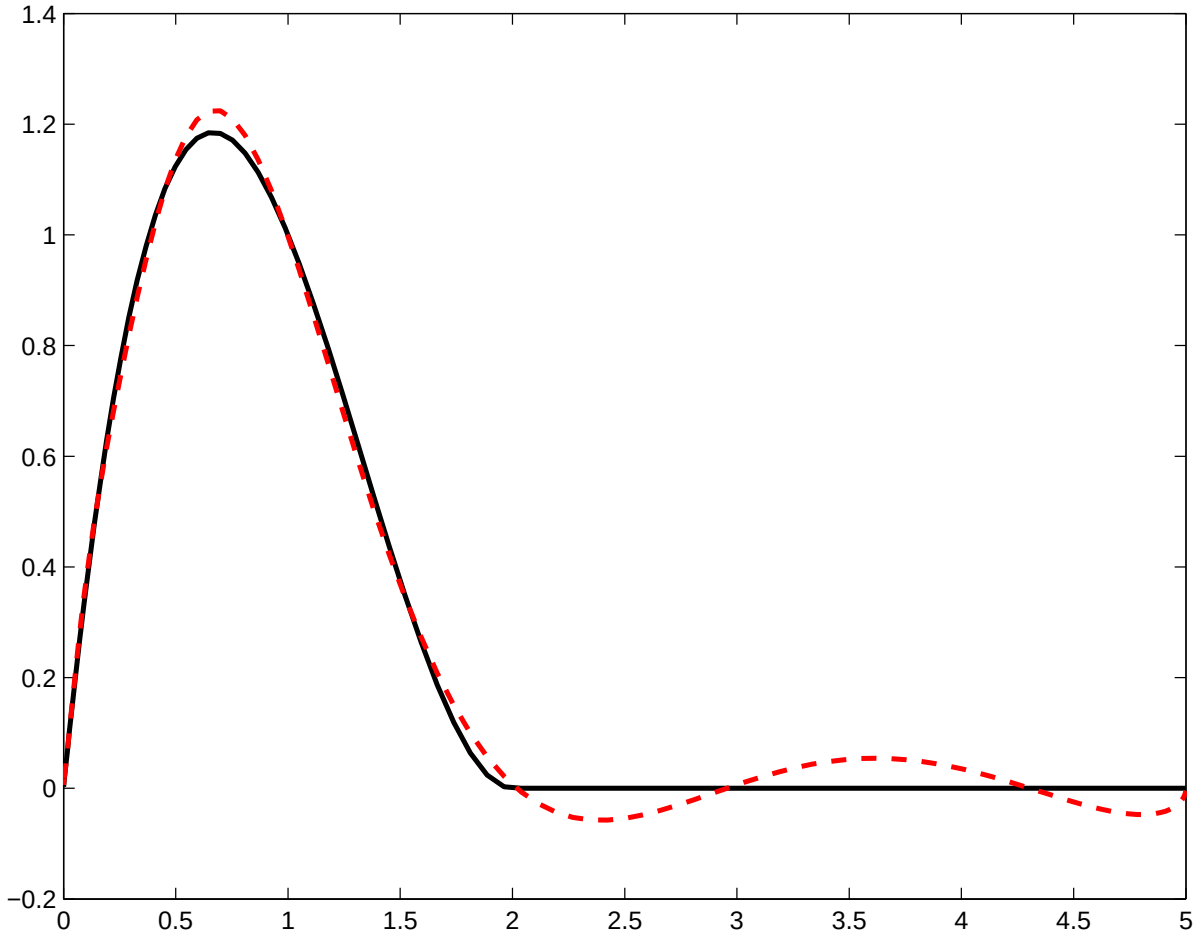
Suppose that we know the noise level, i.e., $\sigma = 10^{-3}$.

Then, we still need to choose the standard deviation (or the variance) of the prior density based on our *a priori* information on the unknown function f. If we believe that the order of magnitude of the values of f is, say, one, a suitable choice for γ could be, e.g., $\gamma = 1$ or $\gamma = 0.5$. (Note that our prior mean is set to zero.)

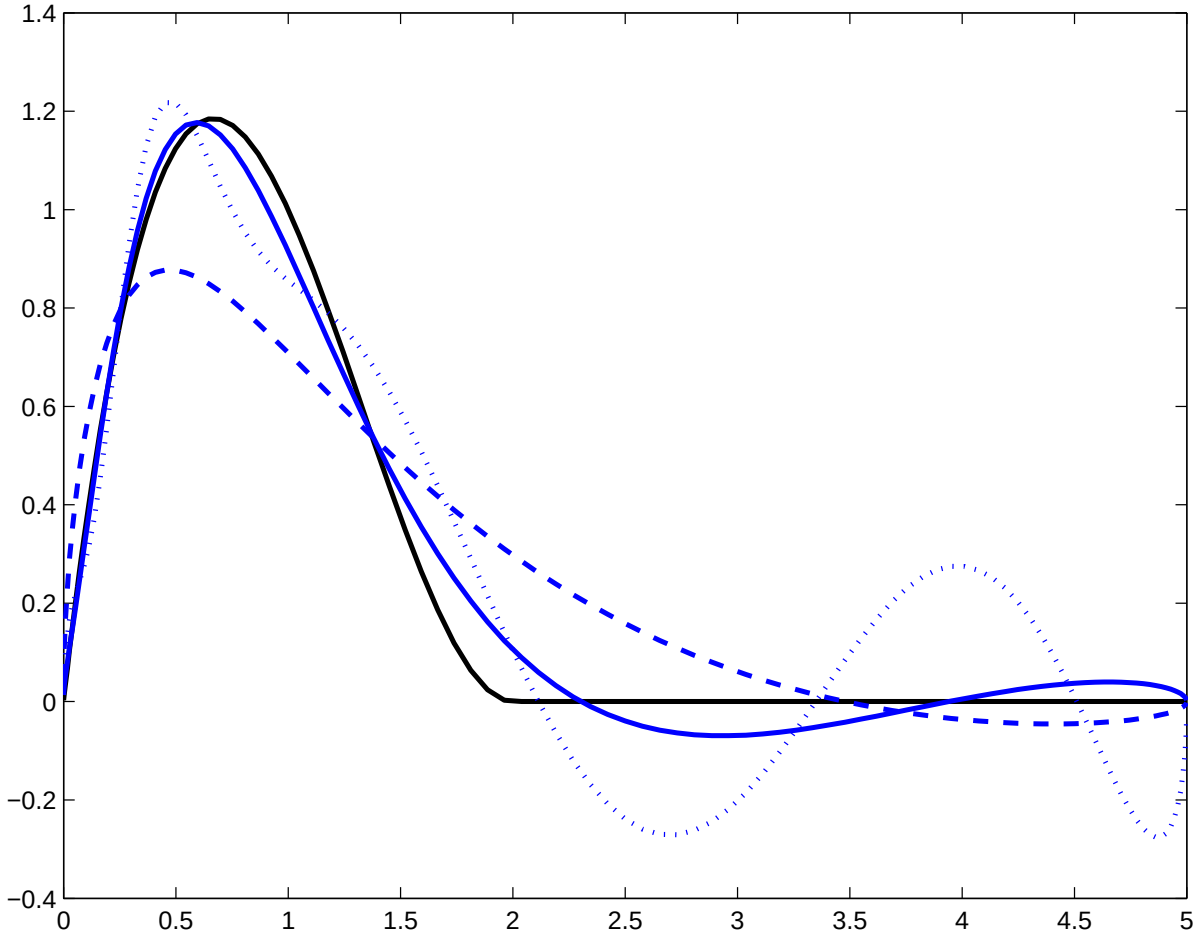
With $\gamma = 0.5$ we get $\delta = \frac{\sigma^2}{\gamma^2} = 4 \cdot 10^{-6}$.

How does the corresponding Tikhonov regularized solution look like?

Tikhonov regularized solution with $\delta = 4 \cdot 10^{-6}$



Traditional Tikhonov with $\delta = \delta_{\mathrm{Mor}} \approx 3.6 \cdot 10^{-5}$ (solid), $\delta = 10^{3} \cdot \delta_{\mathrm{Mor}}$ (slashed) and $\delta = 10^{-3} \cdot \delta_{\mathrm{Mor}}$ (dotted)



The previous test cases presented at the ninth lecture correspond to the following choices of the prior standard deviation:

$$\delta = \delta_{\mathrm{Mor}} \quad \Longrightarrow \quad \gamma = 0.167,$$

$$\delta = 10^3 \cdot \delta_{\mathrm{Mor}} \quad \Longrightarrow \quad \gamma = 0.00527,$$

$$\delta = 10^{-3} \cdot \delta_{\mathrm{Mor}} \quad \Longrightarrow \quad \gamma = 5.27.$$

n-variate Gaussian densities

Definition. *Let*

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

be a positive definite and symmetric matrix, with $\Gamma_{11} \in \mathbb{R}^{k\times k}$, $k < n$, $\Gamma_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, *and* $\Gamma_{21} = \Gamma_{12}^{\mathrm{T}} \in \mathbb{R}^{(n-k)\times k}$. *We define the Schur complement* $\tilde{\Gamma}_{jj}$ *of* Γ_{jj}, $j = 1, 2$, *by the formulas*

$$\tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}, \qquad \tilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}$$

Observe that the definition of Γ implies that Γ_{jj}, $j = 1, 2$, are symmetric, positive definite and, in particular, invertible. In consequence, the Schur complements are well defined and symmetric.

Lemma. *Let Γ be a matrix that satisfies the assumptions of the previous definition. Then, the Schur complements $\tilde{\Gamma}_{jj}$, $j = 1, 2$, are invertible matrices and, furthermore,*

$$\Gamma^{-1} = \begin{bmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \tilde{\Gamma}_{11}^{-1} \end{bmatrix}.$$

Proof: We prove first that the Schur complements are invertible: Consider the determinant of Γ,

$$|\Gamma| = \begin{vmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{vmatrix} \neq 0.$$

By subtracting the first row multiplied by $\Gamma_{21}\Gamma_{11}^{-1}$ from the second one, we find that

$$|\Gamma| = \begin{vmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12} \end{vmatrix} = |\Gamma_{11}||\tilde{\Gamma}_{11}|,$$

implying that $|\tilde{\Gamma}_{11}| \neq 0$. In the same way, we can also show that $|\tilde{\Gamma}_{22}| \neq 0$.

The proof of the second assertion of the lemma follows from the Gaussian elimination: Consider the linear system

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

By solving for x_2 in the second equation, we get

$$x_2 = \Gamma_{22}^{-1}(y_2 - \Gamma_{21}x_1).$$

Substituting this formula into the first equation, then gives us

$$(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21})x_1 = y_1 - \Gamma_{12}\Gamma_{22}^{-1}y_2,$$

or equivalently

$$x_1 = \tilde{\Gamma}_{22}^{-1}y_1 - \tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}y_2,$$

which verifies the first row of claimed representation of Γ^{-1}. The second row of the representation follows by reversing the roles of x_1 and x_2.

Remark: Since Γ is a symmetric matrix, so is Γ^{-1}. In consequence, we have the identity

$$\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} = (\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1})^{\mathrm{T}} = \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}\,.$$

Theorem. *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two Gaussian random variables whose joint probability density $\pi\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ is of the form*

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}\begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}\right).$$

Then, the probability density of X conditioned on $Y = y$, i.e., $\pi(x \,|\, y)\colon \mathbb{R}^n \to \mathbb{R}_+$, is of the form

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2}(x-\bar{x})^{\mathrm{T}} \tilde{\Gamma}_{22}^{-1} (x-\bar{x})\right),$$

where

$$\bar{x} = x_0 + \Gamma_{12}\Gamma_{22}^{-1}(y-y_0).$$

Proof: For simplicity, let us assume that $x_0 = 0$ and $y_0 = 0$.

Due the representation of the joint covariance matrix Γ^{-1} provided by the previous Lemma and the remark that followed, we may write

$$\begin{aligned}\pi(x,y) &\propto \exp\left(-\frac{1}{2}\big(x^{\mathrm{T}}\tilde{\Gamma}_{22}^{-1}x - 2x^{\mathrm{T}}\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}y + y^{\mathrm{T}}\tilde{\Gamma}_{11}^{-1}y\big)\right) \\ &= \exp\left(-\frac{1}{2}\big((x-\Gamma_{12}\Gamma_{22}^{-1}y)^{\mathrm{T}}\tilde{\Gamma}_{22}^{-1}(x-\Gamma_{12}\Gamma_{22}^{-1}y) + c\big)\right),\end{aligned}$$

where $c = y^{\mathrm{T}}(\tilde{\Gamma}_{11}^{-1} - \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1})y$. Hence, it follows that

$$\pi(x\,|\,y) \propto \pi(x,y) \propto \exp\left(-\frac{1}{2}(x-\Gamma_{12}\Gamma_{22}^{-1}y)^{\mathrm{T}}\tilde{\Gamma}_{22}^{-1}(x-\Gamma_{12}\Gamma_{22}^{-1}y)\right),$$

where the proportionality constants depend on y but not on x. This proves the claim. □

Theorem. *Let X and Y be Gaussian random variables with a joint probability density as in the previous theorem. Then, the marginal density of X is*

$$\pi(x) = \int_{\mathbb{R}^m} \pi(x, y) dy \propto \exp\left(-\frac{1}{2} (x - x_0)^{\mathrm{T}} \Gamma_{11}^{-1} (x - x_0) \right).$$

Proof: The proof is slightly more complicated than the previous one. It can be found in the textbook by Kaipio and Somersalo.

Linear inverse problem

Assume that we have a linear model with additive noise,

$$Y = AX + E,$$

where $A \in \mathbb{R}^{m \times n}$ is a known matrix, and $X \in \mathbb{R}^n$ and $Y, E \in \mathbb{R}^m$ are random variables. Assume furthermore that X and E are mutually independent Gaussian variables with probability densities

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1}(x - x_0)\right),$$

and

$$\pi_{\mathrm{noise}}(x) \propto \exp\left(-\frac{1}{2}(e - e_0)^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1}(e - e_0)\right).$$

With this information, we get from the Bayes formula that the posterior distribution of X conditioned on $Y = y$ is

$$\pi(x \,|\, y) \propto \pi_{\mathrm{pr}}(x)\pi(y \,|\, x) = \pi_{\mathrm{pr}}(x)\pi_{\mathrm{noise}}(y - Ax)$$

$$\propto \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x - x_0) - \frac{1}{2}(y - Ax - e_0)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - Ax - e_0)\right)$$

The explicit form of this posterior distribution, i.e., the form that shows the posterior mean and covariance explicitly, can be calculated in a straightforward but tedious manner by 'completing the squares' with respect to x. However, we may also use the first of the two theorems presented on the previous few slides.

Since X and E are Gaussian, so is Y, and we have

$$E\left\{\begin{bmatrix} X \\ Y \end{bmatrix}\right\} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \qquad y_0 = Ax_0 + e_0$$

Furthermore, using the fact that X and E are independent, we deduce that

$$E\left\{(X - x_0)(X - x_0)^{\mathrm{T}}\right\} = \Gamma_{\mathrm{pr}},$$

$$\begin{aligned} E\left\{(Y - y_0)(Y - y_0)^{\mathrm{T}}\right\} &= E\left\{\big(A(X - x_0) + (E - e_0)\big)\big(A(X - x_0) + (E - e_0)\big)^{\mathrm{T}}\right\} \\ &= A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}}, \end{aligned}$$

$$\begin{aligned} E\left\{(X - x_0)(Y - y_0)^{\mathrm{T}}\right\} &= E\left\{(X - x_0)\big(A(X - x_0) + (E - e_0)\big)^{\mathrm{T}}\right\} \\ &= \Gamma_{\mathrm{pr}}A^{\mathrm{T}}. \end{aligned}$$

Hence, we get

$$\operatorname{cov}\begin{bmatrix} X \\ Y \end{bmatrix} = E\left\{ \begin{bmatrix} X - x_0 \\ Y - y_0 \end{bmatrix} \begin{bmatrix} X - x_0 \\ Y - y_0 \end{bmatrix}^{\mathrm{T}} \right\} = \begin{bmatrix} \Gamma_{\mathrm{pr}} & \Gamma_{\mathrm{pr}} A^{\mathrm{T}} \\ A\Gamma_{\mathrm{pr}} & A\Gamma_{\mathrm{pr}} A^{\mathrm{T}} + \Gamma_{\mathrm{noise}} \end{bmatrix}.$$

The joint probability density of X and Y is thus of the form

$$\pi(x, y) \propto \exp\left(-\frac{1}{2} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Gamma_{\mathrm{pr}} & \Gamma_{\mathrm{pr}} A^{\mathrm{T}} \\ A\Gamma_{\mathrm{pr}} & A\Gamma_{\mathrm{pr}} A^{\mathrm{T}} + \Gamma_{\mathrm{noise}} \end{bmatrix}^{-1} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \right).$$

Using the first of the above two theorems, we can thus write the posterior density of X conditioned on $Y = y$.

Theorem. *Assume that $X \in \mathbb{R}^n$ and $E \in \mathbb{R}^m$ are mutually independent Gaussian random variables,*

$$X \sim \mathcal{N}(x_0, \Gamma_{\text{pr}}), \quad E \sim \mathcal{N}(e_0, \Gamma_{\text{noise}})$$

and $\Gamma_{\text{pr}} \in \mathbb{R}^{n \times n}$ and $\Gamma_{\text{noise}} \in \mathbb{R}^{m \times m}$ are positive definite. Assume further that we have a linear model $Y = AX + E$ for a noisy measurement Y, where $A \in \mathbb{R}^{m \times n}$ is a known matrix. Then, the posterior probability density of X given the measurement $Y = y$ is

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2}(x - \bar{x})^{\mathrm{T}} \Gamma_{\text{post}}^{-1} (x - \bar{x})\right),$$

where

$$\bar{x} = x_0 + \Gamma_{\text{pr}} A^{\mathrm{T}} (A \Gamma_{\text{pr}} A^{\mathrm{T}} + \Gamma_{\text{noise}})^{-1} (y - A x_0 - e_0),$$

and

$$\Gamma_{\text{post}} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} A^{\mathrm{T}} (A \Gamma_{\text{pr}} A^{\mathrm{T}} + \Gamma_{\text{noise}})^{-1} A \Gamma_{\text{pr}}.$$

Remark: It holds that

$$\Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{post}} = \Gamma_{\mathrm{pr}} A^{\mathrm{T}} (A \Gamma_{\mathrm{pr}} A^{\mathrm{T}} + \Gamma_{\mathrm{noise}})^{-1} A \Gamma_{\mathrm{pr}},$$

which is a positive semi-definite matrix. Loosely speaking, this means that the prior density is wider than the posterior, i.e., the measurement decreases the uncertainty in the whereabouts of X.

Remark: As already mentioned, the explicit forms of the mean and the covariance of the Gaussian posterior density for this linear model can also be derived directly. This way we get alternative representations for the posterior covariance matrix

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} A)^{-1}$$

and the posterior mean

$$\bar{x} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} A)^{-1} (A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} (y - e_0) + \Gamma_{\mathrm{pr}}^{-1} x_0).$$

Gaussian white noise prior and Tikhonov regularization

Consider the simple *Gaussian white noise prior* case, $X \sim \mathcal{N}(0, \gamma^2 I)$, and assume also that the noise is white noise, i.e., $E \sim (0, \sigma^2 I)$. In this particular case the mean of the posterior distribution given by the above theorem turns into

$$\bar{x} = \gamma^2 A^{\mathrm{T}} (\gamma^2 A A^{\mathrm{T}} + \sigma^2)^{-1} y = A^{\mathrm{T}} (A A^{\mathrm{T}} + \delta I)^{-1} y,$$

where $\delta = \sigma^2 / \gamma^2$.

It can be shown (the seventh exercise session) that this form is equivalent to the Tikhonov regularized solution

$$x_\delta = (A^{\mathrm{T}} A + \delta I)^{-1} A^{\mathrm{T}} y,$$

which is not very surprising, as we have already deduced at the previous lecture that $x_{\mathrm{MAP}} = x_\delta$ for $\delta = \sigma^2 / \gamma^2$ and, on the other hand, $x_{\mathrm{CM}} = x_{\mathrm{MAP}}$ for a Gaussian posterior distribution.

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Fifteenth lecture, March 16, 2011.

Improper Gaussian priors

Motivation: Smoothness priors

Recall from the thirteenth lecture that finding the *maximum a posterior* (MAP) — or *conditional mean* (CM) — estimate for the linear inverse problem

$$Y = AX + E, \qquad Y, E \in \mathbb{R}^m,\ X \in \mathbb{R}^n,$$

where X and E are independent and Gaussian with zero mean,

$$X \sim \mathcal{N}(0, \Gamma), \qquad E \sim \mathcal{N}(0, \sigma^2 I),$$

is equivalent to minimizing the Tikhonov functional

$$T(x) = \|y - Ax\|^2 + \sigma^2 \|Rx\|^2,$$

where R satisfies $\Gamma^{-1} = R^{\mathrm{T}} R$. (The matrix R can be, e.g., the Cholesky factor of the positive definite and symmetric matrix Γ^{-1}.)

Let us then try to work our way in the opposite direction: Consider the corresponding classical linear inverse problem

$$Ax = y,$$

and let us solve it using Tikhonov regularization under the prior knowledge that $x \in \mathbb{R}^n$ represents point values of a smooth function.

We try to incorporate this extra information in the solution process by using a 'smoothness penalty term' for the Tikhonov functional:

$$T(x) = \|y - Ax\|^2 + \delta\|Lx\|^2,$$

where $L \in \mathbb{R}^{k\times n}$ is a discrete approximation of some suitable differential operator.

If you now compare the two Tikhonov functionals on the previous two slides, it seems natural that the Gaussian stochastic extension corresponding to the smoothness penalty approach would be

$$Y = AX + E,$$

with

$$X \sim \mathcal{N}(0, (L^{\mathrm{T}}L)^{-1}), \qquad E \sim \mathcal{N}(0, \sigma^2 I),$$

where $\sigma^2 = \delta$.

Unfortunately, there is a slight flaw in this logic: In order for the inverse $(L^{\mathrm{T}}L)^{-1}$ to exist — and to be positive definite — the matrix $L \in \mathbb{R}^{k \times n}$ needs to be injective, which is not always the case. (As an example, quite often $Lx = 0$ if all elements of x are the same.)

Due to this observation, we will next consider *improper densities* of the form:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}\|L(x-x_0)\|^2\right) = \exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}L^{\mathrm{T}}L(x-x_0)\right),$$

where $L \in \mathbb{R}^{k\times n}$ is a given, possible non-injective matrix.

We will tackle the problem of interpreting such densities as Gaussian priors in three different ways:

1. by introducing a proper density that is 'close' to the considered improper density,
2. by noting that the posterior density may be proper even if the prior is improper, and
3. by using conditioning to update improper priors so that they become proper.

Approximate proper densities

Recall from the first part of the course that any $L \in \mathbb{R}^{k \times n}$ has a singular value decomposition $L = U \Lambda V^{\mathrm{T}}$, where $U \in \mathbb{R}^{k \times k}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and the diagonal matrix $\Lambda \in \mathbb{R}^{k \times n}$ contains the non-negative singular values

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > \lambda_{p+1} = \ldots = \lambda_l = 0, \qquad l := \min(k, n).$$

Moreover, recall that the columns $\{v_1, \ldots, v_n\}$ of V satisfy

$$\mathrm{Ker}(L) = \mathrm{span}\{v_{p+1}, \ldots, v_n\},$$

and let us define $Q = [v_{p+1}, \ldots, v_n] \in \mathbb{R}^{n \times (n-p)}$. In particular, it is easy to see that $QQ^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ is the orthogonal projection onto $\mathrm{Ker}(L)$.

We then define an auxiliary covariance matrix $\Gamma_a \in \mathbb{R}^{n\times n}$ via

$$\Gamma_a \;=\; L^{\dagger}(L^{\dagger})^{\mathrm{T}} + a^2 Q Q^{\mathrm{T}},$$

where $L^{\dagger} \in \mathbb{R}^{n\times k}$ is the pseudoinverse of L and $a > 0$ is an arbitrary (large) scalar.

Lemma. *The covariance matrix Γ_a defined above is positive definite. Moreover, its inverse can be written explicitly as*

$$\Gamma_a^{-1} \;=\; L^{\mathrm{T}} L + \frac{1}{a^2} Q Q^{\mathrm{T}}.$$

Let $x \in \mathbb{R}^n$ be arbitrary and write it in the orthonormal basis $\{v_1, \ldots, v_n\}$, i.e.,

$$x \;=\; \sum_{j=1}^{n} \alpha_j v_j, \qquad \alpha_j \in \mathbb{R},$$

Then,

$$\Gamma_a x = \sum_{j=1}^{p} \frac{\alpha_j}{\lambda_j^2} v_j + a^2 \sum_{j=p+1}^{n} \alpha_j v_j,$$

and thus

$$x^{\mathrm{T}} \Gamma_a x = \sum_{j=1}^{p} \frac{\alpha_j^2}{\lambda_j^2} + a^2 \sum_{j=p+1}^{n} \alpha_j^2 > 0$$

if $x \neq 0$, i.e., Γ_a is positive definite.

Moreover,

$$(L^{\mathrm{T}} L + \frac{1}{a^2} Q Q^{\mathrm{T}}) \Gamma_a x = \sum_{j=1}^{p} \alpha_j v_j + \sum_{j=p+1}^{n} \alpha_j v_j = x,$$

which proves that $\Gamma_a^{-1} = (L^{\mathrm{T}} L + \frac{1}{a^2} Q Q^{\mathrm{T}})$, as x was chosen arbitrarily.

Instead of choosing the improper prior

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}} L^{\mathrm{T}} L(x-x_0)\right),$$

one may consider resorting to the slightly modified version

$$\tilde{\pi}_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}} \Gamma_a^{-1}(x-x_0)\right),$$

which defines a proper Gaussian density because Γ_a is positive definite for any $a > 0$.

Let us next consider in which way these two densities are different; for simplicity assume that $x_0 = 0$.

Let $\mathcal{P} : \mathbb{R}^n \to \mathrm{Ker}(L)^\perp$ be an orthogonal projection, which means, in particular, that $I - \mathcal{P}$ is the orthogonal projection onto $\mathrm{Ker}(L)$, i.e., $I - \mathcal{P} = QQ^{\mathrm{T}}$. Trivial calculations show that

$$\pi_{\mathrm{pr}}(x) \;=\; \pi_{\mathrm{pr}}(\mathcal{P}x), \qquad x \in \mathbb{R},$$

and

$$\tilde{\pi}_{\mathrm{pr}}(x) \;\propto\; \pi_{\mathrm{pr}}(\mathcal{P}x) \exp\left(-\frac{1}{2a^2}\|(I-\mathcal{P})x\|^2\right).$$

In consequence, $\pi_{\mathrm{pr}}(x)$ is constant as a function of the component $(I - \mathcal{P})x$ of x, which makes it an improper prior. Moreover, the functional dependence of $\pi_{\mathrm{pr}}(x)$ and $\tilde{\pi}_{\mathrm{pr}}(x)$ on $\mathcal{P}x$ is the same, but $\tilde{\pi}_{\mathrm{pr}}(x)$ has also a 'density-like' dependence on $(I - \mathcal{P})x$. To sum up, the larger $a > 0$ is, the 'closer' these two densities are to each other.

Proper posteriors corresponding to improper priors

Recall the following theorem from the fourteenth lecture:

Theorem. *Assume that $X \in \mathbb{R}^n$ and $E \in \mathbb{R}^m$ are mutually independent Gaussian random variables, $X \sim \mathcal{N}(x_0, \Gamma_{\text{pr}})$, $E \sim \mathcal{N}(e_0, \Gamma_{\text{noise}})$, and that $\Gamma_{\text{pr}} \in \mathbb{R}^{n \times n}$ and $\Gamma_{\text{noise}} \in \mathbb{R}^{m \times m}$ are positive definite. Assume further that we have a linear model $Y = AX + E$ for a noisy measurement Y, where $A \in \mathbb{R}^{m \times n}$ is a known matrix. Then, the posterior probability density of X given the measurement $Y = y$ is*

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2}(x - \bar{x})^{\mathrm{T}} \Gamma_{\text{post}}^{-1} (x - \bar{x})\right),$$

where

$$\bar{x} = x_0 + \Gamma_{\text{pr}} A^{\mathrm{T}} (A \Gamma_{\text{pr}} A^{\mathrm{T}} + \Gamma_{\text{noise}})^{-1} (y - A x_0 - e_0),$$

and

$$\Gamma_{\text{post}} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} A^{\mathrm{T}} (A \Gamma_{\text{pr}} A^{\mathrm{T}} + \Gamma_{\text{noise}})^{-1} A \Gamma_{\text{pr}}.$$

When dealing with improper prior densities of the form

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}} L^{\mathrm{T}} L(x-x_0)\right),$$

this theorem is unfortunately useless in the construction of the posterior, because the natural candidate for the prior covariance, i.e., $(L^{\mathrm{T}}L)^{-1}$, does not typically exist.

However, recall that we also introduced alternative formulas for the posterior mean and covariance, namely

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A)^{-1},$$

and

$$\bar{x} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A)^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y-e_0) + \Gamma_{\mathrm{pr}}^{-1}x_0).$$

These formulas look more promising as they involve only $\Gamma_{\mathrm{pr}}^{-1}$, not Γ_{pr}.

For simplicity let us only consider the zero mean case:

Theorem. *Consider the linear observation model* $Y = AX + E$, $A \in \mathbb{R}^{m\times n}$, *where* $X \in \mathbb{R}^n$ *and* $E \in \mathbb{R}^m$ *are mutually independent random variables, of which* E *is proper Gaussian,* $E \sim \mathcal{N}(0, \Gamma_{\mathrm{noise}})$. *Let* $L \in \mathbb{R}^{k\times n}$ *be a matrix such that* $\mathrm{Ker}(L) \cap \mathrm{Ker}(A) = \{0\}$. *Then the function*

$$x \mapsto \pi_{\mathrm{pr}}(x)\pi(y \,|\, x) \propto \exp\left(-\frac{1}{2}\left(\|Lx\|^2 + (y - Ax)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - Ax)\right)\right)$$

defines a Gaussian density over $\mathbb{R}^n$, *with the corresponding covariance and mean given by the formulas*

$$\Gamma_{\mathrm{post}} = (L^{\mathrm{T}}L + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A)^{-1}, \qquad \bar{x} = \Gamma_{\mathrm{post}}A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y,$$

respectively.

Proof: Let us denote $G = L^{\mathrm{T}}L + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A \in \mathbb{R}^{n\times n}$ and let $x \in \mathbb{R}^n$ be arbitrary. Because $\Gamma_{\mathrm{noise}}^{-1}$ is positive definite, we have

$$x^{\mathrm{T}}Gx \;=\; \|Lx\|^2 + (Ax)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(Ax) \;\geq\; 0,$$

where the equality holds only if $x \in \mathrm{Ker}(L) \cap \mathrm{Ker}(A) = \{0\}$. In consequence, G is positive definite, meaning that $\Gamma_{\mathrm{post}} = G^{-1}$ is well-defined and also positive definite.

By completing the square with respect to x, the the quadratic functional in the exponent of the posterior density can be written as

$$\begin{aligned}\|Lx\|^2 + (y - Ax)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - Ax) &= x^{\mathrm{T}}Gx - 2x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y + y^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y \\ &= (x - \bar{x})^{\mathrm{T}}G(x - \bar{x}) + c,\end{aligned}$$

where $c \in \mathbb{R}$ depends only on y, not on x, and

$$\bar{x} = G^{-1}A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y = \Gamma_{\mathrm{post}}A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y. \qquad \square$$

If $\mathrm{Ker}(L) \cap \mathrm{Ker}(A) \neq \{0\}$, the 'candidate posterior density' is not a proper probability density. Indeed, it readily follows that

$$\begin{aligned} \pi_{\mathrm{pr}}(x)\pi(y \,|\, x) \;&\propto\; \exp\left(-\frac{1}{2}\left(\|Lx\|^2 + (y - Ax)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - Ax)\right)\right) \\ &=\; \exp\left(-\frac{1}{2}\left(\|L\mathcal{P}x\|^2 + (y - A\mathcal{P}x)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - A\mathcal{P}x)\right)\right), \end{aligned}$$

where $\mathcal{P} : \mathbb{R}^n \to (\mathrm{Ker}(L) \cap \mathrm{Ker}(A))^{\perp}$ is an orthogonal projection. This means that $\pi_{\mathrm{pr}}(x)\pi(y \,|\, x)$ is a constant as a function of the component of x in the direction of the non-trivial subspace $\mathrm{Ker}(L) \cap \mathrm{Ker}(A)$, and thus its integral over the whole $\mathbb{R}^n$ does not attain a finite value.

Using conditioning to create proper priors

Suppose that we would like to have a prior density of the form

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}x^{\mathrm{T}}L^{\mathrm{T}}Lx\right), \qquad x \in \mathbb{R}^n,$$

where $L \in \mathbb{R}^{k\times n}$ is some given matrix. As we have already seen, if L is not injective, such a prior is improper. One technique for obtaining a proper prior based on $\pi_{\mathrm{pr}}(x)$ is fixing the values of some components of x, and then considering π_{pr} as a probability density of the remaining ones.

To this end, we partition x as $x = [(x')^{\mathrm{T}}, (x'')^{\mathrm{T}}]^{\mathrm{T}}$, where, possibly after reordering the components, $x'' \in \mathbb{R}^j$, $0 \le j \le n$, contains the fixed components and $x' \in \mathbb{R}^{n-j}$ carries the unspecified ones.

Let us partition the matrix $L^{\mathrm{T}}L$ accordingly, i.e.,

$$L^{\mathrm{T}}L =: B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where $B_{11} \in \mathbb{R}^{(n-j)\times(n-j)}$ and $B_{22} \in \mathbb{R}^{j\times j}$ are symmetric, and $B_{12} \in \mathbb{R}^{(n-j)\times j}$ and $B_{21} \in \mathbb{R}^{j\times(n-j)}$ satisfy $B_{12} = B_{21}^{\mathrm{T}}$. In what follows, we assume that B_{11} invertible. This can often be achieved by fixing sufficiently many components of x, i.e., by choosing x'' to be extensive enough.

Let us derive the conditional density of X' given $X'' = x''$ properly for once, i.e., in such a way that no constant of proportionality depends on any of the variables at any stage:

Taking into account that this time we have partitioned our original candidate for the *inverse* covariance of X, it follows with some work from the second theorem of the fourteenth lecture that the (improper) marginal density of X'' is

$$\pi(x'') \propto \exp\left(-\frac{1}{2}(x'')^{\mathrm{T}}\tilde{B}_{11}x''\right),$$

where $\tilde{B}_{11} = B_{22} - B_{21}B_{11}^{-1}B_{12} \in \mathbb{R}^{j\times j}$ is the Schur complement of B_{11}. Moreover, it is a straightforward consequence of the partitioning of B that

$$\pi(x', x'') \propto \exp\left(-\frac{1}{2}((x')^{\mathrm{T}}B_{11}x' + 2(x')^{\mathrm{T}}B_{12}x'' + (x'')^{\mathrm{T}}B_{22}x'')\right).$$

Without paying too much attention to the fact that some densities may be improper, we then write

$$\begin{aligned}\pi(x' \,|\, x'') &= \frac{\pi(x', x'')}{\pi(x'')} \\ &\propto \exp\left(-\frac{1}{2}((x')^{\mathrm{T}} B_{11} x' + 2(x')^{\mathrm{T}} B_{12} x'' + (x'')^{\mathrm{T}} B_{21} B_{11}^{-1} B_{12} x'')\right) \\ &= \exp\left(-\frac{1}{2}(x' + B_{11}^{-1} B_{12} x'')^{\mathrm{T}} B_{11} (x' + B_{11}^{-1} B_{12} x'')\right).\end{aligned}$$

NB: One could have obtained this same formula for $\pi(x \,|\, x'')$ by just excluding $\pi(x'')$ and all other multipliers that depend only on x''. At the end, one could have then argued that $\pi(x' \,|\, x'')$ must be Gaussian, and thus the constant of proportionality between $\pi(x' \,|\, x'')$ and the last line above cannot depend on x'', but only on B_{11}. Such argument shows also that the constants of proportionality in the theorems presented at the fourteenth lecture do *not* depend on any of the variables.

To create a prior density that is proper for *all components of* X we may now proceed as follows. We first define a proper Gaussian probability distribution for the variable $X'' \in \mathbb{R}^j$,

$$X'' \sim \mathcal{N}(x_0'', \Gamma''),$$

where $\Gamma'' \in \mathbb{R}^{j \times j}$ is symmetric and positive definite. The corresponding density is denoted by π_0.

Then, we obtain a new candidate for the prior density of X by writing

$$\begin{aligned}
\tilde{\pi}_{\mathrm{pr}}(x', x'') &= \pi(x' \,|\, x'')\pi_0(x'') \\
&\propto \exp\left(-\frac{1}{2}(x' + B_{11}^{-1}B_{12}x'')^{\mathrm{T}} B_{11}(x' + B_{11}^{-1}B_{12}x'')\right) \\
&\quad \times \exp\left(-\frac{1}{2}(x'' - x_0'')^{\mathrm{T}}(\Gamma'')^{-1}(x'' - x_0'')\right) \\
&= \exp\left(-\frac{1}{2}(x - x_0)^{\mathrm{T}}\tilde{\Gamma}_{\mathrm{prior}}^{-1}(x - x_0)\right),
\end{aligned}$$

where the mean $x_0 \in \mathbb{R}^n$ and the covariance $\tilde{\Gamma}_{\text{prior}} \in \mathbb{R}^{n \times n}$ can be obtained relatively easily by completing the squares:

$$x_0 = \begin{bmatrix} -B_{11}^{-1} B_{12} x_0'' \\ x_0'' \end{bmatrix}$$

and

$$\tilde{\Gamma}_{\text{prior}} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{21} B_{11}^{-1} B_{12} + (\Gamma'')^{-1} \end{bmatrix}^{-1}.$$

Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Sixteenth lecture, March 18, 2011.

Exploring non-Gaussian densities

Why sampling is needed?

Remember that the CM estimate and the conditional covariance require solving integration problems involving the posterior density:

$$x_{\mathrm{CM}} = E\{x \mid y\} = \int_{\mathbb{R}^n} x\pi(x \mid y)dx$$

$$\mathrm{cov}(x \mid y) = \int_{\mathbb{R}^n} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}} \pi(x \mid y)dx.$$

In a non-Gaussian case, these integrals cannot typically be expressed in a closed form, and one must thus resort to numerical integration in $\mathbb{R}^n$.

Suppose that our aim is to estimate some quantity of the form

$$I = \int f(x)\pi(x)dx.$$

How about using quadrature rules? In principle, we could approximate

$$I = \int f(x)\pi(x)dx \approx \sum_{j=1}^{N} w_j f(x_j)\pi(x_j),$$

with some suitable weights $\{w_j\}$ and nodal points $\{x_j\}$. Unfortunately, if n is large, such computation is not feasible: For a quadrature rule with k discretization points per dimension, the total number of nodes is $N = k^n$. In addition, the realization of a quadrature rule would require reliable information about the location of the probability density π.

Often it is more advisable to resort to sampling: Draw a large enough sample $\{x_j\}_{j=1}^N$ from the probability distribution corresponding to $\pi(x)$, and use these points to approximate the integral as

$$I = \int f(x)\pi(x)dx = E\{f(X)\} \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j).$$

According to the Law of Large Numbers,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} f(x_j) =: \lim_{N\to\infty} I_N = I$$

almost surely, i.e., the sample average converges almost surely to the expected value. Furthermore, the Central Limit Theorem states that

$$\mathrm{var}(I_N - I) \approx \frac{\mathrm{var}(f(X))}{N},$$

i.e., the discrepancy between I and I_N should go to zero like $1/\sqrt{N}$.

Markov Chain Monte Carlo

Random walk in $\mathbb{R}^n$

Random walk in $\mathbb{R}^n$ is a process of moving around by taking random steps. Elementary random walk:

1. Choose a starting point $x_0 \in \mathbb{R}^n$ and a 'step size' $\sigma > 0$. Set $k = 0$.
2. Draw a random vector $w_{k+1} \sim \mathcal{N}(0, I)$ and set $x_{k+1} = x_k + \sigma w_{k+1}$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your stopping criterion is satisfied.

The location of the random walk at time k is a realization of the random variable X_k, and we have an evolution model

$$X_{k+1} = X_k + \sigma W_{k+1}, \qquad W_{k+1} \sim \mathcal{N}(0, I).$$

The conditional density of X_{k+1}, given $X_k = x_k$, is

$$\pi(x_{k+1} \,|\, x_k) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_k - x_{k+1}\|^2\right) = q(x_k, x_{k+1}).$$

The function q is called the *transition kernel*. Since q does not depend on k, i.e., the step is always distributed in the same way, the kernel is called *time invariant*.

The process above defines a *chain* $\{X_k\}_{k=0}^{\infty}$ of random variables. This chain is a discrete time stochastic process. Note that

$$\pi(x_{k+1} \,|\, x_0, x_1, \ldots, x_k) = \pi(x_{k+1} \,|\, x_k),$$

i.e., the probability distribution of X_{k+1} depends on the past only through the preceding element X_k. A stochastic process with this property is called a *Markov chain*.

Example: Random walk in $\mathbb{R}^2$

A random walk model in $\mathbb{R}^2$:

$$X_{k+1} = X_k + \sigma W_{k+1}, \qquad W_{k+1} \sim \mathcal{N}(0, C), \quad C \in \mathbb{R}^{2\times 2}.$$

Since C is symmetric and positive definite, it has positive eigenvalues and allows an eigenvalue decomposition

$$C = UDU^{\mathrm{T}}.$$

Hence, the inverse of C can be written as

$$C^{-1} = UD^{-1}U^{\mathrm{T}} = (UD^{-1/2})\underbrace{(D^{-1/2}U^{\mathrm{T}})}_{=L},$$

which means that the transition Kernel can in turn be given as

$$q(x_k, x_{k+1}) = \pi(x_{k+1} \,|\, x_k) \propto \exp\left(-\frac{1}{2\sigma^2}\|L(x_k - x_{k+1})\|^2\right).$$

Consequently, the random walk model becomes

$$X_{k+1} = X_k + \sigma L^{-1} \tilde{W}_{k+1}, \qquad \tilde{W}_{k+1} \sim \mathcal{N}(0, I),$$

where we have used the fact that L is the whitening matrix of W_{k+1}.

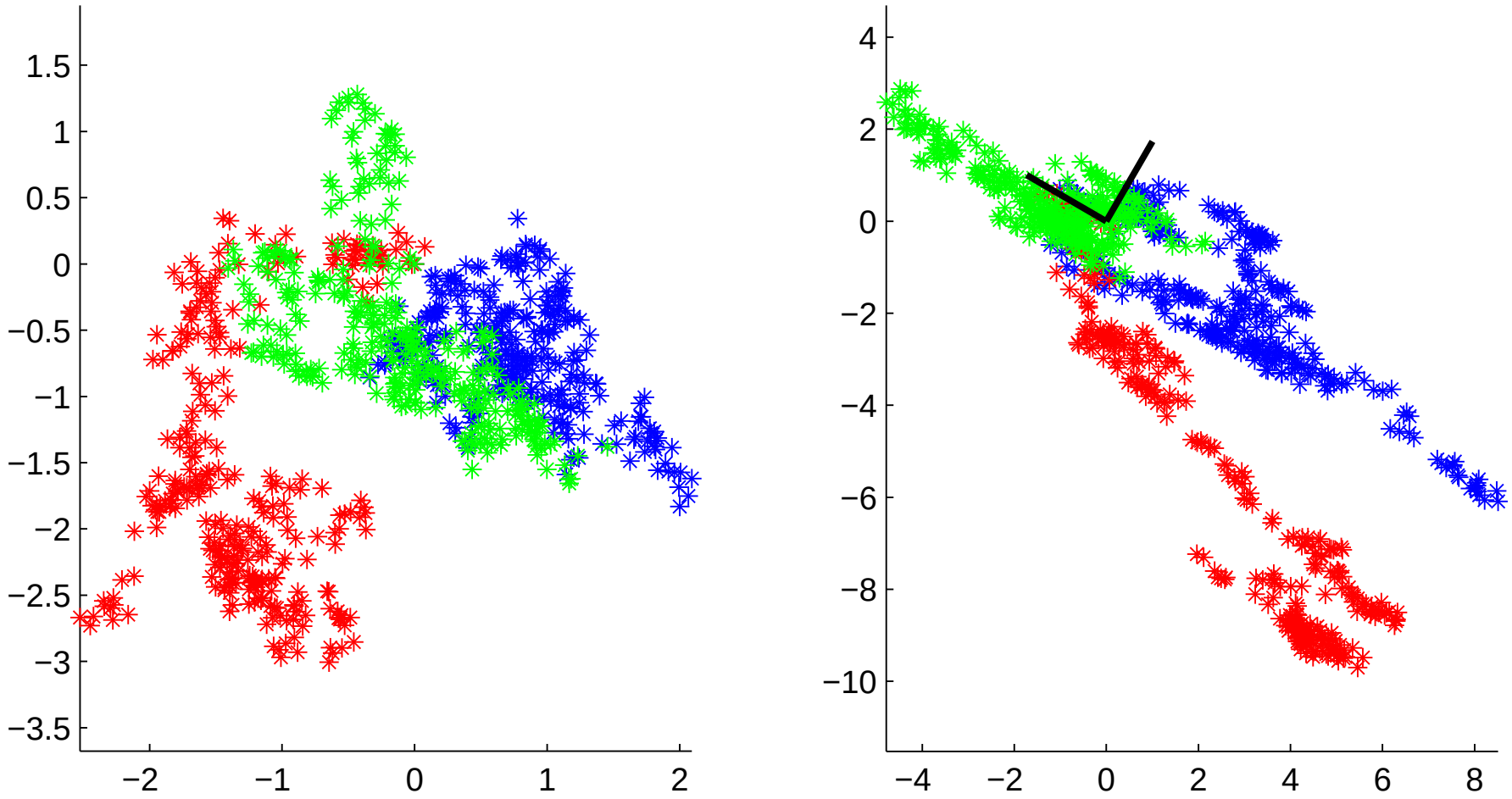
To demonstrate the effect of the covariance matrix, let

$$U = [u^{(1)}, u^{(2)}] = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \theta = \frac{\pi}{3},$$

and

$$D = \mathrm{diag}(s_1^2, s_2^2), \qquad s_1 = 1, \ s_2 = 4.$$

In the light of this random walk model, the random steps should on average have a component about four times larger in the direction of the second eigenvector e_2 than in the direction of the first eigenvector e_1.



On the left, three random walk realizations for $C = I$; on the right, three realizations for C given above. In both cases, $\sigma = 0.1$ and $x_0 = [0, 0]^{\mathrm{T}}$.

How about sampling from a given density $p(x)$?

Assume now that X is a random variable with a probability density $\pi(x) = p(x)$.

Consider an arbitrary transition kernel $q(x, y)$ that we use to generate a new random variable Y given $X = x$, that is,

$$\pi(y \,|\, x) = q(x, y).$$

The probability density of Y is found via marginalization,

$$\pi(y) = \int \pi(y \,|\, x)\pi(x)dx = \int q(x, y)p(x)dx.$$

If the probability density of Y is equal to the probability density of X, i.e.,

$$\int q(x, y)p(x)dx = p(y),$$

we say that p is an *invariant density* of the transition kernel q.

To summarize, if p is an invariant density of the transition kernel q and the random variable X obeys the density p, then the random variable Y defined via the conditional density $\pi(y \mid x) = q(x, y)$ is still distributed according to the density p. Loosely speaking, the transition defined by q does not affect the distribution of X.

This property of invariant densities and corresponding transition kernels can be put to use in sampling.

Theorem. *Let $\{X_k\}_{k=0}^{\infty}$ be a time invariant Markov chain with the transition kernel q, i.e.,*

$$\pi(x_{k+1} \,|\, x_k) \,=\, q(x_k, x_{k+1}).$$

Assume that p is an invariant density of q, and that q satisfies some extra technical conditions (irreducibility and aperiodicity). Then, for all $x_0 \in \mathbb{R}$ and any Borel set $B \in \mathbb{R}^n$, it holds that

$$\lim_{N\to\infty} P\{X_N \in B \mid X_0 = x_0\} \,=\, \int_B p(x)dx.$$

Moreover, for any regular enough function f,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=0}^{N} f(X_j) \,=\, \int_{\mathbb{R}^n} f(x)p(x)dx$$

almost surely.

Proof. Proof is omitted due to obvious reasons. □

Let us try to put the above theorem into practical use. Suppose that we want to sample some probability density p and happen to know that it is invariant with respect to some transition kernel q. Then, we can proceed as follows:

1. Select a starting point x_0 and set $k = 0$.
2. Draw x_{k+1} from $q(x_k, x_{k+1})$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your personal stopping criterion is satisfied.

According to the previous theorem, the sample $\{x_k\}_{k=0}^{N}$ should give a better and better representation of p as N increases.

Hence, we are facing an inverse problem: *Given a probability density p, we would like to find a kernel q such that p is its invariant density.*

Very popular technique for constructing such a transition kernel is the *Metropolis–Hastings* algorithm.

Metropolis–Hastings algorithm

Let us introduce a slightly more general Markov process: If you are currently at some $x \in \mathbb{R}^n$, either

1. stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
2. move away from x using a transition kernel $R(x, y)$ otherwise.

Since R is a transition kernel, the mapping $y \mapsto R(x, y)$ defines a probability density, and thus

$$\int_{\mathbb{R}^n} R(x, y)dy = 1, \qquad \text{for all } x \in \mathbb{R}^n.$$

Denote by $\mathcal{A}$ the event of moving away from x and by $\neg\mathcal{A}$ the event of not moving, meaning that

$$P\{\mathcal{A}\} = 1 - r(x), \qquad P\{\neg\mathcal{A}\} = r(x).$$

What is the density of Y generated by the above strategy, given $X = x$?

Let $B \subset \mathbb{R}^n$ be a Borel set and let us write

$$P\{Y \in B \mid X = x\} = P\{Y \in B \mid X = x, \mathcal{A}\}P\{\mathcal{A}\} + P\{Y \in B \mid X = x, \neg\mathcal{A}\}P\{\neg\mathcal{A}\}.$$

The probability of arriving in B if we happen to move:

$$P\{Y \in B \mid X = x, \mathcal{A}\} = \int_B R(x, y)dy.$$

Arriving in B without moving happens only if $x \in B$, i.e.,

$$P\{Y \in B \mid X = x, \neg\mathcal{A}\} = \chi_B(x) := \begin{cases} 1, & \text{if } x \in B, \\ 0, & \text{if } x \notin B. \end{cases}$$

To sum up, the probability of reaching B from x is

$$P\{Y \in B \mid X = x\} = (1 - r(x)) \int_B R(x, y)dy + r(x)\chi_B(x).$$

Finally, the probability of $Y \in B$ is found through marginalization:

$$\begin{aligned} P\{Y \in B\} &= \int P\{Y \in B \mid X = x\}p(x)dx \\ &= \int p(x) \left(\int_B (1 - r(x))R(x, y)dy \right) dx + \int \chi_B(x)r(x)p(x)dx \\ &= \int_B \left(\int p(x)(1 - r(x))R(x, y)dx \right) dy + \int_B r(x)p(x)dx \\ &= \int_B \left(\int p(x)(1 - r(x))R(x, y)dx + r(y)p(y) \right) dy. \end{aligned}$$

By definition

$$P\{Y \in B\} = \int_B \pi(y)dy,$$

and comparing this with the above formula, we see that the probability density of Y must be

$$\pi(y) = \int p(x)(1 - r(x))R(x, y)dx + r(y)p(y).$$

Our ultimate goal is to find a kernel R and a probability r such that $\pi(y) = p(y)$, that is,

$$p(y) = \int p(x)(1 - r(x))R(x, y)dx + r(y)p(y),$$

or, equivalently,

$$(1 - r(y))p(y) = \int p(x)(1 - r(x))R(x, y)dx.$$

Denote

$$K(x, y) = (1 - r(x))R(x, y),$$

and observe that, since R is a transition kernel,

$$\int K(y, x)dx = (1 - r(y)) \int R(y, x)dx = 1 - r(y).$$

The condition at the bottom of the previous slide can thus be written as

$$\int p(y)K(y, x)dx = \int p(x)K(x, y)dx,$$

which is called the *balance equation*. This condition is satisfied, in particular, if the integrands are equal, i.e.,

$$p(y)K(y, x) = p(x)K(x, y).$$

This condition is known as the *detailed balance equation*. The Metropolis–Hastings algorithm is simply a technique for finding a kernel K that satisfies the detailed version of the balance equation.

Start by selecting a *candidate generating kernel* $q(x, y)$, then define

$$\tilde{\alpha}(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\},$$

and finally set

$$K(x, y) = \tilde{\alpha}(x, y)q(x, y).$$

A simple calculation shows that such K satisfies the detailed balance equation, i.e.,

$$p(y)\tilde{\alpha}(y, x)q(y, x) = p(x)\tilde{\alpha}(x, y)q(x, y).$$

To convince yourself, take note that for any $x, y \in \mathbb{R}^n$ either

$$\tilde{\alpha}(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)} \qquad \text{and} \qquad \tilde{\alpha}(y, x) = 1,$$

or

$$\tilde{\alpha}(x, y) = 1 \qquad \text{and} \qquad \tilde{\alpha}(y, x) = \frac{p(x)q(x, y)}{p(y)q(y, x)}.$$

The actual Metropolis–Hastings algorithm for drawing samples is as follows:

1. Given x, draw y using the transition kernel $q(x, y)$.
2. Calculate the acceptance ratio,
$$\alpha(x, y) := \frac{p(y)q(y, x)}{p(x)q(x, y)}.$$
3. Flip the α-coin: Draw $t \sim \mathrm{Uniform}([0, 1])$. If $\alpha > t$, accept y. Otherwise stay put at x.

Does the above algorithm really work? It is not quite obvious...

Yes, it does. According to our construction, the Markov process introduced at the beginning of this section, i.e., the one involving R and r, is with the choice

$$K(x,y) = (1 - r(x))R(x,y) = \tilde{\alpha}(x,y)q(x,y)$$

such that p is its invariant density. Note, in particular, that for this choice, it holds that

$$P\{\mathcal{A} \text{ and } Y \in B \,|\, X = x\} = (1 - r(x)) \int_B R(x,y)dy = \int_B K(x,y)dy,$$

which is something that the actual algorithm should also satisfy. In other words, everything is OK if for the above introduced algorithm the probability that "the move is accepted and $Y \in B$" under $X = x$ is given by this same formula. (It does not matter what happens to Y if the move is not accepted, because then we do not move in any case.)

For the actual algorithm we have

$$P\{\mathcal{A} \mid Y = y,\, X = x\} = \min\{1, \alpha(x, y)\} = \tilde{\alpha}(x, y)$$

and

$$P\{Y \in B \mid X = x\} = \int_B q(x, y)dy.$$

Hence, it follows in the case of the algorithm that

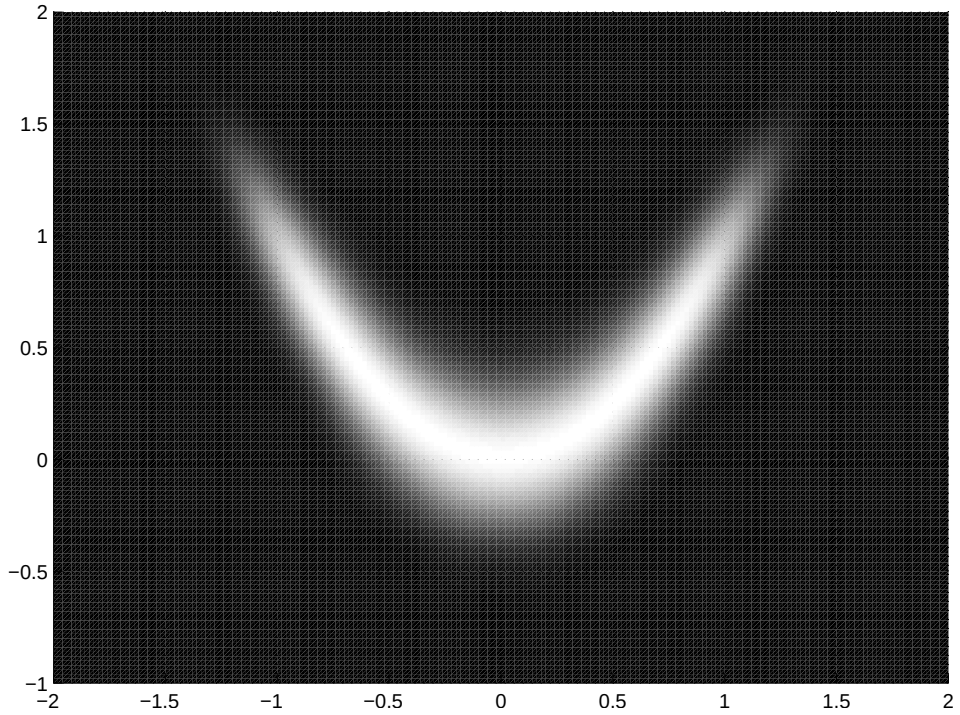
$$P\{\mathcal{A} \text{ and } Y \in B \,|\, X = x\} = \tilde{\alpha}(x, y) \int_B q(x, y)dy = \int_B K(x, y)dy,$$

which means that everything really works as it should.

Example

Consider sampling in $\mathbb{R}^2$ from the density

$$\pi(x) \propto \exp\left(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4\right).$$



We use white noise random walk proposal

$$q(x, y) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{1}{2\gamma^2}\|x - y\|^2\right).$$

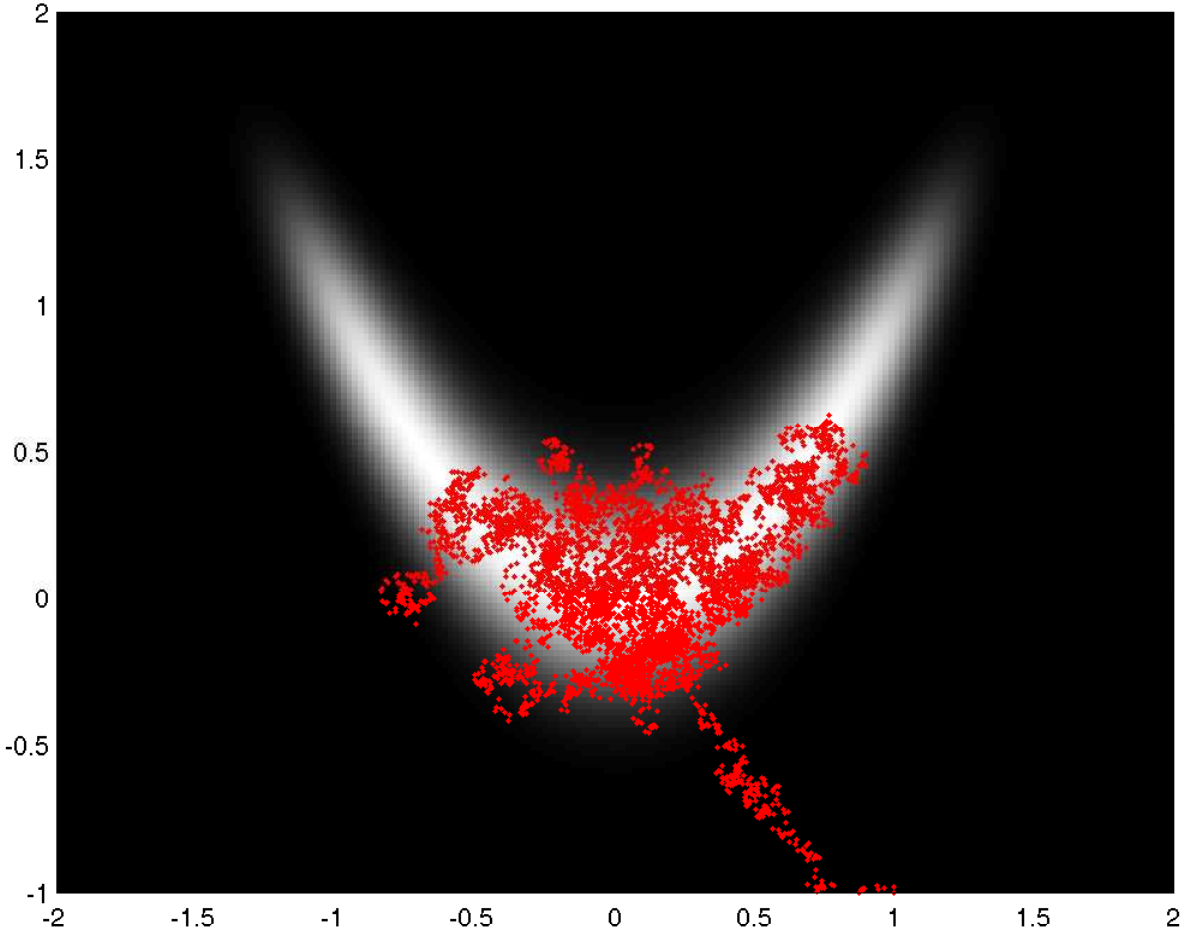
Note that now the transition kernel is symmetric, i.e.,

$$q(x, y) = q(y, x),$$

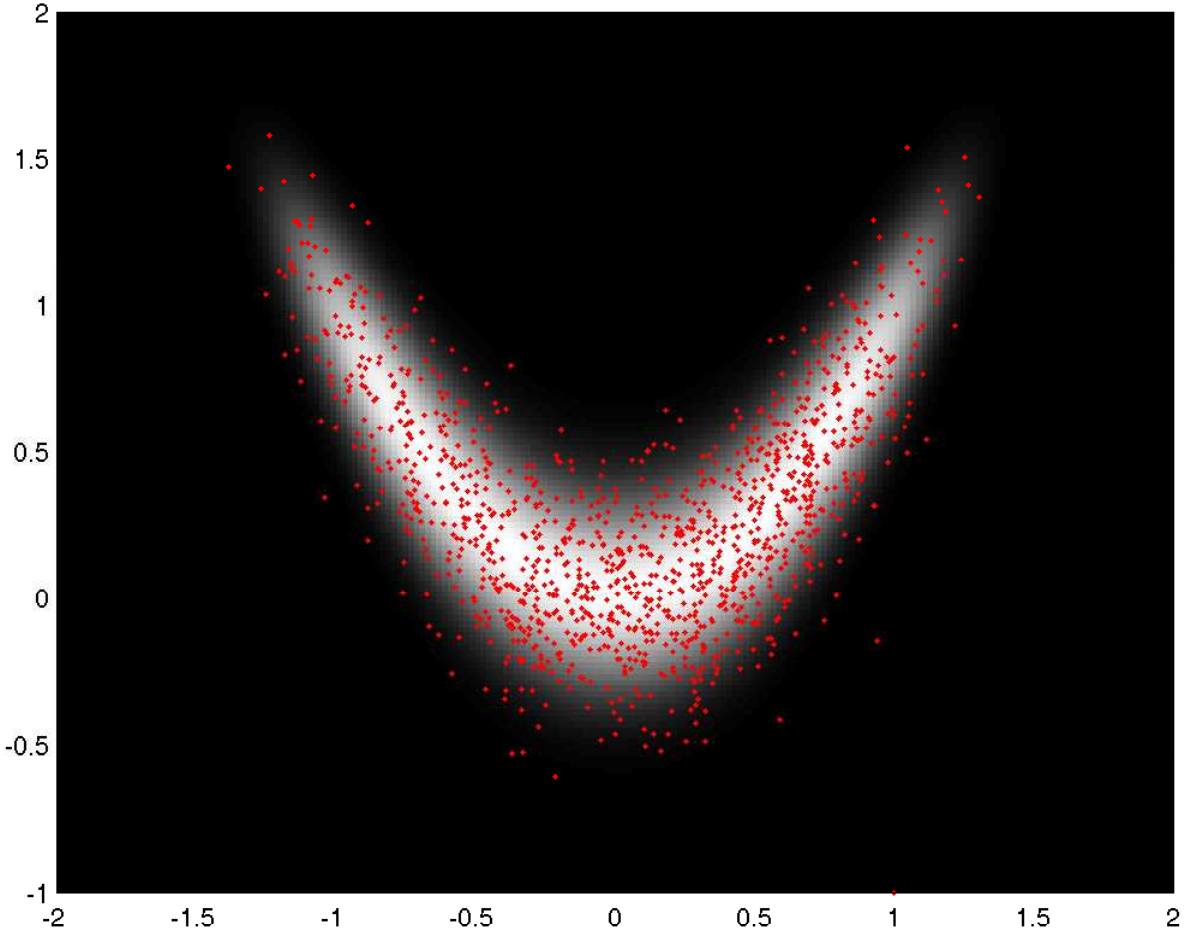
and hence

$$\alpha(x, y) = \frac{\pi(y)}{\pi(x)}.$$

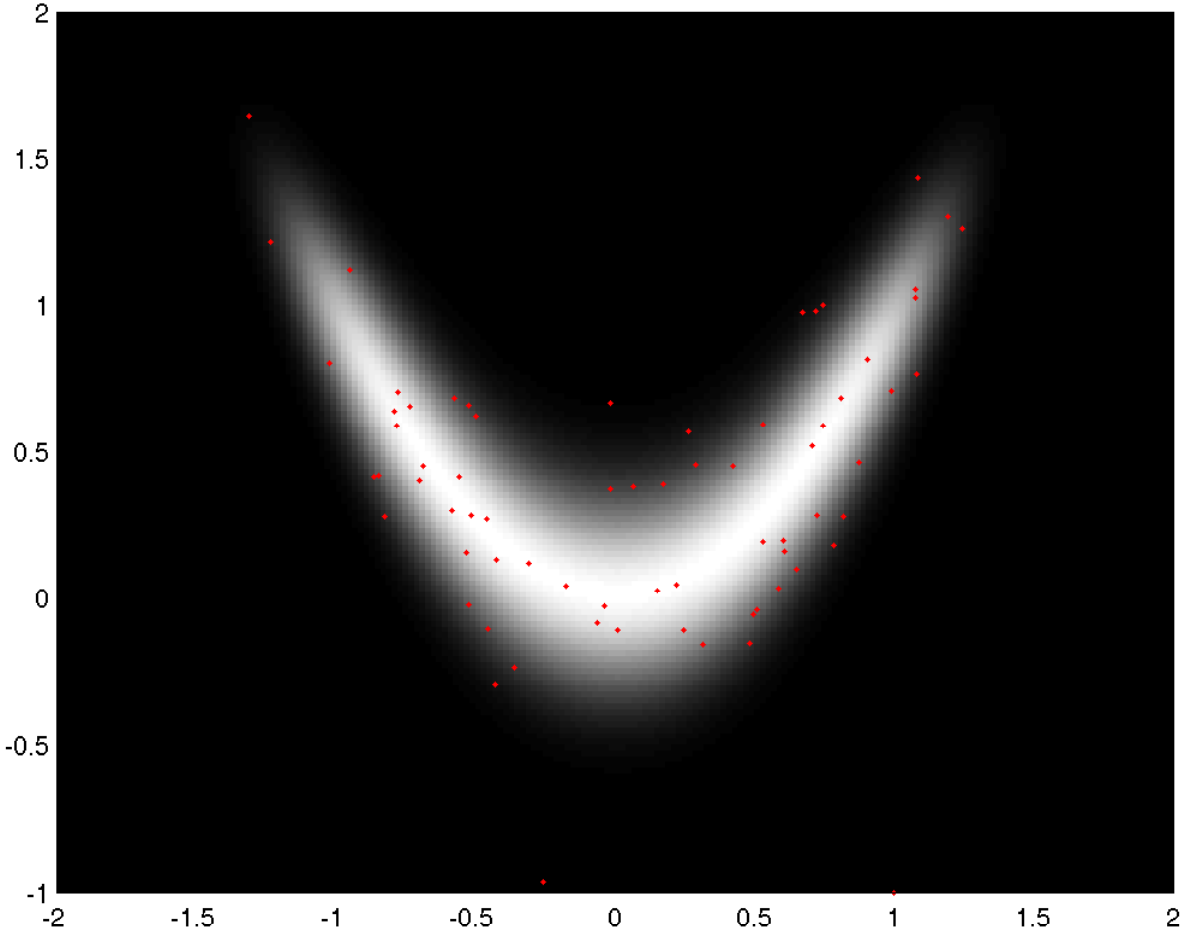
$$\gamma = 0.02$$



$$\gamma = 0.7$$



$\gamma = 4$



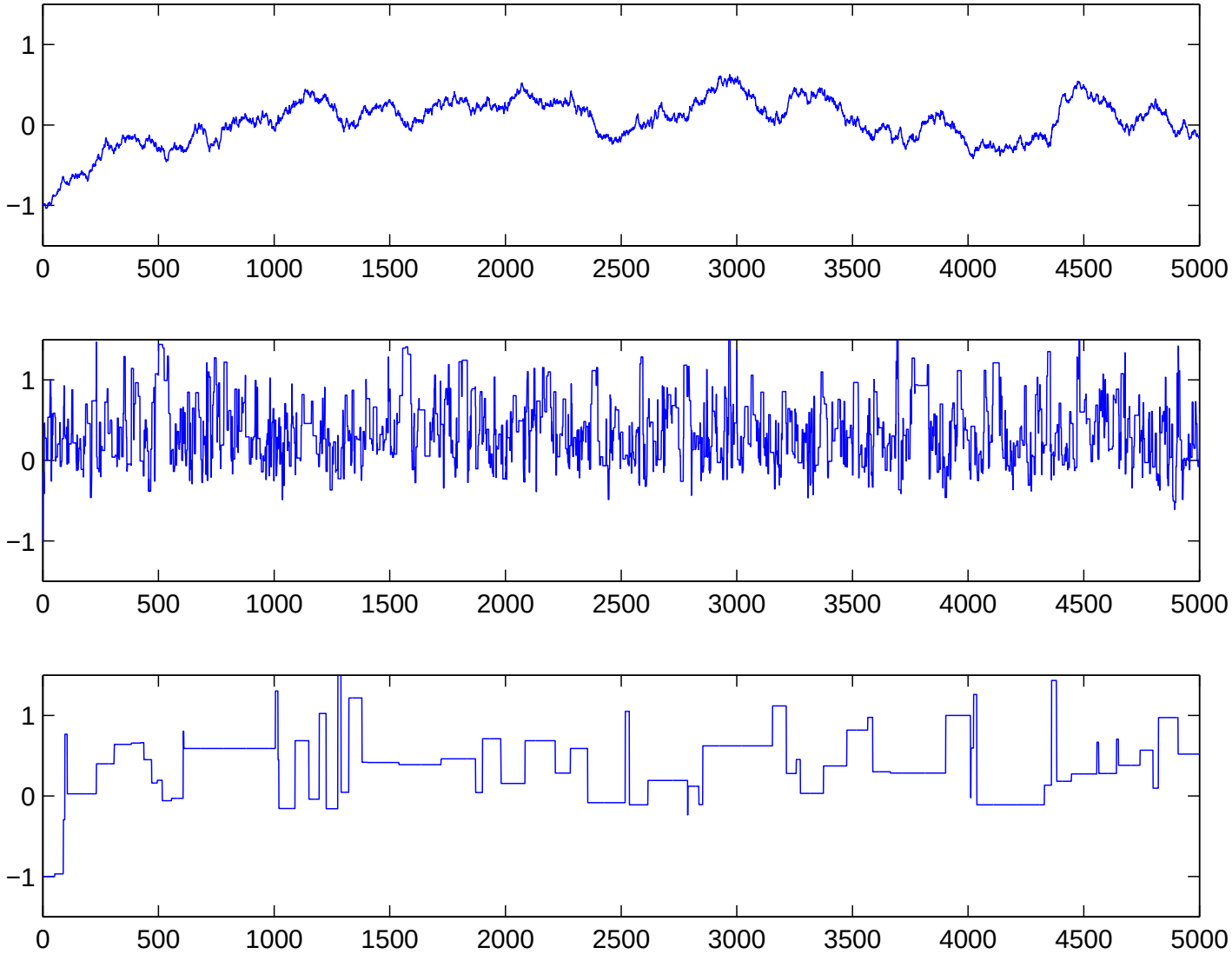
Acceptance rates:

$\gamma = 0.02$: 95.6 %

$\gamma = 0.7$: 24.5 %

$\gamma = 4$: 1.4 %

Sample histories:



Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Seventeenth lecture, March 23, 2011.

Metropolis–Hastings algorithm (continued)

Recall the Metropolis–Hastings algorithm for drawing samples from a given probability density $p : \mathbb{R}^n \to \mathbb{R}_+$.

1. Choose $x_0 \in \mathbb{R}^n$. Set $k = 0$.
2. Given x_k, draw y using the transition kernel $q(x_k, y)$ of your choice.
3. Calculate the acceptance ratio,
$$\alpha(x_k, y) := \frac{p(y)q(y, x_k)}{p(x_k)q(x_k, y)}.$$
4. Flip the α-coin: Draw $t \sim \mathrm{Uniform}([0, 1])$. If $\alpha > t$, set $x_{k+1} = y$. Otherwise, stay put at x_k, i.e., set $x_{k+1} = x_k$.
5. Set $k \leftarrow k + 1$ and return to Step 2, unless your stopping criterion is satisfied.

The constructed sample $\{x_k\}_{k=0}^N$ should represent p if N is large enough.

Adapting the Metropolis-Hastings sampler

With the white noise random walk proposal density (used in the numerical example of the previous lecture), the sampler does not take into account the form of the posterior density.

However, the shape of the density *can* be taken into account when designing the proposal density, in order to minimize the number of 'wasted proposals'. In high-dimensional setting, this becomes especially useful if the posterior density is highly *anisotropic*, i.e., if the posterior is stretched in some directions.

The proposal distribution can be updated while the sampling algorithm moves around the posterior density. This process is often called *adaptation*.

Gibbs sampler

Let us first consider some notational details:

- $I = \{1, 2, \ldots, n\}$ is the index set of $\mathbb{R}^n$.
- $I = \bigcup_{j=1}^{m} I_m$ is a partitioning of the index set into disjoint nonempty subsets.
- The number of elements in I_j is denoted by k_j; $k_1 + \cdots + k_m = n$.
- We partition $\mathbb{R}^n$ as $\mathbb{R}^n = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_m}$, and correspondingly

$$x = [x_{I_1}; \ldots; x_{I_m}] \in \mathbb{R}^n, \qquad x_{I_j} \in \mathbb{R}^{k_j},$$

  where $x_i \in \mathbb{R}$ is a component of the vector x_{I_j} if and only if $i \in I_j$.

In practice, it often holds that $k_j = 1$ for all $j = 1, \ldots, m$, meaning that $m = n$ and x_{I_j} is just the jth component of the original vector x.

Transition kernel for the Gibbs sampler

Suppose that we are still aiming at sampling some given probability density $p : \mathbb{R}^n \to \mathbb{R}_+$, and recall the Markov process considered at the previous lecture: If you are currently at some $x \in \mathbb{R}^n$, either

1. stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
2. move away from x using a transition kernel $R(x, y)$ otherwise.

Recall also that we made the definition

$$K(x, y) \;=\; (1 - r(x))R(x, y).$$

For the Gibbs sampler, we choose $r(x) = 0$ for all $x \in \mathbb{R}^n$, i.e., moving is obligatory, and define

$$K(x, y) \;=\; R(x, y) \;=\; \prod_{i=1}^{m} p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}),$$

where the conditional densities are defined in the natural way based on p, i.e.,

$$p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}) = \frac{p(y_{I_1}, \ldots, y_{I_i}, x_{I_{i+1}}, \ldots, x_{I_m})}{\int_{\mathbb{R}^{k_i}} p(y_{I_1}, \ldots, y_{I_i}, x_{I_{i+1}}, \ldots x_{I_m}))dy_{I_i}}.$$

Such a transition kernel K does not, in general, satisfy the detailed balance equation, i.e.,

$$p(y)K(y,x) \neq p(x)K(x,y),$$

but it satisfies the (standard) balance equation,

$$\int_{\mathbb{R}^n} p(y)K(y,x)dx = \int_{\mathbb{R}^n} p(x)K(x,y)dx,$$

which is a sufficient condition for p being an invariant density of the above introduced Markov process. (See the slides of the previous lecture for the details.)

Proof: Consider first the left-hand side of the balance equation.

Due to the basic properties of probability densities, we have

$$\int_{\mathbb{R}^{k_i}} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) dx_{I_i} = 1$$

for all $i = 1, \ldots, m$. By integrating the kernel $K(y, x)$ over $\mathbb{R}^{k_m}$, we thus get

$$\int_{\mathbb{R}^{k_m}} K(y, x) dx_{I_m} = \int_{\mathbb{R}^{k_m}} \prod_{i=1}^{m} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) dx_{I_m}$$

$$= \prod_{i=1}^{m-1} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) \int_{\mathbb{R}^{k_m}} p(x_{I_m} \,|\, x_{I_1}, \ldots, x_{I_{m-1}}) dx_{I_m}$$

$$= \prod_{i=1}^{m-1} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}).$$

Inductively, by always integrating with respect to the last block of x with respect to which we have not yet integrated, we easily obtain that altogether

$$\int_{\mathbb{R}^n} K(y,x)dx = 1,$$

which in turn implies that

$$\int_{\mathbb{R}^n} p(y)K(y,x)dx = p(y)\int_{\mathbb{R}^n} K(y,x)dx = p(y).$$

Next, we consider the right-hand side of the balance equation. Since $K(x,y)$ is independent of x_{I_1} and due to the definition of marginal probability densities, we have

$$\int_{\mathbb{R}^{k_1}} p(x)K(x,y)dx_{I_1} = K(x,y)\int_{\mathbb{R}^{k_1}} p(x)dx_{I_1} =: K(x,y)p(x_{I_2},\ldots,x_{I_m}).$$

By substituting the definition of K in the above formula, we see that

$$\int_{\mathbb{R}^{k_1}} p(x)K(x,y)dx_{I_1} = K(x,y)p(x_{I_2},\ldots,x_{I_m})$$

$$= \left(\prod_{i=2}^{m} p(y_{I_i} \,|\, y_{I_1},\ldots,y_{I_{i-1}},x_{I_{i+1}},\ldots,x_{I_m})\right) \times p(y_{I_1} \,|\, x_{I_2},\ldots,x_{I_m})p(x_{I_2},\ldots,x_{I_m})$$

$$= \left(\prod_{i=2}^{m} p(y_{I_i} \,|\, y_{I_1},\ldots,y_{I_{i-1}},x_{I_{i+1}},\ldots,x_{I_m})\right) p(y_{I_1},x_{I_2},\ldots,x_{I_m}).$$

Next, we integrate with respect to x_{I_2} over $\mathbb{R}^{k_2}$. By denoting

$$a_i \;=\; p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}), \qquad i = 2, \ldots, m,$$

we may write

$$\int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}^{k_1}} p(x) K(x,y) dx_{I_1} dx_{I_2} = \int_{\mathbb{R}^{k_2}} \prod_{i=2}^{m} a_i \, p(y_{I_1}, x_{I_2}, \ldots, x_{I_m}) dx_{I_2}$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_2} \,|\, y_{I_1}, x_{I_3}, \ldots, x_{I_m}) \int_{\mathbb{R}^{k_2}} p(y_{I_1}, x_{I_2}, \ldots, x_{I_m}) dx_{I_2}$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_2} \,|\, y_{I_1}, x_{I_3}, \ldots, x_{I_m}) p(y_{I_1}, x_{I_3}, \ldots, x_{I_m})$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_1}, y_{I_2}, x_{I_3}, \ldots, x_{I_m}).$$

We can continue inductively integrating over the remaining blocks $x_{I_3}, \ldots, x_{I_m}$ in turns, which eventually results in

$$\int_{\mathbb{R}^n} p(x)K(x,y)dx \;=\; p(y_{I_1}, \ldots, y_{I_m}) \;=\; p(y),$$

and the proof is complete. $\square$

Gibbs sampler algorithm

1. Choose the initial value $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2. Draw the next sample as follows:
   (a) Set $x = x^k$ and $j = 1$.
   (b) Draw $y_{I_j} \in \mathbb{R}^{k_j}$ from the k_j-dimensional distribution $p(y_{I_j} \,|\, y_{I_1}, \ldots, y_{I_{j-1}}, x_{I_{j+1}}, \ldots, x_{I_m})$.
   (c) If $j = m$, set $y = [y_{I_1}; \ldots; y_m]$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (b).
3. Set $x^{k+1} = y$, increase $k \leftarrow k + 1$ and return to step 2, unless the chosen stopping criterion is satisfied.

Single component Gibbs sampler algorithm

1. Choose the initial value $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2. Draw the next sample as follows:
   (a) Set $x = x^k$ and $j = 1$.
   (b) Draw $t \in \mathbb{R}$ from the one-dimensional distribution

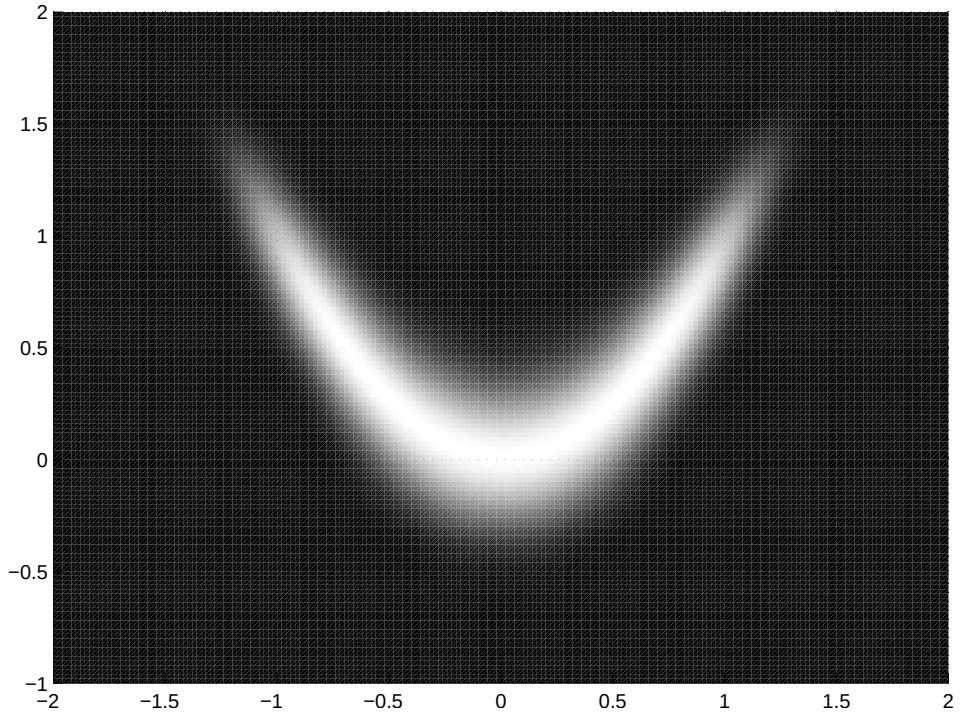
   $$p(t \,|\, y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_n) \propto p(y_1, \ldots, y_{j-1}, t, x_{j+1}, \ldots, x_n)$$

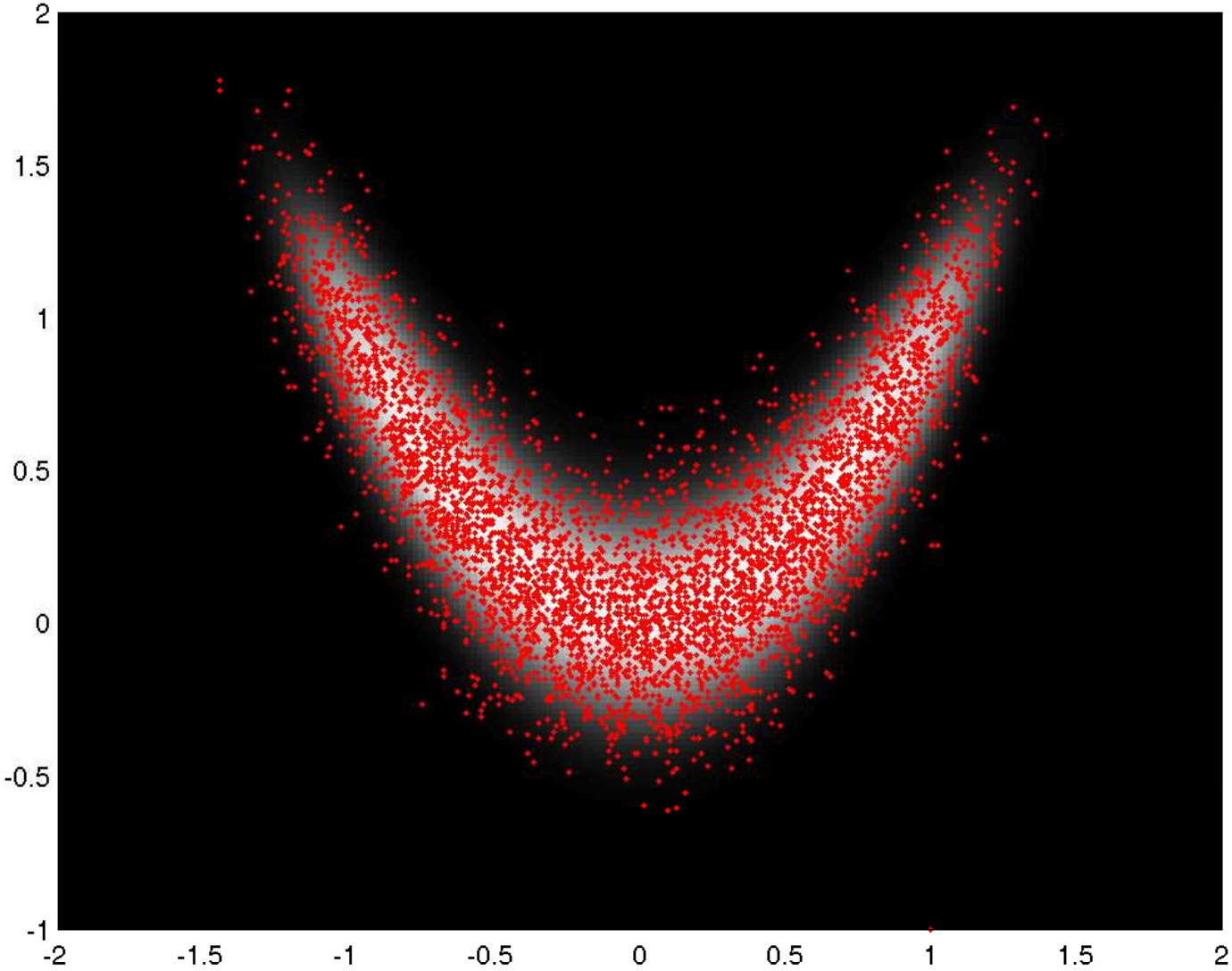
   and set $y_j = t$.
   (c) If $j = n$, set $y = [y_1, \ldots, y_n]^{\mathrm{T}}$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (b).
3. Set $x^{k+1} = y$, increase $k \leftarrow k + 1$ and return to step 2, unless the chosen stopping criterion is satisfied.

Example

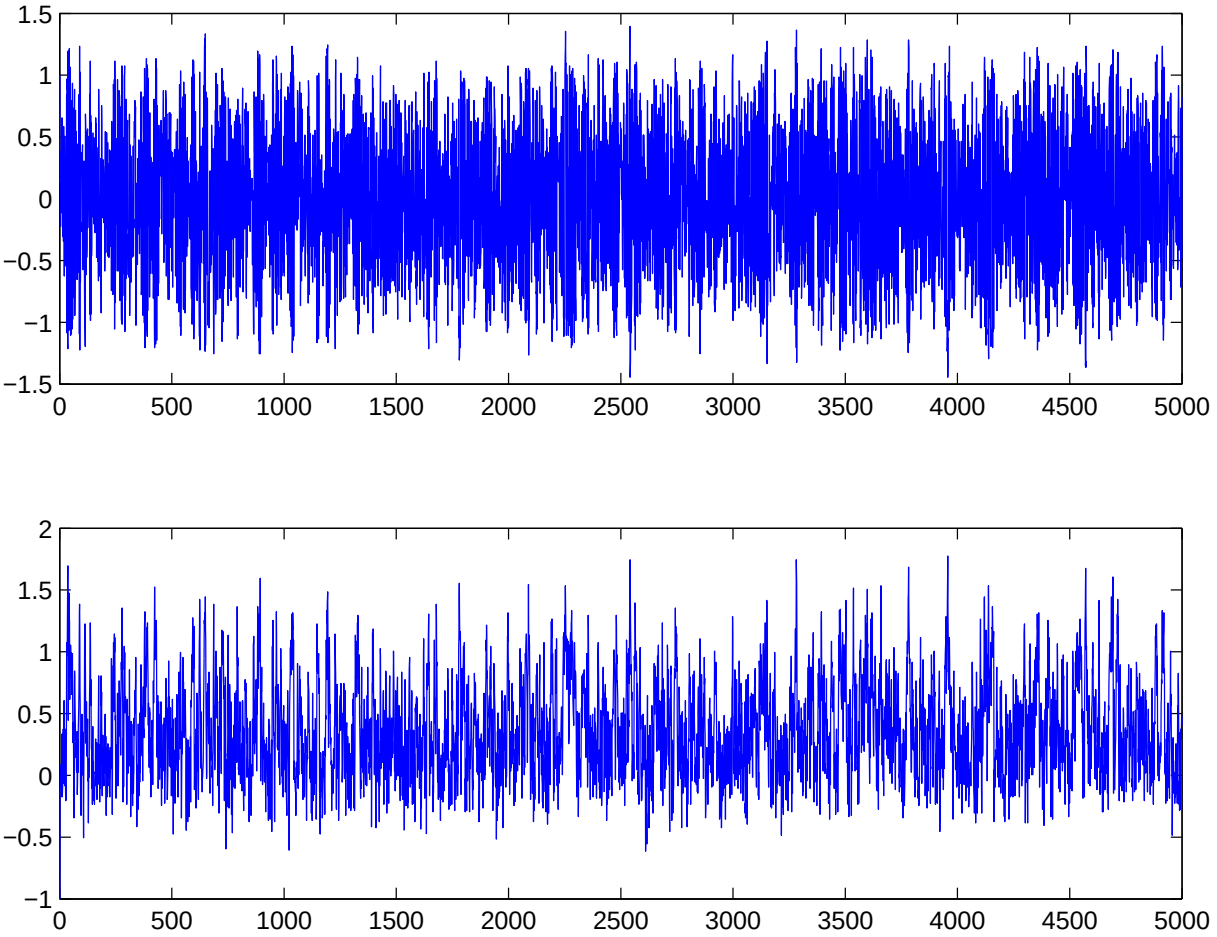
Consider again the density

$$\pi(x) \propto \exp\left(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4\right), \quad x \in \mathbb{R}^2.$$




2
1.5
1
0.5
0
-0.5
-1
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Sample histories for x_1 and x_2:



How to judge the quality of a sample?

Essential questions:

- What sampling strategy and/or proposal distribution works the best?
- Is the sample big enough?

Consider estimates of the form

$$\int f(x)\pi(x)dx = E\{f(X)\} \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j),$$

and recall that the Central Limit Theorem gives some answers regarding the convergence.

Assume that the variables $Y_j = f(X_j) \in \mathbb{R}$ are mutually independent and identically distributed with $E\{Y_j\} = y$ and $\mathrm{var}(Y_j) = \sigma^2$, and define

$$\tilde{Y}_N = \frac{1}{N}\sum_{j=1}^{N} Y_j \qquad \text{and} \qquad Z_N = \frac{\sqrt{N}(\tilde{Y}_N - y)}{\sigma}.$$

Then, $\tilde{Y}_N \to E\{Y\}$ almost surely (LLN). Moreover, Z_N is asymptotically (standard) normally distributed, that is,

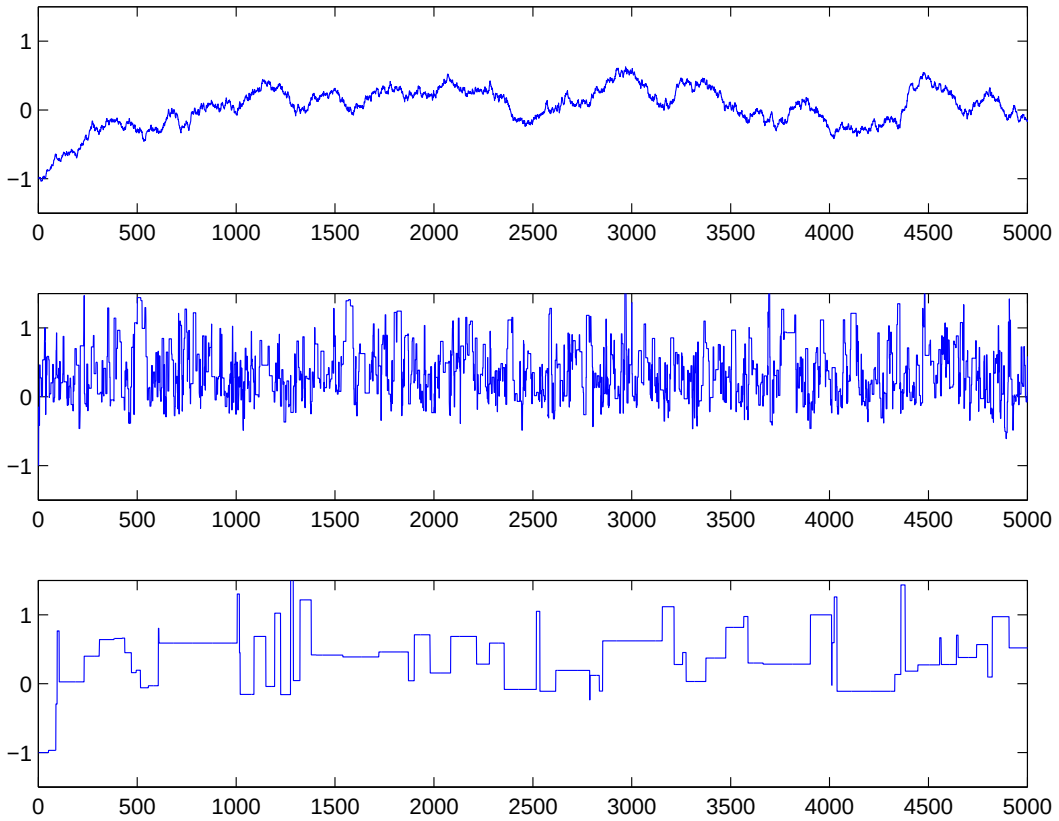
$$\lim_{N\to\infty} P\{Z_n \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}s^2\right) ds.$$

Loosely speaking, the above result says that the approximation error behaves as

$$\frac{1}{N}\sum_{j=1}^{N} f(x_j) - \int f(x)\pi(x)dx \;\approx\; \frac{\sigma}{\sqrt{N}}$$

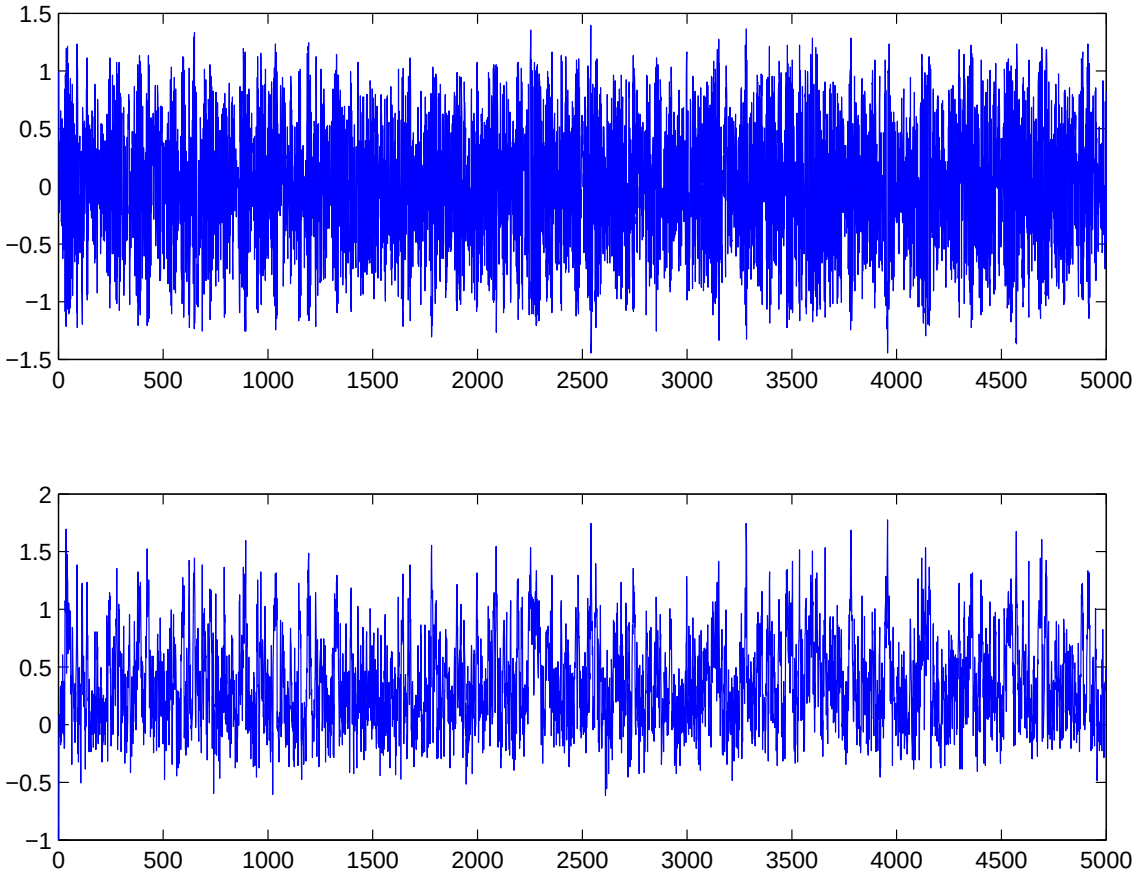
provided that the samples $\{x_j\}$ are *independent*.

Let us have another look at the sample histories corresponding to our standard example. First, the Metropolis–Hastings algorithm for the three choices of γ (the vertical component is plotted):



Clearly, consecutive elements are not independent.

Then, the Gibbs sampler (both components are plotted):

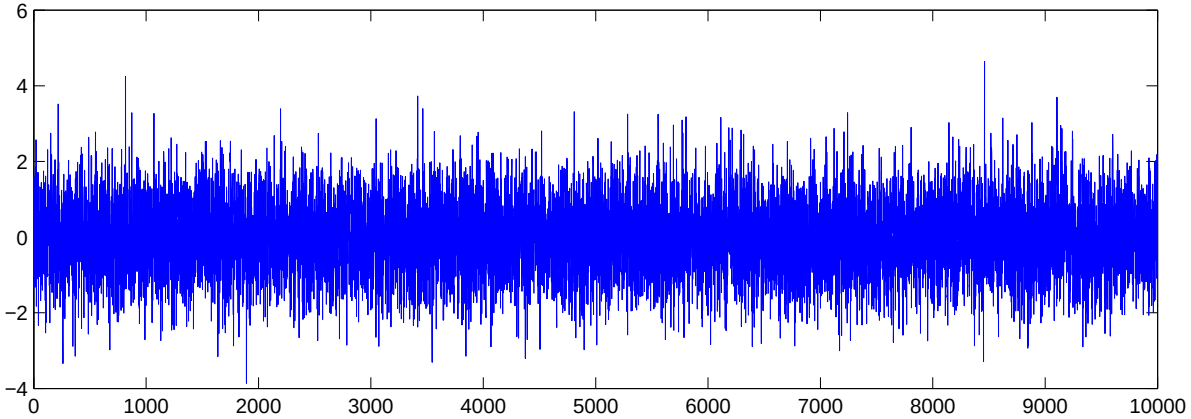


The results are somewhat better, but there is still some correlation between consecutive elements — especially for the vertical component.

If every kth sample point is independent, one might expect the discrepancy to behave as $1/\sqrt{N/k} = \sqrt{k/N}$ instead of $1/\sqrt{N}$. Consequently, one should try to choose the proposal distribution so that the *correlation length* is as small as possible.

Quick visual assessment: Take a look at the sample histories of individual components. How should they look like?

Consider a *white noise* signal, where the sample points are independent and the sample history looks like a "fuzzy worm". This is something one could aim at.



Autocovariance and correlation length

Denote by $f_{\mathrm{c}}(x_j) \in \mathbb{R}$, $j = 1, \ldots, N$, the centered sample points, i.e.,

$$f_{\mathrm{c}}(x_j) = f(x_j) - \frac{1}{N} \sum_{i=1}^{N} f(x_i), \qquad j = 1, \ldots, N.$$

Define the normalized autocovariance of the sample as

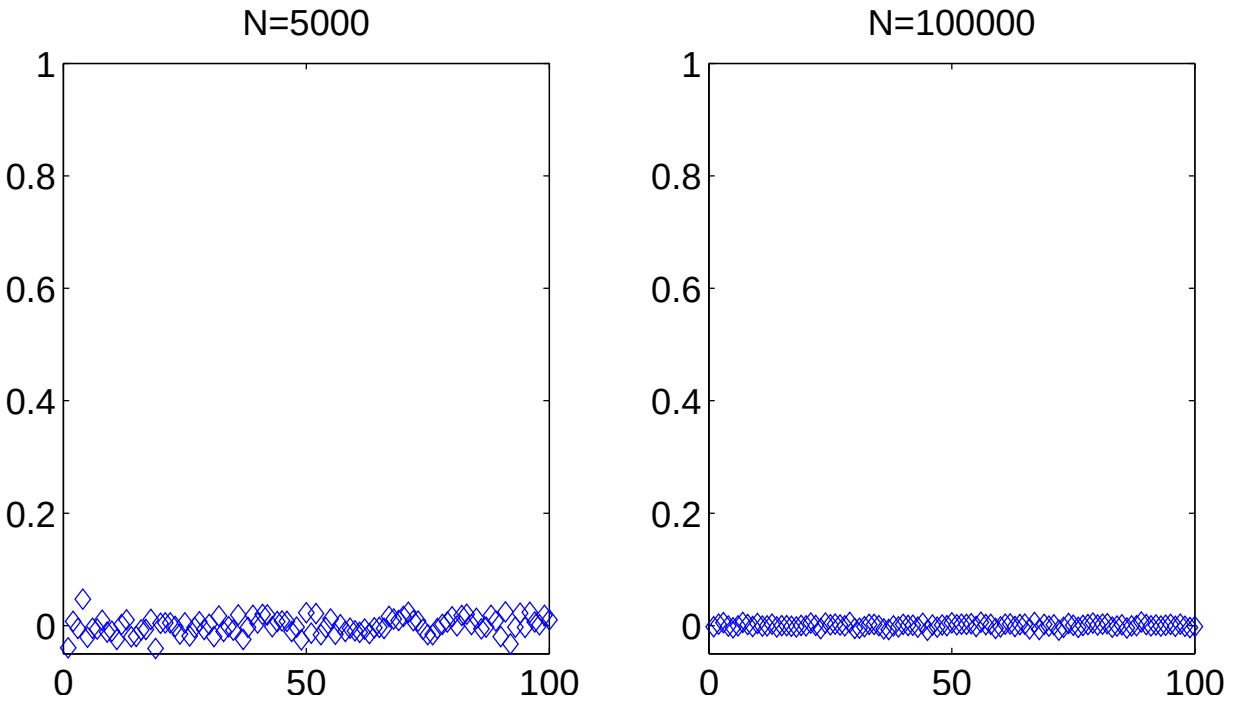
$$\gamma_k^2 = \frac{1}{\gamma_0^2 (N-k)} \sum_{j=1}^{N-k} f_{\mathrm{c}}(x_j) f_{\mathrm{c}}(x_{j+k}), \qquad k \geq 1,$$

where $\gamma_0^2 = \frac{1}{N} \sum_{j=1}^{N} f_{\mathrm{c}}(x_j)^2$ is the mean energy of the signal.

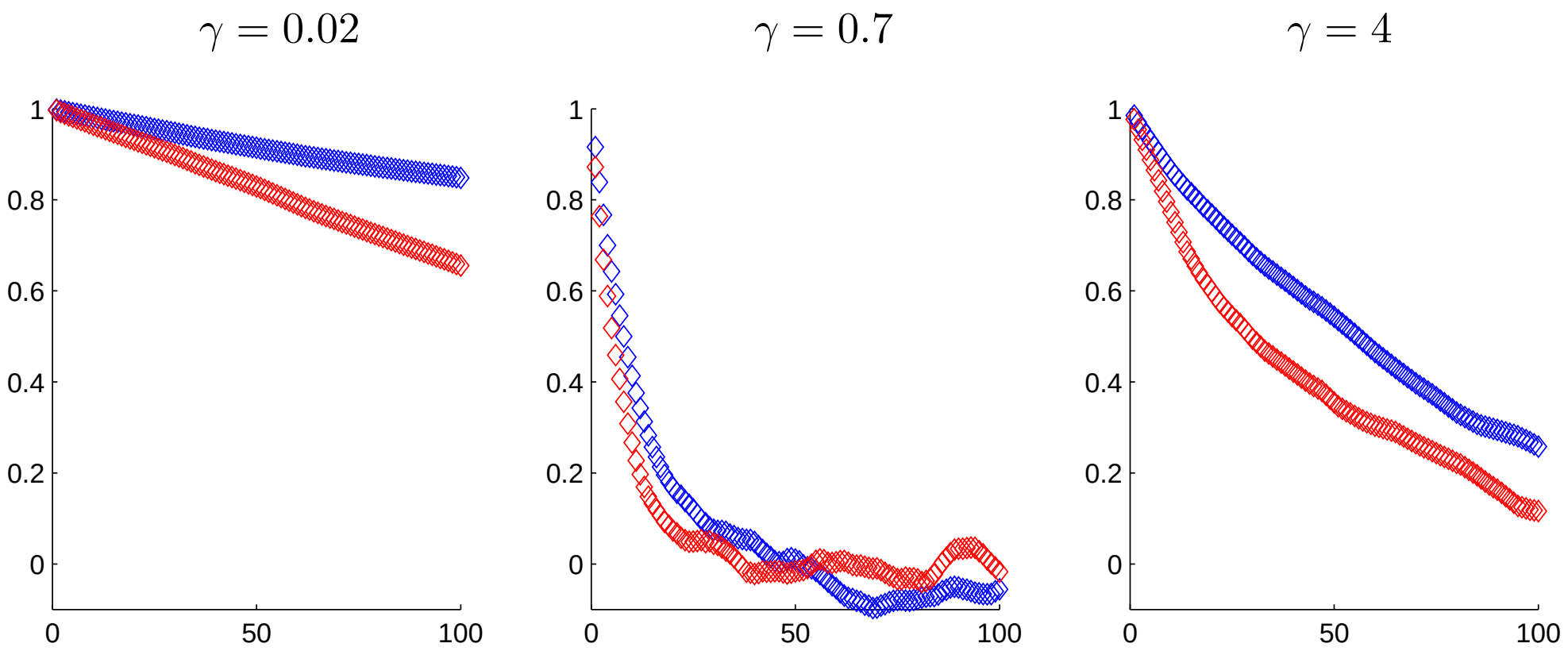
The correlation length of the sample $\{f(x_j)\}_{j=1}^{N}$ can be estimated based on the decay of the normalized autocovariance sequence of the sample.

For a white noise sample, $\gamma_k^2 \approx 0$ for any $k > 0$, where the estimate gets better as the sample, i.e., N, increases.

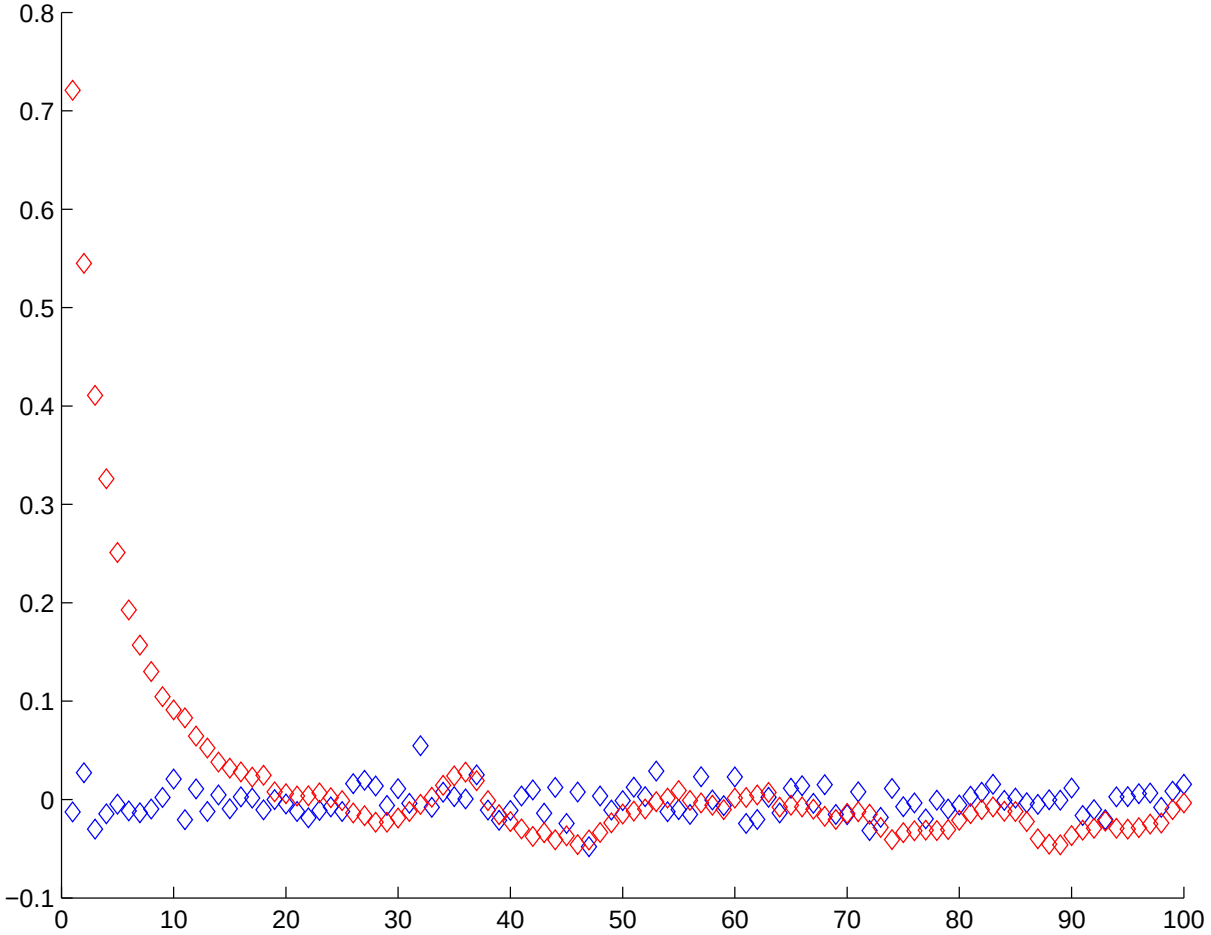
We test this hypothesis by drawing two white noise samples ($N = 5000$ and $N = 100000$) and plotting the function $k \mapsto \gamma_k^2$ in both cases.



Normalized autocovariance sequences for the MH example.



Normalized autocovariances for the Gibbs example; horizontal component in blue and vertical in red.



Computational methods in inverse problems

Jenni Heino, Nuutti Hyvönen,
Matti Leinonen, Stratos Staboulis

nuutti.hyvonen@tkk.fi, matti.leinonen@tkk.fi,
stratos.staboulis@tkk.fi

Eighteenth lecture, March 25, 2011.

Hypermodels

In the statistical framework, the prior densities usually depend on some parameters such as variance or mean. Typically — or at least thus far —, these parameters are assumed to be known.

Some classical regularization methods can be viewed as construction of estimators based on the posterior density (e.g., Tikhonov regularization). The regularization parameter, which corresponds to the parameter that defines the prior distribution, is not assumed to be know, but selected using, e.g., the Morozov discrepancy principle.

What happens if it is not clear how to choose these 'prior parameters' in the statistical framework?

If a parameter is not know, it can be estimated as a part of the statistical inference problem based on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter α which is not assumed to be known. Then we write the prior as a conditional density, that is,

$$\pi_{\mathrm{pr}}(x \,|\, \alpha).$$

Assuming we have a hyperprior for α, i.e.,

$$\pi_{\mathrm{hyper}}(\alpha),$$

we can write the joint distribution of x and α as

$$\pi(x, \alpha) = \pi_{\mathrm{pr}}(x \,|\, \alpha)\pi_{\mathrm{hyper}}(\alpha).$$

Assuming a likelihood model $\pi(y \,|\, x)$ for the measurement data y, we get the posterior density for x and α, given y, from the Bayes formula:

$$\pi(x, \alpha \,|\, y) \propto \pi(y \,|\, x)\pi(x, \alpha) = \pi(y \,|\, x)\pi(x \,|\, \alpha)\pi_{\mathrm{hyper}}(\alpha).$$

In general, the hyperprior density π_{hyper} may depend on some hyperparameter α_0. In such a case, the main reason for the use of a hyperprior model is that the construction of the posterior is assumed to be more robust with respect to fixing a value for the hyperparameter α_0 than fixing a value for α.

Sometimes α_0 can also be treated as a random variable with a respective probability density. Then, we would write

$$\pi_{\text{hyper}}(\alpha \mid \alpha_0),$$

giving rise to nested hypermodels.

Example: Hypermodel for a deconvolution problem

(Adapted from the textbook by Calvetti and Somersalo, Chapter 10)

Consider a one-dimensional deconvolution problem, the goal of which is to estimate a signal $f : [0, 1] \to \mathbb{R}$ from noisy, blurred observations modelled as

$$y_i = g(s_i) = \int_0^1 \mathcal{A}(s_i, t) f(t) dt + e(s_i), \qquad 1 \le i \le m,$$

where $\{s_i\}_{i=1}^m \subset [0, 1]$ are the uniformly distributed measurement points, the blurring kernel is defined to be

$$\mathcal{A}(s, t) = \exp\left(-\frac{1}{2\omega^2}(t - s)^2\right),$$

and the noise is Gaussian, or more precisely $e \sim \mathcal{N}(0, \sigma^2 I)$.

To begin with, we discretize the model as

$$y = Ax + e,$$

where $A \in \mathbb{R}^{m \times n}$ is obtained by approximating the integral with a suitable quadrature rule, and the vector x contains the values of the unknown signal at the discretization points $\{t_j\}_{j=0}^{n}$ that we have chosen to be distributed uniformly over the interval $[0, 1]$. To be more precise,

$$x_j = f(t_j), \qquad t_j = \frac{j}{n}, \qquad 0 \leq j \leq n.$$

For simplicity we assume it is known that $f(0) = x_0 = 0$, and define the actual unknown x to be

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

Assume that as prior information we know that the signal is continuous except for a possible jump discontinuity at a *known* location.

Let us start with a Gaussian first order smoothness prior,

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2\right),$$

where L is a first order finite difference matrix (recall that $x_0 = 0$),

$$L = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

It is easy to see that L is invertible and

$$L^{-1} = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \ddots & \ddots & \\ 1 & \dots & 1 & 1 \end{bmatrix}$$

is a lower triangular matrix. Since $\frac{1}{\gamma}L$ is the whitening matrix of $X \in \mathbb{R}^n$ distributed according to $\pi_{\mathrm{pr}}(x)$ — see the twelfth lecture —, it follows that

$$X = L^{-1}W, \qquad W \sim \mathcal{N}(0, \gamma^2 I).$$

Due to the particular shape of L^{-1}, this relation can alternatively be given as a Markov process:

$$X_j = X_{j-1} + W_j, \qquad W_j \sim \mathcal{N}(0, \gamma^2),\ j = 1, \dots, n, \quad X_0 = 0.$$

Next, we aim at fine-tuning the the above smoothness prior so that it allows a jump discontinuity over the interval $[t_{k-1}, t_k]$.

To this end, we modify the above Markov model (only) at $j = k$ by setting

$$X_k = X_{k-1} + W_k, \qquad W_k \sim \mathcal{N}\left(0, \frac{\gamma^2}{\delta^2}\right),$$

where $\delta < 1$ is a parameter controlling the variance of W_k, i.e., the expected size of the jump.

Let us walk the the above steps backwards: It is easy to see that this new Markov process can alternatively be given as

$$X \;=\; L^{-1}(D^{1/2})^{-1}W, \qquad W \sim \mathcal{N}(0, \gamma^2 I),$$

where

$$D^{1/2} = \mathrm{diag}(1, 1, \ldots, \delta, \ldots, 1, 1) \in \mathbb{R}^{n \times n}$$

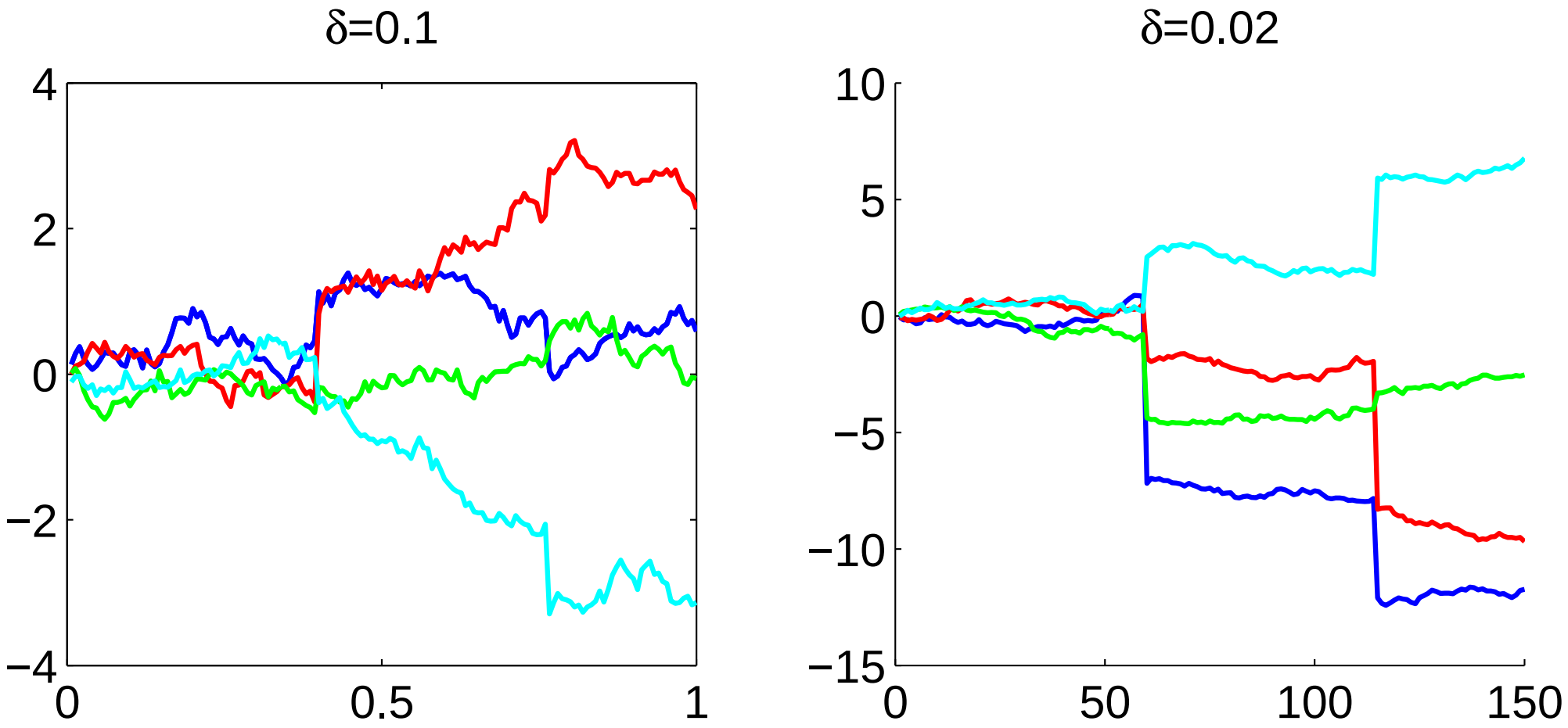
is defined so that $(D^{1/2})^{-1}$ scales the kth component of W by $1/\delta$.

In consequence, after the above modification in the kth step of the Markov process *defining* X, the random variable $D^{1/2}LX$ is distributed according to $\mathcal{N}(0, \gamma^2 I)$, and thus we have introduced the fine-tuned 'jump prior'

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|D^{1/2}Lx\|^2\right).$$

Let us draw samples from this kind of a prior density. We set $n = 150$ and $\gamma = 0.1$, meaning that we expect increments of the order 0.1 at most of the subintervals. As an exception, at two known locations $t \approx 0.4$ and $t \approx 0.8$ we use $\delta < 1$ at the corresponding diagonal element of $D^{1/2}$, in anticipation of a jump of the order $\gamma/\delta = 0.1/\delta$.

Random draws from the jump discontinuity prior with two different values of δ.



As the additive noise was assumed to be Gaussian, the likelihood density corresponding to the considered measurement is

$$\pi(y \,|\, x) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2\right),$$

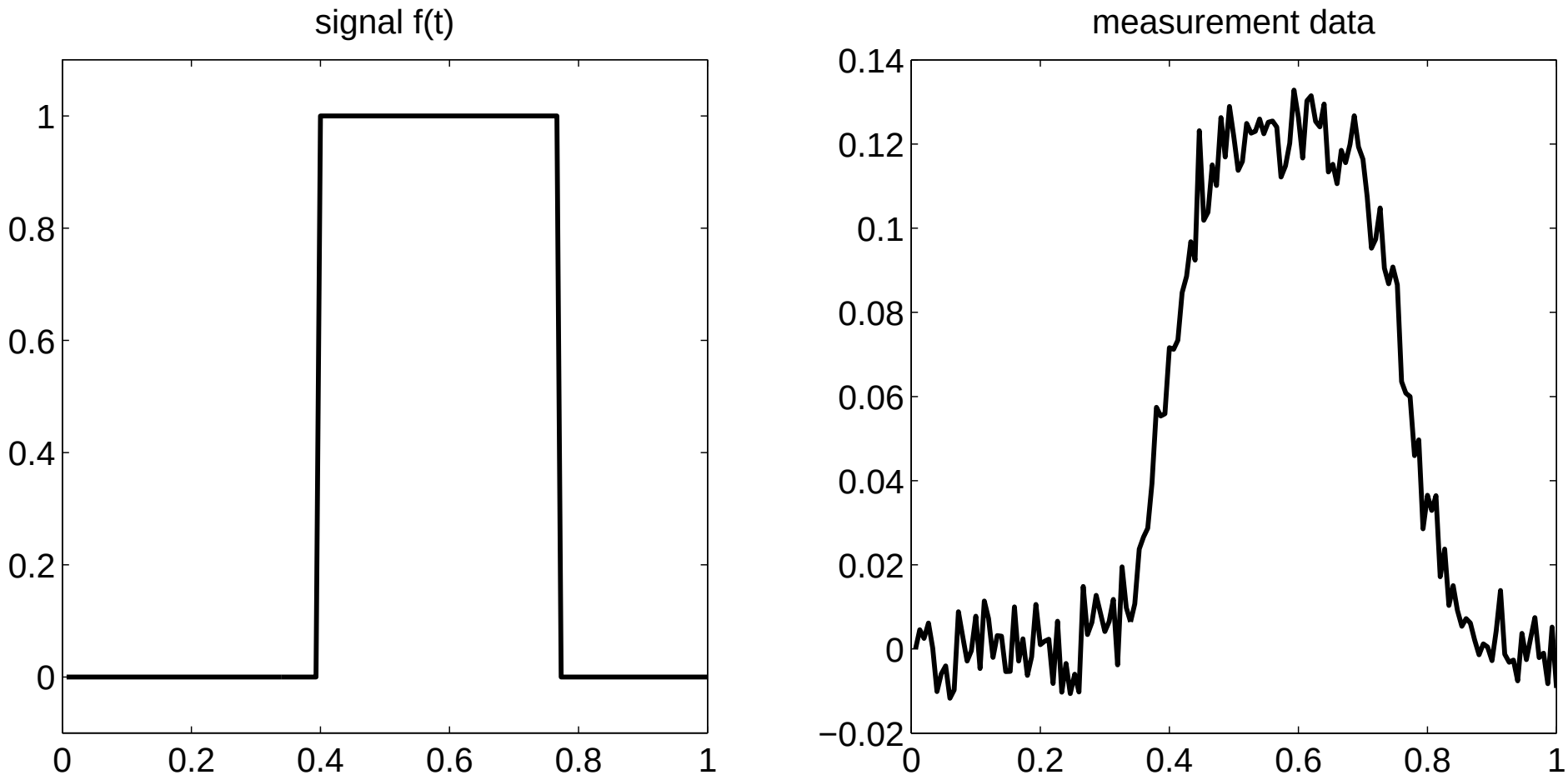
and due to the Bayes formula, the posterior density can thus be written as

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2 - \frac{1}{2\gamma^2}\|D^{1/2}Lx\|^2\right).$$

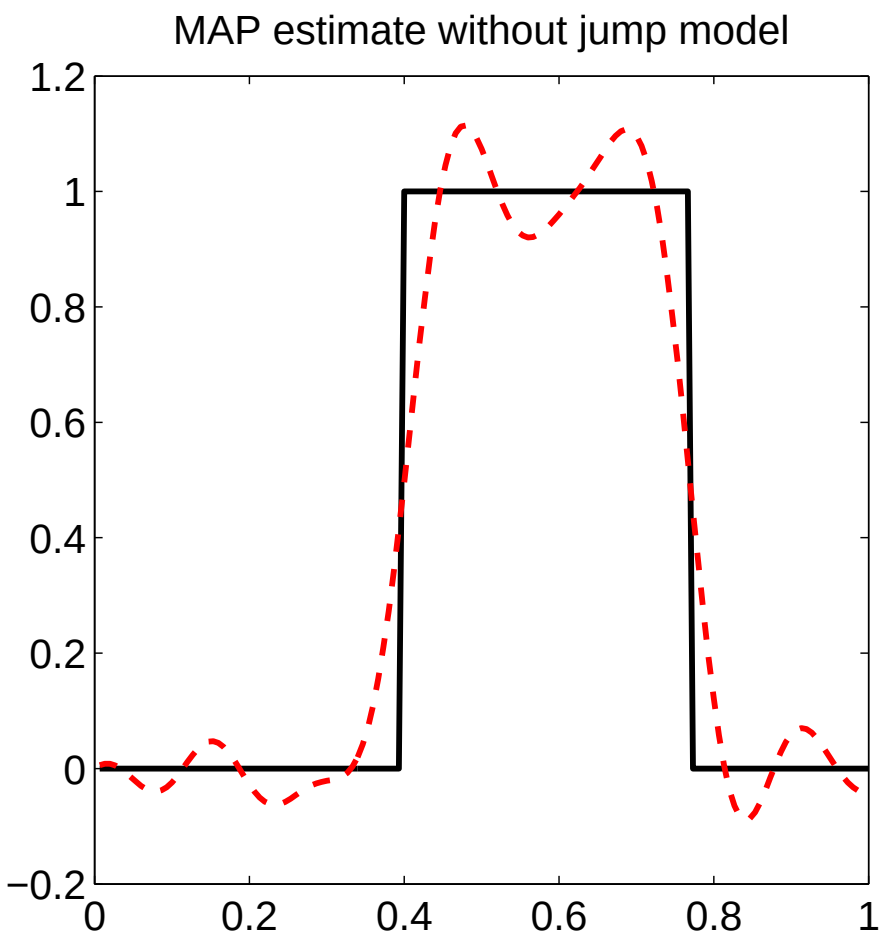
Using the results for Gaussian densities from previous lectures, the mean of the posterior, which is also the MAP and the CM estimate, can be written explicitly as

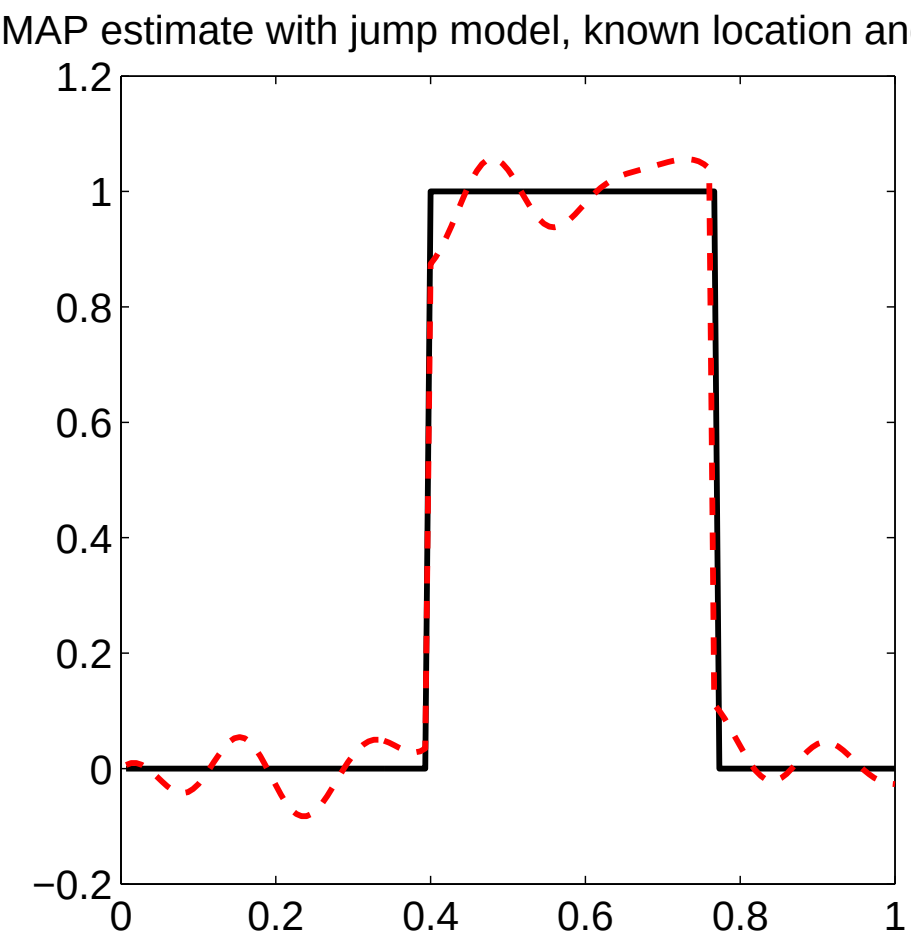
$$x_{\mathrm{CM}} = x_{\mathrm{MAP}} = \left(\frac{\sigma^2}{\gamma^2}L^{\mathrm{T}}(D^{1/2})^{\mathrm{T}}D^{1/2}L + A^{\mathrm{T}}A\right)^{-1} A^{\mathrm{T}}y.$$

The original signal $f(t)$ and the measurement data ($\omega \approx 0.05$):

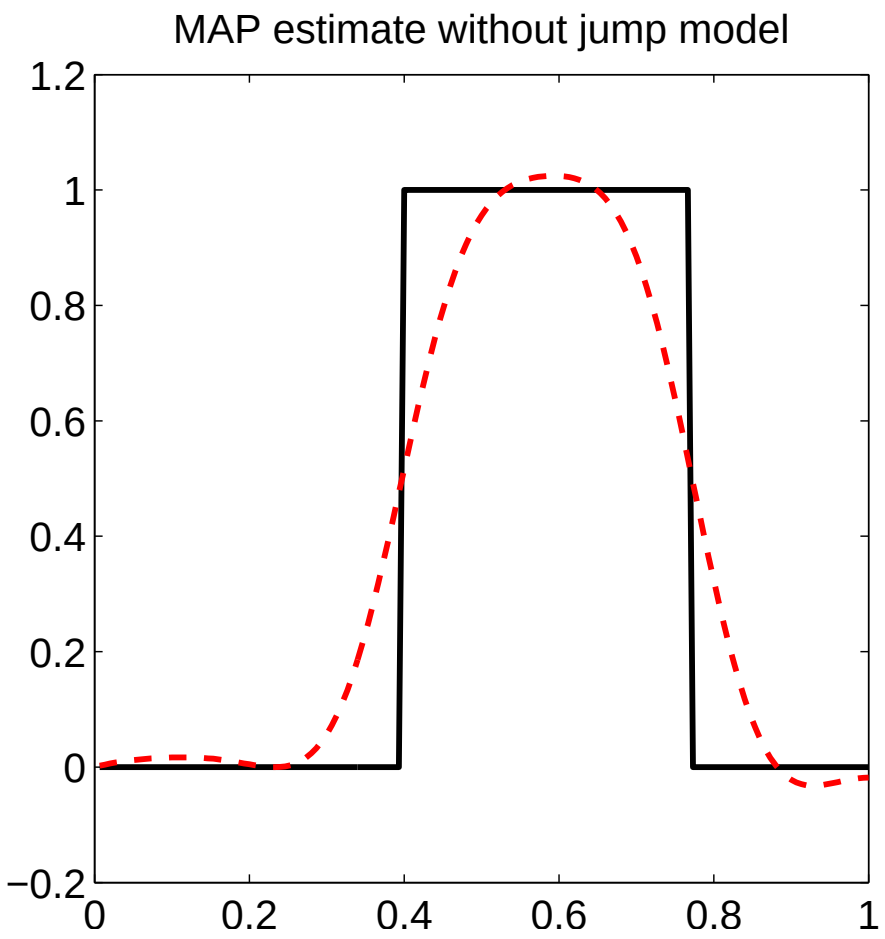


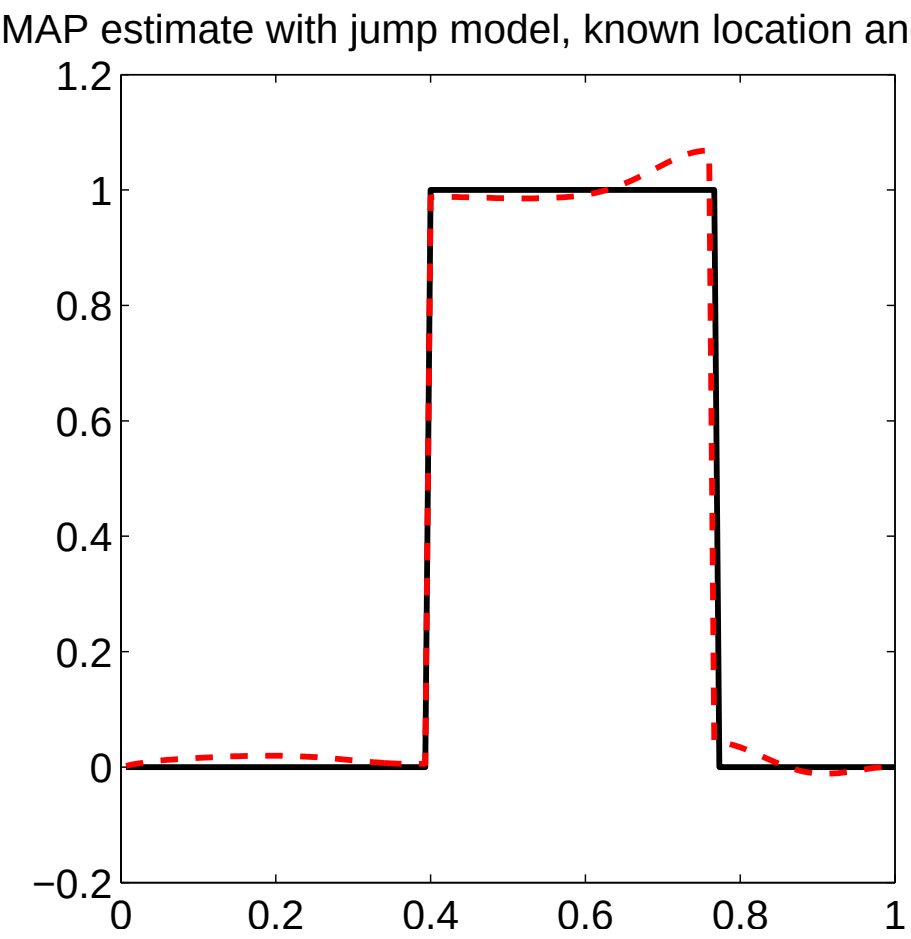
Posterior estimates for f without the discontinuity model (i.e., with the mere first order smoothness prior) and with the discontinuity model with known locations and jump sizes ($\gamma = 0.1$):





Next we choose $\gamma = 0.01$ that corresponds to increments of the order of 0.01 at each subinterval, and scale δ accordingly so that it is in accordance with jump sizes of the order 1.





Assume next that the locations and expected sizes of the jumps are not known, but we expect *a slowly varying signal that could have a few jumps at unknown locations*.

We modify the Markov model to allow different increments at different positions:

$$X_j = X_{j-1} + W_j, \qquad W_j \sim \mathcal{N}\left(0, \frac{1}{\theta_j}\right), \quad \theta_j > 0, \quad j = 1, \ldots, n.$$

The corresponding prior model can be obtained in the same way as above:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}\|D^{1/2}Lx\|^2\right),$$

where this time around

$$D^{1/2} = \mathrm{diag}(\theta_1^{1/2}, \theta_2^{1/2}, \ldots, \theta_n^{1/2}).$$

If we knew the vector $\theta = [\theta_1, \ldots, \theta_n]^{\mathrm{T}}$, we could proceed as previously.

If $\theta \in \mathbb{R}^n$ is not know, it can be considered as a random variable and its estimation can be included as a part of the inference problem. To this end, we need to write the conditional density

$$\pi_{\mathrm{pr}}(x \,|\, \theta).$$

In this case, the normalizing constant of the density $\pi_{\mathrm{pr}}(x \,|\, \theta)$ is no longer a constant, but depends on the random variable θ and thus *cannot* be ignored.

Recall the probability density of a n-variate Gaussian distribution:

$$\pi(z) \,=\, \left(\frac{1}{(2\pi)^n \det(\Gamma)}\right)^{1/2} \exp\left(-\frac{1}{2} z^{\mathrm{T}} \Gamma^{-1} z\right),$$

where the mean is assumed to be zero.

In our case, $\Gamma = (L^{\mathrm{T}} D L)^{-1}$, where $D = \mathrm{diag}(\theta) \in \mathbb{R}^{n \times n}$. Recall that the determinant of a triangular matrix is the product of its diagonal elements, meaning that $\det(L) = \det(L^{\mathrm{T}}) = 1$. Moreover, the determinant of an inverse matrix is the inverse of the determinant of the original matrix. Hence, it holds that

$$\det(\Gamma)^{-1} \;=\; \det(L^{\mathrm{T}} D L) = \det(L^{\mathrm{T}}) \det(D) \det(L) = \prod_{j=1}^{n} \theta_j,$$

and the properly normalized density becomes

$$\begin{aligned} \pi_{\mathrm{pr}}(x \,|\, \theta) \;&=\; \left(\frac{\prod_{j=1}^{n} \theta_j}{(2\pi)^n} \right)^{1/2} \exp\left(-\frac{1}{2} \| D^{1/2} L x \|^2 \right) \\ &=\; \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \| D^{1/2} L x \|^2 + \frac{1}{2} \sum_{j=1}^{n} \log \theta_j \right) \Bigg). \end{aligned}$$

Next we need to choose a hyperprior density for θ. Qualitatively, we should allow some components of θ to deviate strongly from the 'average'.

We decide to use an ℓ_1-type impulse prior with a positivity constraint:

$$\pi_{\text{hyper}}(\theta) \propto \pi_+(\theta) \exp\left(-\frac{\gamma}{2}\sum_{j=1}^{n}\theta_j\right)$$

where $\pi_+(\theta)$ is one if all components of θ are positive, and zero otherwise, and $\gamma > 0$ is a hyperparameter.

The posterior distribution can then be written as

$$\pi(x, \theta \,|\, y) \propto \pi(y \,|\, x)\pi(x, \theta) = \pi(y \,|\, x)\pi(x \,|\, \theta)\pi_{\text{hyper}}(\theta)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2 - \frac{1}{2}\|D^{1/2}Lx\|^2 - \frac{\gamma}{2}\sum_{j=1}^{n}\theta_j + \frac{1}{2}\sum_{j=1}^{n}\log\theta_j \right)$$

if all components of θ are positive, and $\pi(x, \theta \,|\, y) = 0$ otherwise. It is straightforward to see that the corresponding MAP estimate is the minimizer of the functional

$$F(x, \theta) = \left\| \begin{bmatrix} \frac{1}{\sigma}A \\ D^{1/2}L \end{bmatrix} x - \begin{bmatrix} \frac{1}{\sigma}y \\ 0 \end{bmatrix} \right\|^2 + \gamma\sum_{j=1}^{n}\theta_j - \sum_{j=1}^{n}\log\theta_j.$$

over $(x, \theta) \in \mathbb{R}^n \times \mathbb{R}^n_+$.

We apply a two stage minimization algorithm:

Choose some initial guesses for x and θ. Then, repeat the following two steps until convergence is achieved:

1. Keep θ fixed and update x to be the least squares solution of

$$\begin{bmatrix} \frac{1}{\sigma}A \\ D^{1/2}L \end{bmatrix} x = \begin{bmatrix} \frac{1}{\sigma}y \\ 0 \end{bmatrix},$$

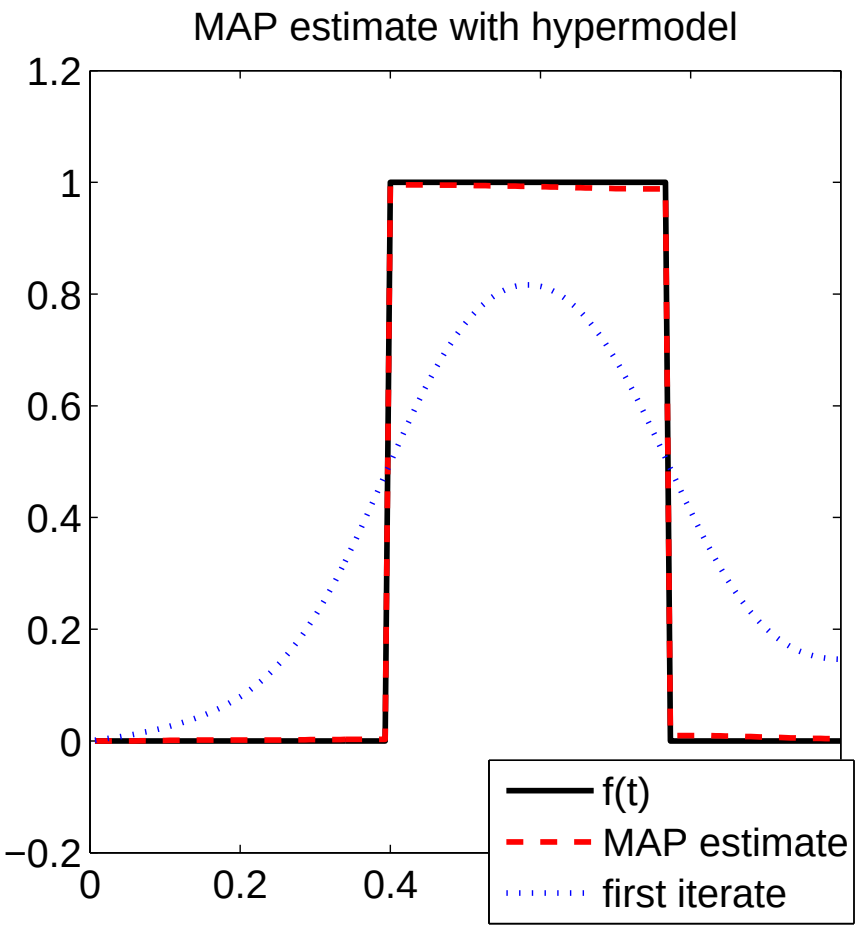
   where $D = \mathrm{diag}(\theta)$.

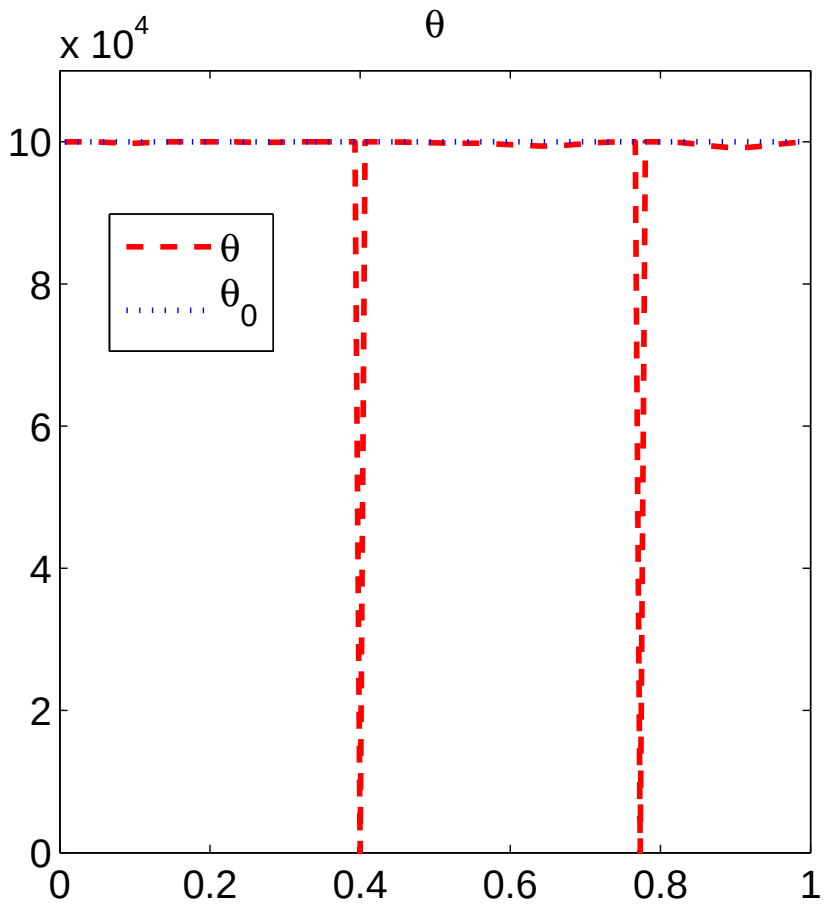
2. Fix x and update θ by minimizing $F(x, \cdot)$ with respect to the second variable. An easy calculation shows that this minimizer can be given componentwise as

$$\theta_j = \frac{1}{w_j^2 + \gamma}, \qquad j = 1, \ldots, n,$$

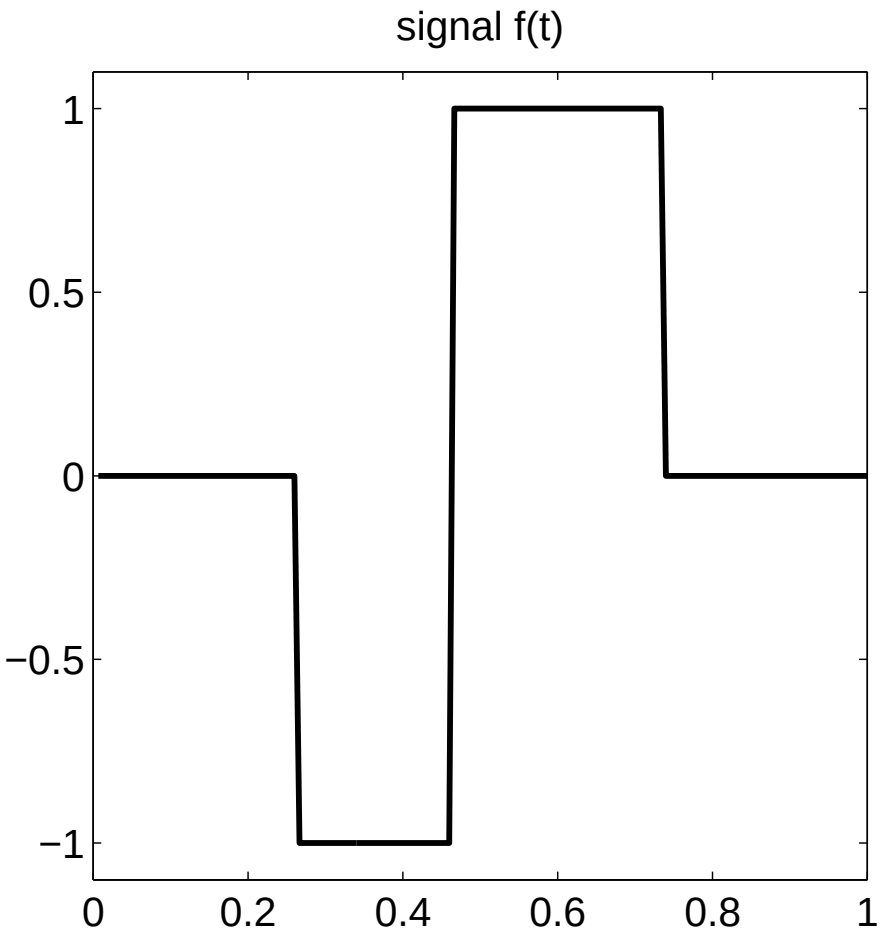
   where $w = Lx \in \mathbb{R}^n$ is the vector of increments corresponding to x.

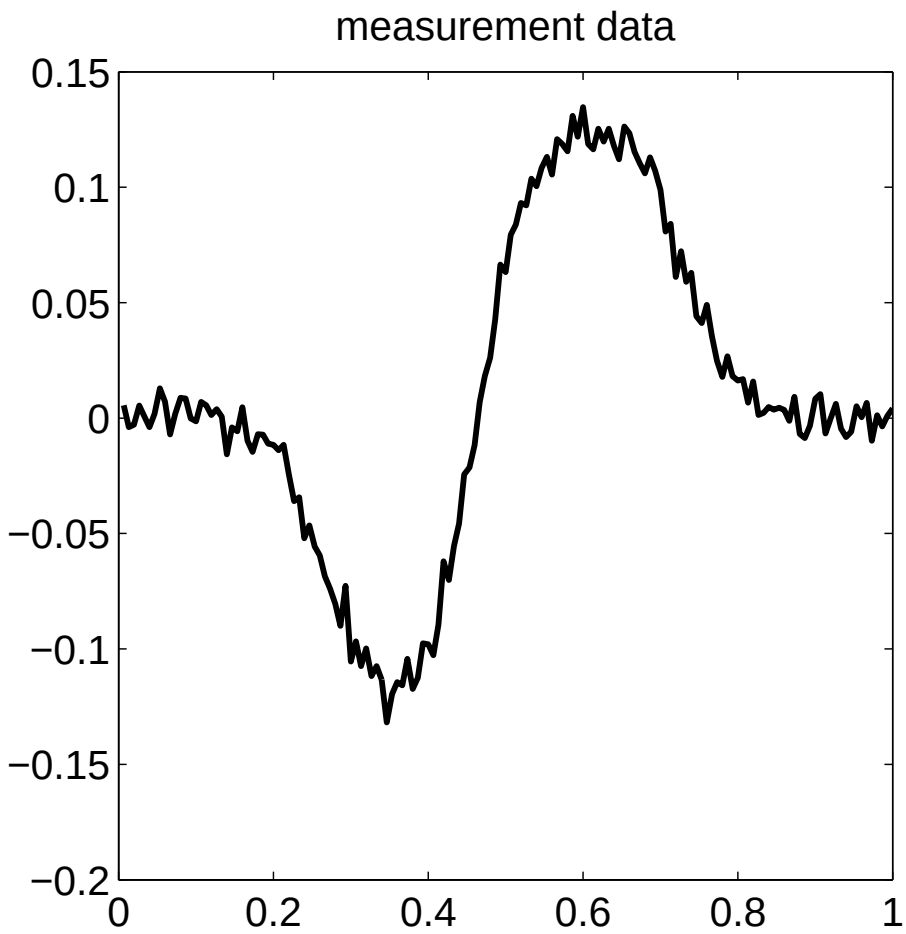
MAP estimates for x and θ provided by the above alternating algorithm with $\gamma = 10^{-5}$ and the initial guesses $x_0 = 0$ and $\theta_{0,j} = 1/\gamma$, $j = 1, \ldots, n$. The data is the same as depicted on page 448.



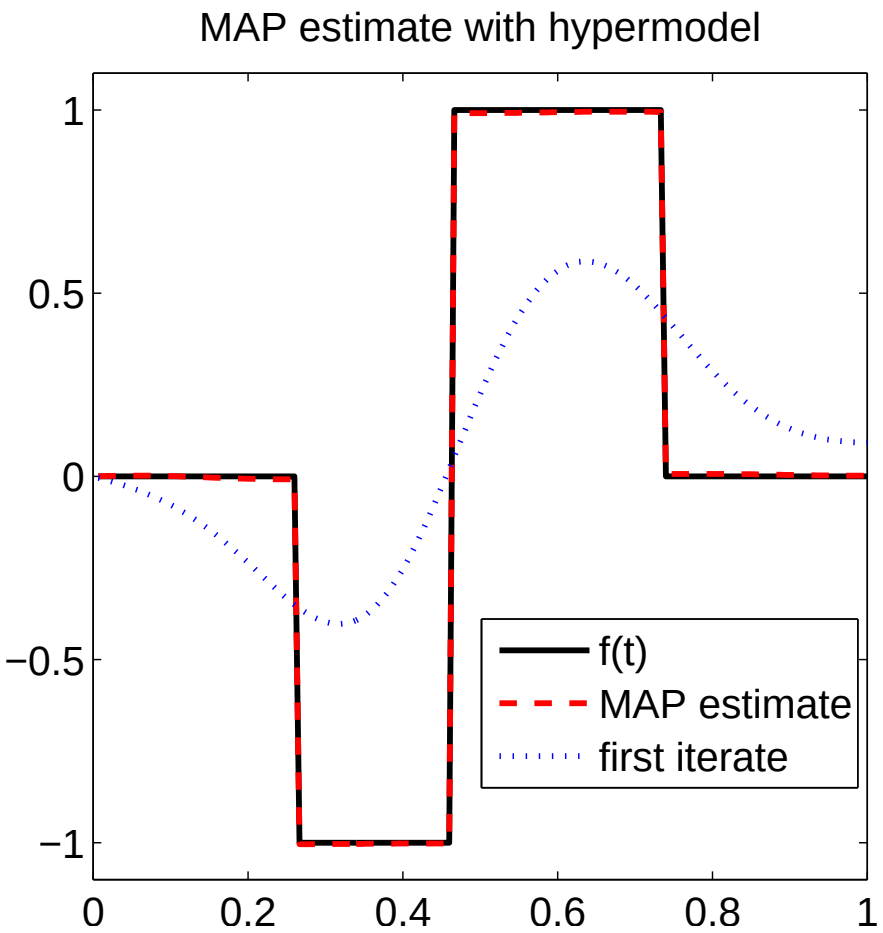


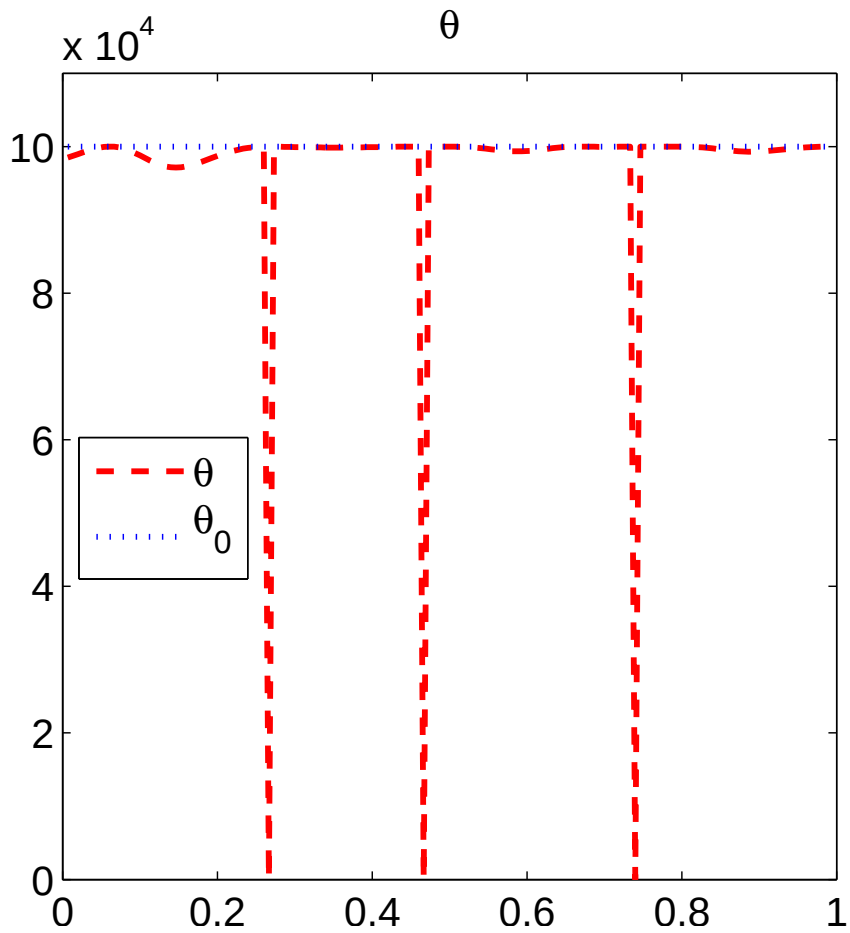
Another example: The original signal $f(t)$ and the measurement data.





MAP estimates for x and θ provided by the above alternating algorithm with $\gamma = 10^{-5}$ and the initial guesses $x_0 = 0$ and $\theta_{0,j} = 1/\gamma$, $j = 1, \ldots, n$.





The End.