Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

First lecture, April 17, 2023

Practical matters

- Lectures on Mondays at 10:15-12:00 in A6/025/026 by Vesa Kaarnioja.
- Exercises on Tuesdays at 10:15-12:00 in A6/007/008 by Vesa Kaarnioja starting next week.
- Weekly exercises published after each lecture. Please return your written solutions to Vesa either by email (vesa.kaarnioja@fu-berlin.de) or at the beginning of the exercise session in the following week.
- The conditions for completing this course are *successfully completing and submitting at least 60% of the course's exercises* and *successfully passing the course exam*.

Course contents

- The first part of the course will cover classical variational regularization methods. We will follow Chapters 1–4 in
  - J. Kaipio and E. Somersalo (2005). Statistical and Computational Inverse Problems. Springer, New York, NY.
- Second part of the course will cover Bayesian inverse problems. We will follow the texts
  - D. Sanz-Alonso, A. M. Stuart, and A. Taeb (2018). Inverse Problems and Data Assimilation. https://arxiv.org/abs/1810.06191
  - J. Kaipio and E. Somersalo (2005). Statistical and Computational Inverse Problems. Springer, New York, NY.
  - D. Calvetti and E. Somersalo (2007). Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing. Springer, New York, NY.

What is an inverse problem?

- **Forward problem:** Given known causes (initial conditions, material properties, other model parameters), determine the effects (data, measurements).
- **Inverse problem:** Observing the effects (noisy data), recover the cause.

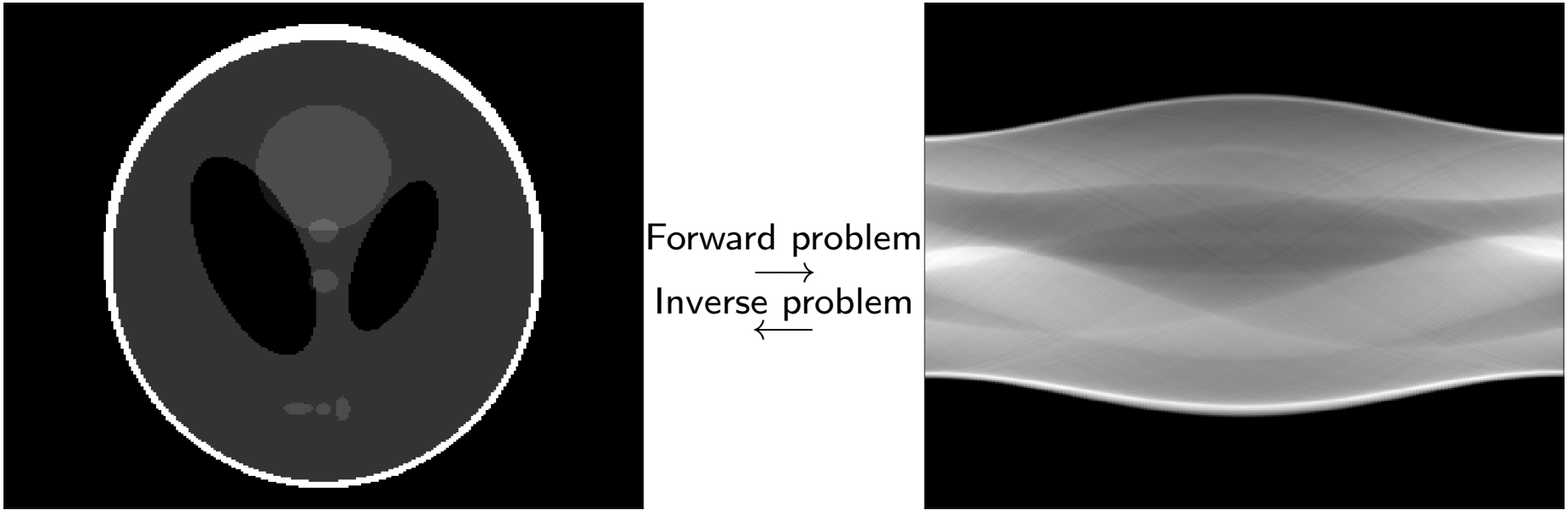
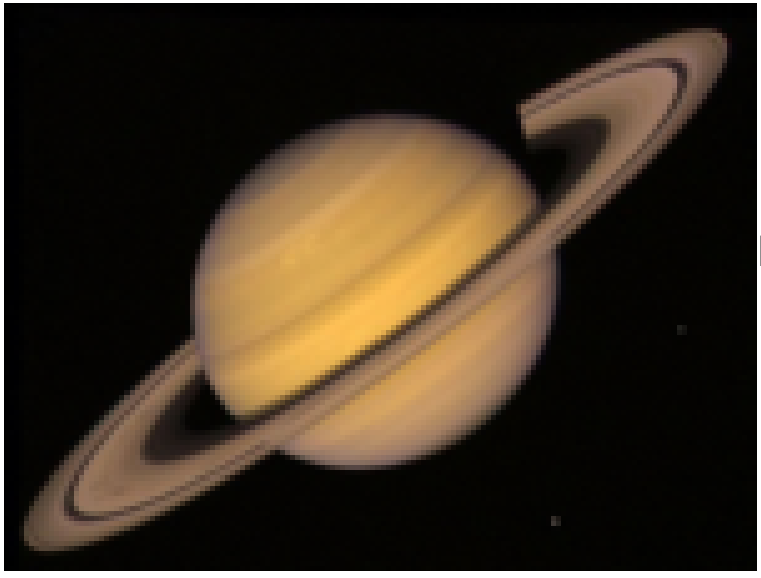


Figure: Computerized tomography (CT)



Forward problem $\longrightarrow$

Inverse problem $\longleftarrow$

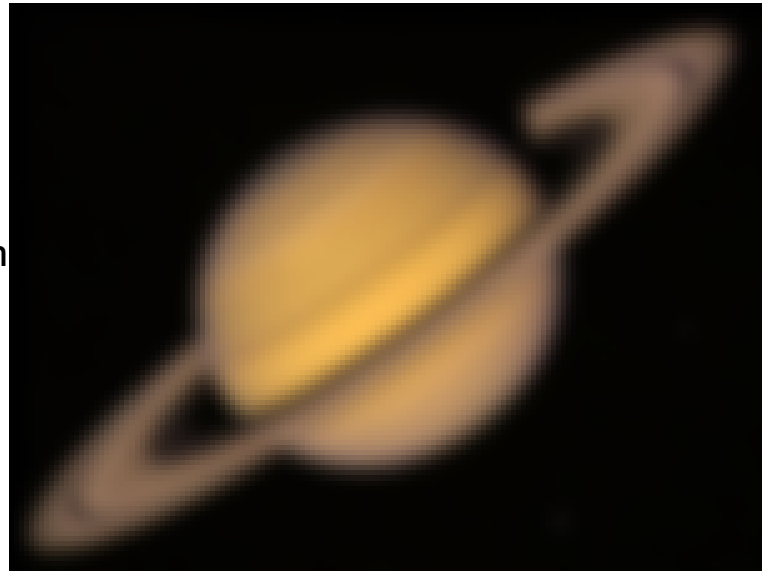


Figure: Image deblurring (deconvolution)

$$y = (K * f)(x) = \int_{\mathbb{R}^2} K(x - x')f(x')\,\mathrm{d}x'$$

Introduction: What is an inverse problem?

We consider the indirect measurement of an unknown physical quantity $x \in X$. The measurement $y \in Y$ is related to the unknown by a physical or mathematical *model*

$$y = F(x), \tag{1}$$

where $F\colon X \to Y$ is called the *forward mapping*.

- Computing y for a given x is called the *forward problem*.
- Finding x for a given measurement y (the *data*) is called the *inverse problem*.

The inverse problem is often ill-posed, making it more difficult than the corresponding direct problem.

A problem is called *well-posed (in the sense of Hadamard)*, if

(a) a solution exists,

(b) the solution is unique, and

(c) the solution depends continuously on the data.

If one or more of these conditions are violated, the problem is called *ill-posed*.

Some examples of ill-posed inverse problems are X-ray tomography, image deblurring, the inverse heat equation, and electrical impedance tomography (EIT).

The ill-posedness of an inverse problem poses a challenge because usually, errors are present in the measurements. Incorporating these into model (1) in the form of additive *noise* η leads to a more realistic model

$$y = F(x) + \eta.$$

The violation of the above conditions leads to various difficulties.

- If condition (a) is violated, i.e., if the image $\mathrm{Ran}(F)$ of F does not cover the whole space Y, then there may not exist a solution to $F(x) = y$ for noisy data $y = F(x^\dagger) + \eta$ created by a ground truth $x^\dagger$, although a solution exists for noise free data $y = F(x^\dagger)$, since η does not need to lie in $\mathrm{Ran}(F)$.
- If condition (c) is violated, then the solution to $F(x) = y$ for noisy data $y = F(x^\dagger) + \eta$ may be far away from the solution for noise free data $y = F(x^\dagger)$, even if F is invertible and the noise η is small, due to the discontinuity of F^{-1}.

Example.
The deblurring (or deconvolution) problem of recovering an input signal x from an observed signal y (possibly contaminated by noise) occurs in many imaging as well as image and signal processing applications. The mathematical model is

$$y(t) = \underbrace{\int_{-\infty}^{\infty} a(t-s)x(s)\mathrm{d}s}_{=:(a*x)(t)},$$

where the function a is known as the blurring kernel.

If $\widehat{a}$ is "nice", we can use the Fourier transform together with the convolution theorem to solve the problem analytically:

$$y(t) = (a * x_{\text{exact}})(t) \quad \Leftrightarrow \quad \widehat{y}(\xi) = \widehat{a}(\xi)\widehat{x}_{\text{exact}}(\xi) \quad \Leftrightarrow \quad \widehat{x}_{\text{exact}}(\xi) = \frac{\widehat{y}(\xi)}{\widehat{a}(\xi)}$$

$$\Leftrightarrow \quad x_{\text{exact}}(t) = \mathcal{F}^{-1}\left\{\frac{\widehat{y}}{\widehat{a}}\right\}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{it\xi}\frac{\widehat{y}(\xi)}{\widehat{a}(\xi)}\mathrm{d}\xi.$$

Here, x_{exact} denotes the solution to this problem with *exact, noiseless data*.

However, if we can only observe noisy measurements, we must consider

$$y(t) = (a * x)(t) + \eta(t) \quad \Leftrightarrow \quad \widehat{y}(\xi) = \widehat{a}(\xi)\widehat{x}(\xi) + \widehat{\eta}(\xi).$$

The solution formula from the previous slide gives (in the Fourier side)

$$\widehat{x}(\xi) = \frac{\widehat{y}(\xi)}{\widehat{a}(\xi)} = \widehat{x}_{\text{exact}}(\xi) + \frac{\widehat{\eta}(\xi)}{\widehat{a}(\xi)};$$

then we apply the inverse Fourier transform on both sides. However, this reconstruction might not be well-defined and it is typically not stable, i.e., it does not depend continuously on the data y. The kernel a usually decreases exponentially (or has compact support). A typical example is a Gaussian kernel

$$a(t) = \frac{1}{2\pi\alpha^2} \exp\left(-\frac{t^2}{2\alpha^2}\right)$$

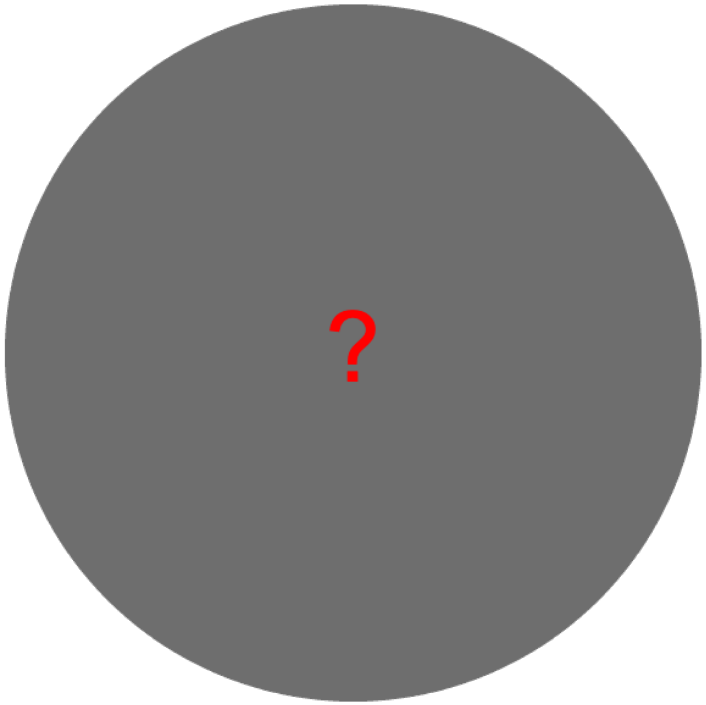
for some $\alpha > 0$.

By the Plancherel theorem, $\widehat{a} \in L^2(\mathbb{R})$ and

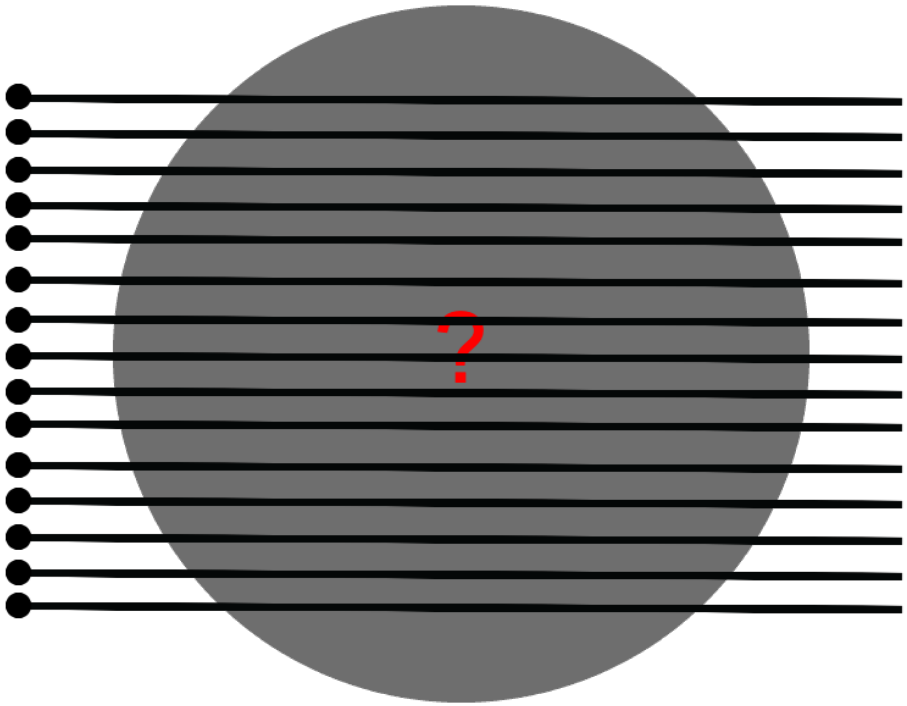
$$\int_{-\infty}^{\infty} |a(t)|^2 \mathrm{d}t = \int_{-\infty}^{\infty} |\widehat{a}(\xi)|^2 \mathrm{d}\xi$$

if $a \in L^2(\mathbb{R})$. This implies in particular that $\widehat{a}(\xi) \to 0$ as $|\xi| \to \infty$. As a consequence, high frequencies $\widehat{\eta}(\xi)$ of the noise get amplified arbitrarily strong in $\widehat{x}$. Thus, even the presence of small noise can lead to large changes in the reconstruction.

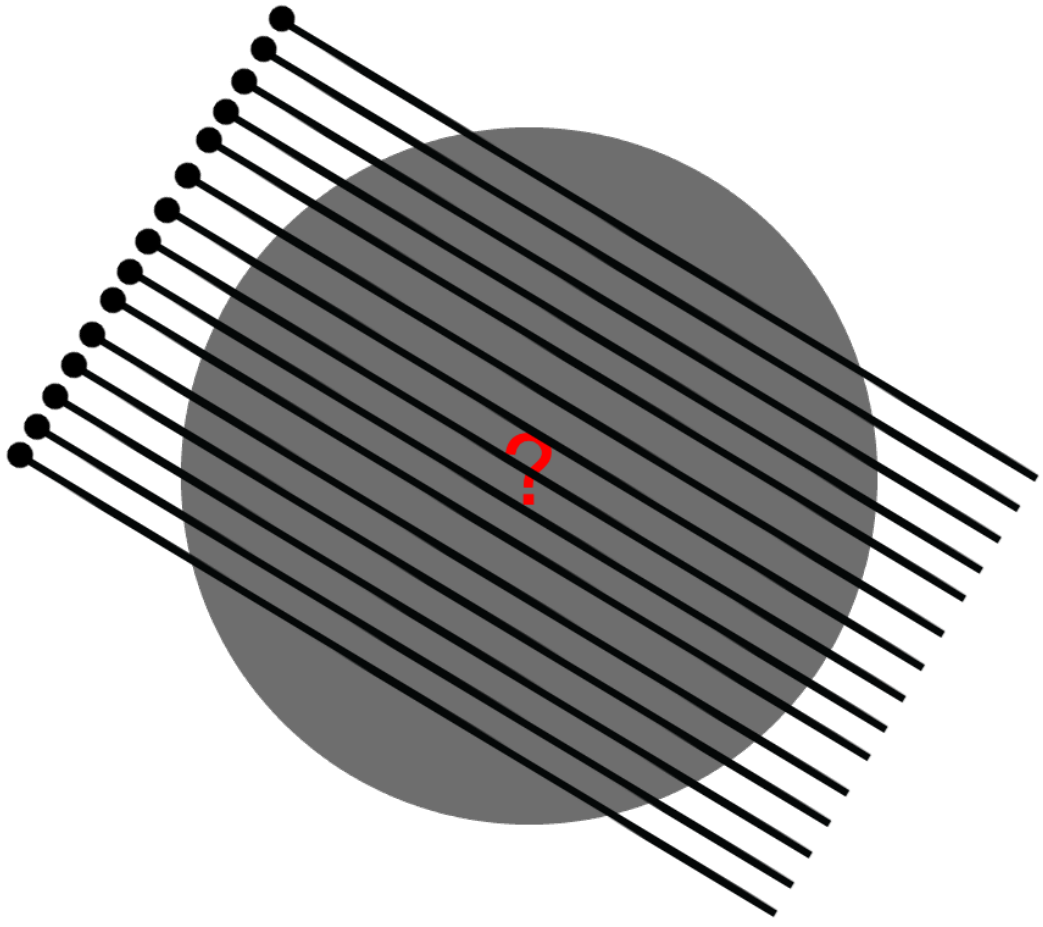
Case study: parallel-beam X-ray tomography



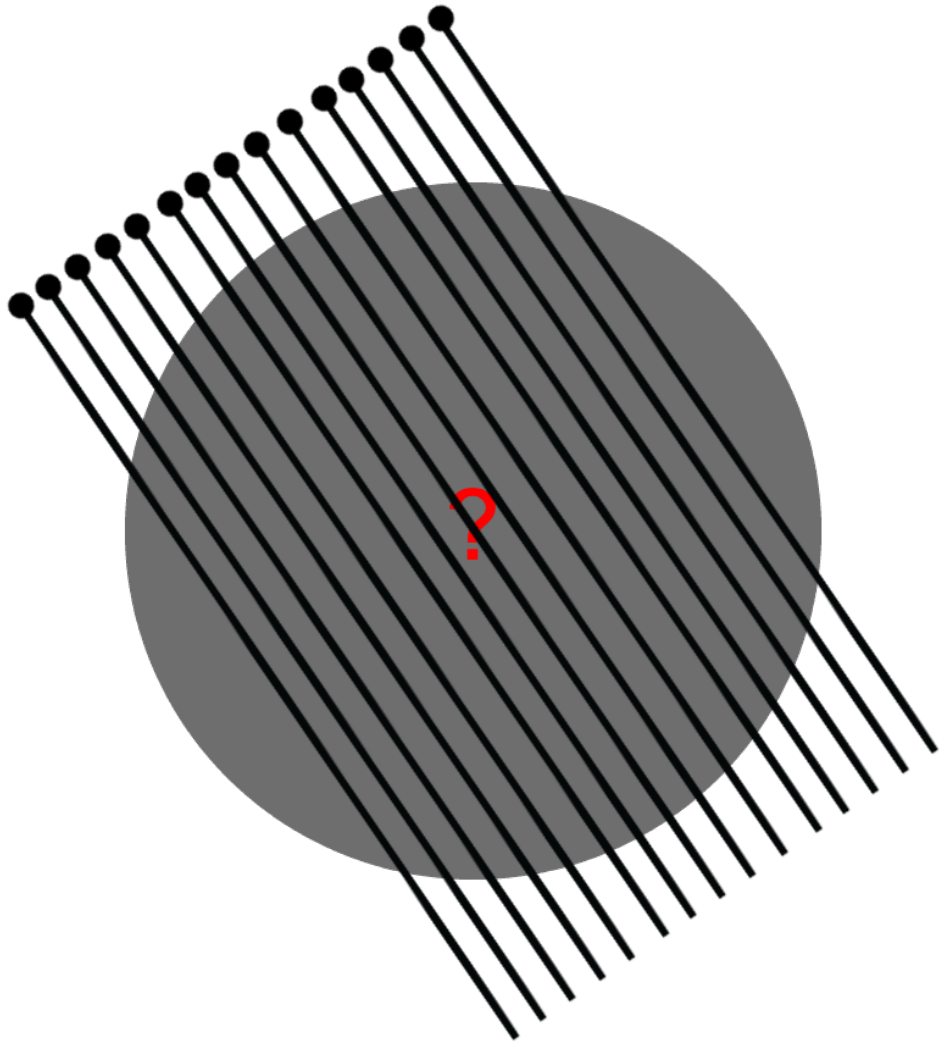
Case study: parallel-beam X-ray tomography



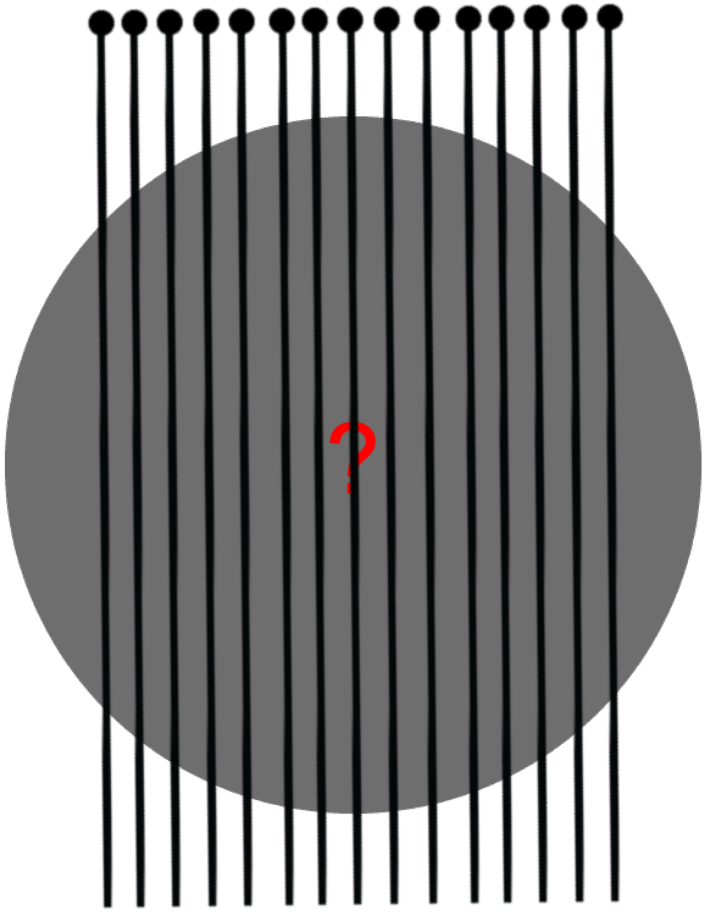
Case study: parallel-beam X-ray tomography



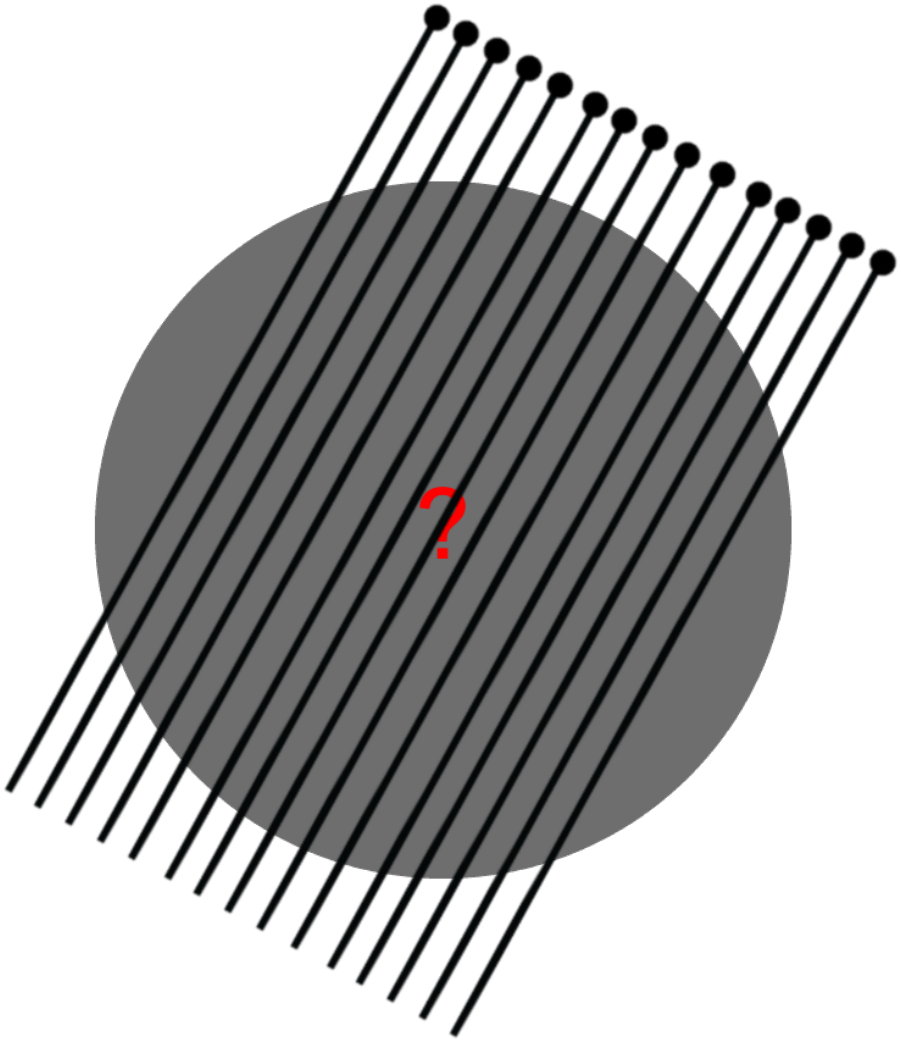
Case study: parallel-beam X-ray tomography



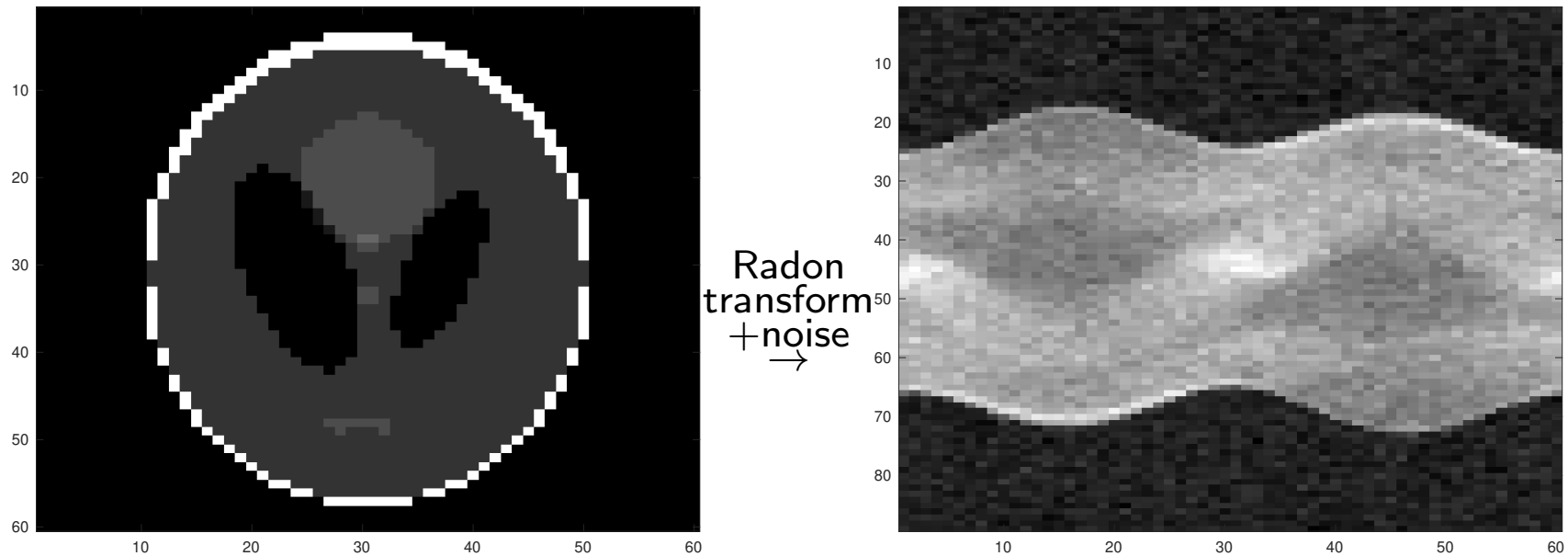
Case study: parallel-beam X-ray tomography



Case study: parallel-beam X-ray tomography



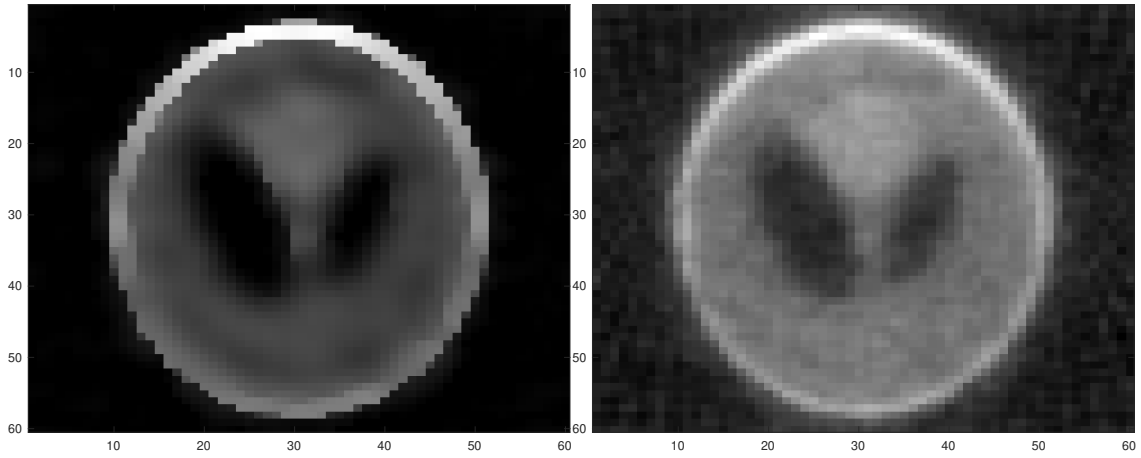
Let us consider the following phantom (botton left), which we use to simulate measurements taken from 60 angles contaminated with 5 % Gaussian noise (sinogram on the bottom right). Inverse problem: use the sinogram data (X-ray images taken from the different directions) to reconstruct the internal structure of the physical body (i.e., the phantom).



Technical (but important) note: to avoid the so-called inverse crime, the measurements for the inversion on the following page were generated using a higher resolution phantom.

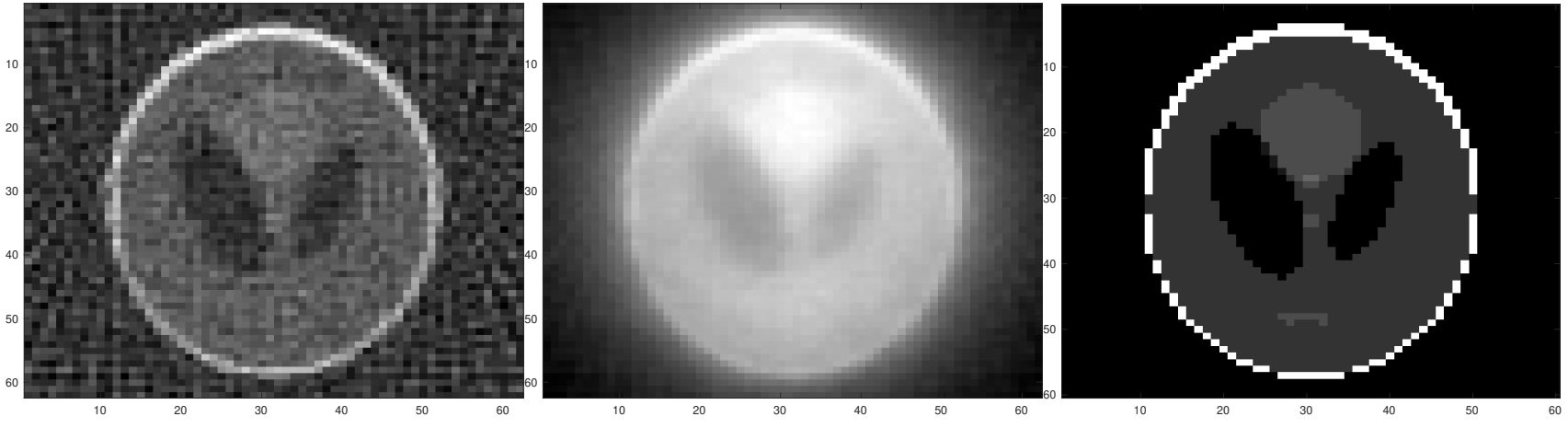
Formation of a CT sinogram (Samuli Siltanen):
https://www.youtube.com/watch?v=q7Rt_OY_7tU

Reconstructions $\arg\min_x \{\|Ax - m\|^2 + \mathcal{R}(x)\}$ from noisy measurements m with some selected penalty terms $\mathcal{R}$ are given immediately below.



Left: reconstruction with total variation regularization. Right: same with Tikhonov regularization.

Some other reconstructions for comparison (and the target phantom).

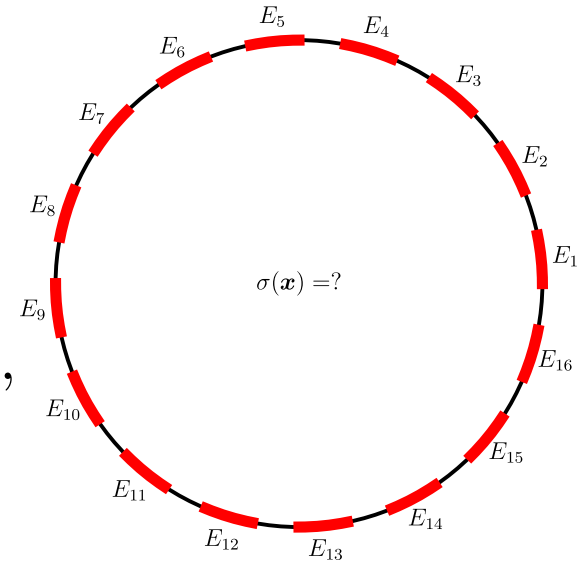


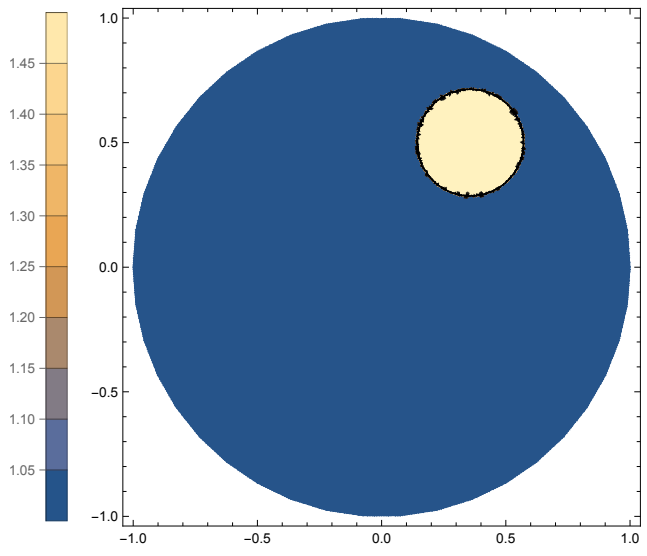
Left: filtered back projection. Middle: unfiltered back projection. Right: ground truth.

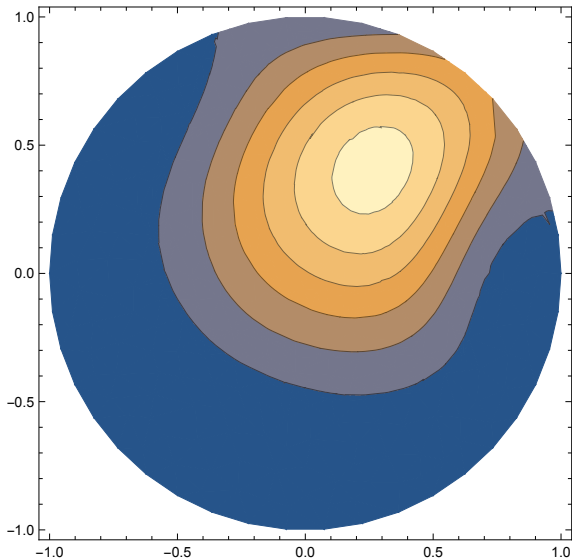
Electrical impedance tomography

Use measurements of current and voltage collected at electrodes covering part of the boundary to infer the interior conductivity of an object/body.

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 & \text{in } D, \\ \sigma \frac{\partial u}{\partial \boldsymbol{n}} = 0 & \text{on } \partial D \setminus \bigcup_{k=1}^{L} \overline{E_k}, \\ u + z_k \sigma \frac{\partial u}{\partial \boldsymbol{n}} = U_k & \text{on } E_k, \ k \in \{1, \ldots, L\}, \\ \int_{E_k} \sigma \frac{\partial u}{\partial \boldsymbol{n}} \, \mathrm{d}S = I_k, & k \in \{1, \ldots, L\}, \end{cases}$$







- Successful solution of inverse problems requires specially designed algorithms that can tolerate errors in measured data.
- How to incorporate all possible prior and expert knowledge about the possible solutions when solving inverse problems?
- The statistical approach to inverse problems aims to quantify how uncertainty in the data or model affects the solutions obtained in problems.

Preliminary functional analysis

Inner product space

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

- $\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ for all $x_1, x_2, y \in X$ and $a, b \in \mathbb{R}$;
- $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in X$;
- $\langle x, x \rangle \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.

A mapping $\langle \cdot, \cdot \rangle$ satisfying these conditions is called an *inner product*.

Example

i) $\mathbb{R}^n = \{(x_1, \ldots, x_n) \mid x_k \in \mathbb{R}\}$. Then the inner product is the Euclidean dot product

$$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k, \quad x = (x_1, \ldots, x_n),\ y = (y_1, \ldots, y_n).$$

ii) Let $X = C([a, b]) = \{f \mid f : [a, b] \to \mathbb{R} \text{ is continuous}\}$ and define

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,\mathrm{d}x.$$

Then this is an inner product on $C([a, b])$.

iii) Let $X = \ell^2(\mathbb{R}) = \left\{(z_k)_{k=1}^{\infty} \mid \sum_{k=1}^{\infty} |z_k|^2 < \infty\right\}$. Then $\ell^2(\mathbb{R})$ is an inner product space when

$$\langle x, y \rangle = \sum_{k=1}^{\infty} x_k y_k, \quad x = (x_1, x_2, \ldots),\ y = (y_1, y_2, \ldots).$$

Definition

A real vector space X is a *normed space* if there exists a mapping $\|\cdot\| : X \to \mathbb{R}$ satisfying

- $\|ax\| = |a|\|x\|$ for all $a \in \mathbb{R}$ and $x \in X$;
- $\|x\| \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

If X is an inner product space, then it is a normed space in a canonical way with the induced norm $\|\cdot\| : X \to \mathbb{R}$ defined by

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad x \in X.$$

The first two postulates follow immediately from the properties of inner product spaces, the triangle inequality follows from the Cauchy–Schwarz inequality.

Proposition (Cauchy–Schwarz inequality)

If $(X, \langle \cdot, \cdot \rangle)$ *is an inner product space, then*

$$|\langle x, y \rangle| \leq \|x\|\|y\| \quad \textit{for all } x, y \in X.$$

Proof. Let $x, y \in X$ and $t \in \mathbb{R}$. If $x = 0$ or $y = 0$, then the claim is trivial. Suppose that $x \neq 0 \neq y$. Then

$$0 \leq \langle x + ty, x + ty \rangle = \|x\|^2 + 2t\langle x, y \rangle + t^2\|y\|^2.$$

This is a second degree polynomial w.r.t. t with at most 1 real root. Hence,

$$\begin{aligned} \text{discriminant} \leq 0 \quad &\Leftrightarrow \quad 4|\langle x, y \rangle|^2 - 4\|x\|^2\|y\|^2 \leq 0 \\ &\Leftrightarrow \quad |\langle x, y \rangle|^2 \leq \|x\|^2\|y\|^2. \end{aligned}$$

Note that if $y = ax$, $a \in \mathbb{R}$, then discriminant $= 0$ and Cauchy–Schwarz holds with equality. □

The triangle inequality is an immediate consequence of Cauchy–Schwarz:

$$\begin{aligned} \|x + y\|^2 = \langle x + y, x + y \rangle &= \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x, y \rangle| \leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\ &= (\|x\| + \|y\|)^2 \quad \text{for all } x, y \in X. \end{aligned}$$

For our purposes, having an inner product is not enough. We need to know that these spaces are also *complete* normed spaces.

Definition (Cauchy sequence)

A sequence $(x_k)_{k=1}^{\infty}$ of elements of $(X, \|\cdot\|)$ is called a *Cauchy sequence* if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|x_m - x_n\| < \varepsilon.$$

Definition (Complete space)

A normed space $(X, \|\cdot\|)$ is *complete* if all Cauchy sequences in X converge to an element of X.

Definition (Banach space)

A normed space $(X, \|\cdot\|)$ which is complete with respect to $\|\cdot\|$ is a *Banach space*.

Definition (Hilbert space)

An inner product space $(H, \langle\cdot,\cdot\rangle)$ which is complete with respect to $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ defined by the inner product is a *Hilbert space*.

Example

i) $\mathbb{R}^n$ and $\ell^2(\mathbb{R})$ are complete.

ii) $C([a,b])$ is *not* complete w.r.t. the norm

$$\|f\|^2 = \int_a^b |f(x)|^2 \,\mathrm{d}x.$$

Let $a = -1$, $b = 1$, and define

$$f_n(x) := \begin{cases} 0, & -1 \le x < 0, \\ nx, & 0 \le x \le \frac{1}{n}, \\ 1, & \frac{1}{n} < x \le 1. \end{cases}$$

Then f_n is continuous, and if $H(x) = \chi_{[0,1]}(x) = \begin{cases} 0, & -1 \le x \le 0, \\ 1, & 0 < x \le 1, \end{cases}$ we have

$$\int_{-1}^1 |f_n(x) - H(x)|^2 \,\mathrm{d}x = \int_0^{1/n} |nx - 1|^2 \,\mathrm{d}x = \int_0^{1/n} (n^2x^2 - 2nx + 1) \,\mathrm{d}x$$

$$= \left[\frac{n^2x^3}{3} - nx^2 + x \right]_{x=0}^{x=1/n} = \frac{1}{3n} - \frac{1}{n} + \frac{1}{n} = \frac{1}{3n} \xrightarrow{n\to\infty} 0.$$

We have $\|f_n - H\| \to 0$, but $H \notin C([-1,1])$.

However, note that $C([a,b])$ *is* complete w.r.t. the sup-norm $\|f\|_\infty = \sup_{a\le x\le b} |f(x)|$, but $\|\cdot\|_\infty \neq \|\cdot\|$ and *there is no inner product inducing* $\|\cdot\|_\infty$*-norm*.

Bounded linear operators in Hilbert spaces

Definition

Let X and Y be normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. A linear operator $A\colon X \to Y$ is said to be *bounded* if there exists $C > 0$ such that

$$\|Ax\|_Y \leq C\|x\|_X \quad \text{for all } x \in X.$$

Lemma

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces. Then a linear operator $A\colon X \to Y$ is bounded iff

$$\|A\| := \|A\|_{X\to Y} := \sup_{\|x\|_X \leq 1} \|Ax\|_Y < \infty. \qquad \text{(operator norm)}$$

Proof. "$\Rightarrow$" If there is $C > 0$ s.t. $\|Ax\|_Y \leq C\|x\|_X$ for all $x \in X$, then clearly $\|A\| = \sup_{\|x\|_X \leq 1} \|Ax\|_Y \leq C$.
"$\Leftarrow$" Let $\|A\| < \infty$. Since $\|\frac{x}{\|x\|_X}\|_X = 1$ for all $x \neq 0$, from the linearity of A we infer

$$\frac{\|Ax\|_Y}{\|x\|_X} = \|A(\frac{x}{\|x\|_X})\|_Y \leq \|A\| \quad \text{for all } x \in X.$$

This implies the important estimate

$$\|Ax\|_Y \leq \|A\|\|x\|_X \quad \text{for all } x \in X.$$

A linear operator is continuous precisely when it is bounded.

Proposition

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces and $A\colon X \to Y$ a linear operator. Then the following are equivalent:

(i) *A is a bounded operator;*

(ii) *A is continuous (in X);*

(iii) *A is continuous at one point $x_0 \in X$.*

Proof. (i) $\Rightarrow$ (ii): if $x, y \in X$ and $\varepsilon > 0$, then

$$\|x - y\|_X \leq \frac{\varepsilon}{\|A\|} =: \delta \quad \Rightarrow \quad \|Ax - Ay\|_Y \overset{A \text{ linear}}{=} \|A(x-y)\|_Y \leq \|A\|\|x-y\|_X \leq \varepsilon.$$

(ii) $\Rightarrow$ (iii): trivial.
(iii) $\Rightarrow$ (i): let A be continuous at $x_0 \in X$. By definition, there exists $\delta > 0$ such that

$$\|y - x_0\|_X \leq \delta \quad \Rightarrow \quad \|Ay - Ax_0\|_Y \leq 1.$$

If $x \in X$ is such that $\|x\|_X \leq \delta$, then by taking $y = x + x_0$:

$$\|Ax\|_Y = \|A(x + x_0) - Ax_0\|_Y \leq 1.$$

On the other hand, for any $\|x\|_X \leq 1$, there holds $\|\delta x\|_X = \delta\|x\|_X \leq \delta$ and thus

$$\delta\|Ax\|_Y = \|A(\delta x)\|_Y \leq 1, \quad \text{i.e.,} \quad \|Ax\|_Y \leq \frac{1}{\delta} \quad \text{for all } \|x\|_X \leq 1.$$

Therefore $\|A\| \leq \frac{1}{\delta}$, meaning that A is bounded. □

Let H be a real Hilbert space.

Definition

Two elements $x, y \in H$ are said to be *orthogonal* if $\langle x, y \rangle = 0$.

Let $M \subset H$ be a subset. The orthogonal complement of M in H is defined as

$$M^{\perp} := \{y \in H \mid \langle x, y \rangle = 0 \quad \text{for all } x \in M\}.$$

We state the following easy consequences.

Lemma

For any subset $M \subset H$, $M^{\perp}$ is a closed subspace of H and $M \subset (M^{\perp})^{\perp}$.

Lemma

If M is a subspace of H, then $(M^{\perp})^{\perp} = \overline{M}$.
If M is a closed *subspace of H, then $(M^{\perp})^{\perp} = M$.*

Proposition (Hilbert projection theorem)

Let M be a nonempty, closed, and convex[†] subset of a real Hilbert space H. Then there exists a unique element $x_0 \in M$ satisfying

$$\|x_0\| \leq \|x\| \quad \textit{for all } x \in M.$$

Proof. Let $\delta = \inf\{\|x\| \mid x \in M\}$. We use the parallelogram identity $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$ applied to vectors $u = \frac{1}{2}x$ and $v = \frac{1}{2}y$, $x, y \in M$, to obtain

$$\frac{1}{4}\|x-y\|^2 = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \left\|\frac{x+y}{2}\right\|^2.$$

Due to convexity $\frac{1}{2}(x+y) \in M$, so

$$\|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 - 4\delta^2 \quad \text{for all } x, y \in M. \tag{2}$$

Existence: let $(x_k)_{k=1}^{\infty} \subset M$ s.t. $\|x_k\| \xrightarrow{k\to\infty} \delta$. Substituting $x \leftarrow x_n$ and $y \leftarrow x_m$ in (2) yields $\|x_n - x_m\|^2 \leq 2\|x_n\|^2 + 2\|x_m\|^2 - 4\delta^2$, since $\frac{1}{2}(x_n + x_m) \in M$ for all n, m. Thus $\|x_n - x_m\| \to 0$ as $n, m \to \infty$. $(x_k)_{k=1}^{\infty}$ is Cauchy in the Hilbert space H, so there exists $x_0 := \lim_{k\to\infty} x_k \in H$. Since $\|\cdot\|$ is continuous, $\|x_0\| = \lim_{k\to\infty} \|x_k\| = \delta$. Since M is closed and $(x_k)_{k=1}^{\infty} \subset M$, the limit $x_0 \in M$.

Uniqueness: If $\|x\| = \|y\| = \delta \Rightarrow \|x-y\|^2 \leq 0$ by (2) and so $x = y$. □

[†] $tx + (1-t)y \in M$ for all $x, y \in M$, $t \in (0,1)$.

Corollary

Let H be a real Hilbert space, M a nonempty, closed, and convex subset of H, and $x \in H$. Then there exists a unique element $y_0 \in M$ such that

$$\|x - y_0\| \leq \|x - y\| \quad \text{for all } y \in M.$$

Proof. The set $x - M := \{x - y \mid y \in M\}$ is closed and convex, and $\min\{\|x - y\| \mid x - y \in x - M\} = \min\{\|x - y\| \mid y \in M\}$. The claim follows from the previous result. □

Proposition (Orthogonal decomposition)

If M is a closed subspace of a real Hilbert space H, then

$$H = M \oplus M^{\perp},$$

which means that every element $y \in H$ can be uniquely represented as

$$y = x + x^{\perp}, \quad x \in M,\ x^{\perp} \in M^{\perp}.$$

Proof. It suffices to prove that $M \cap M^{\perp} = \{0\}$ and $M + M^{\perp} = H$.
• If $x \in M \cap M^{\perp}$, then $0 = \langle x, x \rangle = \|x\|^2$ (i.e., $x \perp x$) so $x = 0$.
$\therefore$ $M \cap M^{\perp} = \{0\}$.
• Let $x \in H$. The Hilbert projection theorem guarantees that there exists a unique $y_0 \in M$ such that

$$\|x - y_0\| \leq \|x - y\| \quad \text{for all } y \in M. \tag{3}$$

Let $x_0 = x - y_0$ so that $x = y_0 + x_0 \in M + x_0$. It remains to show that $x_0 \in M^{\perp}$.
The inequality (3) can be written as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x - M.$$

Since $y_0 \in M$ and M is a vector space, $y_0 + M = M$ and $M = -M$ which implies $x - M = x + M = y_0 + x_0 + M = x_0 + M$. The previous inequality can be recast as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x_0 + M \quad \Leftrightarrow \quad \|x_0\| \leq \|x_0 + z\| \quad \text{for all } z \in M.$$

This statement is true if and only if $\langle x_0, z \rangle = 0$ for all $z \in M$. Therefore $x_0 \in M^{\perp}$.

Let M be a closed subspace. The orthogonal decomposition implies that every element $y \in H$ can be uniquely represented as

$$y = x + x^{\perp},\ x \in M,\ x^{\perp} \in M^{\perp}.$$

Lemma

Let $M \subset H$ be a closed subspace. The mapping $P_M : H \to M$, $y \mapsto x$, is an orthogonal projection, i.e., $P_M^2 = P_M$ and $\mathrm{Ran}(P_M) \perp \mathrm{Ran}(I - P_M)$. It satisfies the following properties:

- *P_M is linear;*
- *$\|P_M\| = 1$ if $M \neq \{0\}$;*
- *$I - P_M = P_{M^{\perp}}$;*
- *$\|y - P_M y\| \leq \|y - z\|$ for all $z \in M$;*
- *$y \in M \ \Rightarrow\ P_M y = y,\ (I - P_M)y = 0$;*
  $y \in M^{\perp} \Rightarrow\ P_M y = 0,\ (I - P_M)y = y$;
- *$\|y\|^2 = \|P_M y\|^2 + \|(I - P_M)y\|^2$ (Pythagoras).*

Proof. Omitted; see for example [Rudin, Real and Complex Analysis, pp. 34–35]. □

Example

Let H_1 and H_2 be real Hilbert spaces and let $A\colon H_1 \to H_2$ be a continuous linear operator.

The kernel (or null space) of operator A is defined as

$$\mathrm{Ker}(A) := \{x \in H_1 \mid Ax = 0\}.$$

The range (or image) of operator A is defined as

$$\mathrm{Ran}(A) := \{y \in H_2 \mid y = Ax,\ x \in H_1\}.$$

Then we have the following:

- $\mathrm{Ker}(A)$ is a *closed* subspace of H_1, and $\mathrm{Ran}(A)$ is a subspace of H_2.
- $H_1 = \mathrm{Ker}(A) \oplus (\mathrm{Ker}(A))^{\perp}$.
- $H_2 = \overline{\mathrm{Ran}(A)} \oplus (\mathrm{Ran}(A))^{\perp}$.

Proposition (Riesz representation theorem)

Let H be a real Hilbert space. If $A\colon H \to \mathbb{R}$ is a bounded linear functional, i.e., A is linear and there exists $C > 0$ such that

$$|A(x)| \leq C\|x\| \quad \textit{for all } x \in H,$$

then there exists a unique *$y \in H$ such that*

$$A(x) = \langle x, y \rangle \quad \textit{for all } x \in H.$$

Proof. If $A \equiv 0$, then $y = 0$ and this is unique. Suppose $A \neq 0$ and let

$$M := \mathrm{Ker}(A) = \{x \in H \mid A(x) = 0\}.$$

Since A is continuous, M is a *closed* subspace of H. Furthermore, by the orthogonal decomposition $H = M \oplus M^{\perp}$, our assumption $A \neq 0$ implies that $M \neq H \Rightarrow M^{\perp} \neq \{0\}$.

Let $x \in H$ and $z \in M^{\perp}$ with $\|z\| = 1$. Define

$$u := A(x)z - A(z)x.$$

Then

$$A(u) = A(x)A(z) - A(z)A(x) = 0.$$

meaning that $u \in M$. In particular $\langle u, z\rangle = \langle A(x)z - A(z)x, z\rangle = 0$ and

$$\begin{aligned} A(x) = A(x)\underbrace{\langle z, z\rangle}_{=\|z\|^2=1} &= \langle A(x)z, z\rangle \\ &= \langle A(z)x, z\rangle = A(z)\langle x, z\rangle = \langle x, zA(z)\rangle. \end{aligned}$$

$\therefore$ The element $y = zA(z)$ satisfies $A(x) = \langle x, y\rangle$.
To prove uniqueness, suppose that there exist $y_1, y_2 \in H$ such that

$$A(x) = \langle x, y_1\rangle = \langle x, y_2\rangle.$$

Then $\langle x, y_1 - y_2\rangle = 0$ for all $x \in H$. Choose $x = y_1 - y_2$. Then

$$0 = \langle y_1 - y_2, y_1 - y_2\rangle = \|y_1 - y_2\|^2 \quad \Leftrightarrow \quad y_1 = y_2.$$

Adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A\colon H_1 \to H_2$ is a bounded linear operator. Then there exists a unique bounded linear operator $A^*\colon H_2 \to H_1$, called the adjoint *of A, satisfying $\langle Ax, y\rangle_{H_2} = \langle x, A^*y\rangle_{H_1}$. Moreover, $\|A\|_{H_1\to H_2} = \|A^*\|_{H_2\to H_1}$.*

Proof. Let $y \in H_2$ and consider $T_y\colon H_1 \to \mathbb{R}$, $x \mapsto \langle Ax, y\rangle_{H_2}$. Clearly, T_y is linear and bounded so by the Riesz representation theorem there exists a *unique* $z \in H_1$ s.t.

$$\langle Ax, y\rangle_{H_2} = T_y(x) = \langle x, z\rangle_{H_1} \quad \text{for all } x \in H_1.$$

Define $A^*y := z$.

- Let $a, b \in \mathbb{R}$ and $y_1, y_2 \in H_2$. Linearity follows from $\langle x, A^*(ay_1 + by_2)\rangle = \langle Ax, ay_1 + by_2\rangle = a\langle Ax, y_1\rangle + b\langle Ax, y_2\rangle = a\langle x, A^*y_1\rangle + b\langle x, A^*y_2\rangle = \langle x, aA^*y_1 + bA^*y_2\rangle$. Since $x \in H_1$ was arbitrary, $A^*(ay_1 + by_2) = aA^*y_1 + bA^*y_2$.
- $\|A^*\|_{H_2\to H_1} = \sup_{\|y\|_{H_2}\le 1} \|A^*y\|_{H_1} \overset{(*)}{=} \sup_{\|y\|_{H_2}\le 1} \sup_{\|x\|_{H_1}\le 1} |\langle A^*y, x\rangle|$
  $= \sup_{\|y\|_{H_2}\le 1} \sup_{\|x\|_{H_1}\le 1} |\langle y, Ax\rangle| \overset{(*)}{=} \sup_{\|x\|_{H_1}\le 1} \|Ax\|_{H_2} = \|A\|_{H_1\to H_2} < \infty.$ □

$^{(*)}$ Let $\Lambda \in \mathcal{L}(H, K)$, H, K Hilbert spaces. Cauchy–Schwarz: $\sup_{\|y\|_K\le 1} |\langle \Lambda x, y\rangle_K| \le \|\Lambda x\|_K$.
Other direction: $\sup_{\|y\|_K\le 1} |\langle \Lambda x, y\rangle_K| \ge |\langle \Lambda x, \frac{1}{\|\Lambda x\|_K}\Lambda x\rangle|_K = \|\Lambda x\|_K$.
$\therefore \|\Lambda x\|_K = \sup_{\|y\|_K\le 1} |\langle \Lambda x, y\rangle_K|$.

Some properties of the adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A, B\colon H_1 \to H_2$ are bounded linear operators. Then

(i) $\|A^*A\|_{H_1\to H_1} = \|A\|^2_{H_1\to H_2}$,

(ii) $A^{**} = A$, *where* $A^{**} = (A^*)^*$;

(iii) $(c_1A + c_2B)^* = c_1A^* + c_2B^*$, $c_1, c_2 \in \mathbb{R}$.

Proof. (i) Let $x \in H_1$, $\|x\|_{H_1} = 1$. By the Cauchy–Schwarz inequality,

$$\|Ax\|^2_{H_2} = \langle Ax, Ax\rangle_{H_2} = \langle x, A^*Ax\rangle_{H_1} \le \|A^*Ax\|_{H_1} \;\Rightarrow\; \|A\|^2_{H_1\to H_2} \le \|A^*A\|_{H_1\to H_1}.$$

Other direction: $\|A^*A\| \le \|A^*\| \cdot \|A\| = \|A\|^2$ (previous slide and exercise of week 1).
(ii) If $x \in H_1$ and $y \in H_2$, then

$$\langle A^{**}x, y\rangle_{H_2} = \langle x, A^*y\rangle_{H_1} = \langle A^*y, x\rangle_{H_1} = \langle y, Ax\rangle_{H_2} = \langle Ax, y\rangle_{H_2}.$$

Hence $\langle A^{**}x - Ax, y\rangle_{H_2} = 0$ for all $y \in H_2 \Rightarrow A^{**}x = Ax$ for all $x \in H_1 \Rightarrow A^{**} = A$.
(iii) Let $x \in H_1$ and $y \in H_2$. Then

$$\begin{aligned}\langle (c_1A + c_2B)^*y, x\rangle_{H_1} &= \langle y, (c_1A + c_2B)x\rangle_{H_2} = c_1\langle y, Ax\rangle_{H_2} + c_2\langle y, Bx\rangle_{H_2}\\ &= c_1\langle A^*y, x\rangle_{H_1} + c_2\langle B^*y, x\rangle_{H_1} = \langle (c_1A^* + c_2B^*)y, x\rangle_{H_1}.\end{aligned}$$

Similarly to the previous part, we conclude that $(c_1A + c_2B)^* = c_1A^* + c_2B^*$. □

Self-adjoint operators

Definition

Let H be a Hilbert space. A bounded, linear operator $A\colon H \to H$ is called *self-adjoint* if $A^* = A$, i.e.,

$$\langle Ax, y\rangle = \langle x, Ay\rangle \quad \text{for all } x, y \in H.$$

Example

Let H be a Hilbert space and let $A, B\colon H \to H$ be bounded, linear, self-adjoint operators. Then

(i) $A + B$ is self-adjoint.

(ii) if $c \in \mathbb{R}$, then cA is self-adjoint.

(iii) if $AB = BA$, then AB is self-adjoint.

Parts (i) and (ii) follow immediately from part (iii) on the previous slide. If $x, y \in H$, then

$$\langle ABx, y\rangle = \langle BAx, y\rangle = \langle Ax, By\rangle = \langle x, ABy\rangle \quad \Rightarrow \quad (AB)^* = AB.$$

Example

Let H be a real Hilbert space and $M \subset H$ a closed subspace. Then the orthogonal projections $P_M\colon H \to M$ and $I - P_M =: P_{M^\perp}\colon H \to M^\perp$ are self-adjoint.

Compact operators

Definition

Let H_1 and H_2 be real Hilbert spaces. A bounded linear operator $K : H_1 \to H_2$ is compact if the sets $\overline{K(U)} \subset H_2$ are compact for every bounded set $U \subset H_1$.

The following characterization will be useful.

Characterization

Let H_1 and H_2 be real Hilbert spaces. A bounded linear operator $K : H_1 \to H_2$ is compact if and only if $(Kx_j)_{j=1}^{\infty} \subset H_2$ contains a convergent subsequence for every bounded sequence $(x_j)_{j=1}^{\infty} \subset H_1$.

Let H, H_1, and H_2 be Hilbert spaces. We have the following properties:

- All linear maps to finite-dimensional spaces are compact.
- If $A, B : H_1 \to H_2$ are compact, then $A + B$ is compact.
- If $K : H_1 \to H_2$ is compact, then
  - AK is compact for all bounded and linear $A : H_2 \to H$.
  - KB is compact for all bounded and linear $B : H \to H_1$.
- If $K_n : H_1 \to H_2$ are compact operators and $K : H_1 \to H_2$ is a bounded, linear operator such that $\|K_n - K\| \xrightarrow{n \to \infty 0} 0$, then K is compact.
- If $K : H_1 \to H_2$ is compact, then so is $K^* : H_2 \to H_1$.

Proposition

Let H_1 and H_2 be real Hilbert spaces and $A\colon H_1 \to H_2$ a continuous linear operator. Then

$$H_1 = \mathrm{Ker}(A) \oplus (\mathrm{Ker}(A))^\perp = \mathrm{Ker}(A) \oplus \overline{\mathrm{Ran}(A^*)},$$
$$H_2 = \overline{\mathrm{Ran}(A)} \oplus (\mathrm{Ran}(A))^\perp = \overline{\mathrm{Ran}(A)} \oplus \mathrm{Ker}(A^*).$$

Proof. $H_1 = \mathrm{Ker}(A) \oplus (\mathrm{Ker}(A))^\perp$ and $H_2 = \overline{\mathrm{Ran}(A)} \oplus (\overline{\mathrm{Ran}(A)})^\perp = \overline{\mathrm{Ran}(A)} \oplus (\mathrm{Ran}(A))^\perp$ follow immediately from the previous discussion.[†] The claim

$$(\mathrm{Ran}(A))^\perp = \mathrm{Ker}(A^*) \tag{4}$$

follows immediately by observing that $x \in \mathrm{Ker}(A^*)$ iff

$$0 = \langle A^* x, y\rangle = \langle x, Ay\rangle \quad \text{for all } y \in H_1.$$

The claim $(\mathrm{Ker}(A))^\perp = \overline{\mathrm{Ran}(A^*)}$ follows by applying (4) with A replaced by A^*. □

[†] Here we use the fact that $\overline{X}^\perp = X^\perp$ for any subspace X of H; see exercise 1.

Appendix: some auxiliary results

Let X and Y be normed spaces. We denote

$$\mathcal{L}(X, Y) := \{A \mid A\colon X \to Y \text{ is bounded and linear}\}.$$

Proposition

If Y is complete, then $\mathcal{L}(X, Y)$ is complete w.r.t. operator norm (i.e., it is a Banach space).

Proof. Let $x \in X$ and assume that $A_k \in \mathcal{L}(X, Y)$, $k \in \mathbb{N}$, is a Cauchy sequence. Then for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|A_m - A_n\| < \frac{\varepsilon}{\|x\|_X}.$$

Especially,

$$\|A_m x - A_n x\|_Y \leq \|A_m - A_n\| \|x\|_X < \varepsilon \quad \text{when } m, n > N,$$

so $(A_k x)$ is a Cauchy sequence in Y and therefore the limit

$$A(x) := \lim_{k \to \infty} A_k x$$

exists.

It is easy to see that $A(x) := \lim_{k\to\infty} A_k x$ is linear. It is also bounded: there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|A_m - A_n\| < 1.$$

Fix $m > N$. Then for all $n > m$,

$$\|A_n\| < 1 + \|A_m\|$$

and thus

$$\|A_n x\|_Y \leq (1 + \|A_m\|)\|x\|_X.$$

But $\|Ax\|_Y = \lim_{n\to\infty} \|A_n x\|_Y \leq (1 + \|A_m\|)\|x\|_X$. Therefore A is bounded.
Finally, we need to show that $\|A_n - A\| \to 0$ as $n \to \infty$. Since we assumed $(A_k)_{k=1}^{\infty}$ to be Cauchy, let $\varepsilon > 0$ be s.t. for $m, n > N$, there holds $\|A_m - A_n\| < \varepsilon$. Then

$$\|(A - A_n)x\|_Y = \lim_{m\to\infty} \|A_m x - A_n x\|_Y \leq \varepsilon\|x\|_X \quad \text{for all } x \in X$$

$$\Rightarrow \quad \|A - A_n\| < \varepsilon.$$

Hence $\|A - A_n\| \to 0$ as $n \to \infty$. □

If $X = H_1$ and $Y = H_2$ are Hilbert spaces, then $\mathcal{L}(H_1, H_2)$ is a complete normed space.

In general, $\mathcal{L}(H_1, H_2)$ is *not* a Hilbert space even when both H_1 and H_2 are. However, in the special case $\mathcal{L}(H, \mathbb{R})$ it turns out that indeed one can associate an inner product that induces the operator norm $\|\cdot\|$ – meaning that $\mathcal{L}(H, \mathbb{R})$ is a Hilbert space! This is a consequence of the Riesz representation theorem (details omitted).

Basic properties of vector-valued series

Definition

Let E be a normed space and $(x_k) \subset E$. Define the n^{th} partial sum $S_n := \sum_{k=1}^{n} x_k$. If there exists an element $S \in E$ such that $\lim_{n\to\infty} \|S - S_n\| = 0$, then we say that the series $\sum_{k=1}^{\infty} x_k$ is *convergent* (in E) and denote

$$S = \sum_{k=1}^{\infty} x_k.$$

Moreover, we say that the series $\sum_{k=1}^{\infty} x_k$ is *absolutely convergent* if $\sum_{k=1}^{\infty} \|x_k\| < \infty$.

Proposition

The normed space E is a Banach space iff every absolutely convergent series $\sum_{k=1}^{\infty} x_k$ is convergent in E.

Theorem (Generalized Pythagorean theorem)

Let (e_k) be an orthonormal sequence in Hilbert space H and let $(\lambda_k) \subset \mathbb{R}$. Then

$$\sum_{k=1}^{\infty} \lambda_k e_k \text{ is convergent} \quad \text{iff} \quad \sum_{k=1}^{\infty} |\lambda_k|^2 < \infty.$$

In this case, we have

$$\left\| \sum_{k=1}^{\infty} \lambda_k e_k \right\|^2 = \sum_{k=1}^{\infty} |\lambda_k|^2.$$

Neumann series: "Sufficiently small perturbations of the identity are still invertible"

The following result is a well-known generalization of the geometric series formula, named after 19^{th} century mathematician Carl Neumann.

Theorem (Neumann series)

Let H be a real Hilbert space and let $A \in \mathcal{L}(H) := \mathcal{L}(H, H)$ be such that $\|A\| < 1$. Then $I - A$ is invertible in $\mathcal{L}(H)$ with

$$(I - A)^{-1} = I + A + \cdots + A^n + \cdots = \sum_{k=0}^{\infty} A^k,$$

and this series converges in operator norm.

Proof. Let $B_{m,n} := \sum_{k=m}^{n} A^k$, $m < n$. Since $\|A\| < 1$, we have

$$\|B_{m,n}\| \leq \sum_{k=m}^{n} \|A\|^k = \|A\|^m \sum_{k=0}^{m-n} \|A\|^k = \|A\|^m \frac{1 - \|A\|^{n-m+1}}{1 - \|A\|} \xrightarrow{m,n\to\infty} 0.$$

$\therefore$ The partial sums $\sum_{k=0}^{n} A^k$ form a Cauchy sequence in $\mathcal{L}(H)$.

Since H is a Hilbert space, $\mathcal{L}(H)$ is a Banach space and the limit

$$B := \lim_{n\to\infty} \sum_{k=0}^{n} A^k \in \mathcal{L}(H)$$

exists. We need to prove that $(I - A)B = I = B(I - A)$. Let

$$B_n := I + A + \cdots + A^n.$$

Then

$$(I - A)B_n = I - A^{n+1},$$
$$B_n(I - A) = I - A^{n+1},$$

and since $\|A\| < 1$, $\|A^{n+1}\| \leq \|A\|^{n+1} \xrightarrow{n\to\infty} 0$, we thus obtain

$$I - A^{n+1} \xrightarrow{n\to\infty} I \quad \text{in } \mathcal{L}(H)$$

and

$$(I - A)B = \lim_{n\to\infty} (I - A)B_n = I = \lim_{n\to\infty} B_n(I - A) = B(I - A). \quad \square$$

Theorem (Bessel's inequality)

Let H be a real Hilbert space and let (e_n) be an orthonormal sequence in H. Then

$$\sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \leq \|x\|^2 \quad \textit{for all } x \in H.$$

Especially $\lim_{n\to\infty}\langle x, e_n\rangle = 0$.

Proof. Let $k \in \mathbb{N}$. Noting that

$$\left\langle x - \sum_{n=1}^{k}\langle x, e_n\rangle e_n, e_j \right\rangle = \langle x, e_j\rangle - \sum_{n=1}^{k}\langle x, e_n\rangle\langle e_n, e_j\rangle = \langle x, e_j\rangle - \langle x, e_j\rangle = 0$$

for all $j \in \{1, \ldots, k\}$, we deduce that $x - \sum_{n=1}^{k}\langle x, e_n\rangle e_n \perp \sum_{n=1}^{k}\langle x, e_n\rangle e_n$ (recall that the orthogonal complement is a subspace). By the Pythagorean theorem,

$$\|x\|^2 = \left\|x - \sum_{n=1}^{k}\langle x, e_n\rangle e_n\right\|^2 + \left\|\sum_{n=1}^{k}\langle x, e_n\rangle e_n\right\|^2 \geq \left\|\sum_{n=1}^{k}\langle x, e_n\rangle e_n\right\|^2 = \sum_{n=1}^{k}|\langle x, e_n\rangle|^2.$$

Letting $k \to \infty$ yields the assertion. □

Lax–Milgram lemma

Proposition (Lax–Milgram lemma)

Let H be a real Hilbert space and let $B\colon H \times H \to \mathbb{R}$ be a bilinear mapping[†] *with $C, c > 0$ such that*

$$|B(u,v)| \le C\|u\| \cdot \|v\| \quad \text{for all } u, v \in H, \qquad \text{(boundedness)}$$

$$B(u,u) \ge c\|u\|^2 \quad \text{for all } u \in H. \qquad \text{(coercivity)}$$

Let $F\colon H \to \mathbb{R}$ be a bounded linear mapping. Then there exists a unique element $u \in H$ satisfying

$$B(u,v) = F(v) \quad \text{for all } v \in H.$$

and

$$\|u\| \le \frac{1}{c}\|F\|.$$

[†] $B(u+v,w) = B(u,w) + B(v,w)$, $B(au,v) = aB(u,v)$, $B(u,v+w) = B(u,v) + B(u,w)$, $B(u,av) = aB(u,v)$ for all $u, v, w \in H$ and $a \in \mathbb{R}$.

Proof. 1) Let $v \in H$ be fixed. Then the mapping

$$T : w \mapsto B(v, w),\ H \to \mathbb{R},$$

is bounded and linear. It follows from the Riesz representation theorem that there exists a unique element $a \in H$ with

$$Tw = \langle a, w \rangle \quad \text{for all } w \in H.$$

Let us define the mapping $A : H \to H$ by setting

$$Av = a.$$

Then

$$B(v, w) = \langle Av, w \rangle \quad \text{for all } v, w \in H.$$

2) We show that the mapping $A\colon H \to H$ is linear and bounded. Clearly,

$$\begin{aligned}\langle A(c_1 v_1 + c_2 v_2), w\rangle &= B(c_1 v_1 + c_2 v_2, w)\\ &= c_1 B(v_1, w) + c_2 B(v_2, w)\\ &= \langle c_1 A v_1 + c_2 A v_2, w\rangle\end{aligned}$$

for all $w \in H$, so $A(c_1 v_1 + c_2 v_2) = c_1 A v_1 + c_2 A v_2$. Moreover,

$$\begin{aligned}\|Av\|^2 &= \langle Av, Av\rangle\\ &= B(v, Av)\\ &\leq C\|v\|\|Av\|\end{aligned}$$

which implies that

$$\|Av\| \leq C\|v\|.$$

3) We show that

$$\begin{cases} A \text{ is one-to-one,} \\ \mathrm{Ran}(A) = AH \text{ is closed in } H. \end{cases}$$

We begin by noting that

$$c\|v\|^2 \leq B(v,v) = \langle Av, v\rangle \leq \|Av\|\|v\|$$

and thus

$$\|Av\| \geq c\|v\| \quad \text{for all } v \in H. \tag{5}$$

Especially

$$Av = Aw \Rightarrow A(v-w) = 0 \Rightarrow 0 = \|A(v-w)\| \geq c\|v-w\| \geq 0 \Rightarrow v = w$$

so A is one-to-one.

To see that $\mathrm{Ran}(A)$ is closed, let $y_j = Ax_j \in \mathrm{Ran}(A)$. The goal is to show that $y := \lim_{j\to\infty} y_j \in \mathrm{Ran}(A)$. We observe that

$$\lim_{j,k\to\infty} \|x_j - x_k\| \overset{(5)}{\leq} \lim_{j,k\to\infty} \frac{1}{c}\|y_j - y_k\| = 0,$$

i.e., $(x_j)_{j=1}^{\infty}$ is Cauchy and $x := \lim_{j\to\infty} x_j \in H$ exists by completeness. Moreover,

$$\lim_{j\to\infty} \|Ax_j - Ax\| \leq \lim_{j\to\infty} \|A\|\|x_j - x\| \leq C \lim_{j\to\infty} \|x_j - x\| = 0$$

and therefore

$$y = \lim_{j\to\infty} Ax_j = Ax \in \mathrm{Ran}(A).$$

4) We show that $\mathrm{Ran}(A) = H$. We prove this by contradiction: suppose that $\mathrm{Ran}(A) = \overline{\mathrm{Ran}(A)} \neq H$. Then there exists $w \in \mathrm{Ran}(A)^{\perp}$, $w \neq 0$.[†] This implies that

$$\|w\|^2 \leq \frac{1}{c} B(w, w) = \frac{1}{c} \langle Aw, w \rangle = 0,$$

i.e., $w = 0$. This contradiction shows that $\mathrm{Ran}(A) = H$. Therefore $A \colon H \to H$ is a continuous bijection.
5) Existence of a solution. We use the Riesz representation theorem: since $F \colon H \to \mathbb{R}$ is linear and continuous, there exists $b \in H$ such that

$$F(v) = \langle b, v \rangle \quad \text{for all } v \in H.$$

Define $u := A^{-1} b$. Hence

$$\begin{aligned} Au = b \quad &\Leftrightarrow \quad \langle Au, v \rangle = \langle b, v \rangle \quad \text{for all } v \in H \\ &\Leftrightarrow \quad B(u, v) = F(v) \quad \text{for all } v \in H. \end{aligned}$$

[†]Since $(\mathrm{Ran}(A)^{\perp})^{\perp} = \overline{\mathrm{Ran}(A)} \neq H \Rightarrow (\mathrm{Ran}(A))^{\perp} \neq \{0\}$.

6) Uniqueness. Suppose that

$$B(u_1, w) = F(w) \quad \text{for all } w \in H,$$
$$B(u_2, w) = F(w) \quad \text{for all } w \in H.$$

Let $u := u_1 - u_2$. By linearity,

$$B(u, w) = 0 \quad \text{for all } w \in H.$$

The coercivity of B implies that

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = 0$$

so that $u = 0$, i.e., $u_1 = u_2$.

7) *A priori bound*. If $B(u, w) = F(w)$ for all $w \in H$, then by setting $w = u$ we obtain

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = \frac{1}{c} F(u) \leq \frac{1}{c} \|F\| \|u\|$$

which immediately yields

$$\|u\| \leq \frac{1}{c} \|F\|.$$

Density argument

Lemma

Let X, Y be Banach spaces and let $Z \subset X$ be a dense subspace. If $T : Z \to Y$ is a linear mapping such that

$$\|Tx\|_Y \leq C\|x\|_X, \quad x \in Z, \tag{6}$$

then there exists a unique extension $\widetilde{T} : X \to Y$ with $\widetilde{T}|_Z = T$ and

$$\|\widetilde{T}x\|_Y \leq C\|x\|_X, \quad x \in X. \tag{7}$$

Moreover, if (6) holds with equality, then so does (7).

Proof. Let $x \in X$. Because $Z \subset X$ is dense, there exists a sequence $(z_k)_{k=1}^{\infty} \subset Z$ s.t. $\|z_k - x\|_X \xrightarrow{k\to\infty} 0$. Let $\varepsilon > 0$. Since $(z_k)_{k=1}^{\infty}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ s.t.

$$m, n \geq N \quad \Rightarrow \quad \|z_m - z_n\|_X < \frac{\varepsilon}{C}.$$

Then there holds

$$\|Tz_m - Tz_n\|_Y = \|T(z_m - z_n)\|_Y \leq C\|z_m - z_n\|_X < \varepsilon,$$

which means that $(Tz_k)_{k=1}^{\infty}$ is a Cauchy sequence in Y. Since Y is complete, there exists $y := \lim_{k\to\infty} Tz_k$. Hence we may define $\widetilde{T} : X \to Y$ by setting $\widetilde{T}(x) = y$.

We begin by showing that $\widetilde{T}$ is well-defined. Let $(z_k)_{k=1}^{\infty}$, $(\widetilde{z}_k)_{k=1}^{\infty}$ be two sequences in Z s.t. $z_k, \widetilde{z}_k \xrightarrow{k\to\infty} x$ in X. Then

$$\|Tz_k - T\widetilde{z}_k\|_Y = \|T(z_k - \widetilde{z}_k)\|_Y \le C\|z_k - \widetilde{z}_k\| \le C\|z_k - x\| + C\|\widetilde{z}_k - x\| \overset{k\to\infty}{\to} 0.$$

Recalling that $\widetilde{T}(x) := \lim_{k\to\infty} Tz_k$, we obtain

$$\|T\widetilde{z}_k - \widetilde{T}(x)\| \le \|T\widetilde{z}_k - Tz_k\| + \|Tz_k - \widetilde{T}(x)\| \overset{k\to\infty}{\to} 0,$$

showing that $\widetilde{T}$ is well-defined.
Next we show that $\widetilde{T}$ is linear. Let $x, \widetilde{x} \in X$ and $a, b \in \mathbb{R}$. Let $Z \ni z_k \xrightarrow{k\to\infty} x$ and $Z \ni \widetilde{z}_k \xrightarrow{k\to\infty} \widetilde{x}$. Now $ax + b\widetilde{x} \in X$ and $Z \ni az_k + b\widetilde{z}_k \to ax + b\widetilde{x}$. Thus

$$\widetilde{T}(ax + b\widetilde{x}) = \lim_{k\to\infty} T(az_k + b\widetilde{z}_k) = a\lim_{k\to\infty} Tz_k + b\lim_{k\to\infty} T\widetilde{z}_k = a\widetilde{T}x + b\widetilde{T}x,$$

since the limit is linear.[†]
Since the norm is continuous,

$$\|\widetilde{T}x\| = \|\lim_{k\to\infty} Tx_k\| = \lim_{k\to\infty} \|Tx_k\| \le C\lim_{k\to\infty} \|x_k\| = C\|x\|.$$

Finally, $\widetilde{T}|_Z = T$ holds by construction and the uniqueness of the limit $Tz_k \to y$ ensures that there cannot exist another mapping $L\colon X \to Y$ s.t. $L|_Z = T$ and $\|Lx\| \le C\|x\|$. □

[†]Let $y := \lim_{k\to\infty} Tz_k$ and $\widetilde{y} := \lim_{k\to\infty} T\widetilde{z}_k$.
Then $\|T(az_k + b\widetilde{z}_k) - ay - b\widetilde{y}\| \le a\|Tz_k - y\| + b\|T\widetilde{z}_k - \widetilde{y}\| \to 0$.
Hence $\lim_{k\to\infty} T(az_k + b\widetilde{z}_k) = a\lim_{k\to\infty} Tz_k + b\lim_{k\to\infty} T\widetilde{z}_k$.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Second lecture, April 24, 2023

Practical matters

- **Monday May 1** (next week) is a public holiday
  $\rightarrow$ **no lecture on May 1!**
- We will have a bonus live-coding lecture on **Tuesday May 2** about Computerized Tomography in place of the usual exercise session (this material will not be essential to the course).
- The deadline for the second exercise sheet will be moved to **Tuesday May 9**. Note that tomorrow's exercise session will happen as planned.

Spectral theory of compact operators

Let E be a (complex) Banach space and $A\colon E \to E$ a bounded linear operator. The *spectrum* of operator A is denoted by

$$\sigma(A) := \{\lambda \in \mathbb{C} \mid \lambda I - A \text{ does not have an inverse}\}.$$

Proposition

Let H be a real Hilbert space and $A\colon H \to H$ a bounded linear operator. Then

$$\sup\{|\lambda| : \lambda \in \sigma(A)\} \leq \|A\|.$$

Proof. Let $|\lambda| > \|A\|$. Then $\lambda I - A = \lambda(I - \frac{1}{\lambda}A)$, where $\|\frac{1}{\lambda}A\| < 1$. Thus $I - \frac{1}{\lambda}A$ is invertible (its inverse can be expressed as a Neumann series), and therefore the operator $\lambda I - A$ is always invertible for all $|\lambda| > \|A\|$. □

Lemma

The eigenvalues of a self-adjoint operator $A\colon H \to H$ are real-valued.

Proof. If $Ax = \lambda x$, with $x \neq 0$, then[†]

$$\lambda\langle x, x\rangle = \langle Ax, x\rangle = \langle x, A^*x\rangle = \langle x, \lambda x\rangle = \overline{\lambda}\langle x, x\rangle \quad \Rightarrow \quad \lambda = \overline{\lambda} \in \mathbb{R} \quad \square$$

[†]If the scalar field of an inner product space is complex, then recall that the inner product needs to satisfy $\langle x, y\rangle = \overline{\langle y, x\rangle}$.

Lemma

Let H be a real Hilbert space and let $A\colon H \to H$ be a self-adjoint operator. Then

$$\|A\| = \sup_{\|x\|=1} |\langle Ax, x\rangle|.$$

Proof. Let us denote $\alpha := \sup\{|\langle Ax, x\rangle| : \|x\| = 1\}$.
"$\geq$" By Cauchy–Schwarz, $|\langle Ax, x\rangle| \leq \|A\|$ for $\|x\| = 1$, and thus $\alpha \leq \|A\|$.
"$\leq$" Using $A^* = A$, we obtain the identity

$$\begin{aligned}
&\langle A(x+y), x+y\rangle - \langle A(x-y), x-y\rangle \\
&= \cancel{\langle Ax, x\rangle} + \langle Ax, y\rangle + \langle Ay, x\rangle + \cancel{\langle Ay, y\rangle} - \cancel{\langle Ax, x\rangle} + \langle Ax, y\rangle + \langle Ay, x\rangle - \cancel{\langle Ay, y\rangle} \\
&= 4\langle Ax, y\rangle \quad \text{for all } x, y \in H.
\end{aligned}$$

Let $x, y \in H$ be such that $\|x\| = 1 = \|y\|$. Using the inequality $|\langle Av, v\rangle| \leq \alpha\|v\|^2$ for all $v \in H$ and the parallelogram rule (exercise 1), we obtain

$$\begin{aligned}
4\langle Ax, y\rangle \leq |\langle A(x+y), x+y\rangle| + |\langle A(x-y), x-y\rangle| &\leq \alpha(\|x+y\|^2 + \|x-y\|^2) \\
&= 2\alpha(\|x\|^2 + \|y\|^2) = 4\alpha.
\end{aligned}$$

Let $\lambda = \operatorname{sign}\langle Ax, y\rangle$. Then $|\langle Ax, y\rangle| = \lambda\langle Ax, y\rangle = \langle A(\lambda x), y\rangle \leq \alpha$
$\Rightarrow \|A\| = \sup_{\|x\|=1} \sup_{\|y\|=1} |\langle Ax, y\rangle| \leq \alpha.$ □

If A is a compact operator, then there exists an element in H which satisfies the following.

Lemma

Let H be a real Hilbert space and let $A\colon H \to H$ be a compact, self-adjoint operator. Then

$$\|A\| = |\langle Ax_0, x_0\rangle| \quad \text{for some } x_0 \in H,\ \|x_0\| = 1. \tag{1}$$

Moreover, x_0 is an eigenvector of A, $Ax_0 = \lambda_0 x$ with $|\lambda_0| = \|A\|$.

Proof. Suppose that $A \neq 0$. By the previous lemma,

$$\|A\| = \sup\{|\langle Ax, x\rangle| : \|x\| = 1\},$$

and thus there exists a sequence $(x_n) \subset \{x \in H : \|x\| = 1\}$ such that $|\langle Ax_n, x_n\rangle| \overset{n\to\infty}{\to} \|A\|$, i.e., $\langle Ax_n, x_n\rangle \overset{n\to\infty}{\to} \lambda_0$, where $\lambda_0 \in \{-\|A\|, \|A\|\}$. Now

$$0 \leq \|Ax_n - \lambda_0 x_n\|^2 = \|Ax_n\|^2 + \lambda_0^2\|x_n\|^2 - 2\lambda_0\langle Ax_n, x_n\rangle \leq \lambda_0^2 + \lambda_0^2 - 2\lambda_0\langle Ax_n, x_n\rangle \overset{n\to\infty}{\to} 0.$$

By compactness of A, there exists a subsequence (x_{n_j}) of (x_n) and a limit $x_0 \in H$ such that $Ax_{n_j} \to x_0$. Since $Ax_{n_j} - \lambda_0 x_{n_j} \to 0$, then $\lambda x_{n_j} \to x_0$, $\|x_0\| = 1$, and $Ax_0 = \lambda_0 x_0$. □

Theorem (Spectral theorem for compact, self-adjoint operators)

Let H be a real Hilbert space and let $A\colon H \to H$ be a compact, self-adjoint operator. Then

- *each $\lambda \in \sigma(A) \setminus \{0\}$ is an eigenvalue of A;*
- *0 is the only limit point of $\sigma(A)$, i.e., if there are an infinite number of eigenvalues $(\lambda_n) \subset \mathbb{R}$, then $\lim_n \lambda_n = 0$;*
- *the eigenvectors $(u_n) \subset H$ form an orthonormal sequence such that*

$$Ax = \sum_n \lambda_n \langle x, u_n \rangle u_n.$$

Proof. We have already established that there exists $u_0 \in H$ s.t. $Au_0 = \lambda_0 u_0$, $|\lambda_0| = \|A\|$ and $\|u_0\| = 1$. Define $H_1 := \{u_0\}^\perp$. If $y \in H_1$, then

$$\langle Ay, u_0 \rangle = \langle y, Au_0 \rangle = \lambda_1 \langle y, u_0 \rangle = 0,$$

which means that $A|_{H_1} : H_1 \to H_1$ is a compact, self-adjoint operator.

By (1), there exists $u_1 \in H_1$ such that

$$\|A|_{H_1}\| = |\langle u_1, Au_1\rangle|$$

with $Au_1 = \lambda_1 u_1$, where $|\lambda_1| \leq |\lambda_0|$ and $\langle u_0, u_1\rangle = 0$.

Next, let $H_2 := \{u_0, u_1\}^\perp$. As before, $A|_{H_2} : H_2 \to H_2$ is a compact, self-adjoint operator and (1) again implies that there exists $u_2 \in H_2$ such that $Au_2 = \lambda_2 u_2$, where $|\lambda_2| \leq |\lambda_1| \leq |\lambda_0|$ and $\|u_2\| = 1$.

Proceeding inductively, we obtain $H_n := \{u_0, \ldots, u_{n-1}\}^\perp \subset H_{n-1}$, where $A|_{H_n} : H_n \to H_n$ is compact and self-adjoint, $|\lambda_n| = \|A|_{H_n}\|$, $|\lambda_n| \leq |\lambda_{n-1}| \leq \cdots \leq |\lambda_0|$ and $Au_n = \lambda_n u_n$ for some $u_n \in H_n$, $\|u_n\| = 1$.

If $\dim \mathrm{Ran}(A) = \infty$, we claim that $|\lambda_n| \to 0$ as $n \to \infty$. Since $u_k \perp u_j$ whenever $j \neq k$, we deduce that

$$|\lambda_j|^2 + |\lambda_k|^2 = \|\lambda_k u_k - \lambda_j u_j\|^2 = \|Au_k - Au_j\|^2.$$

Note that (λ_j^2) is convergent as a bounded, monotonic sequence. Since (u_j) is bounded and A is compact, (Au_j) contains a convergent subsequence – and hence it contains a Cauchy subsequence. This implies that (λ_j^2) contains a subsequence which converges to 0. Since (λ_j^2) is a convergent sequence, it follows that $\lim_{j\to\infty} \lambda_j = 0$.

Let $M := \operatorname{span}\{u_n \mid n \in \mathbb{N}\}^\perp$. The previous discussion implies that $A|_M = 0$. Let $H_\infty := \overline{\operatorname{span}\{u_n \mid u_n \in \mathbb{N}\}}$. By the orthogonal decomposition $H = M \oplus H_\infty$, the orthogonal projection $P\colon H \to H_\infty$ can be written as

$$Px = \sum_n \langle x, u_n\rangle u_n, \quad x \in H \qquad \text{(proof left as an exercise)}$$

and therefore

$$Ax = APx = A\left(\sum_n \langle x, u_n\rangle u_n\right) = \sum_n \langle x, u_n\rangle Au_n = \sum_n \lambda_n \langle x, u_n\rangle u_n,$$

as desired.

Finally, to see that each $\lambda \in \sigma(A) \setminus \{0\}$ is an eigenvalue, suppose that $\lambda \notin \overline{\{\lambda_n \mid n \in \mathbb{N}\}} \cup \{0\}$. Then there exists $\delta > 0$ such that $|\lambda - \lambda_n| > \delta$ for all $n \in \mathbb{N}$ and $|\lambda| > \delta$. If $Q\colon H \to M$ is an orthogonal projection, then ´

$$(\lambda I - A)^{-1}x = \sum_n \frac{1}{\lambda - \lambda_n}\langle x, u_n\rangle u_n + \frac{1}{\lambda}Qx, \quad x \in H,$$

is bounded by the previous discussion, i.e., $\lambda \notin \sigma(A)$. □

Our goal is to obtain a spectral expansion for all compact operators $A\colon H_1 \to H_2$. To begin with, note that if $A\colon H_1 \to H_2$ is a compact operator, then $A^*A\colon H_1 \to H_1$ is compact and self-adjoint since

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

Note in addition that the eigenvalues of A^*A are nonnegative: if $A^*Av_n = \lambda_n v_n$, $\|v_n\|_{H_1} = 1$, then

$$\lambda_n = \lambda_n \|v_n\|_{H_1}^2 = \langle A^*Av_n, v_n\rangle_{H_1} = \|Av_n\|_{H_2}^2 \geq 0.$$

In particular, we can write down the eigendecomposition

$$A^*Ax = \sum_n \lambda_n \langle x, v_n\rangle_{H_1} v_n,$$

where $(v_n) \subset H_1$ is an orthonormal sequence of eigenvectors.

Lemma

Let H_1 and H_2 be real Hilbert spaces and let $A\colon H_1 \to H_2$ be a compact operator. Then there exist orthonormal sequences $(v_n) \subset H_1$ and $(w_n) \subset H_2$ such that

$$Av_n = \sqrt{\lambda_n} w_n \quad \text{and} \quad A^* w_n = \sqrt{\lambda_n} v_n, \tag{2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ are the nonzero eigenvalues of A^*A. Define $|A|\colon H_1 \to H_2$ by setting $|A|x = \sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n$. Then

$$|A| \text{ is compact and } |A|^*|A| = A^*A.$$

Proof. Let $(v_n) \subset H_1$ denote the orthonormal sequence of eigenfunctions of A^*A, i.e.,

$$A^*Av_n = \lambda_n v_n$$

and define a second sequence by

$$w_n = \frac{1}{\sqrt{\lambda_n}} Av_n.$$

Straightforward computations show that (2) holds as well as $\langle w_n, w_n \rangle_{H_2} = 1$ and $\langle w_n, w_m \rangle_{H_2} = 0$ whenever $n \neq m$.

Next, let us show that $|A|\colon H_1 \to H_2$ is compact. It follows from the generalized Pythagorean theorem and Bessel's inequality that

$$\begin{aligned}
&\left\| |A|x - \sum_{n=1}^{m} \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n \right\|_{H_2}^2 = \left\| \sum_{n=m+1}^{\infty} \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n \right\|_{H_2}^2 \\
&= \sum_{n=m+1}^{\infty} |\lambda_n| |\langle x, v_n \rangle_{H_1}|^2 \leq \sup_{n \geq m+1} |\lambda_n| \cdot \|x\|^2 \\
&\leq \sup_{n \geq m+1} |\lambda_n| \quad \text{for all } \|x\|_{H_1} \leq 1.
\end{aligned}$$

Thus $\| |A| - \sum_{n=1}^{m} \sqrt{\lambda_n} \langle \cdot, v_n \rangle_{H^1} w_n \| \leq \sup_{n \geq m+1} \sqrt{\lambda_n} \to 0$ as $m \to \infty$. Since the operators $x \mapsto \langle x, v_n \rangle_{H_1} w_n$ have 1-dimensional range, they are compact. Moreover, finite sums $\sum_{n=1}^{m} \sqrt{\lambda_n} \langle \cdot, v_n \rangle_{H_1} w_n$ of compact operators are compact and, in consequence, their limiting operator $|A|$ is compact. (See, e.g., properties of compact operators from the lecture notes of week 1.)

Finally, we wish to show that $|A|^*|A| = A^*A$. It is not difficult to check that

$$|A|^* = \sum_n \sqrt{\lambda_n} \langle \cdot, w_n \rangle_{H_2} v_n.$$

Let $x \in H_1$. A direct computation then reveals that

$$\begin{aligned}
|A|^*|A|x &= |A|^* \left(\sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n \right) \\
&= \sum_m \sqrt{\lambda_m} \left\langle \sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n, w_m \right\rangle_{H_2} v_m \\
&= \sum_{m,n} \sqrt{\lambda_m \lambda_n} \langle x, v_n \rangle_{H_1} \langle w_n, w_m \rangle_{H_2} v_m \\
&= \sum_n \lambda_n \langle x, v_n \rangle_{H_1} v_n = A^*Ax,
\end{aligned}$$

where we used $\langle w_n, w_n \rangle_{H_2} = 1$ and $\langle w_n, w_m \rangle_{H_2} = 0$ whenever $n \neq m$. □

Proposition (Polar decomposition)

Let H_1 and H_2 be real Hilbert spaces, $A\colon H_1 \to H_2$ a compact operator, and let $|A|$ be defined as before. Then there exists a bounded, linear operator $U\colon H_2 \to H_2$ such that

$$A = U|A|,$$

where $\|Ux\|_{H_2} = \|x\|_{H_2}$ for all $x \in \overline{\mathrm{Ran}(|A|)}$ and $Uy = 0$ for all $y \in \mathrm{Ran}(|A|)^{\perp}$.

Proof. If $x \in H_1$, then

$$\||A|x\|_{H_2}^2 = \langle |A|x, |A|x\rangle_{H_2} = \langle x, |A|^*|A|x\rangle_{H_1} = \langle x, A^*Ax\rangle_{H_1} = \langle Ax, Ax\rangle_{H_2} = \|Ax\|_{H_2}^2.$$

We can define a linear mapping $U\colon \mathrm{Ran}(|A|) \to \mathrm{Ran}(A)$ by setting $U(|A|x) = Ax$ for $x \in H_1$. Since the above formula implies

$$\|U(|A|x)\| = \|Ax\| = \||A|x\| \quad \text{for all } |A|x \in \mathrm{Ran}(|A|),$$

there exists a unique extension $U\colon \overline{\mathrm{Ran}(|A|)} \to \overline{\mathrm{Ran}(A)}$ s.t. $\|Ux\| = \|x\|$ for all $x \in \overline{\mathrm{Ran}(|A|)}$. Finally, since we have the orthogonal decomposition $H_2 = \overline{\mathrm{Ran}(|A|)} \oplus \mathrm{Ran}(|A|)^{\perp}$, we can set $Uy = 0$ for all $y \in \mathrm{Ran}(|A|)^{\perp}$. □

Theorem (Singular value decomposition of compact operators)

Let H_1 and H_2 be real Hilbert spaces and let $A\colon H_1 \to H_2$ be a compact operator. Then there exists a (possibly finite) sequence of positive real numbers $(\lambda_n) \subset \mathbb{R}$ with $\lim_{n\to\infty} \lambda_n = 0$ and (possibly finite) orthonormal sequences $(v_n) \subset H_1$ and $(u_n) \subset H_2$ such that

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle_{H_1} u_n, \quad x \in H_1.$$

Proof. The operator $A^*A\colon H_1 \to H_1$ is compact and self-adjoint. Let

$$A^*Ax = \sum_n \lambda_n \langle x, v_n \rangle_{H_1} v_n, \quad x \in H_1,$$

be its eigendecomposition. Moreover, let

$$|A|x = \sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n, \quad x \in H_1,$$

be defined as before. Then using the polar decomposition:

$$Ax = U|A|x = U\left(\sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} w_n \right) = \sum_n \sqrt{\lambda_n} \langle x, v_n \rangle_{H_1} \underbrace{U(w_n)}_{=:u_n}.$$

Since U is an isometry in $\overline{\mathrm{Ran}(|A|)}$, $\langle w_n, w_m \rangle_{H_2} = \langle U(w_n), U(w_m) \rangle_{H_2}$ (exercise 1). Therefore (u_n) is also an orthonormal sequence. □

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Third lecture, May 2, 2023

Numerical example: X-ray tomography

As an application, we consider X-ray tomography and describe here the construction of the tomography matrix. We will return to this example on Tuesday May 30 when we will discuss total variation regularization for X-ray tomography.

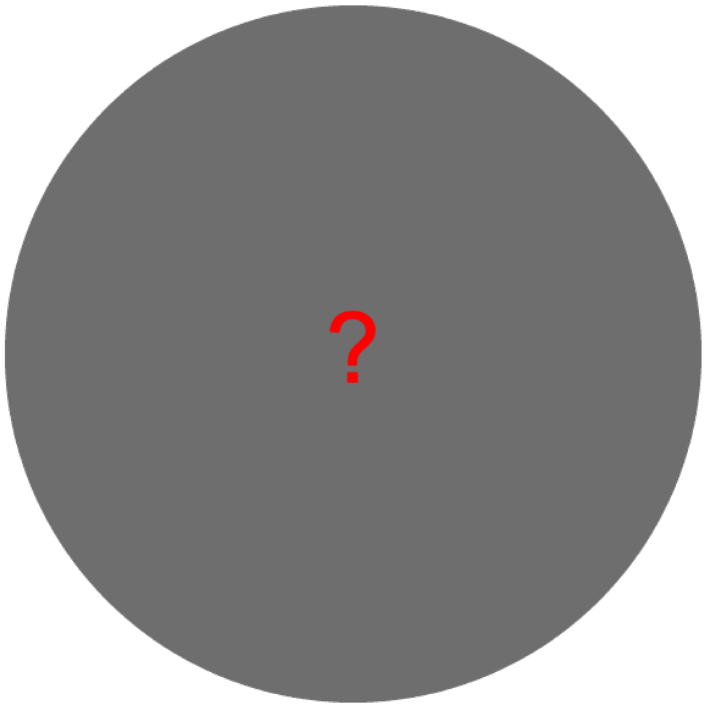
The following content follows roughly the material presented in the following monographs.


J. Kaipio and E. Somersalo. Statistical and Computational Inverse Problems. 2005.

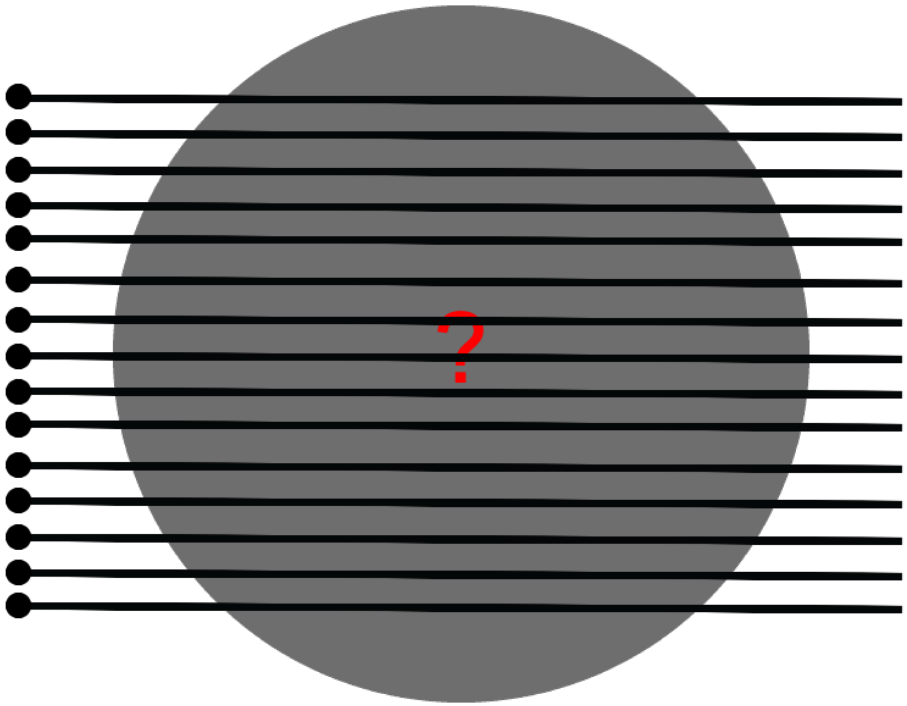

J. L. Mueller and S. Siltanen. Linear and Nonlinear Inverse Problems with Practical Applications. 2012.

ASTRA Toolbox for 2D and 3D tomography:
https://www.astra-toolbox.com/

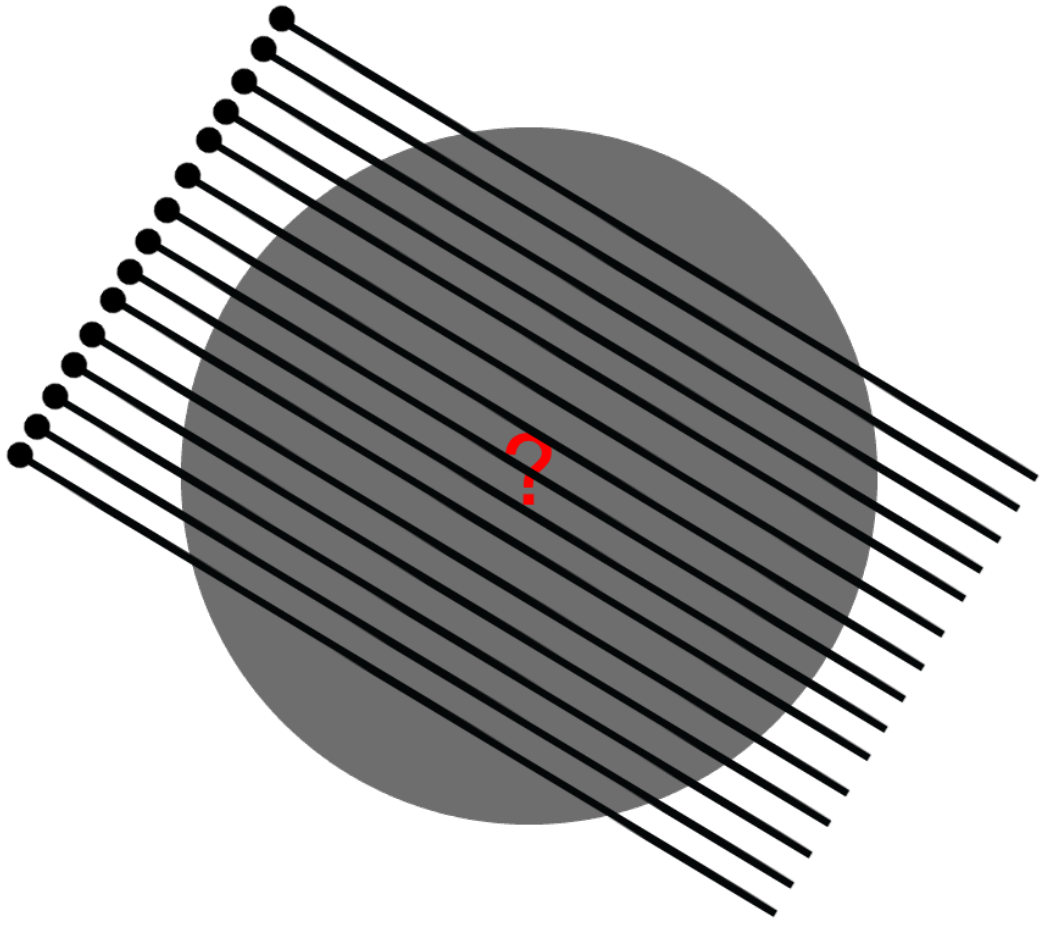
Parallel-beam X-ray tomography



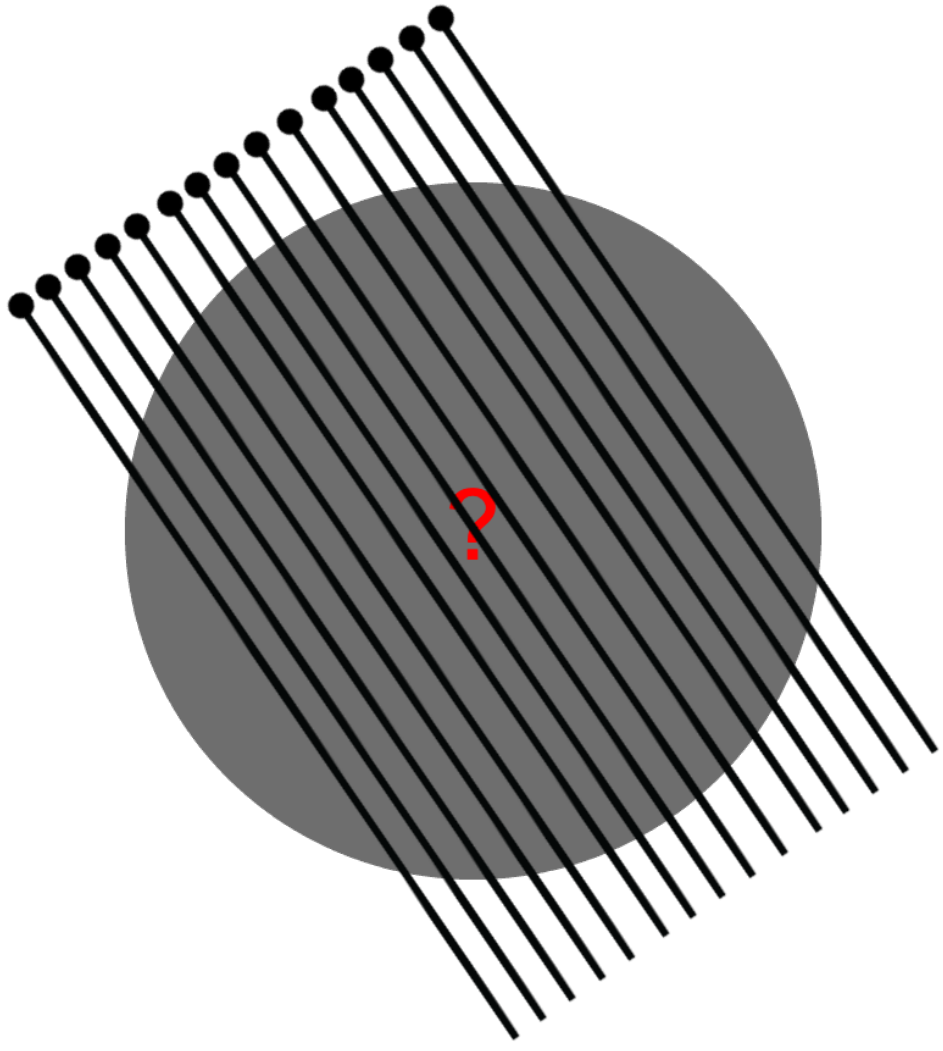
Parallel-beam X-ray tomography



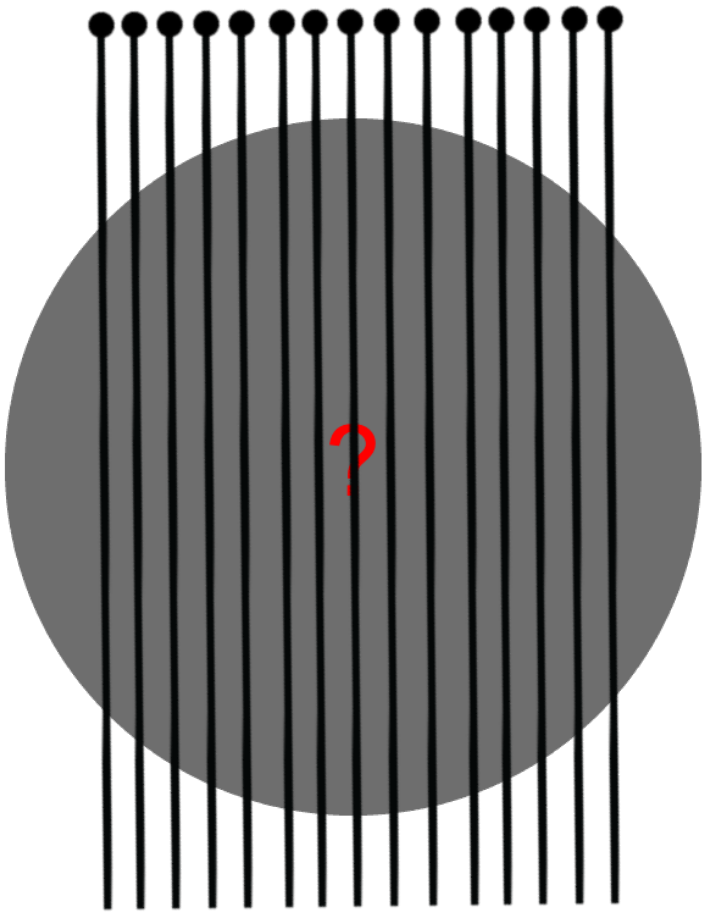
Parallel-beam X-ray tomography



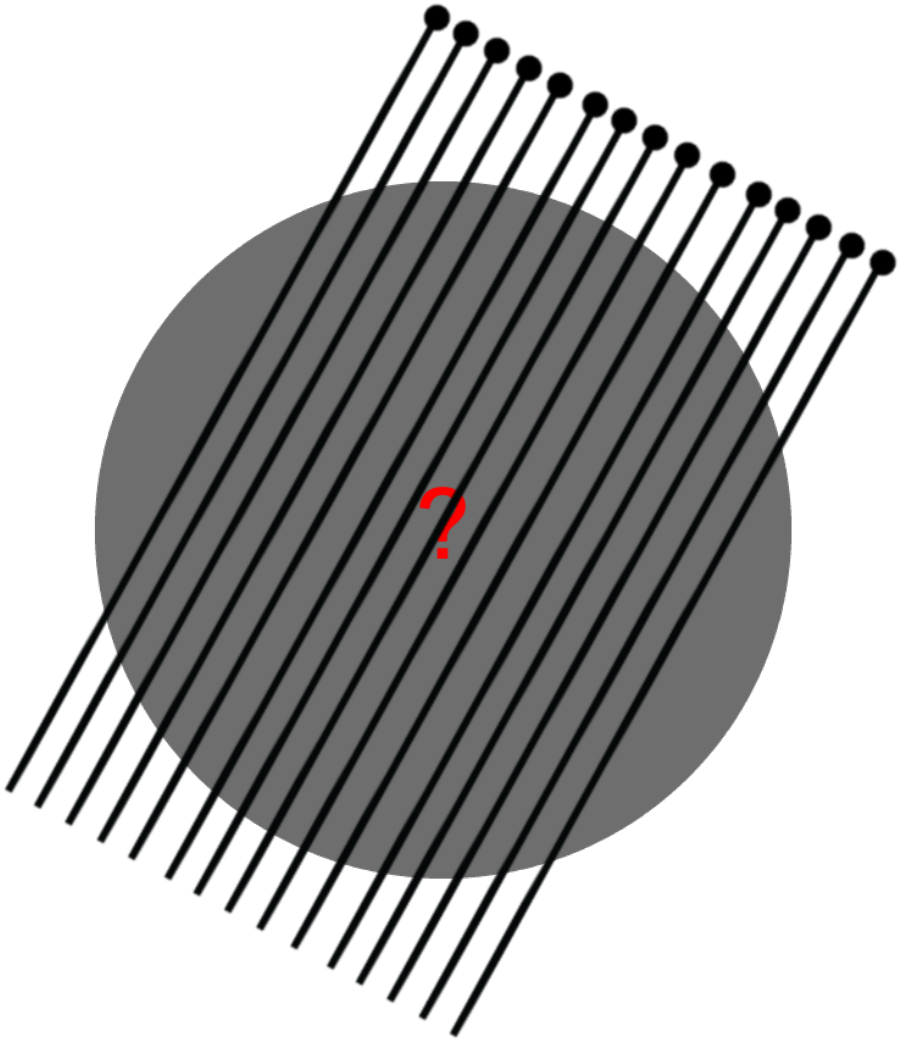
Parallel-beam X-ray tomography



Parallel-beam X-ray tomography



Parallel-beam X-ray tomography



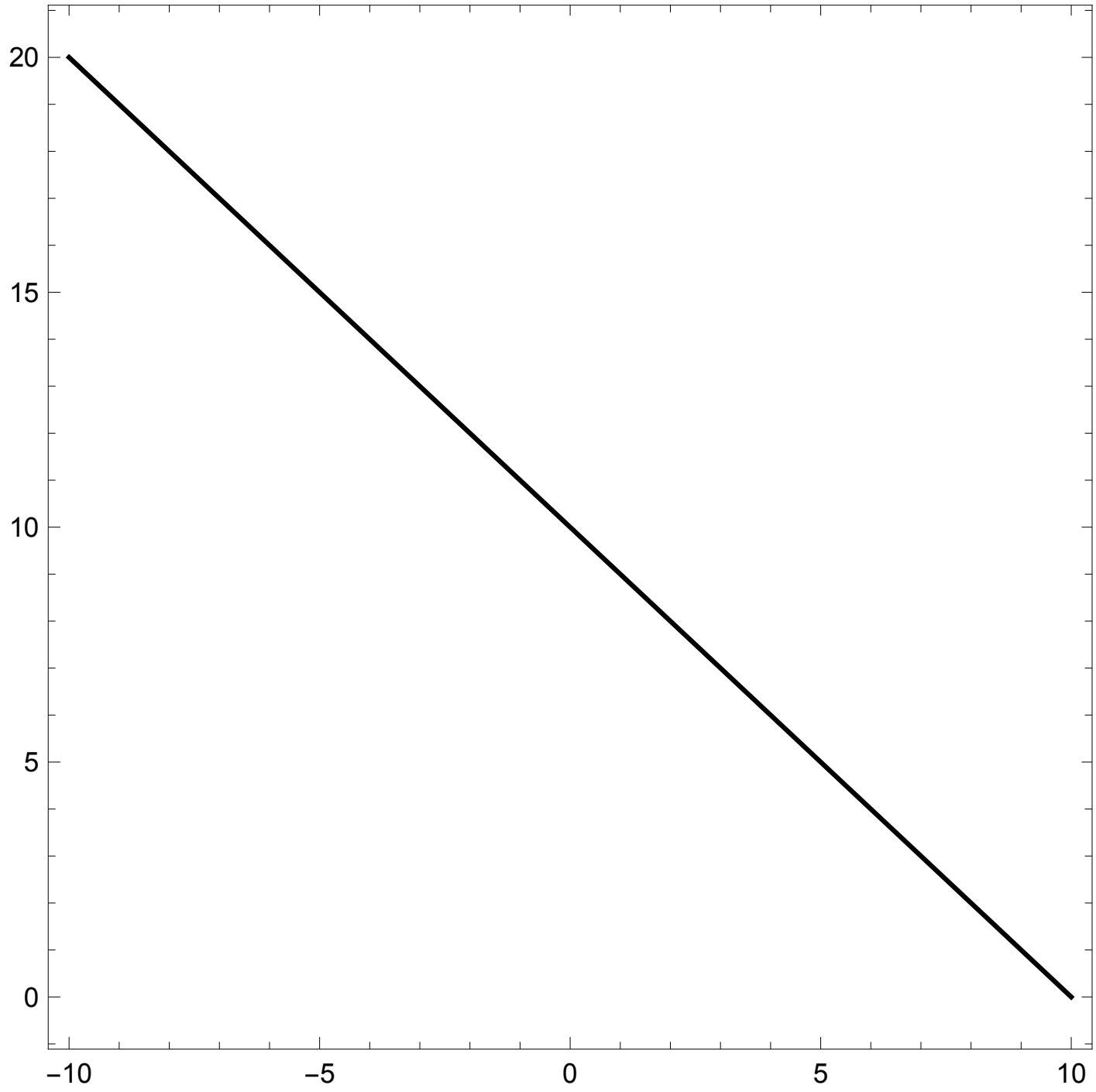
Radon transform in $\mathbb{R}^2$

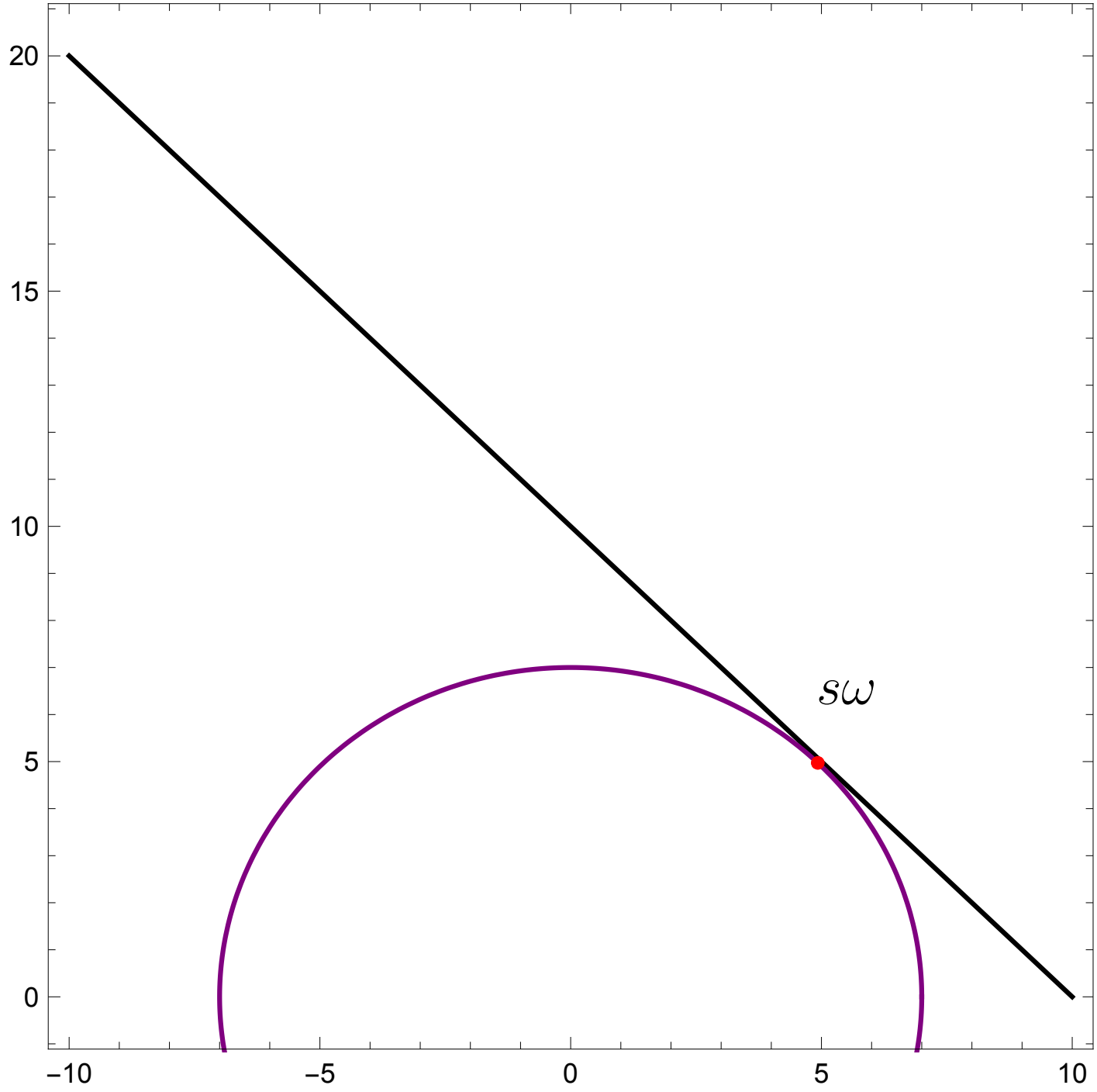
Let L be a straight line in $\mathbb{R}^2$.

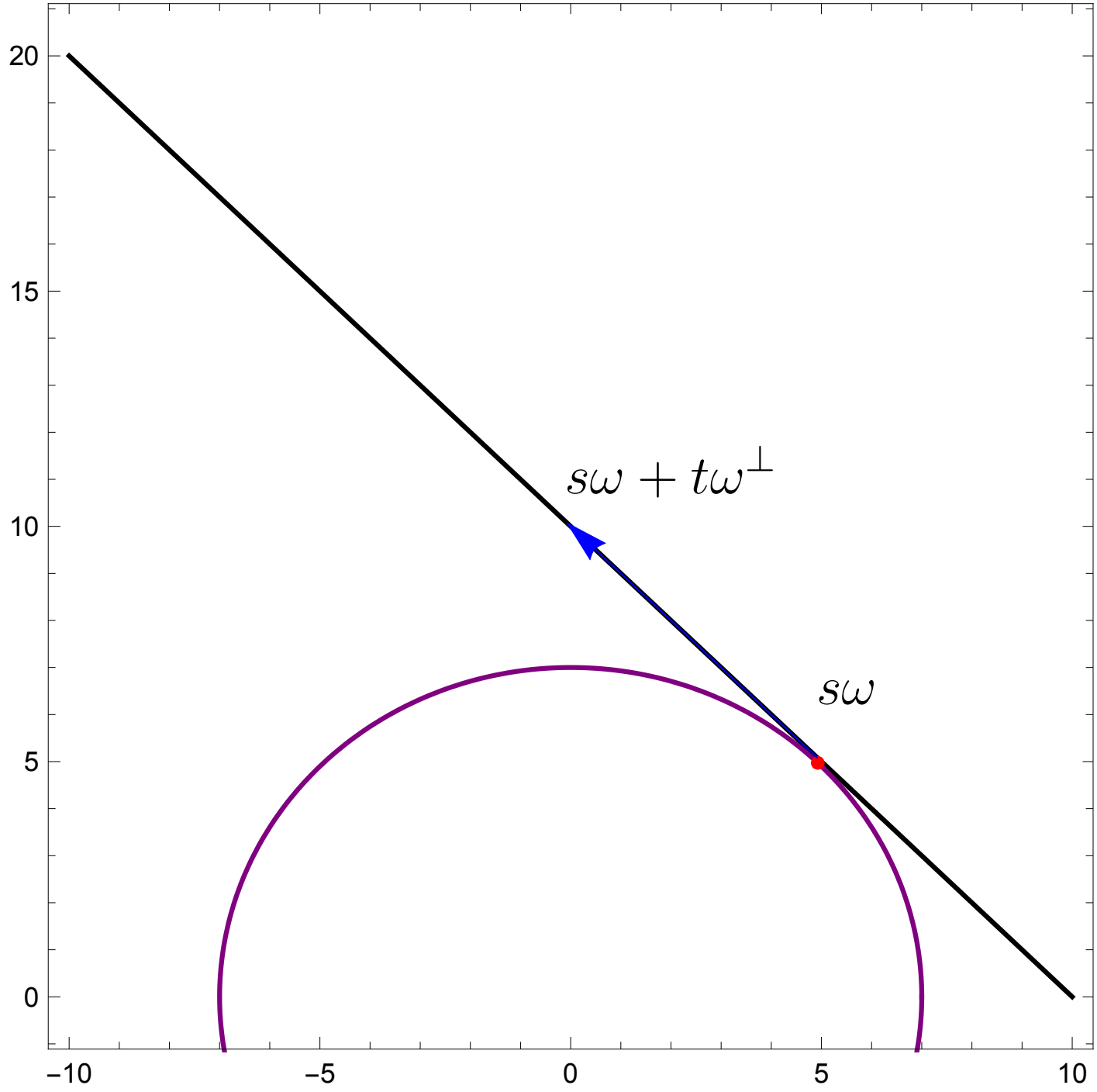
Any line in $\mathbb{R}^2$ can be parameterized as

$$L = \{s\omega + t\omega^{\perp};\ t \in \mathbb{R}\} \quad \text{for some } s \in \mathbb{R} \text{ and } \omega \in S^1,$$

where $\omega^{\perp} \perp \omega$.


20
15
10
5
0
−10
−5
0
5
10


20
15
10
5
0
−10
−5
0
5
10
$s\omega$


20
15
10
5
0
−10
−5
0
5
10
$s\omega + t\omega^{\perp}$
$s\omega$

Radon transform in $\mathbb{R}^2$

Let L be a straight line in $\mathbb{R}^2$.

Any line in $\mathbb{R}^2$ can be parameterized as

$$L = \{s\omega + t\omega^{\perp};\ t \in \mathbb{R}\} \quad \text{for some } s \in \mathbb{R} \text{ and } \omega \in S^1,$$

where $\omega^{\perp} \perp \omega$.

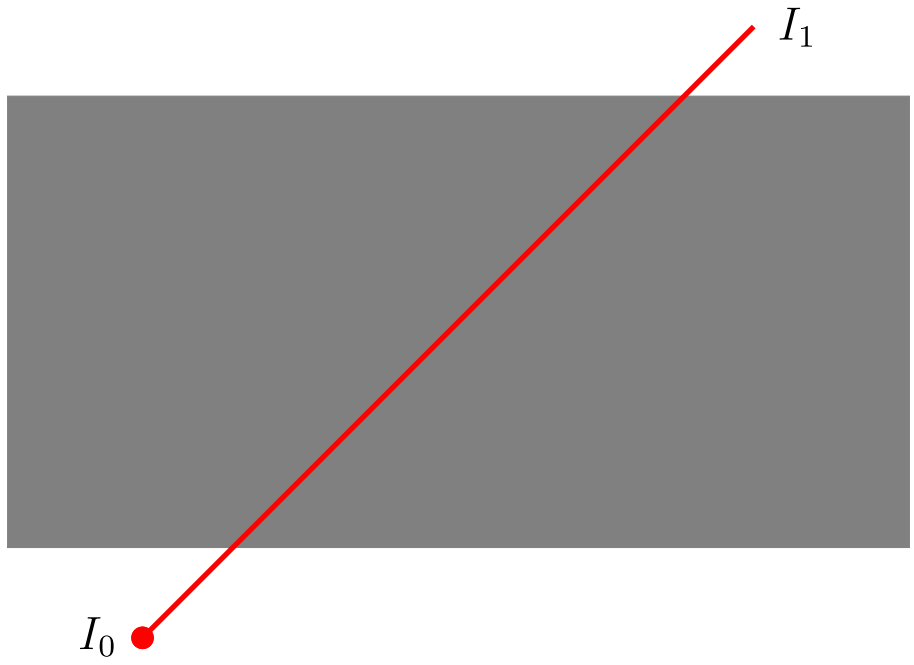
Writing $\omega = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$, we get

$$L = L(s,\theta) = \left\{ s \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + t \begin{bmatrix} \sin\theta \\ -\cos\theta \end{bmatrix};\ t \in \mathbb{R} \right\}, \quad s \in \mathbb{R} \text{ and } \theta \in [0,\pi).$$

The *Radon transform* of a continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ on L is defined as

$$\mathcal{R}f(L) = \int_L f(\boldsymbol{x})\,|\mathrm{d}\boldsymbol{x}| = \int_{-\infty}^{\infty} f(s\cos\theta + t\sin\theta, s\sin\theta - t\cos\theta)\,\mathrm{d}t.$$

Let f be a nonnegative function modeling X-ray attenuation (density) inside a physical body.



Beer–Lambert law:

$$\mathcal{R}f(L) = \log \frac{I_0}{I_1}.$$

$f_{9,0}$	$f_{9,1}$	$f_{9,2}$	$f_{9,3}$	$f_{9,4}$	$f_{9,5}$	$f_{9,6}$	$f_{9,7}$	$f_{9,8}$	$f_{9,9}$
$f_{8,0}$	$f_{8,1}$	$f_{8,2}$	$f_{8,3}$	$f_{8,4}$	$f_{8,5}$	$f_{8,6}$	$f_{8,7}$	$f_{8,8}$	$f_{8,9}$
$f_{7,0}$	$f_{7,1}$	$f_{7,2}$	$f_{7,3}$	$f_{7,4}$	$f_{7,5}$	$f_{7,6}$	$f_{7,7}$	$f_{7,8}$	$f_{7,9}$
$f_{6,0}$	$f_{6,1}$	$f_{6,2}$	$f_{6,3}$	$f_{6,4}$	$f_{6,5}$	$f_{6,6}$	$f_{6,7}$	$f_{6,8}$	$f_{6,9}$
$f_{5,0}$	$f_{5,1}$	$f_{5,2}$	$f_{5,3}$	$f_{5,4}$	$f_{5,5}$	$f_{5,6}$	$f_{5,7}$	$f_{5,8}$	$f_{5,9}$
$f_{4,0}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$	$f_{4,5}$	$f_{4,6}$	$f_{4,7}$	$f_{4,8}$	$f_{4,9}$
$f_{3,0}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{3,5}$	$f_{3,6}$	$f_{3,7}$	$f_{3,8}$	$f_{3,9}$
$f_{2,0}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{2,5}$	$f_{2,6}$	$f_{2,7}$	$f_{2,8}$	$f_{2,9}$
$f_{1,0}$	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	$f_{1,6}$	$f_{1,7}$	$f_{1,8}$	$f_{1,9}$
$f_{0,0}$	$f_{0,1}$	$f_{0,2}$	$f_{0,3}$	$f_{0,4}$	$f_{0,5}$	$f_{0,6}$	$f_{0,7}$	$f_{0,8}$	$f_{0,9}$

Let us consider the computational domain $[-1,1]^2$. We divide this region into $n \times n$ pixels and approximate the density by a piecewise constant function with constant value

$$f_{i,j} \text{ in pixel } P_{i,j}$$

for $i,j \in \{0,\ldots,n-1\}$.

$$P_{i,j} := \{(x,y);\ -1+2\,\tfrac{j}{n} < x < -1+2\,\tfrac{j+1}{n},\ -1+2\,\tfrac{i}{n} < y < -1+2\,\tfrac{i+1}{n}\}$$

x_{90}	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	x_{99}
x_{80}	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	x_{88}	x_{89}
x_{70}	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	x_{77}	x_{78}	x_{79}
x_{60}	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}	x_{67}	x_{68}	x_{69}
x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

It is convenient to reshape the matrix/image $(f_{i,j})$ into a vector x of length n^2 so that

$$x_{in+j} = f_{i,j}, \quad i,j \in \{0,\ldots,n-1\}.$$

The image on the left illustrates the new numbering corresponding to the pixels.

Note that `x = f.reshape((n*n,1))` and `f = x.reshape((n,n))`.
(In MATLAB: `x = f(:)` and `f = reshape(x,n,n)`).

Measurement model

Let us consider a measurement setup where we take X-ray measurements of an object using K X-rays $L(s_0, \theta), \ldots, L(s_{K-1}, \theta)$ taken at angles $\theta \in \{\theta_0, \ldots, \theta_{M-1}\}$. The total number of X-rays is $Q = MK$.

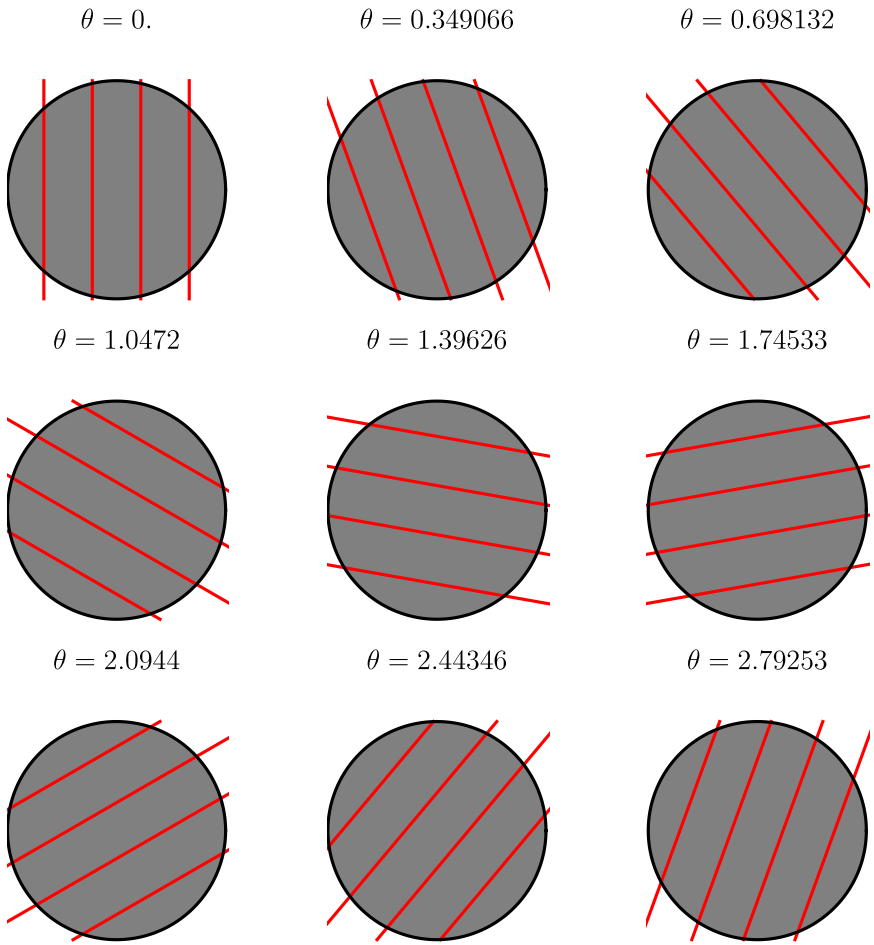
For brevity, let us write $L_{mK+k} := L(s_k, \theta_m)$ for $k \in \{0, \ldots, K-1\}$ and $m \in \{0, \ldots, M-1\}$.

The measurement model is

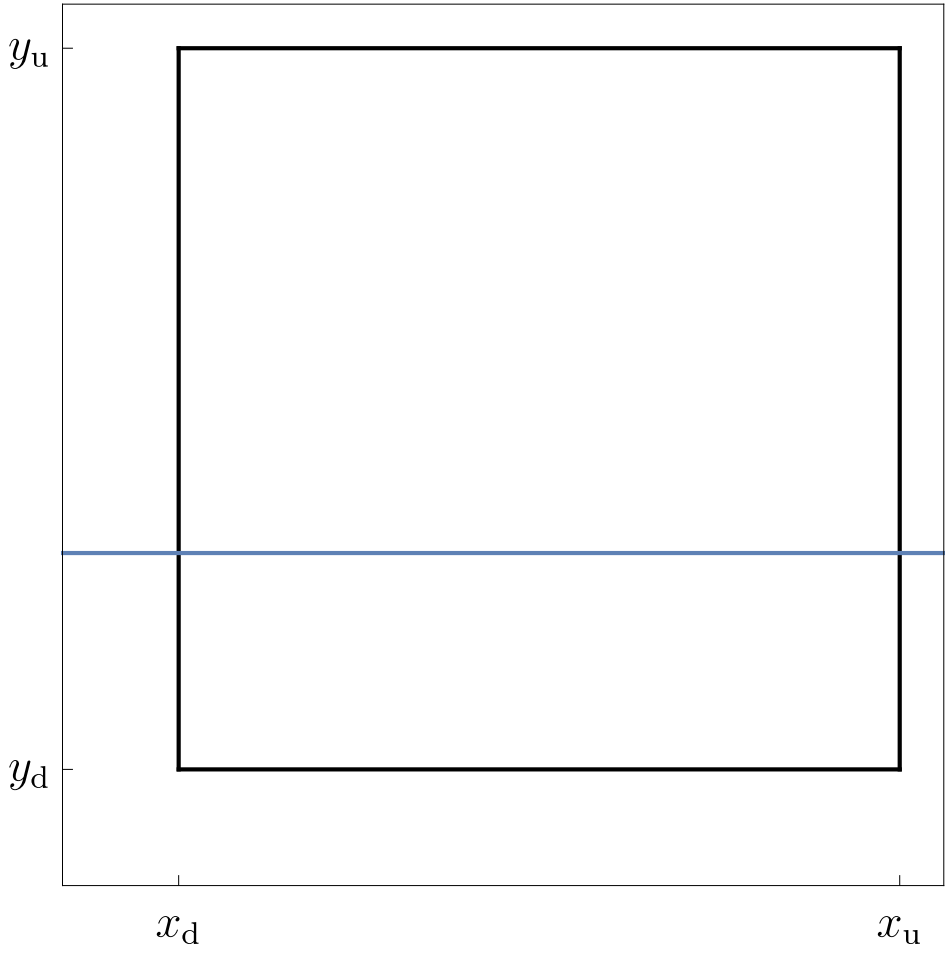
$$y = \begin{bmatrix} \int_{L_0} f(\boldsymbol{x}) |\mathrm{d}\boldsymbol{x}| \\ \vdots \\ \int_{L_{Q-1}} f(\boldsymbol{x}) |\mathrm{d}\boldsymbol{x}| \end{bmatrix} + \eta \approx \begin{bmatrix} \sum_{j=0}^{n^2-1} A_{0,j} x_j \\ \vdots \\ \sum_{j=0}^{n^2-1} A_{Q-1,j} x_j \end{bmatrix} + \eta = Ax + \eta,$$

where $A \in \mathbb{R}^{Q \times n^2}$ and $A_{i,j}$ is the distance that ray L_i travels through pixel j. Here, x is a vector containing the (piecewise constant) densities within each pixel and η is measurement noise.

$$L_{mK+k} = \left\{ s_k \begin{bmatrix} \cos\theta_m \\ \sin\theta_m \end{bmatrix} + t \begin{bmatrix} \sin\theta_m \\ -\cos\theta_m \end{bmatrix};\ t \in \mathbb{R} \right\}, \quad \begin{array}{l} k = 0, \ldots, K-1, \\ m = 0, \ldots, M-1. \end{array}$$



Pixel-by-pixel construction of the tomography matrix A
(See the files `tomodemo.py`/`tomodemo.m` on the course page!)



Case $\cos\theta = 0$ and $\sin\theta = 1$:

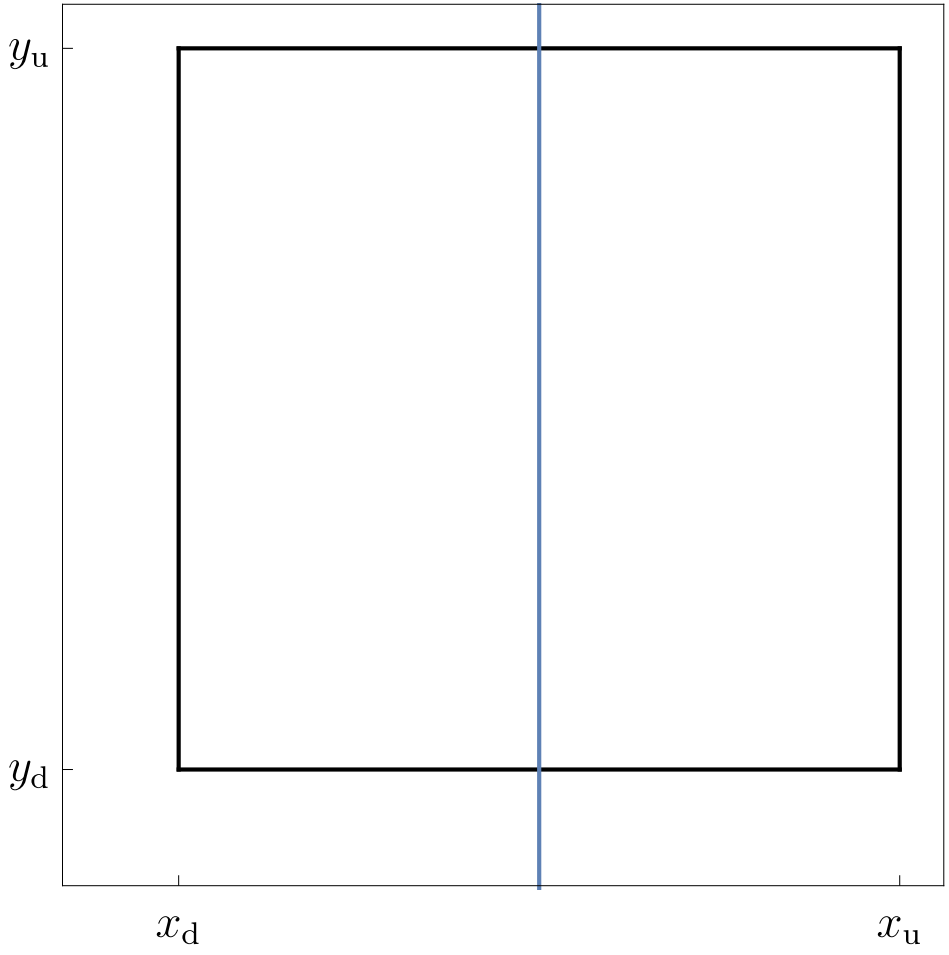
$$\begin{bmatrix} x_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} \leq \begin{bmatrix} s\cos\theta + t\sin\theta \\ s\sin\theta - t\cos\theta \end{bmatrix} < \begin{bmatrix} x_\mathrm{u} \\ y_\mathrm{u} \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} x_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} \leq \begin{bmatrix} t \\ s \end{bmatrix} < \begin{bmatrix} x_\mathrm{u} \\ y_\mathrm{u} \end{bmatrix}.$$

The distance that ray L_m travels through pixel k is

$$A_{m,k} = \int_{L_m} \chi_k \, |\mathrm{d}\boldsymbol{x}| = \int_{\substack{x_\mathrm{d} \leq t < x_\mathrm{u} \\ y_\mathrm{d} \leq s < y_\mathrm{u}}} \mathrm{d}t = \begin{cases} x_\mathrm{u} - x_\mathrm{d} & \text{if } y_\mathrm{d} \leq s < y_\mathrm{u}, \\ 0 & \text{otherwise.} \end{cases}$$

N.B. In here and in the following, $\chi_k = \chi_k(\boldsymbol{x})$ denotes the characteristic function of the k^{th} pixel. In the above illustration, the pixel is denoted by the rectangle $[x_\mathrm{d}, x_\mathrm{u}) \times [y_\mathrm{d}, y_\mathrm{u})$.



Case $\cos\theta = 1$ and $\sin\theta = 0$:

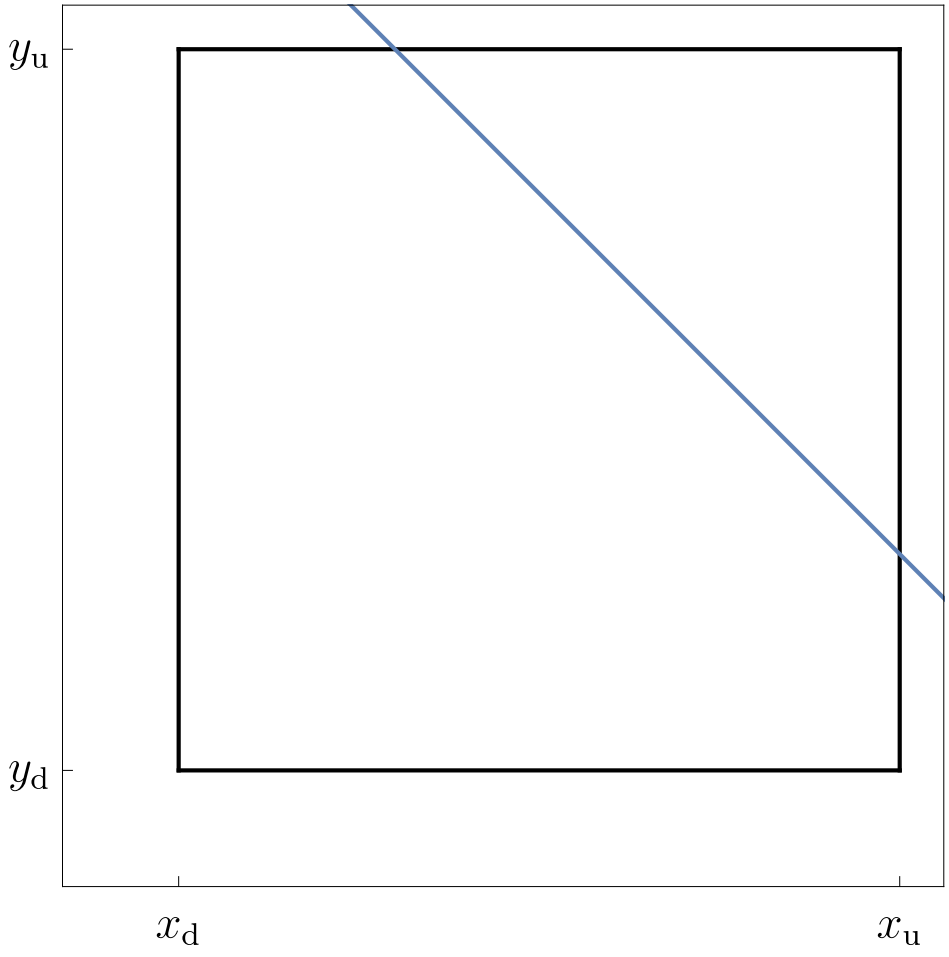
$$\begin{bmatrix} x_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} \leq \begin{bmatrix} s\cos\theta + t\sin\theta \\ s\sin\theta - t\cos\theta \end{bmatrix} < \begin{bmatrix} x_\mathrm{u} \\ y_\mathrm{u} \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} x_\mathrm{d} \\ y_\mathrm{d} \end{bmatrix} \leq \begin{bmatrix} s \\ -t \end{bmatrix} < \begin{bmatrix} x_\mathrm{u} \\ y_\mathrm{u} \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} x_\mathrm{d} \\ -y_\mathrm{u} \end{bmatrix} < \begin{bmatrix} s \\ t \end{bmatrix} \leq \begin{bmatrix} x_\mathrm{u} \\ -y_\mathrm{d} \end{bmatrix}.$$

The distance that ray L_m travels through pixel k is

$$A_{m,k} = \int_{L_m} \chi_k \, |\mathrm{d}\boldsymbol{x}| = \int_{\substack{-y_\mathrm{u} < t \leq -y_\mathrm{d} \\ x_\mathrm{d} < s \leq x_\mathrm{u}}} \mathrm{d}t = \begin{cases} y_\mathrm{u} - y_\mathrm{d} & \text{if } x_\mathrm{d} < s \leq x_\mathrm{u}, \\ 0 & \text{otherwise.} \end{cases}$$



Case $\cos\theta > 0$:

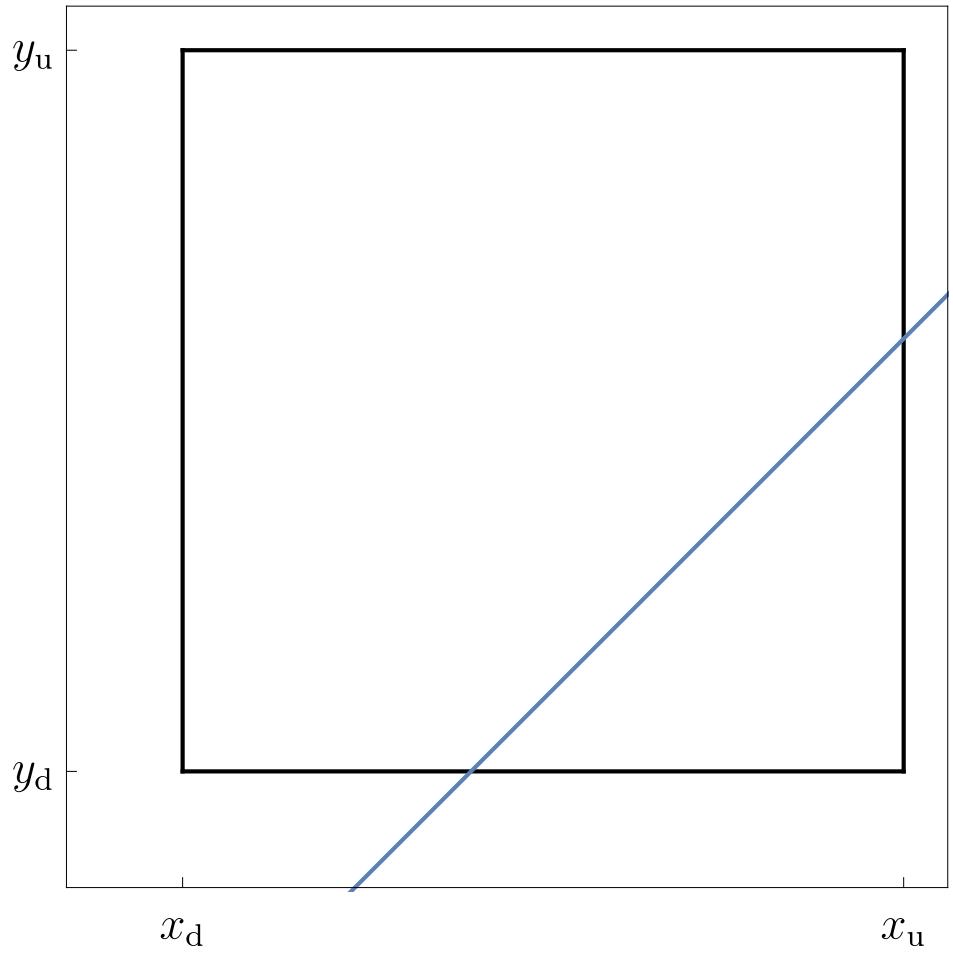
$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} < \begin{bmatrix} s\cos\theta + t\sin\theta \\ s\sin\theta - t\cos\theta \end{bmatrix} < \begin{bmatrix} x_u \\ y_u \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} \frac{x_d - s\cos\theta}{\sin\theta} \\ \frac{s\sin\theta - y_u}{\cos\theta} \end{bmatrix} < \begin{bmatrix} t \\ t \end{bmatrix} < \begin{bmatrix} \frac{x_u - s\cos\theta}{\sin\theta} \\ \frac{s\sin\theta - y_d}{\cos\theta} \end{bmatrix}.$$

The distance that ray L_m travels through pixel k is

$$A_{m,k} = \int_{L_m} \chi_k \, |\mathrm{d}\boldsymbol{x}| = \int_{\max\left\{\frac{x_d - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_u}{\cos\theta}\right\} < t < \min\left\{\frac{x_u - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_d}{\cos\theta}\right\}} \mathrm{d}t$$

$$= \left(\min\left\{ \frac{x_u - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_d}{\cos\theta} \right\} - \max\left\{ \frac{x_d - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_u}{\cos\theta} \right\} \right)_+ .$$



Case $\cos\theta < 0$:

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} < \begin{bmatrix} s\cos\theta + t\sin\theta \\ s\sin\theta - t\cos\theta \end{bmatrix} < \begin{bmatrix} x_u \\ y_u \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} \frac{x_d - s\cos\theta}{\sin\theta} \\ s\sin\theta - y_u \end{bmatrix} < \begin{bmatrix} t \\ t\cos\theta \end{bmatrix} < \begin{bmatrix} \frac{x_u - s\cos\theta}{\sin\theta} \\ s\sin\theta - y_d \end{bmatrix}$$

$$\overset{!}{\Leftrightarrow} \begin{bmatrix} \frac{x_d - s\cos\theta}{\sin\theta} \\ \frac{s\sin\theta - y_d}{\cos\theta} \end{bmatrix} < \begin{bmatrix} t \\ t \end{bmatrix} < \begin{bmatrix} \frac{x_u - s\cos\theta}{\sin\theta} \\ \frac{s\sin\theta - y_u}{\cos\theta} \end{bmatrix}.$$

The distance that ray L_m travels through pixel k is

$$A_{m,k} = \int_{L_m} \chi_k \, |\mathrm{d}\boldsymbol{x}| = \int_{\max\left\{\frac{x_d - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_d}{\cos\theta}\right\} < t < \min\left\{\frac{x_u - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_u}{\cos\theta}\right\}} \mathrm{d}t$$

$$= \left(\min\left\{ \frac{x_u - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_u}{\cos\theta} \right\} - \max\left\{ \frac{x_d - s\cos\theta}{\sin\theta}, \frac{s\sin\theta - y_d}{\cos\theta} \right\} \right)_+ .$$

Discussion

Tomography problems can be classified into three classes based on the nature of the measurement data:

- Full angle tomography
  - Sufficient number of measurements from all angles $\rightarrow$ not a very ill-posed problem.
- Limited angle tomography
  - Data collected from a restricted angle of view $\rightarrow$ reconstructions very sensitive to measurement error and it is not possible to reconstruct the object perfectly (even with noiseless data). Applications include, e.g., dental imaging.
- Sparse data tomography
  - The data consist of only a few projection images, possibly from any direction $\rightarrow$ extremely ill-posed inverse problem and prior knowledge necessary for successful reconstructions. (E.g., minimizing a patient's radiation dose.)

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fourth lecture, May 8, 2023

Fredholm equation and its solvability

Separable Hilbert space

A Hilbert space is said to be *separable* if (and only if) there exists a *countable orthonormal basis* $\{\psi_j\}_{j=1}^{\infty}$ of H with respect to the inner product $\langle \cdot, \cdot \rangle_H$, that is,

$$\langle \psi_j, \psi_k \rangle_H = \delta_{j,k} \quad \text{and} \quad \left\| f - \sum_{j=1}^{\ell} \langle f, \psi_j \rangle_H \psi_j \right\|_H \xrightarrow{\ell \to \infty} 0 \quad \text{for all } f \in H.$$

This last condition is usually written as

$$f = \sum_{j=1}^{\infty} \langle f, \psi_j \rangle_H \psi_j.$$

Note that $\sum_{j=1}^{\ell} \langle f, \psi_j \rangle_H \psi_j$ is precisely the orthogonal projection onto the subspace spanned by $\psi_1, \ldots, \psi_\ell$.

Fredholm equation

Let us formalize the problem that we will concentrate on during the first part of the course.

Let H_1 and H_2 be separable real Hilbert spaces and let $A\colon H_1 \to H_2$ be a *compact* linear operator. We are interested in finding $x \in H_1$ such that

$$y = Ax,$$

where $y \in H_2$ is given. Recall that compact operators are the closure of finite-dimensional operators (loosely speaking: matrices) in the operator topology.

Examples:

- $H_1 = H_2 = L^2(a, b)$.
- $H_1 = \mathbb{R}^n$, $H_2 = \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times n}$.

Singular value decomposition of a compact operator

Let us assume that H_1 and H_2 are separable real Hilbert spaces and let $A\colon H_1 \to H_2$ be a compact linear operator.

Then there exist (possibly countably infinite) orthonormal sets of vectors $\{v_n\} \subset H_1$ and $\{u_n\} \subset H_2$, and a sequence of positive numbers $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq 0$ with $\lim_{n\to\infty} \lambda_n = 0$ in the countably infinite case such that

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n \quad \text{for all } x \in H_1. \tag{1}$$

In particular, since H_1 and H_2 are separable, we have

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

The system (λ_n, v_n, u_n) is called a *singular system* of A, and (1) is a *singular value decomposition* (SVD) of A.

Singular value decomposition of matrices: $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$

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

$$y = Ax$$

is a matrix equation with $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$.

Since this operator has finite rank ($\mathrm{rank}(A) := \dim \mathrm{Ran}(A) < \infty$), we have

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

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 $\lambda_j \geq \lambda_{j+1}$.

It is possible to complete the sequences of (orthonormal) singular vectors $\{v_j\}_{j=1}^{p} \subset \mathbb{R}^n$ and $\{u_j\}_{j=1}^{p} \subset \mathbb{R}^m$ with complementary orthonormal vectors $\{v_j\}_{j=p+1}^{n}$ and $\{u_j\}_{j=p+1}^{m}$ such that $\{v_j\}_{j=1}^{n}$ forms an orthonormal basis for $\mathbb{R}^n$ and $\{u_j\}_{j=1}^{m}$ forms an orthonormal basis for $\mathbb{R}^m$. This can be done, e.g., using the Gram–Schmidt process.

Define the matrices

$$V = [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n},$$
$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m \times m}.$$

Due to the orthonormality of $\{v_j\}$ and $\{u_j\}$, the matrices V and U are orthogonal:

$$V^{\mathrm{T}} V = V V^{\mathrm{T}} = I \quad \text{and} \quad U^{\mathrm{T}} U = U U^{\mathrm{T}} = I.$$

Next, we define the matrix $\Lambda \in \mathbb{R}^{m \times n}$ as follows:

$$\Lambda = \left(\begin{array}{ccc|c} \lambda_1 & & & \\ & \ddots & & O_{m \times (n-m)} \\ & & \lambda_m & \end{array} \right) \quad \text{if } m < n,$$

$$\Lambda = \left(\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \\ \hline & O_{(m-n) \times n} & \end{array} \right) \quad \text{if } m > n,$$

and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ if $m = n$.

It is simple to check that

$$Ax = \sum_{j=1}^{p} \lambda_j u_j v_j^{\mathrm{T}} x = U \Lambda V^{\mathrm{T}} x \quad \text{for all } x \in \mathbb{R}^n,$$

which yields the matrix singular value decomposition (SVD)

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

In Python: `numpy.linalg.svd`. *In MATLAB:* `svd`.

Note that in the matrix SVD, the singular values $\{\lambda_j\}_{j=1}^{\min\{m,n\}}$ are *non-negative* and

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

where $p = \mathrm{rank}(A) = \max_{1\le k\le \min\{m,n\}}\{k \mid \lambda_k > 0\}$.

Solvability of $y = Ax$

Let us assume that H_1 and H_2 are separable real Hilbert spaces and let $A\colon H_1 \to H_2$ be a compact linear operator. Let $P\colon H_2 \to \overline{\mathrm{Ran}(A)}$ be an orthogonal projection. This can be represented using the singular system of A as

$$Py = \sum_n \langle y, u_n \rangle u_n.$$

Theorem

Let $A\colon H_1 \to H_2$ be a compact operator with the singular system (λ_n, v_n, u_n). The equation $y = Ax$ has a solution iff

$$y = Py \quad \text{and} \quad \underbrace{\sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 < \infty}_{\text{“Picard criterion”}}.$$

In this case, the solution is of the form

$$x = x_0 + \sum_n \frac{1}{\lambda_n} \langle y, u_n \rangle v_n \quad \text{for arbitrary } x_0 \in \mathrm{Ker}(A).$$

Proof. "⇒" Suppose that $y = Ax$ has a solution $x \in H_1$. This implies that $y \in \mathrm{Ran}(A)$ (thus $y = Py$) and, moreover,

$$\langle y, u_j \rangle = \langle Ax, u_j \rangle = \langle x, A^* u_j \rangle = \lambda_j \langle x, v_j \rangle$$

$$\Rightarrow \sum_n \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2 = \sum_n |\langle x, v_n \rangle|^2 \overset{\text{Bessel inequ.}}{\leq} \|x\|^2 < \infty.$$

"⇐" Next, suppose that $y = Py$ and the Picard criterion hold and define $x := x_0 + \sum_n \lambda_n^{-1} \langle y, u_n \rangle v_n$, where $x_0 \in \mathrm{Ker}(A)$ is arbitrary. We obtain

$$Ax = Ax_0 + \sum_n \frac{1}{\lambda_n} \langle y, u_n \rangle A v_n = \sum_n \langle y, u_n \rangle u_n = Py = y. \quad \square$$

Remark. In the above proof, it is helpful to note that if A has the SVD

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n,$$

then its adjoint A^* has the SVD

$$A^* y = \sum_n \lambda_n \langle y, u_n \rangle v_n.$$

Note that for any $x \in H_1$, we have

$$\|Ax - y\|^2 = \|Ax - Py\|^2 + \|(I - P)y\|^2 \geq \|(I - P)y\|^2.$$

Hence, if y has a nonzero component in the subspace orthogonal to the range of A (which may happen if y is contaminated by noise), the equation $Ax = y$ cannot be satisfied exactly. Thus, the best we can do is to solve the projected equation

$$Ax = PAx = Py.$$

However, there is in general no guarantee that the Picard criterion

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

is satisfied for a general $y \in H_2$ if $\operatorname{rank}(A) = \dim \operatorname{Ran}(A) = \infty$.

Truncated singular value decomposition (TSVD)

To recap: the best we can do is to solve the projected equation

$$Ax = Py.$$

However, the solution exists iff the very restrictive Picard criterion holds.

We begin by considering one of the simplest regularization techniques for linear inverse problems. By restricting the range of P onto a finite-dimensional subspace of the range of A, we obtain a well-defined approximation to the above problem.

Truncated singular value decomposition (TSVD)

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

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

By 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

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

Note that $k \leq \mathrm{rank}(A)$ if $\mathrm{rank}(A) < \infty$.

It follows that the problem

$$Ax = P_k y \tag{2}$$

is always solvable. Taking on both sides the inner product with u_n, we find that

$$\lambda_n \langle x, v_n \rangle = \begin{cases} \langle y, u_n \rangle, & 1 \leq n \leq k \\ 0, & n > k. \end{cases}$$

Hence the solutions to (2) are given by

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

for any $x_0 \in \mathrm{Ker}(A)$. Observe that since for increasing k,

$$\|Ax_k - Py\|^2 = \|(P - P_k)y\|^2 \xrightarrow{k \to \infty} 0,$$

the residual of the projected equation can be made arbitrarily small.

Finally, to remove the ambiguity of the sought solution due to the possible noninjectivity of A, we select $x_0 = 0$. This choice minimizes the norm of x_k since, by orthogonality,

$$\|x_k\|^2 = \|x_0\|^2 + \sum_{n=1}^{k} \frac{1}{\lambda_n^2} |\langle y, u_n \rangle|^2.$$

Definition

Let H_1 and H_2 be separable real Hilbert spaces and let $A\colon H_1 \to H_2$ be a compact linear operator with a singular system (λ_n, v_n, u_n). By the truncated SVD approximation (TSVD) of the problem $Ax = y$, we mean the problem of finding $x \in H_1$ such that

$$Ax = P_k y, \quad x \perp \mathrm{Ker}(A)$$

for some $k \geq 1$.

Theorem

The solution to the TSVD problem has a unique solution x_k, called the truncated SVD (TSVD) solution, given by

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

The TSVD solution satisfies

$$\|Ax_k - y\|^2 = \|(I - P)y\|^2 + \|(P - P_k)y\|^2 \xrightarrow{k\to\infty} \|(I - P)y\|^2.$$

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

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

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

where $P_k : \mathbb{R}^m \to \mathrm{span}\{u_1, \ldots, u_k\}$ is an orthogonal projection, is given by

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

where A has the SVD $A = U\Lambda V^{\mathrm{T}}$ and we define

$$\Lambda_k^{\dagger} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & & & & \vdots \\ \vdots & & \ddots & & & \\ & & & 1/\lambda_k & & \\ & & & & 0 & \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & & & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n\times m},$$

where $\lambda_1 \ge \cdots \ge \lambda_p > 0$ are the singular values of A (i.e., diagonal of Λ).

Moore–Penrose pseudoinverse of matrices

For the largest possible cut-off $k = p = \operatorname{rank}(A)$, 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 that $x^\dagger = A^\dagger y$ is the solution of the projected (matrix) equation

$$Ax = Py,$$

where $P\colon \mathbb{R}^m \to \operatorname{Ran}(A)$ is the orthogonal projection.

The solution $x^\dagger = A^\dagger y$ is called the *minimum norm solution* of the problem $y = Ax$ since

$$\|A^\dagger y\| = \min\{\|x\| : \|Ax - y\| = \|(I - P)y\|\},$$

where P is the projection onto the range of A. The minimum norm solution is the solution that minimizes the residual error and has the minimum norm.

Since the smallest singular value λ_p is extremely small in inverse problems, the use of the pseudoinverse is usually very sensitive to inaccuracies in the data y.

Spectral regularization using TSVD, i.e., discarding singular values below a certain threshold from the forward model, is a simple and popular technique used to render linear problems less ill-posed while improving the noise robustness of the numerical inversion procedure.

However, obtaining the singular values and vectors for large system matrices is usually very slow.

Numerical experiment: TSVD for X-ray tomography

Let us consider the inverse problem of recovering the attenuation coefficient (density) of an object given a set of X-ray measurements. Recall from last week that the mathematical model can be expressed as

$$y = Ax,$$

where $y \in \mathbb{R}^Q$ denotes the (noisy) measurements for Q X-rays, $A \in \mathbb{R}^{Q \times n^2}$ is the projection matrix subject to an $n \times n$ pixel discretization of the computational domain, and $x \in \mathbb{R}^{n^2}$ denotes the (piecewise constant) discretization of the unknown attenuation inside the object of interest.

The data y can be reshaped into an $n \times n$ array, which is a graphical representation of the X-ray measurements (sinogram). The unknown can be reshaped into an $n \times n$ image of the density of the imaged object.

Let us use TSVD to solve this inverse problem for *real-life measurement data*. We use the FIPS open dataset of carved cheese available at https://doi.org/10.5281/zenodo.1254210

The files `DataFull_128x15.mat` and `DataLimited_128x15.mat` contain sparse angle and limited angle tomography measurements, respectively. The data has been collected using 15 projections spanning either the full 360° circle in the first dataset, and 15 projections spanning a limited 90° angle of view in the second dataset. The computational domain is a 128×128 pixel grid in both cases. Each file contains a projection matrix `A` and a sinogram measurement matrix `m`.

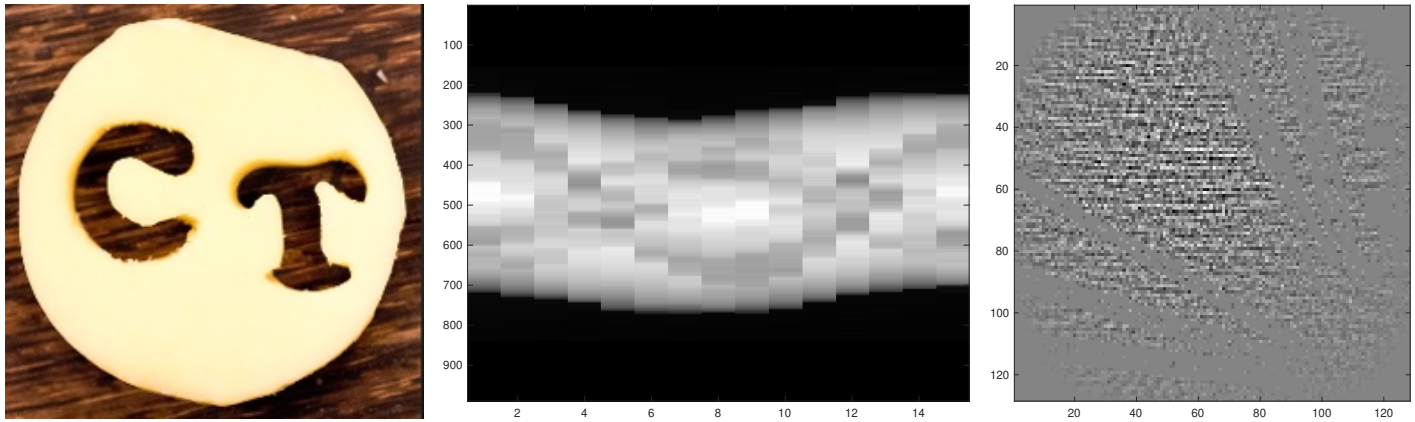
By defining `y = m.reshape((m.size,1))` (in MATLAB: `y = m(:)`), the unknown `x` can be solved from the linear equation

$$\mathtt{y} = \mathtt{A}\mathtt{x}.$$

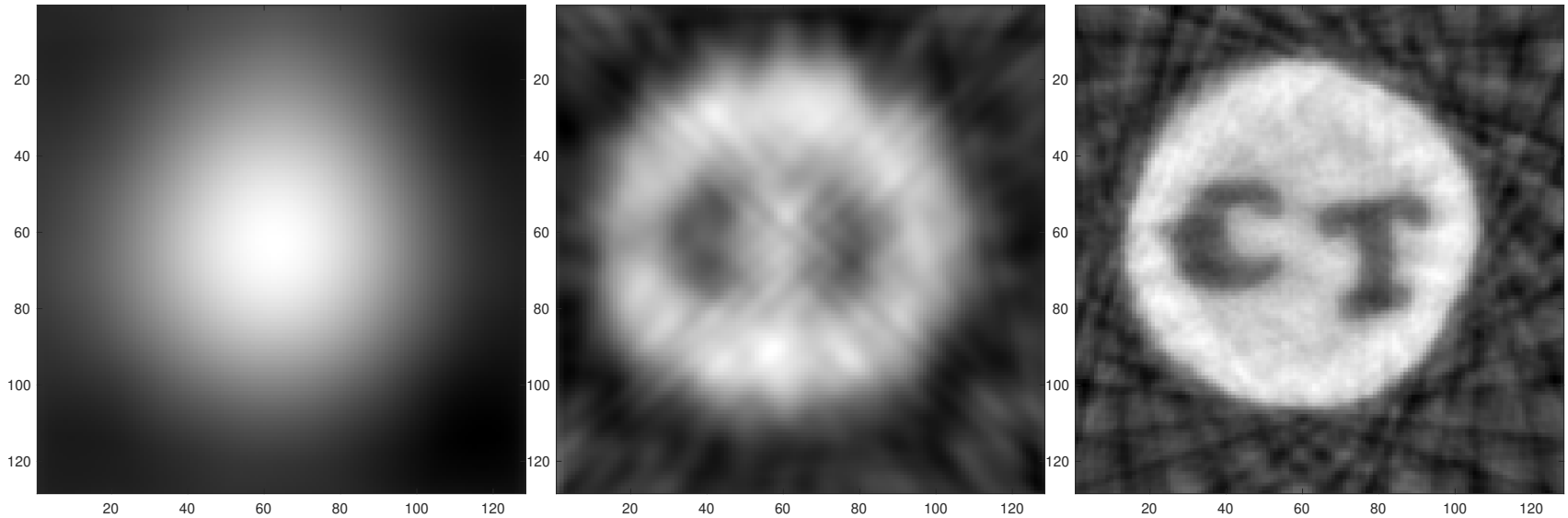
The reconstruction is the image `x.reshape((128,128))` (in MATLAB: `reshape(x,128,128)`).

See the files `tomo_tsvd.py` / `tomo_tsvd.m` on the course webpage!

TSVD for sparse angle tomography data

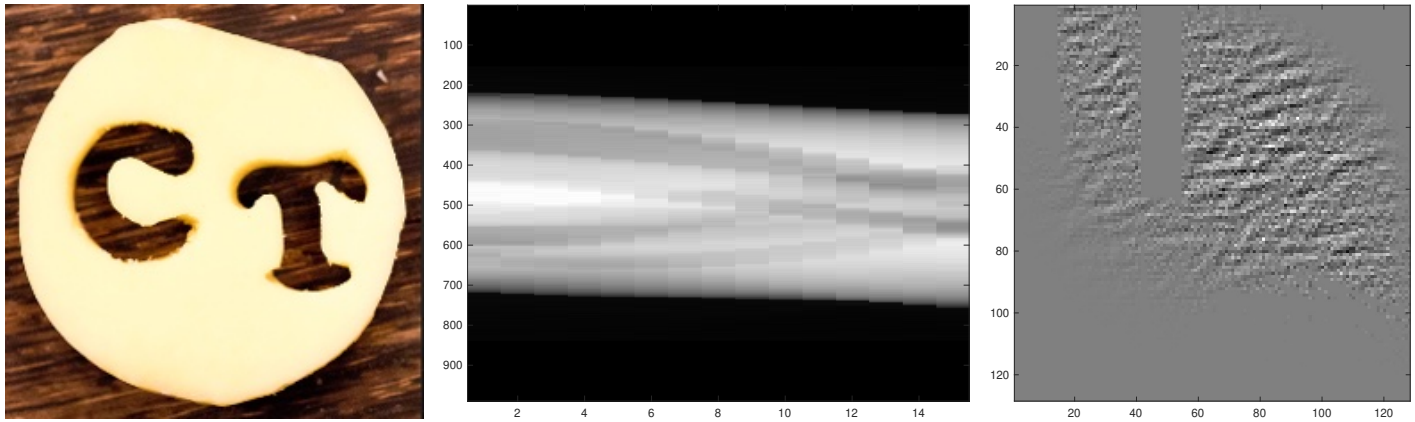


Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: naïve reconstruction without any regularization.

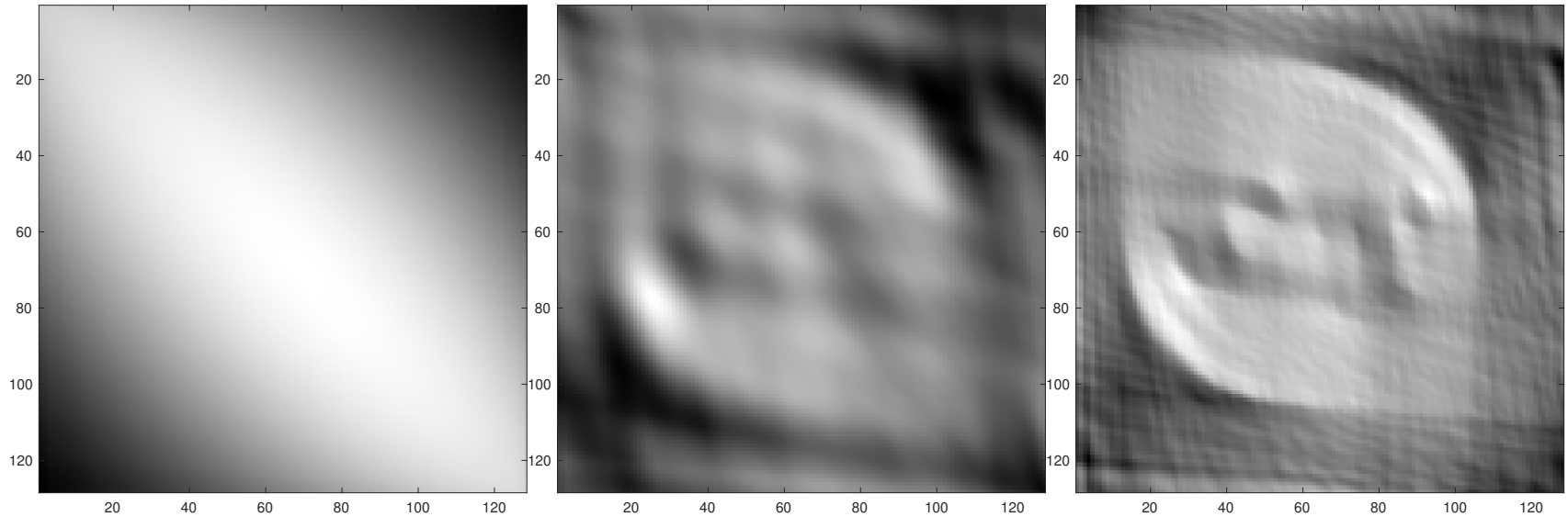


TSVD reconstructions with spectral cut-off $k = 10$ (left), $k = 100$ (middle), and $k = 1000$ (right).

TSVD for limited angle tomography data



Left: the actual object. Middle: sinogram data for limited angle tomography. Right: naïve reconstruction without any regularization.



TSVD reconstructions with spectral cut-off $k = 10$ (left), $k = 100$ (middle), and $k = 1000$ (right).

In summary (matrix case $H_1 = \mathbb{R}^n$, $H_2 = \mathbb{R}^m$): Let the SVD of matrix $A \in \mathbb{R}^{m\times n}$ be given by

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

where $\Lambda \in \mathbb{R}^{m\times n}$ has the non-negative singular values $\{\lambda_j\}_{j=1}^{\min\{m,n\}}$ on its diagonal and $V \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ are orthogonal matrices.[†]

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

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

where

$$\Lambda_k^{\dagger} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 & & \cdots & 0 \\ 0 & 1/\lambda_2 & & & & & \vdots \\ \vdots & & \ddots & & & & \\ & & & 1/\lambda_k & & & \\ & & & & 0 & & \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & & & & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

The matrix $A^{\dagger} = V\Lambda_p^{\dagger} U^{\mathrm{T}}$ is called the *Moore–Penrose pseudoinverse* of A.

[†]This means that the columns $\{v_j\}_{j=1}^{n}$ of V form an orthonormal basis for $\mathbb{R}^n$, and similarly the columns $\{u_j\}_{j=1}^{m}$ of U are an orhonormal basis of $\mathbb{R}^m$.

Morozov discrepancy principle

The implementation of TSVD raises the question: how to choose the spectral cut-off parameter k? If k is too small, the TSVD operator loses information about the forward operator. On the other hand, if k is chosen too large, then the forward operator becomes ill-conditioned and sensitive to measurement noise.

If the noise level of the data is known (or can be estimated), then one of the simplest criteria is to choose the spectral parameter as large as possible without fitting the solution to noise.

Morozov discrepancy principle

Let H_1 and H_2 be separable real Hilbert spaces and $A\colon H_1 \to H_2$ a compact linear operator.

How to choose the spectral cut-off index $k \geq 1$ in the TSVD problem

$$Ax = P_k y \quad \text{and} \quad x \perp \operatorname{Ker}(A)?$$

There is a rule of thumb called the *Morozov discrepancy principle:*

Suppose that the data $y \in H_2$ is a noisy approximation of noiseless "exact" data $y_0 \in H_2$. While y_0 is unknown to us, we may have an estimate on the noise level, e.g.,

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

We choose the smallest $k \geq 1$ such that the residual satisfies

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

Intuitively, this means that we cannot expect the approximate solution to yield a smaller residual than the measurement error without fitting the solution to noise.

Q: *When does an index $k \geq 1$ satisfying $\|y - Ax_k\| \leq \varepsilon$ exist?*
A: When $\varepsilon > \|Py - y\|$ and $\operatorname{rank}(A) = \infty$, it follows from $\overline{\operatorname{Ran}(A)} = \operatorname{Ran}(P) \perp \operatorname{Ran}(I - P)$ that

$$\|Ax_k - y\|^2 = \|Ax_k - Py + Py - y\|^2 = \|Ax_k - Py\|^2 + \|(P - I)y\|^2$$
$$= \sum_{n=k+1}^{\infty} |\langle y, u_n \rangle|^2 + \|(P - I)y\|^2 \xrightarrow{k \to \infty} \|Py - y\|^2.$$

Due to the properties of the orthogonal projection, $\|Py - y\| = \inf_{z \in \operatorname{Ran}(A)} \|z - y\|$, so this is the best we can do. (Note however that there is no guarantee that prevents $\|x_k\|$ from blowing up as $k \to \infty$.)

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

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

One should usually avoid choosing the spectral cut-off to be this large in practice.

Numerical example: backward heat equation

Let us consider the backward heat equation:

$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi), \end{cases}$$

where $f : (0,\pi) \to \mathbb{R}$ is the initial heat distribution.

Forward problem: Given initial data $f : (0,\pi) \to \mathbb{R}$, determine the heat distribution $u(\cdot, T)$ at time $T > 0$.

Inverse problem: Reconstruct the initial state f based on noisy measurements of $u(\cdot, T)$ at time $T > 0$.

Let us consider a simple discretization of the PDE

$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi). \end{cases}$$

Let $x_j = jh$ for $j = 0, \ldots, 100$, where $h = \pi/100$ is the step size.

Zero Dirichlet boundary conditions imply that $u(x_0,t) = u(x_{100},t) = 0$.

The spatial second derivative can be discretized using the stencils

$$\partial_x^2 u(x_1,t) = \frac{-2u(x_1,t) + u(x_2,t)}{h^2} + \mathcal{O}(h^2),$$
$$\partial_x^2 u(x_j,t) = \frac{u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)}{h^2} + \mathcal{O}(h^2) \quad \text{for } j = 2, \ldots, 98,$$
$$\partial_x^2 u(x_{99},t) = \frac{u(x_{98},t) - 2u(x_{99},t)}{h^2} + \mathcal{O}(h^2).$$

Denote $U(t) = (U_j(t))_{j=1}^{99} = (u(x_j,t))_{j=1}^{99}$ and $F = (f(x_j))_{j=1}^{99}$.

$$\frac{\partial}{\partial t}\begin{bmatrix} U_1(t) \\ U_2(t) \\ U_3(t) \\ \vdots \\ U_{98}(t) \\ U_{99}(t) \end{bmatrix} = \underbrace{\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & & \ddots & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix}}_{=:B} \begin{bmatrix} U_1(t) \\ U_2(t) \\ U_3(t) \\ \vdots \\ U_{98}(t) \\ U_{99}(t) \end{bmatrix}.$$

After spatial discretization, our PDE has been transformed into the initial value problem

$$\dot{U}(t) = BU(t), \quad U(0) = F.$$

At time $t = T > 0$, the discretized heat distribution $U := U(T)$ is given by

$$U = AF,$$

where $A = \mathrm{e}^{TB} \in \mathbb{R}^{99\times 99}$ and

$$\mathrm{e}^{M} := \sum_{k=0}^{\infty} \frac{1}{k!} M^k$$

is the *matrix exponential* (`scipy.linalg.expm` in Python / `expm` in MATLAB).

A note on simulating measurement data and inverse crimes

When simulating measurement data, one should take care not to use the same computational model for inversion as the one which was used to generate the measurements in the first place. This would lead to unreasonably good reconstructions, since this is akin to multiplying a matrix with its own inverse. This is known as an *inverse crime*. (Similar concerns also apply to non-linear problems.)

With real-life measurement data, we do not have worry about this phenomenon – measurements that come from nature are automatically independent of any computational model we end up using for practical inverse problems simulations.

A popular technique to avoid committing an inverse crime is using a higher resolution computational model to generate the measurements and interpolating the simulated data onto a coarser grid, where we plan to carry out the actual computational inversion. Another good option is to use an analytic solution, if one is readily available. We will use this technique with the heat equation.

The forward problem of the heat equation

$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi), \end{cases}$$

has the classical series solution

$$u(x,t) = \sum_{n=1}^{\infty} \hat{f}_n \, \mathrm{e}^{-n^2 t} \sin(nx),$$

where the coefficients $\hat{f}_n$ are the Fourier sine series coefficients of the initial heat distribution f satisfying

$$f(x) = \sum_{n=1}^{\infty} \hat{f}_n \sin(nx), \quad \hat{f}_n = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(nx) \, \mathrm{d}x.$$

Let us fix the ground truth

$$f(x) = \begin{cases} 1 & \text{if } x \in [1,2], \\ 0 & \text{if } x \in (0,1) \cup (2,\pi). \end{cases}$$

It is easy to see that the Fourier sine coefficients are given by

$$\hat{f}_n = \frac{2}{n\pi}(\cos n - \cos 2n).$$

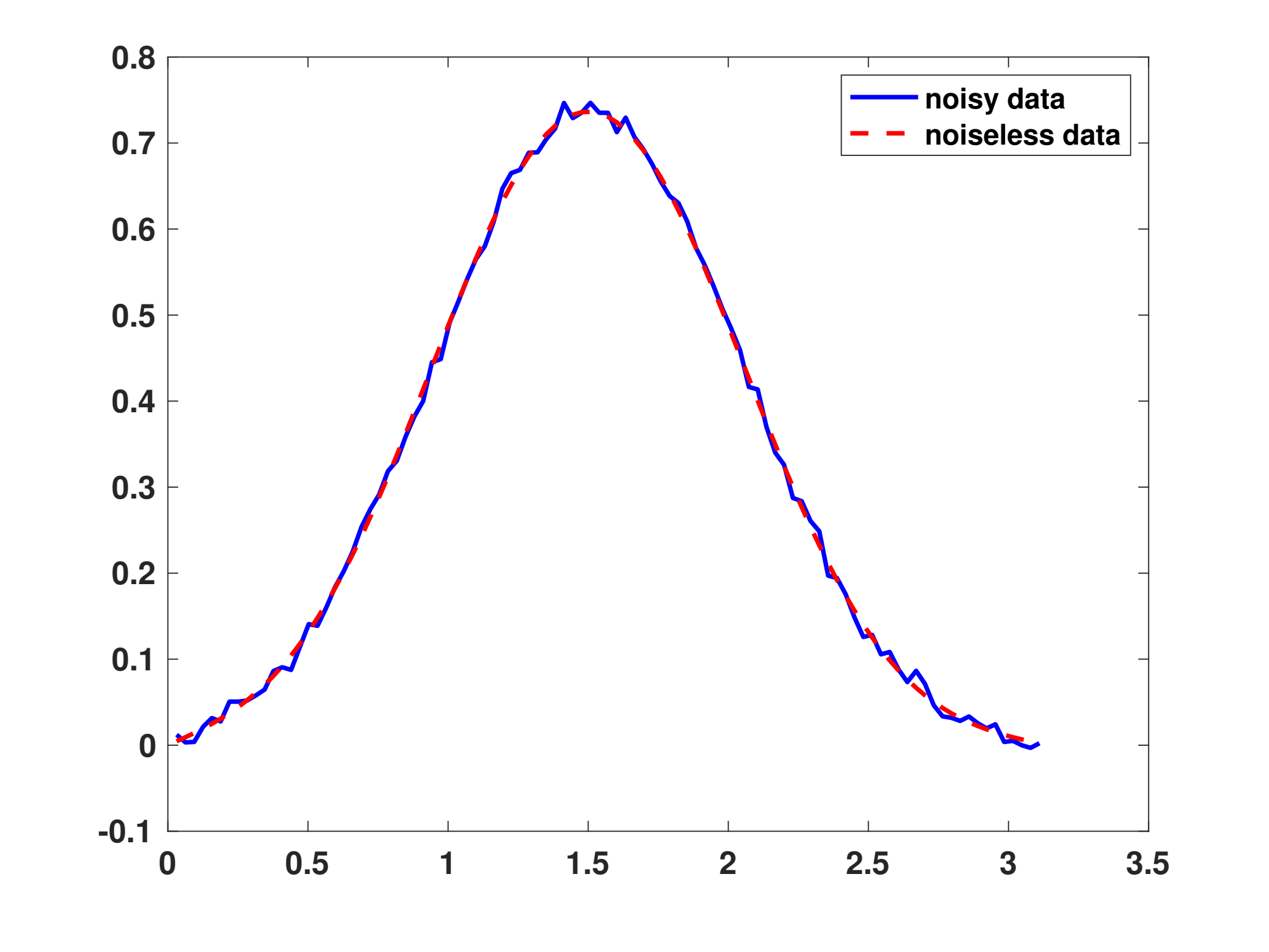
Let us plug these into the forward solution at time $t = T > 0$

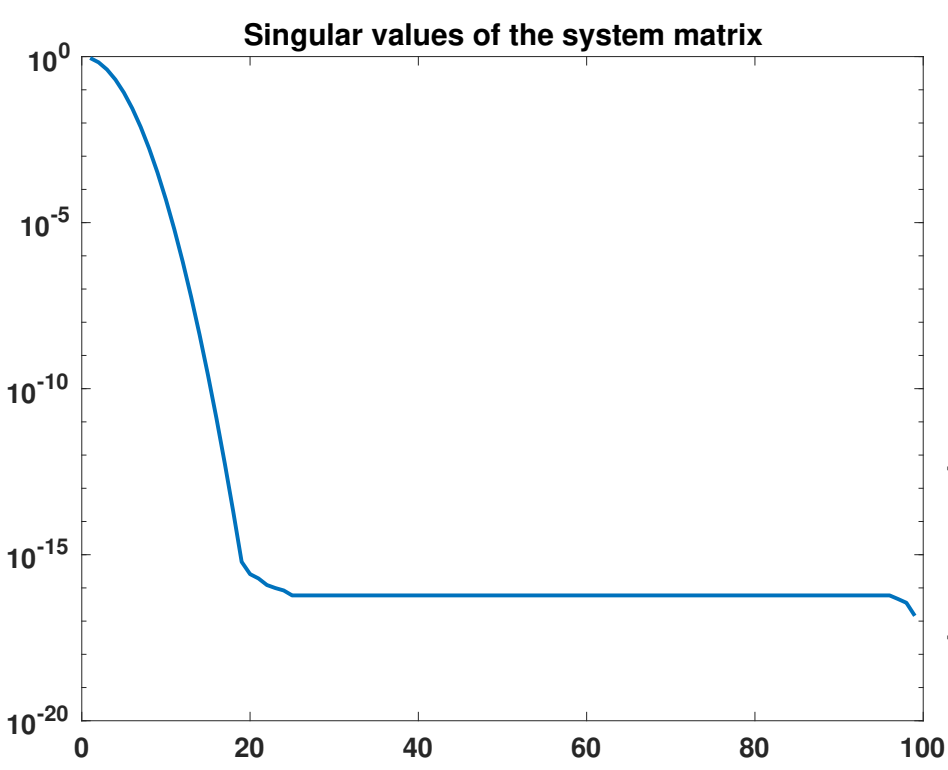
$$u(x_j, T) = \sum_{n=1}^{\infty} \hat{f}_n \, \mathrm{e}^{-n^2 T} \sin(n x_j), \quad j = 1, \ldots, 99,$$

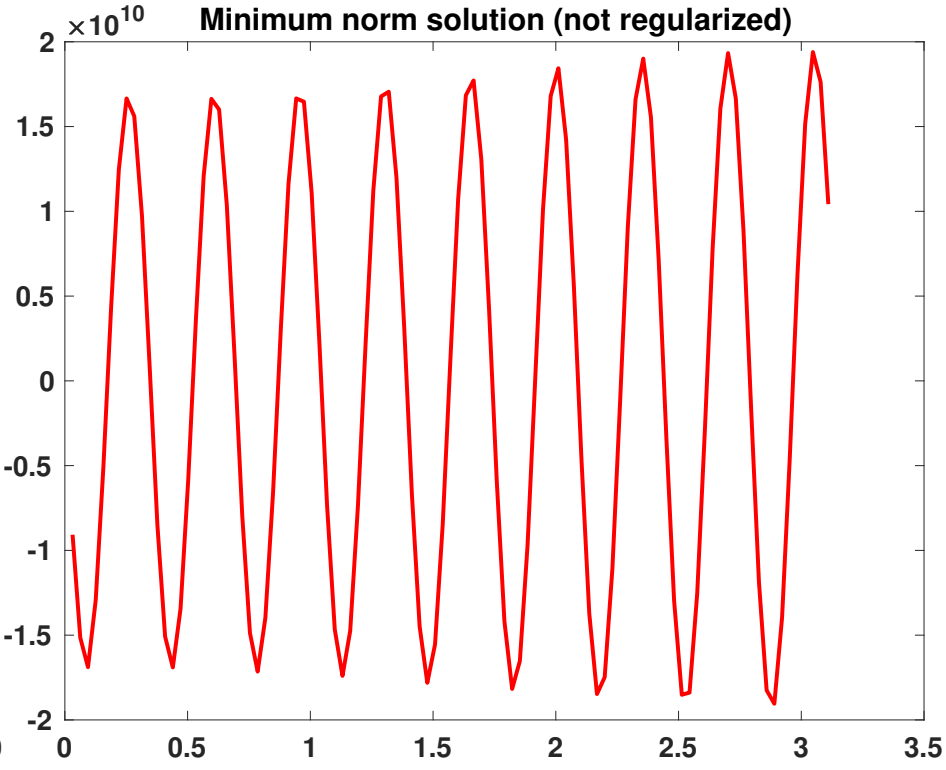
and add some simulated measurement noise!

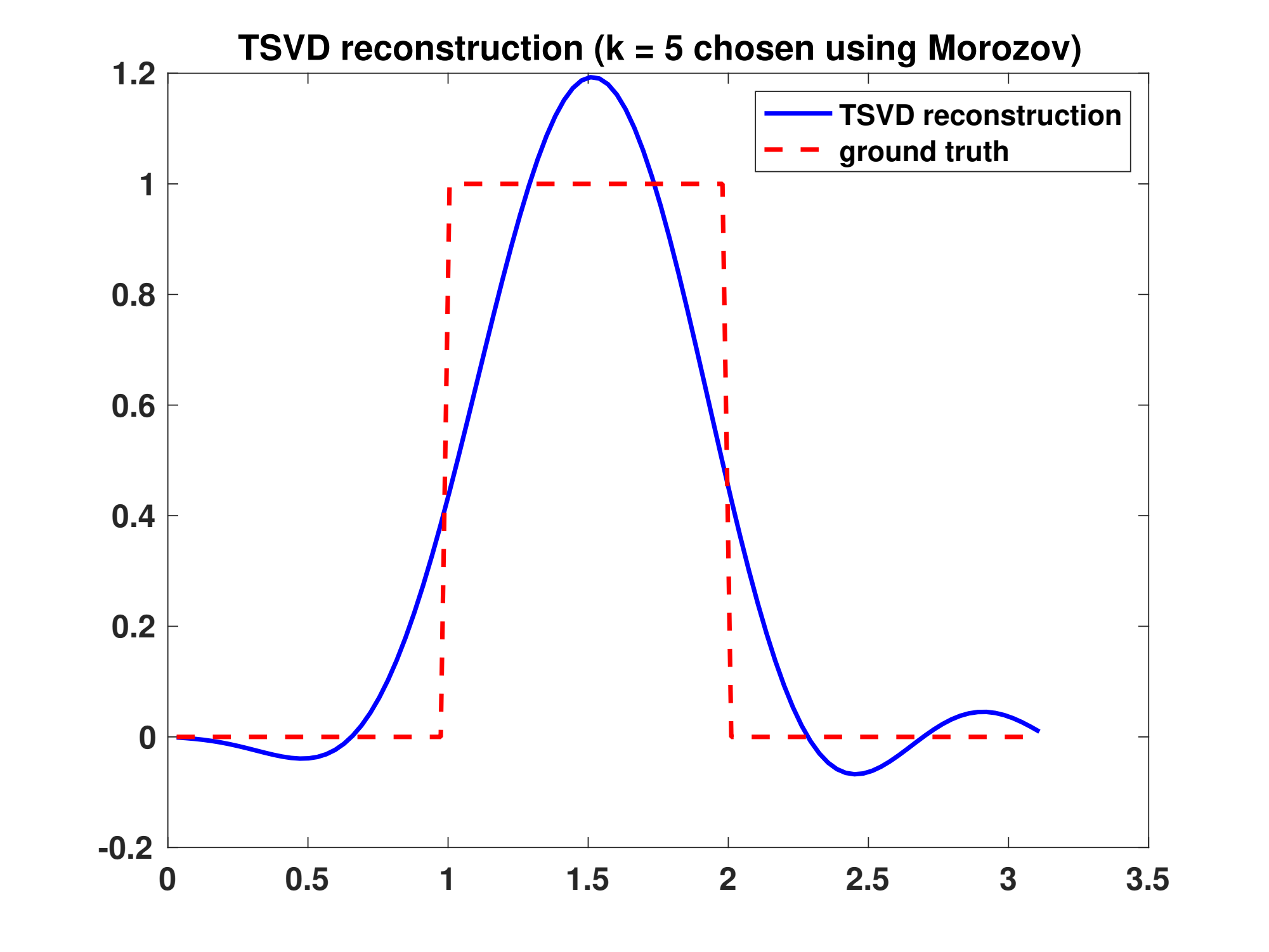
We assume that the data $U(T) \in \mathbb{R}^{99}$ at time $T = 0.1$ is contaminated with mean-zero Gaussian noise with standard deviation 0.01, and that the discrepancy between the measured data and the underlying "exact" data equals the square root of the expected value of the squared norm of the noise vector, i.e.,

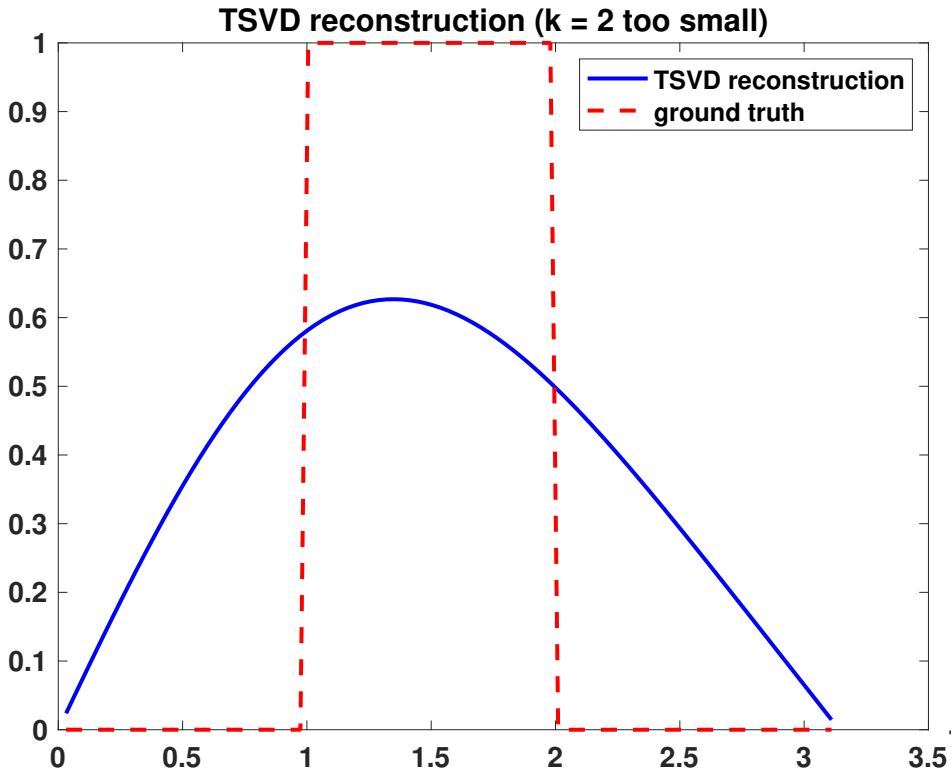
$$\varepsilon = \sqrt{99 \cdot 0.01^2} \approx 0.0995.$$

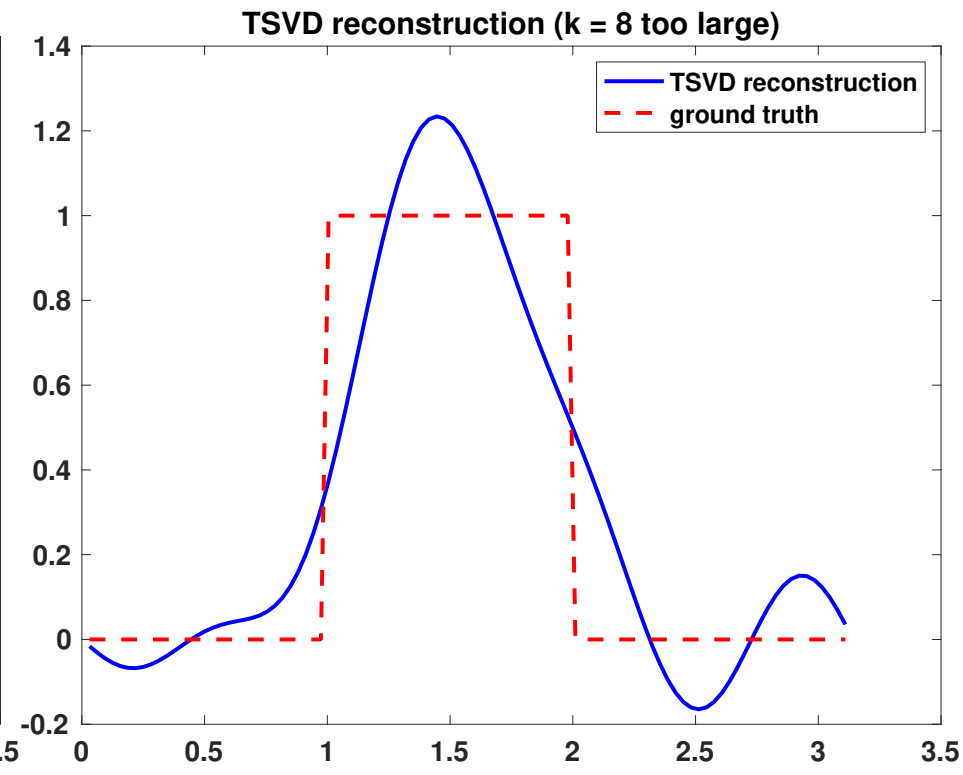

noisy data
noiseless data
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0
0.5
1
1.5
2
2.5
3
3.5


Singular values of the system matrix
10^0
10^-5
10^-10
10^-15
10^-20
0
20
40
60
80
100


Minimum norm solution (not regularized)
×10^10
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
0.5
1
1.5
2
2.5
3
3.5


TSVD reconstruction (k = 5 chosen using Morozov)
TSVD reconstruction
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5





See the files `heateq_tsvd.py` / `heateq_tsvd.m` on the course webpage!

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fifth lecture, May 15, 2023

Tikhonov regularization

Tikhonov regularization

The sequence of TSVD solutions $\{x_k\}$ minimizes the norm of the residual

$$\|Ax - y\|$$

as k tends to $\mathrm{rank}(A)$. Unfortunately, when inverse/ill-posed problems are considered, it may also happen that

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

In consequence, it appears reasonable to try minimizing the residual and the norm of the solution *simultaneously*.

Definition

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

Let $A\colon H_1 \to H_2$ be a compact linear operator with the singular system (λ_n, v_n, u_n). Then the Tikhonov regularized solution exists, is unique, and is given by the formula

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

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

Remark. The Tikhonov regularized solution can be obtained without knowing the SVD of A by solving x_δ from $(A^*A + \delta I)x_\delta = A^*y$.

Proof. We make use of the *Lax–Milgram lemma*:

Lemma (Lax–Milgram)

Let H be a Hilbert space, and let $B\colon H \times H \to \mathbb{R}$ be a bilinear quadratic form such that

$$|B(x,y)| \leq C\|x\|\|y\| \quad \text{for all } x, y \in H,$$
$$B(x,x) \geq c\|x\|^2 \quad \text{for all } x \in H$$

for some constants $0 < c \leq C < \infty$. Then there exists a unique linear boundedly invertible operator $T\colon H \to H$ such that

$$B(x,y) = \langle x, Ty \rangle \quad \text{for all } y \in H,$$
$$\|T\| \leq C \quad \text{and} \quad \|T^{-1}\| \leq \frac{1}{c}.$$

In our case, we define the bilinear operator $B(x,y) := \langle x, (A^*A + \delta I)y \rangle$ and observe that $|B(x,y)| \leq (\|A\|^2 + \delta)\|x\|\|y\|$ (boundedness) and $B(x,x) = \langle x, (A^*A + \delta I)x \rangle = \|Ax\|^2 + \delta\|x\|^2 \geq \delta\|x\|^2$ (coercivity).
$\therefore (A^*A + \delta I)^{-1}$ exists such that $\|(A^*A + \delta I)^{-1}\| \leq \frac{1}{\delta}$. In particular, $x_\delta = (A^*A + \delta I)^{-1}A^*y$ is well-defined.

Recall that $Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n$ and $A^* y = \sum_n \lambda_n \langle y, u_n \rangle v_n$. Especially,

$$A^* A x = \sum_n \lambda_n^2 \langle x, v_n \rangle v_n.$$

Since $H_1 = \mathrm{Ker}(A) \oplus \mathrm{Ker}(A)^\perp$, we can write

$$x_\delta = P x_\delta + Q x_\delta = \sum_n \langle x_\delta, v_n \rangle v_n + Q x_\delta,$$

where $P \colon H_1 \to \mathrm{Ker}(A)^\perp = \overline{\mathrm{span}\{v_n\}}$ and $Q \colon H_1 \to \mathrm{Ker}(A)$ are orthogonal projections. Thus

$$(A^* A + \delta I) x_\delta = A^* y \quad \Leftrightarrow \quad \sum_n (\lambda_n^2 + \delta) \langle x_\delta, v_n \rangle v_n + Q x_\delta = \sum_n \lambda_n \langle y, u_n \rangle v_n.$$

Equating terms yields that $Q x_\delta = 0$ and

$$(\lambda_n^2 + \delta) \langle x_\delta, v_n \rangle = \lambda_n \langle y, u_n \rangle \quad \Leftrightarrow \quad \langle x_\delta, v_n \rangle = \frac{\lambda_n}{\lambda_n^2 + \delta} \langle y, u_n \rangle,$$

as desired.

Finally, to show that x_δ minimizes the quadratic functional $F_\delta(x) = \|Ax - y\|^2 + \delta\|x\|^2$, consider

$$x = x_\delta + z,$$

where $z \in H_1$ is arbitrary. Now

$$\begin{aligned} F_\delta(x) &= F_\delta(x_\delta + z) = F_\delta(x_\delta) + \langle z, (A^*A + \delta I)x_\delta - A^*y\rangle + \langle z, (A^*A + \delta I)z\rangle \\ &= F_\delta(x_\delta) + \langle z, (A^*A + \delta I)z\rangle, \end{aligned}$$

by definition of x_δ.The last term is nonnegative and vanishes only if $z = 0$. This proves the claim. □

Morozov discrepancy principle for Tikhonov regularization

Suppose that the measurement $y \in H_2$ is a noisy version of some underlying "exact" data $y_0 \in H_2$, and that

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

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

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

It turns out that there is a unique regularization parameter satisfying this condition if

$$\|y - Py\| < \varepsilon < \|y\|,$$

where $P \colon H_2 \to \overline{\mathrm{Ran}(A)}$ is an orthogonal projection.

Properties of the Tikhonov regularized solution

Theorem

Let $A\colon H_1 \to H_2$ be a compact linear operator with the singular system (λ_n, v_n, u_n). Let $P\colon H_2 \to \overline{\mathrm{Ran}(A)}$ be an orthogonal projection. Then we have the following:

(i) $\delta \mapsto \|Ax_\delta - y\|$ *is a strictly increasing function of* $\delta > 0$.

(ii) $\|Py - y\| = \lim_{\delta \to 0+} \|Ax_\delta - y\| \le \|Ax_\delta - y\| \le \lim_{\delta \to \infty} \|Ax_\delta - y\| = \|y\|$.

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

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

as $\delta \to 0+$.

Corollary

The equation $\|Ax_\delta - y\| = \varepsilon$ *has a unique solution* $\delta = \delta(\varepsilon)$ *iff* $\|(I - P)y\| < \varepsilon < \|y\|$.

Interpretation: $\|(I - P)y\| < \varepsilon$ means that any component in the data y orthogonal to the range of A must be due to noise; $\varepsilon < \|y\|$ means that the error level should not exceed the signal level.

Proof. Suppose that the operator A has the SVD

$$Ax = \sum_n \lambda_n \langle x, v_n \rangle u_n.$$

Then $Av_n = \lambda_n u_n$, the orthogonal projection $P\colon H_2 \to \overline{\mathrm{Ran}(A)}$ is

$$Py = \sum_n \langle y, u_n \rangle u_n,$$

and the Tikhonov regularized solution x_δ and its image under A are

$$x_\delta = \sum_n \frac{\lambda_n}{\lambda_n^2 + \delta} \langle y, u_n \rangle v_n \quad \Rightarrow \quad Ax_\delta = \sum_n \frac{\lambda_n^2}{\lambda_n^2 + \delta} \langle y, u_n \rangle u_n.$$

(i) It follows that

$$\begin{aligned} \|Ax_\delta - y\|^2 &= \|Ax_\delta - Py\|^2 + \|(I - P)y\|^2 \\ &= \sum_n \left(\frac{\lambda_n^2}{\lambda_n^2 + \delta} - 1 \right)^2 |\langle y, u_n \rangle|^2 + \|(I - P)y\|^2 \\ &= \sum_n \left(\frac{\delta}{\lambda_n^2 + \delta} \right)^2 |\langle y, u_n \rangle|^2 + \|(I - P)y\|^2. \end{aligned}$$

We arrived at

$$\|Ax_\delta - y\|^2 = \sum_n \left(\frac{\delta}{\lambda_n^2 + \delta}\right)^2 |\langle y, u_n\rangle|^2 + \|(I - P)y\|^2.$$

For each term of the sum,

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\left(\frac{\delta}{\lambda_n^2 + \delta}\right)^2 = \frac{2\delta\lambda_n^2}{(\lambda_n^2 + \delta)^3} > 0,$$

implying that the mapping $\delta \mapsto \|Ax_\delta - y\|^2$ is strictly increasing.

(ii) It is easy to see that

$$\|Ax_\delta - y\|^2 = \sum_n \left(\frac{\delta}{\lambda_n^2 + \delta}\right)^2 |\langle y, u_n\rangle|^2 + \|(I - P)y\|^2 \xrightarrow{\delta\to 0+} \|(I - P)y\|^2,$$

$$\begin{aligned}\|Ax_\delta - y\|^2 &= \sum_n \left(\frac{\delta}{\lambda_n^2 + \delta}\right)^2 |\langle y, u_n\rangle|^2 + \|(I - P)y\|^2 \\ &\xrightarrow{\delta\to\infty} \|Py\|^2 + \|(I - P)y\|^2 = \|y\|^2.\end{aligned}$$

(iii) Let $Py \in \mathrm{Ran}(A)$. This implies that there exists $x \in \mathrm{Ker}(A)^{\perp}$ such that $Ax = Py$; this is the minimum norm solution

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

for which it can be shown that

$$x_\delta = \sum_n \frac{\lambda_n}{\lambda_n^2 + \delta} \langle y, u_n \rangle v_n \xrightarrow{\delta \to 0+} \sum_n \frac{1}{\lambda_n} \langle y, u_n \rangle v_n = x. \quad \square$$

Remark. In parts (ii) and (iii), one should take care when interchanging the order of the limit and the summation, i.e., justifying the steps

$$\lim_{\lambda \to 0+} \sum_n = \sum_n \lim_{\lambda \to 0+} \quad \text{and} \quad \lim_{\lambda \to \infty} \sum_n = \sum_n \lim_{\lambda \to \infty} .$$

Standard techniques involve the monotone convergence theorem and the dominated convergence theorem (note that these apply to infinite series as well as integrals). In part (iii), it is helpful to observe that $x_\delta \xrightarrow{\delta \to 0+} x$ iff $\langle x_\delta, \phi \rangle \xrightarrow{\delta \to 0+} \langle x, \phi \rangle$ for all $\phi \in H_1$ and $\|x_\delta\| \xrightarrow{\delta \to 0+} \|x\|$.

Tikhonov regularization with matrices

Consider the special case $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^m$ corresponding to the matrix equation $y = Ax$. The Tikhonov functional takes the special form

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

The minimizer can be found by solving the least squares problem

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

or, equivalently,

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

In MATLAB, this can be implemented simply as follows:

```
K = [A;sqrt(delta)*eye(n)];
z = [y; zeros(n,1)];
xdelta = K\z;
```

In Python, e.g., `scipy.linalg.lstsq` can be used to obtain the least squares solution. For sparse matrices, e.g.,

```
xdelta =
scipy.sparse.linalg.lsqr(A,y,damp=numpy.sqrt(delta))[0].
```

Numerical example: backward heat equation

Let us revisit the backward heat equation from earlier:

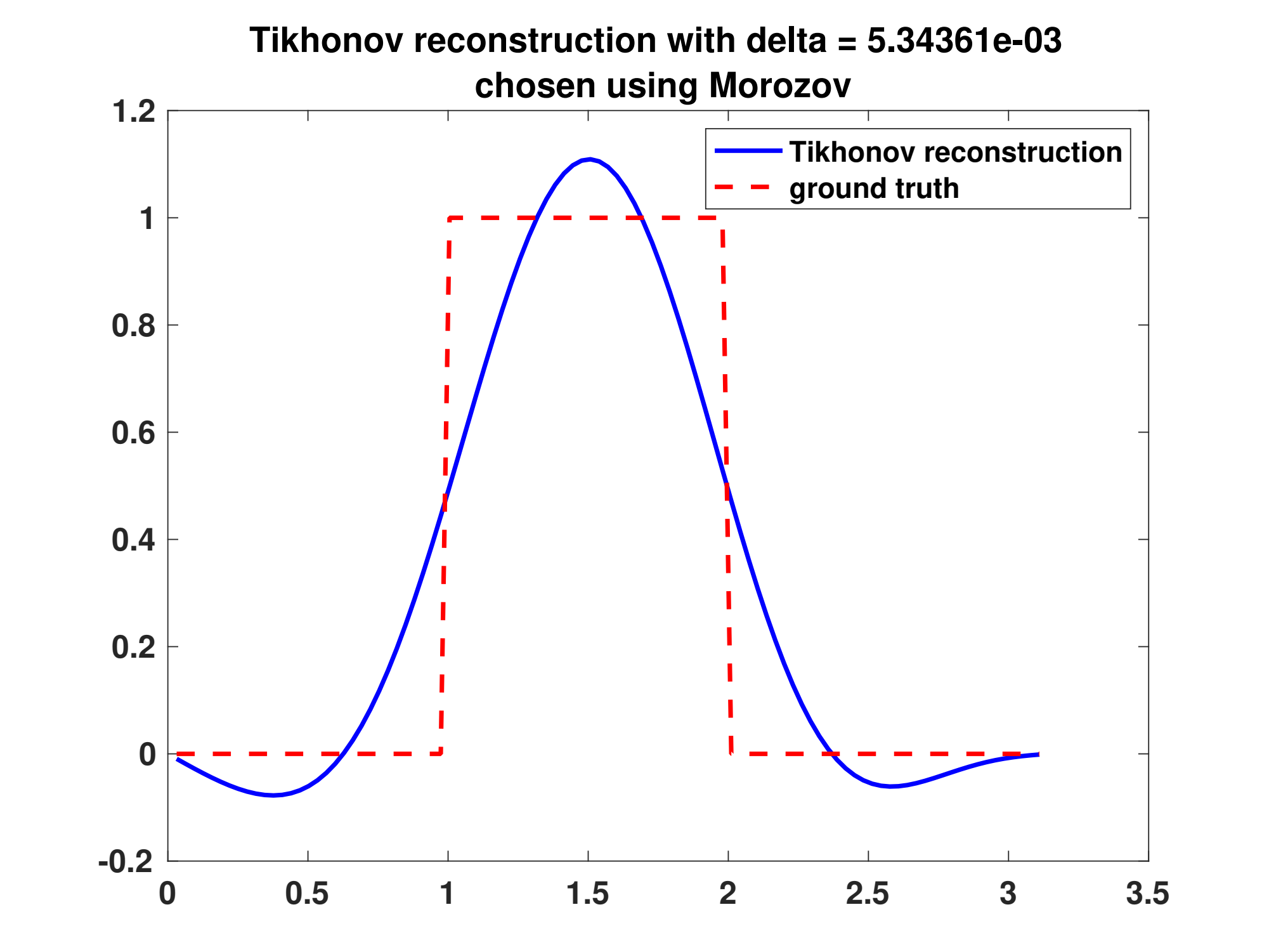
$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi), \end{cases}$$

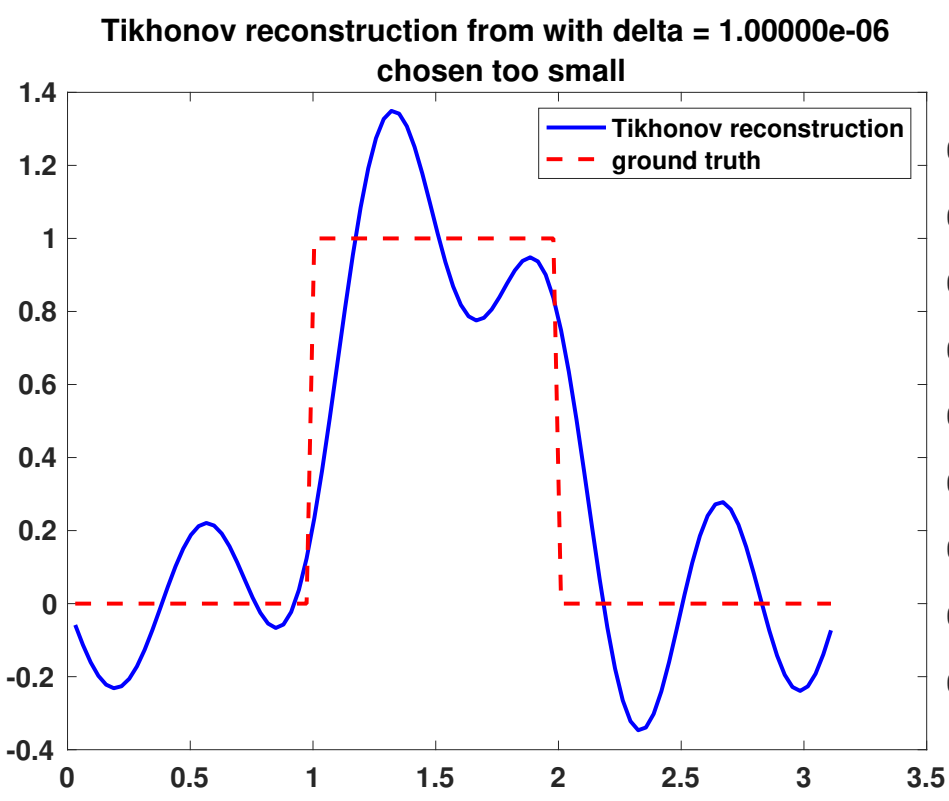
where $f : (0,\pi) \to \mathbb{R}$ is the initial heat distribution.

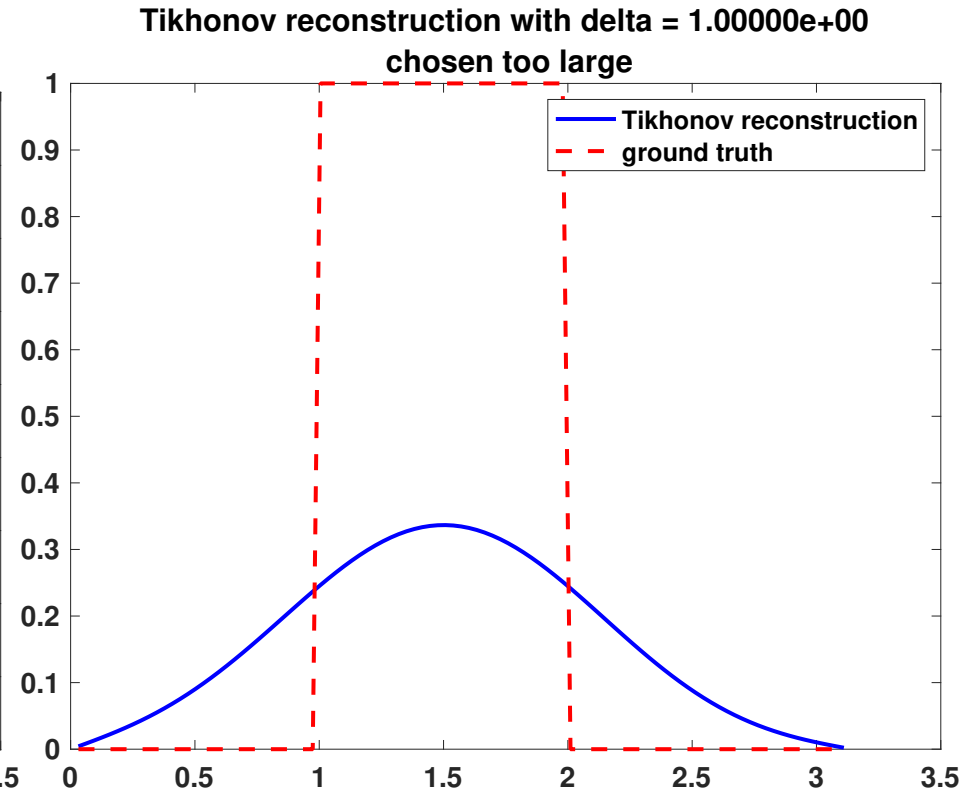
We reconstruct the initial state f based on noisy measurements of $u(\cdot, T)$ at time $T > 0$ using Tikhonov regularization.

We assume that the data $U(T) \in \mathbb{R}^{99}$ at time $T = 0.1$ is contaminated with mean-zero Gaussian noise with standard deviation 0.01, and that the discrepancy between the measured data and the underlying "exact" data equals the square root of the expected value of the squared norm of the noise vector, i.e.,

$$\varepsilon = \sqrt{99 \cdot 0.01^2} \approx 0.0995.$$


Tikhonov reconstruction with delta = 5.34361e-03
chosen using Morozov
Tikhonov reconstruction
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5





See the files `heateq_tikhonov.py` / `heateq_tikhonov.m` on the course webpage!

Tikhonov regularization for nonlinear problems

Unlike the TSVD, Tikhonov regularization can be generalized to nonlinear problems as well. Consider a *nonlinear* operator $A\colon H_1 \to H_2$ and the problem

$$y = A(x).$$

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

Definition

The function $A\colon H_1 \to H_2$ is called *Fréchet differentiable* at $x_0 \in H_1$ if there exists a continuous linear operator $A'_{x_0} : H_1 \to H_2$ such that

$$A(x+h) = A(x) + A'_{x_0} h + W_{x_0}(z),$$

where $\|W_{x_0}(z)\| \leq \varepsilon(x_0, z)\|z\|$ and the functional $z \mapsto \varepsilon(x_0, z)$ tends to zero as $z \to 0$.

The linear operator A'_{x_0} is called the *Fréchet derivative* of A at x_0.

We are interested in minimizing

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

Since F_δ is no longer quadratic, it is unclear whether a unique minimizer exists and typically the minimizer cannot be given by an explicit formula even it exists.

Let A be Fréchet differentiable. The linearization of A around a given point x_0 leads to the approximation of the functional F_δ,

$$\begin{aligned} F_\delta(x) \approx \widetilde{F}_\delta(x; x_0) &= \|A(x_0) + A'_{x_0}(x - x_0) - y\|^2 + \delta\|x\|^2 \\ &= \|A'_{x_0}(x) - g(y, x_0)\|^2 + \delta\|x\|^2, \end{aligned}$$

where $g(y, x_0) := y - A(x_0) + A'_{x_0}(x_0)$.

From the previous discussion on the linear case, we know that the minimizer of $\widetilde{F}_\delta(x; x_0)$ is given by

$$x = ((A'_{x_0})^* A_{x_0} + \delta I)^{-1} (A'_{x_0})^* g(y, x_0).$$

Minimization strategy with step size control

It may happen that the solution of the linearized problem does not reflect adequately the nonlinearities of the original function. A better strategy is to implement some form of step size control. For example, we might design the following iterative method.

1. Pick an initial guess x_0 and set $k = 0$.

   Repeat:

   2. Calculate the Fréchet derivative A'_{x_0}.
   3. Determine

      $$x = ((A'_{x_k})^* A'_{x_k} + \delta I)^{-1} (A'_{x_k})^* g(y, x_k), \quad g(y, x_k) = y - A(x_k) + A'_{x_k} x_k,$$

      and define $\Delta x = x - x_k$.
   4. Find step size $s > 0$ by minimizing the function

      $$f(s) = \|A(x_k + s\Delta x) - y\|^2 + \|x_k + s\Delta x\|^2.$$

   5. Set $x_{k+1} = x_k + s\Delta x$ and increase $k \leftarrow k + 1$.

   until convergence.

Remarks on nonlinear Tikhonov regularization

- In practice, evaluating A'_{x_k} is often the most difficult part.
- For finite-dimensional operators, the Fréchet derivative is simply the Jacobi matrix.
- Depending on the nature of the nonlinearity, one might also consider more "specialized" optimization methods (e.g., Gauss–Newton algorithm, Levenberg–Marquardt algorithm...).

More general penalty terms

A more general way of defining the Tikhonov functional is

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

where $G\colon H_1 \to \mathbb{R}_{\geq 0}$ 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 it.

One typical way of defining G is

$$G(x) = \|L(x - x_0)\|^2,$$

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 "feasible" solutions: Lx is some property that is known to be relatively close to the reference value Lx_0 for all reasonable solutions. (In the 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 $G(x) = \|L(x - x_0)\|^2$ is approximately as easy as for the standard penalty term.

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

$$F_\delta(x) = \left\| \begin{bmatrix} A \\ \sqrt{\delta}L \end{bmatrix} x - \begin{bmatrix} y \\ \sqrt{\delta}Lx_0 \end{bmatrix} \right\|^2.$$

Assuming that the singular values of K are bounded suitably far away from zero, the Tikhonov solution can be computed in MATLAB as

```
K = [A; sqrt(delta)*L];
z = [y; sqrt(delta)*L*x0];
xdelta = K\z;
```

In Python, e.g., `scipy.linalg.lstsq` can be used to solve the equivalent least squares problem $\begin{bmatrix} A \\ \sqrt{\delta}L \end{bmatrix} x = \begin{bmatrix} y \\ \sqrt{\delta}Lx_0 \end{bmatrix}$. For sparse matrices, e.g.,

```
K = scipy.sparse.vstack((A,np.sqrt(delta)*L))
z = np.hstack((y,np.sqrt(delta)*L@x0))
xdelta = scipy.sparse.linalg.lsqr(K,z)[0]
```

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Sixth lecture, May 22, 2023

Practical matters

- **Monday May 29** (next week) is a public holiday
  $\rightarrow$ **no lecture on May 29!**
- We will have a bonus live-coding lecture on **Tuesday May 30** about total variation regularization in place of the usual exercise session (this material will not be essential to the course).
- The deadline for the fifth exercise sheet will be moved to **Tuesday June 6**. Note that tomorrow's exercise session will happen as planned.

Regularization by truncated iterative methods

Regularization by truncated iterative methods

For simplicity, we will only consider the case when

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

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

- Iterative methods attempt to solve (1) by finding successive approximations for the solution starting from some initial guess.
- Typically, the computation of such iterations involves multiplications by A and its adjoint, but not explicit computation of inverse operators. (Direct methods, such as *Gaussian elimination*, produce 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 (e.g., in the order of millions), in which case direct methods are prohibitively expensive. Iterations are especially useful if multiplications by A are cheap: for example, if A is sparse or it contains some other structure (e.g., it is a multi-diagonal matrix arising from finite difference or finite element approximation of an elliptic PDE).

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

Banach fixed point iteration

Let E be a Banach space and $S \subset E$. Consider a mapping, not necessarily linear, $T : E \to E$. We say that S is an *invariant set* for T if $T(S) \subset S$, that is,

$$T(x) \in S \quad \text{for all } x \in S.$$

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

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

Finally, a vector $x \in E$ is called a *fixed point* of T if

$$T(x) = x.$$

Theorem (Banach fixed point theorem)

Let E be a Banach space and $S \subset E$ a closed invariant set for the (possibly nonlinear) mapping $T : E \to E$. Assume further that T is a contraction in S. Then there exists a unique fixed point $x \in S$ such that $T(x) = x$. Furthermore, this fixed point can be found by the fixed point iteration

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

for any $x_0 \in S$.

Proof. Let $T : E \to E$ be a mapping, $S \subset E$ a closed invariant set such that $T(S) \subset S$, and let T be a contraction in S,

$$\|T(x) - T(y)\| \leq \kappa \|x - y\| \quad \text{for all } x, y \in S,$$

with $\kappa < 1$. For all $j > 1$, we have

$$\|x_{j+1} - x_j\| = \|T(x_j) - T(x_{j-1})\| \leq \kappa \|x_j - x_{j-1}\|.$$

Inductively, it follows that

$$\|x_{j+1} - x_j\| \leq \kappa^{j-1} \|x_2 - x_1\|.$$

For any $n, k \in \mathbb{N}$, we have

$$\begin{aligned}
\|x_n - x_k\| &\leq \sum_{j=1}^{\max\{n,k\}-\min\{n,k\}} \|x_{\min\{n,k\}+j} - x_{\min\{n,k\}+j-1}\| \\
&\leq \sum_{j=1}^{\max\{n,k\}-\min\{n,k\}} \kappa^{\min\{n,k\}+j-2}\|x_2 - x_1\| \\
&\leq \frac{\kappa^{\min\{n,k\}-1}}{1-\kappa}\|x_2 - x_1\| \xrightarrow{n,k\to\infty} 0,
\end{aligned}$$

where we used the formula for the geometric series. Therefore (x_j) is a Cauchy sequence and thus convergent (since E is a Banach space and thus complete). The limit is in S since S is closed.

Finally, as a contraction, T is (Lipschitz) continuous and we have that

$$x = \lim_{k\to\infty} x_k = \lim_{k\to\infty} T(x_{k-1}) = T\Big(\lim_{k\to\infty} x_{k-1}\Big) = T(x),$$

as desired. □

Landweber–Fridman iteration

Landweber–Fridman iteration

Instead of considering the original equation

$$Ax = y,$$

let us consider the normal equation

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

Recall that $x \in \mathbb{R}^n$ satisfies the normal equation iff it minimizes the residual

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

Moreover, there exists a unique element of $\mathbb{R}^n$, given by $x^\dagger := A^\dagger y$, which satisfies the normal equation and $x^\dagger \in \mathrm{Ker}(A)^\perp$ (the minimum norm solution).

Let us define the affine mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ by

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

Note that any solution of the normal equation $A^{\mathrm{T}}Ax = A^{\mathrm{T}}y$ is a fixed point of T.

If β is small enough, then there is only one fixed point of T in $\mathrm{Ker}(A)^{\perp}$, precisely $x^{\dagger}$, and it can be reached by the fixed point iteration if $x_0 = 0$.

Theorem

Let λ_1 be the largest singular value of matrix A and let $0 < \beta < 2/\lambda_1^2$ be fixed. Then the fixed point iteration

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

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

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

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

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

Recall that A and its transpose can be written using the SVD of A as

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

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

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

Let $x, z \in S$. Then $x - z \in S$ and

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

Since λ_1 is the largest singular value, it follows that

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

Hence

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

In consequence,

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

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

Finally, recall that $x^\dagger = A^\dagger y$ belongs to $S = \mathrm{Ker}(A)^\perp$ and it satisfies the normal equation. Since $x_0 = 0$ is in S (it is orthogonal to all vectors), the fixed point iteration starting from x_0 converges to $x^\dagger$. □

Regularization properties of Landweber–Fridman

In what follows, we will assume that $0 < \beta < 2/\lambda_1^2$.

In the exercises, it will be shown that the k^{th} iterate of the Landweber–Fridman iteration can be written explicitly as

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

Since we assumed $|1 - \beta\lambda_j^2| < 1$, then

$$(1 - \beta\lambda_j^2)^k \xrightarrow{k\to\infty} 0.$$

This is what one would expect since

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

While $k \in \mathbb{N}$ is finite, the coefficients appearing in the series representation

$$x_k = \sum_{j=1}^{p} \frac{1}{\lambda_j}(1 - (1 - \beta\lambda_j^2)^k)(u_j^{\mathrm{T}} y)v_j \tag{2}$$

satisfy

$$\frac{1}{\lambda_j}(1 - (1 - \beta\lambda_j^2)^k) = \frac{1}{\lambda_j}\left(1 - \sum_{\ell=0}^{k} \binom{k}{\ell}(-1)^\ell \beta^\ell \lambda_j^{2\ell}\right)$$

$$= \frac{1}{\lambda_j} \sum_{\ell=1}^{k} \binom{k}{\ell}(-1)^{\ell+1} \beta^\ell \lambda_j^{2\ell} = \sum_{\ell=1}^{k} \binom{k}{\ell}(-1)^{\ell+1} \beta^\ell \lambda_j^{2\ell-1},$$

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

In consequence, while k is "small enough", no coefficient of $(u_j^{\mathrm{T}} y)v_j$ in (2) is so large that the component of the measurement noise in the direction u_j is amplified in an uncontrolled manner. (Compare with Tikhonov regularization, where the corresponding coefficients are $\lambda_j/(\lambda_j^2 + \delta)$.)

Discrepancy principle for Landweber–Fridman iteration

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

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

The Morozov discrepancy principle for the Landweber–Fridman iteration is analogous to the truncated SVD: choose the smallest $k \geq 0$ such that the residual satisfies

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

Q: *When does an index $k \geq 1$ satisfying $\|y - Ax_k\| \leq \varepsilon$ exist?*
A: When $\varepsilon > \|Py - y\| = \|y - A(A^\dagger y)\| = \|y - Ax^\dagger\|$, where $P = AA^\dagger$ is the orthogonal projection onto $\mathrm{Ran}(A)$ (cf. 3^{rd} exercises) and $x^\dagger = A^\dagger y$ is the minimum norm solution. Since the sequence $(x_k)_{k=0}^\infty$ converges to $x^\dagger$, for any $\varepsilon > \|y - Ax^\dagger\|$, there exists $k = k_\varepsilon \in \mathbb{N}$ such that

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

By the reverse triangle inequality

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

From this, we deduce that $\|y - Ax_k\| \leq \varepsilon$ as desired.

Conjugate gradient method

Krylov subspace methods

Krylov subspace methods are iterative solvers for (large scale) matrix equations of the form $Ax = y$, $A \in \mathbb{R}^{n \times n}$. In general terms, the solution vector $x \in \mathbb{R}^n$ is approximated as a linear combination of vectors of the form u, Au, A^2u, ..., with some given $u \in \mathbb{R}^n$. If multiplication by A is cheap – for example, when A is sparse – Krylov subspace methods can be particularly efficient.

We consider only the most well-known Krylov subspace method, the conjugate gradient method. It is worth mentioning that other methods in this class include, e.g., the generalized minimum residual method (GMRES) and the biconjugate gradient method (BiCG).

Assumptions on A and A-dependent inner product

In what follows, we assume that the system matrix $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite:

$$A^{\mathrm{T}} = A \quad \text{and} \quad u^{\mathrm{T}} A u > 0 \quad \text{for all } u \in \mathbb{R}^n \setminus \{0\}.$$

Note that this implies that A is injective.[†] By the fundamental theorem of linear algebra, A is invertible. Furthermore, the inverse $A^{-1} \in \mathbb{R}^{n \times n}$ is also symmetric and positive definite.

We define

$$\langle u, v \rangle_A := u^{\mathrm{T}} A v \quad \text{and} \quad \|u\|_A := \sqrt{\langle u, u \rangle_A}.$$

Since A was assumed to be symmetric and positive definite, it is straightforward to check that $\langle \cdot, \cdot \rangle_A \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defines an inner product on $\mathbb{R}^n$. In consequence, $\| \cdot \|_A \colon \mathbb{R}^n \to \mathbb{R}$ is a norm.

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

$$\langle s_i, s_j \rangle_A = s_i^{\mathrm{T}} A s_j = 0 \quad \text{whenever } i \neq j,$$

i.e., they are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_A$.

[†] $Ax = Ay \Rightarrow A(x - y) = 0 \Rightarrow (x - y)^{\mathrm{T}} A (x - y) = 0 \Rightarrow x - y = 0.$

Error, residual, and minimization problem

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

$$Ax = y$$

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

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

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

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

Since $\|\cdot\|_A$ is a norm, $\phi(x) \geq 0$ for all $x \in \mathbb{R}^n$ and

$$\phi(x) = 0 \quad \Leftrightarrow \quad e = 0 \quad \Leftrightarrow \quad x = x_*.$$

Minimizing ϕ is equivalent to solving $Ax = y$.

The conjugate gradient method is an iterative scheme which, at each step of the iteration, returns $x_{k+1} = \arg\min_{x \in \mathcal{S}_k} \phi(x)$, where

$$\mathcal{S}_k := \{x \in \mathbb{R}^n \mid x = x_0 + c_0 s_0 + \cdots + c_k s_k,\ c_0, \ldots, c_k \in \mathbb{R}\}$$

is a hyperplane determined by a sequence of vectors $s_0, \ldots, s_k \in \mathbb{R}^n$.

Starting from an initial guess $x_0 \in \mathbb{R}^n$, the successive iterates are given by

$$x_{k+1} = x_k + \alpha_k s_k, \quad k = 0, 1, 2, \ldots.$$

Define the *residual* $r_k = y - Ax_k$ corresponding to iterate x_k and let $s_0 = r_0$ be the *initial search direction*. Then the parameters are

$$\alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k} \quad \text{for } k \geq 0, \qquad (\text{“step size”})$$

$$s_k = r_k + \beta_{k-1} s_{k-1},\ \beta_{k-1} = -\frac{s_{k-1}^{\mathrm{T}} A r_k}{s_{k-1}^{\mathrm{T}} A s_{k-1}} \text{ for } k \geq 1. \ (\text{“search direction”})$$

We proceed to show that the search directions defined by the above recursion are A-conjugate (and thus linearly independent) and the iterates x_{k+1} obtained using this algorithm are minimizers of the functional $\phi(x)$ on the hyperplanes $\mathcal{S}_k$. Note especially that $\mathcal{S}_{n-1} = \mathbb{R}^n$, so an exact solution (up to rounding errors) is achieved in at most n iteration steps.

Step 1: If $s_0, \ldots, s_k$ are A-conjugate, then $r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\}$.

Now $x_{k+1} = x_k + \alpha_k s_k = x_{k-1} + \alpha_{k-1} s_{k-1} + \alpha_k s_k = \cdots = x_0 + \sum_{j=0}^{k} \alpha_j s_j$ and $r_{k+1} = y - Ax_{k+1} = y - Ax_0 - \sum_{j=0}^{k} \alpha_j A s_j = r_0 - \sum_{j=0}^{k} \alpha_j A s_j$.
Let $\ell \in \{0, \ldots, k\}$. Then

$$
\begin{aligned}
r_{k+1}^{\mathrm{T}} s_\ell &= \left(r_0 - \sum_{j=0}^{k} \alpha_j A s_j \right)^{\mathrm{T}} s_\ell && (A^{\mathrm{T}} = A) \\
&= r_0^{\mathrm{T}} s_\ell - \sum_{j=0}^{k} \alpha_j s_j^{\mathrm{T}} A s_\ell && (s_j^{\mathrm{T}} A s_\ell = 0 \text{ for } j \neq \ell) \\
&= r_0^{\mathrm{T}} s_\ell - \alpha_\ell s_\ell^{\mathrm{T}} A s_\ell && (\alpha_\ell = \tfrac{s_\ell^{\mathrm{T}} r_\ell}{s_\ell^{\mathrm{T}} A s_\ell}) \\
&= r_0^{\mathrm{T}} s_\ell - s_\ell^{\mathrm{T}} r_\ell && (r_\ell = r_0 - \textstyle\sum_{j=0}^{\ell-1} \alpha_j A s_j) \\
&= r_0^{\mathrm{T}} s_\ell - s_\ell^{\mathrm{T}} r_0 + \sum_{j=0}^{\ell-1} \alpha_j s_\ell^{\mathrm{T}} A s_j \\
&= 0,
\end{aligned}
$$

as desired.

Step 2: $s_0, \ldots, s_k$ are A-conjugate and linearly independent.

By induction with respect to $k \in \mathbb{N}_0$. If $k = 0$, then $\{s_0\}$ is trivially A-conjugate. Suppose that the claim has been proved for some $k \in \mathbb{N}_0$; we show that $s_{k+1}^{\mathrm{T}} A s_j = 0$ for all $j \in \{0, \ldots, k\}$.
Let $j \in \{0, \ldots, k\}$. Then

$$s_{k+1}^{\mathrm{T}} A s_j = (r_{k+1} + \beta_k s_k)^{\mathrm{T}} A s_j = r_{k+1}^{\mathrm{T}} A s_j + \beta_k s_k^{\mathrm{T}} A s_j.$$

If $0 \leq j \leq k-1$, then the above expression vanishes by the previous slide and the induction hypothesis. Let $j = k$. Then

$$\begin{aligned} s_{k+1}^{\mathrm{T}} A s_k &= r_{k+1}^{\mathrm{T}} A s_k + \beta_k s_k^{\mathrm{T}} A s_k && \left(\beta_k = -\frac{s_k^{\mathrm{T}} A r_{k+1}}{s_k^{\mathrm{T}} A s_k}\right) \\ &= 0, \end{aligned}$$

as desired. For the linear dependence, write $c_0 s_0 + \cdots + c_k s_k = 0$ for some undetermined coefficients $c_0, \ldots, c_k \in \mathbb{R}$. For any $\ell \in \{0, \ldots, k\}$, multiplying from the left by $s_\ell^{\mathrm{T}} A$ yields

$$c_0 s_\ell^{\mathrm{T}} A s_0 + \cdots + c_k s_\ell^{\mathrm{T}} A s_k = 0 \Rightarrow c_\ell s_\ell^{\mathrm{T}} A s_\ell = 0 \overset{\substack{x^{\mathrm{T}} A x = 0 \\ \text{iff } x = 0}}{\Rightarrow} c_\ell = 0$$

as desired.

Step 3: $h_* = \underset{h \in \mathbb{R}^{k+1}}{\arg\min} \phi(x_0 + S_k h)$ iff $h_* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0$, where $x_0 \in \mathbb{R}^n$, $r_0 = y - A x_0$, $S_k = [s_0, \ldots, s_k]$, and $s_0, \ldots, s_k \in \mathbb{R}^n$ are lin. independent.

We first verify that the expression $(S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0$ is well-defined by showing that $S_k^{\mathrm{T}} A S_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is invertible. By the positive definiteness of A,

$$S_k^{\mathrm{T}} A S_k z = 0 \quad \Rightarrow z^{\mathrm{T}} S_k^{\mathrm{T}} A S_k z = 0 \quad \Rightarrow \|S_k z\|_A^2 = 0 \quad \Rightarrow S_k z = 0,$$

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

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

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

thus (recall that $\phi(x) = r^{\mathrm{T}} A^{-1} r$ for $r = y - Ax$)

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

We obtained

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

The Hessian of $h \mapsto \phi(x_0 + S_k h)$ is $2 S_k^{\mathrm{T}} A S_k$, which is positive definite since

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

where equality holds iff $S_k u = 0 \Leftrightarrow u = 0$. Hence $h \mapsto \phi(x_0 + S_k h)$ is convex, and we can find its unique minimizer by solving the zero point of its gradient:

$$\begin{aligned} 0 = \nabla_h \phi(x_0 + S_k h) &= 2 S_k^{\mathrm{T}} A S_k h - 2 S_k^{\mathrm{T}} r_0 \\ \Leftrightarrow \quad h &= (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0. \end{aligned}$$

Step 4: Let $x_0 \in \mathbb{R}^n$ be the initial guess and $S_k = [s_0, \ldots, s_k]$, where $s_0, \ldots, s_k \in \mathbb{R}^n$ are the conjugate gradient search directions. The conjugate gradient iterates satisfy $x_{k+1} = \arg\min_{h \in \mathbb{R}^{k+1}} \phi(x_0 + S_k h)$.

Let $a_j = (\alpha_0, \ldots, \alpha_j)^{\mathrm{T}} \in \mathbb{R}^{j+1}$, where $\alpha_i = \frac{s_i^{\mathrm{T}} r_i}{s_i^{\mathrm{T}} A s_i}$ are the line search parameters of the conjugate gradient method. Then

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

The residual corresponding to x_j is

$$r_j = y - A x_j = (y - A x_0) - A S_{j-1} a_{j-1} = r_0 - A S_{j-1} a_{j-1}$$

and hence

$$s_j^{\mathrm{T}} r_j = s_j^{\mathrm{T}} r_0 - s_j^{\mathrm{T}} A S_{j-1} a_{j-1} = s_j^{\mathrm{T}} r_0 - \underbrace{s_j^{\mathrm{T}} [A s_0, \ldots, A s_{j-1}]}_{=0} a_{j-1},$$

since $s_j^{\mathrm{T}} A s_i = 0$, $i < j$, due to A-conjugacy. Therefore

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

The line search parameters can be written as

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

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

$$\begin{aligned}(S_k^{\mathrm{T}} A S_k)^{-1} &= \mathrm{diag}(s_0^{\mathrm{T}} A s_0, \ldots, s_k^{\mathrm{T}} A s_k)^{-1} \\ &= \mathrm{diag}\left(\frac{1}{s_0^{\mathrm{T}} A s_0}, \ldots, \frac{1}{s_k^{\mathrm{T}} A s_k}\right).\end{aligned}$$

Especially, this means that the minimizer h^* of $\phi(x_0 + S_k h)$ over the hyperplane $\mathcal{S}_k$ is given by

$$h^* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0 = \mathrm{diag}\left(\frac{1}{s_0^{\mathrm{T}} A s_0}, \ldots, \frac{1}{s_k^{\mathrm{T}} A s_k}\right) \begin{bmatrix} s_0^{\mathrm{T}} r_0 \\ \vdots \\ s_k^{\mathrm{T}} r_0 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{bmatrix} = a_k.$$

In consequence, $x_{k+1} = x_0 + S_k a_k = x_0 + S_k h_*$.

Remark. In the conjugate gradient method, the search directions are given by $s_0 = r_0$ and

$$s_k = r_k + \beta_{k-1} s_{k-1}, \quad k \geq 1,$$

where $r_k = y - Ax_k$. Note that $\operatorname{span}\{s_0, \ldots, s_k\} = \operatorname{span}\{r_0, \ldots, r_k\}$.

Especially, the conjugate gradient iterate x_{k+1} satisfies

$$\begin{aligned} x_{k+1} = \operatorname*{arg\,min}_{x \in x_0 + \operatorname{span}\{s_0, \ldots, s_k\}} \|x - x_*\|_A^2 &= \operatorname*{arg\,min}_{x \in x_0 + \operatorname{span}\{r_0, \ldots, r_k\}} \|x - x_*\|_A^2 \\ &= \operatorname*{arg\,min}_{x \in x_0 + \mathcal{K}_k} \|x - x_*\|_A^2, \end{aligned}$$

where the *search space* $\mathcal{K}_k := \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\}$ is precisely the k^{th} Krylov subspace of A with the initial vector $r_0 = y - Ax_0$. Some basic properties of Krylov subspaces:

- $A(\mathcal{K}_k) \subset \mathcal{K}_{k+1}$.
- $\mathcal{K}_{k-1} \subset K_k$ (Krylov subspaces are nested).
- $\dim \mathcal{K}_k \leq k$ (dimension of the k^{th} Krylov subspace is at most k).
- $\dim \mathcal{K}_k \leq \dim \mathcal{K}_{k-1} + 1$ (dimension of the successive Krylov space is at most one higher than that of the former).

The conjugate gradient algorithm is usually presented in slightly different form. Assuming that the iteration has not yet converged at the iterate x_k, we can deduce the following formulae for $\alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k}$ and $\beta_k = -\frac{s_k^{\mathrm{T}} A r_{k+1}}{s_k^{\mathrm{T}} A s_k}$.

Simplifying α_k: Since $r_k \perp s_{k-1}$, we have that

$$s_k^{\mathrm{T}} r_k = (r_k + \beta_{k-1} s_{k-1})^{\mathrm{T}} r_k = \|r_k\|^2 \quad \Rightarrow \quad \alpha_k = \frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k}. \tag{3}$$

Simplifying β_k: since $r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\} \ni r_k$ and $r_{k+1} = r_k - \alpha_k A s_k$, then

$$\|r_{k+1}\|^2 = r_{k+1}^{\mathrm{T}}(r_k - \alpha_k A s_k) \overset{(3)}{=} -\frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k} r_{k+1}^{\mathrm{T}} A s_k = \beta_k \|r_k\|^2$$

and thus

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

This leads to the "standard form" of the method.

Pseudocode for the conjugate gradient algorithm

Given: symmetric, positive definite system matrix $A \in \mathbb{R}^{n \times n}$, data $y \in \mathbb{R}^n$.

1. Choose initial guess $x_0 \in \mathbb{R}^n$.
2. Set $k = 0$, $r_0 = y - Ax_0$, $s_0 = r_0$;

   Repeat until the chosen stopping rule is satisfied:

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

   end

Numerical example

Let us consider minimization with the *steepest descent* directions

$$s_k = -\nabla\phi(x_k) = 2(y - Ax_k), \quad k = 0, 1, \ldots . \tag{4}$$

In general, the convergence of the sequence $\{x_k\}$ toward the global minimizer $x_* = A^{-1}y$ can be fairly slow. We demonstrate this with the following example.

Let

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

Now

$$\phi(x) = x_1^2 + 5x_2^2.$$

We plot the level contours of ϕ and the sequence $\{x_k\}_{k=0}^5$ starting from $x_0 = (1, 0.3)^{\mathrm{T}}$. The true solution $x_* = (0, 0)^{\mathrm{T}}$ is marked with a blue cross.

We also illustrate minimization over the hyperplanes $\mathcal{S}_0$ and $\mathcal{S}_1$, i.e., $x_0 + S_0 h_*$ and $x_0 + S_1 h_*$ with $S_0 = [s_0] \in \mathbb{R}^{2\times 1}$ and $S_1 = [s_0, s_1] \in \mathbb{R}^{2\times 2}$, where s_0 and s_1 were computed using the sequential method (4).

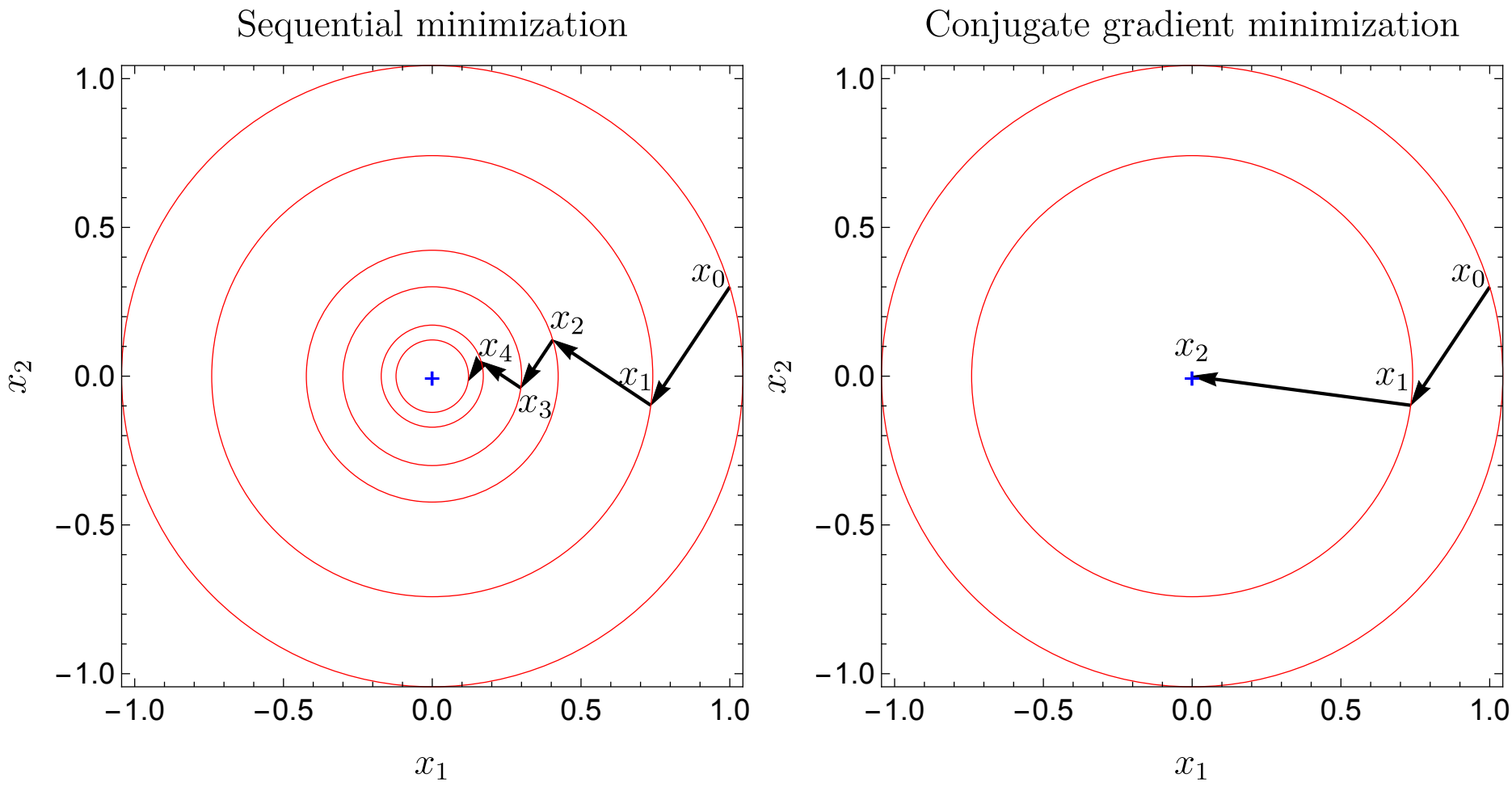


Figure: Left: Minimization using steepest descent search directions $s_k = -\nabla\phi(x_k)$. Right: In the linear case, the conjugate gradient method iteratively finds the optima over the hyperplanes $\mathcal{S}_1$ and $\mathcal{S}_2$. The CG method converges to the actual solution $x_* = (0, 0)^{\mathrm{T}}$ (marked with a blue cross) in $n = 2$ iterations (which equals the dimensionality of the ambient space $\mathbb{R}^2$).

Conjugate gradient method for inverse problems

According to the previous construction, if the conjugate gradient method is applied to the equation

$$Ax = y,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, an exact solution (up to rounding errors) is achieved in at most n iteration steps, i.e., $x_n = x_* = A^{-1}y$. However, the algorithm typically converges satisfactorily much quicker. A (pessimistic) convergence rate is proved in the first exercise of week 4.

With ill-posed problems, one should be more cautious and terminate the iterations well before convergence to avoid fitting the solution to noise. In fact, since the conjugate gradient method often converges very fast, one should be extremely cautious.

Let us 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.

- If $m = n$ and there is some available prior information suggesting that A is, at least in theory, positive (semi-)definite, one can apply the conjugate gradient algorithm directly on the original equation.
- More generally, one may still consider the normal equation

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

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

The system matrix $A^{\mathrm{T}}A = (A^{\mathrm{T}}A)^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ is symmetric and

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

Thus 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 $A^{\mathrm{T}}y$.[†]

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

$$\|y - Ax_k\| \leq \varepsilon$$

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

[†]Small remark on implementation: matrix-matrix products are typically far more expensive to compute than matrix-vector products. For example, instead of computing expressions like `residual = A'*y - A'*A*x0` when implementing the conjugate gradient method in MATLAB, one should use parentheses to parse the computation like `residual = A'*y - A'*(A*x0)`. Similarly `residual = A.T@y - A.T@(A@x0)` in Python.

Numerical example: backward heat equation revisited

Let us revisit the backward heat equation:

$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi), \end{cases}$$

where $f : (0,\pi) \to \mathbb{R}$ is the initial heat distribution.

Inverse problem: Reconstruct the initial state f based on noisy measurements of $u(\cdot, T)$ at time $T > 0$.

Let $x_j = jh$, $j = 0, \ldots, 100$ with $h = \pi/100$, and denote $U(t) = (U_j(t))_{j=1}^{99}$ and $F = (f(x_j))_{j=1}^{99}$. At time $t = T > 0$, the discretized heat distribution $U := U(T)$ is given by

$$U = AF,$$

where $A = \mathrm{e}^{TB} \in \mathbb{R}^{99\times 99}$ and $B = h^{-2}\mathrm{tridiag}(1,-2,1) \in \mathbb{R}^{99\times 99}$.

As ground truth, we take

$$f(x) = \begin{cases} 1 & \text{if } x \in [1,2], \\ 0 & \text{if } x \in (0,1) \cup (2,\pi). \end{cases}$$

We assume that the simulated data $U = U(T) \in \mathbb{R}^{99}$ at time $T = 0.1$ is contaminated with mean-zero Gaussian noise with standard deviation 0.01, and that the discrepancy between the measured data and the underlying "exact" data equals the square root of the expected value of the squared norm of the noise vector, i.e.,

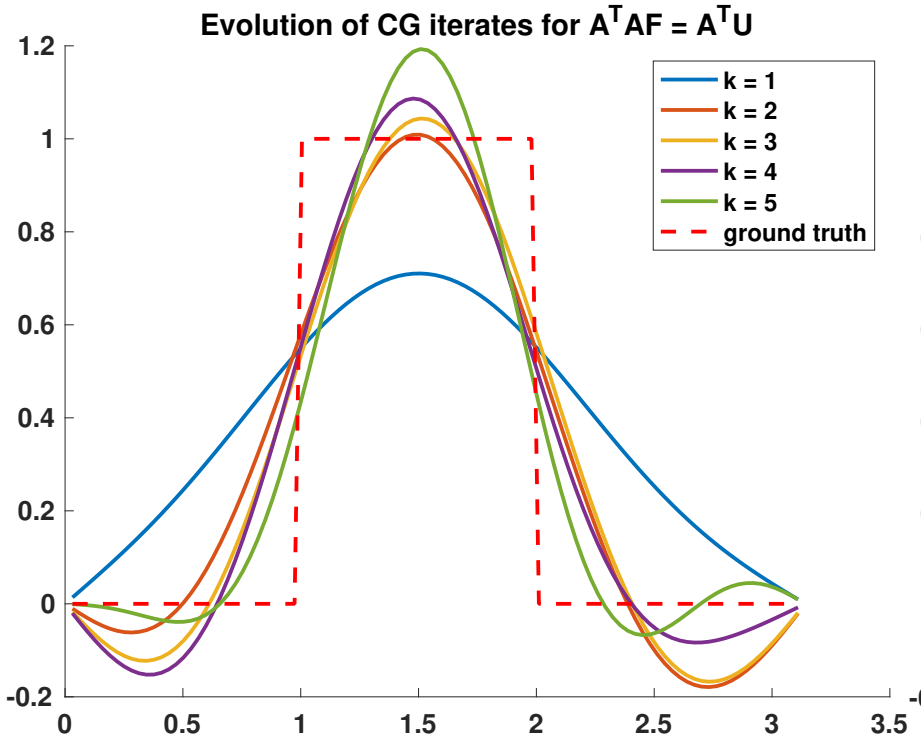
$$\varepsilon = \sqrt{99 \cdot 0.01^2} \approx 0.0995.$$

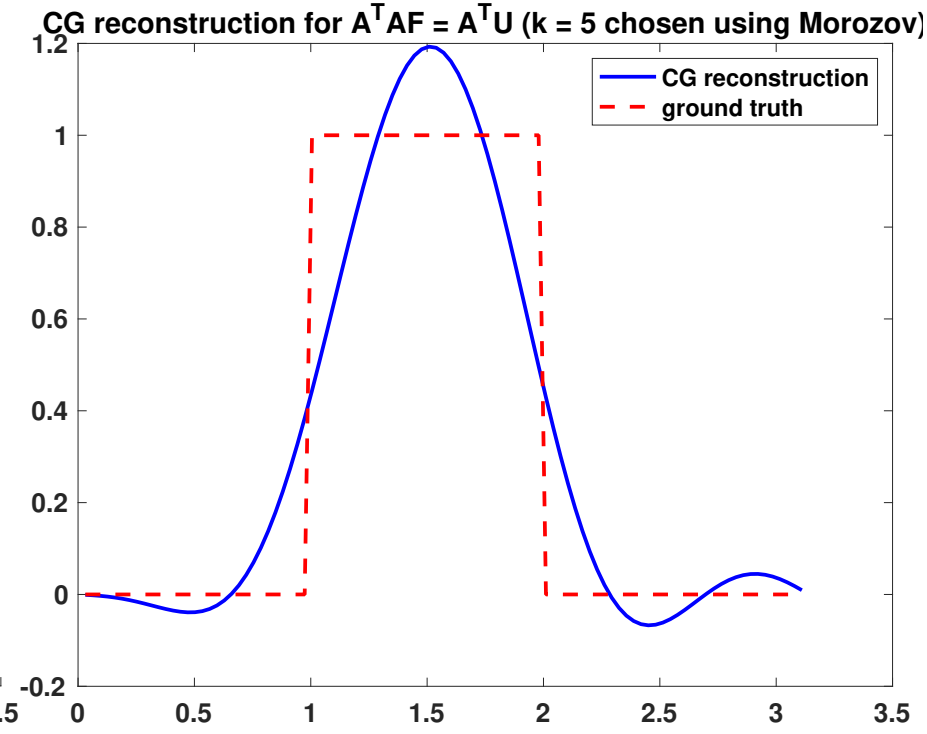
We use the conjugate gradient method to solve the normal equation

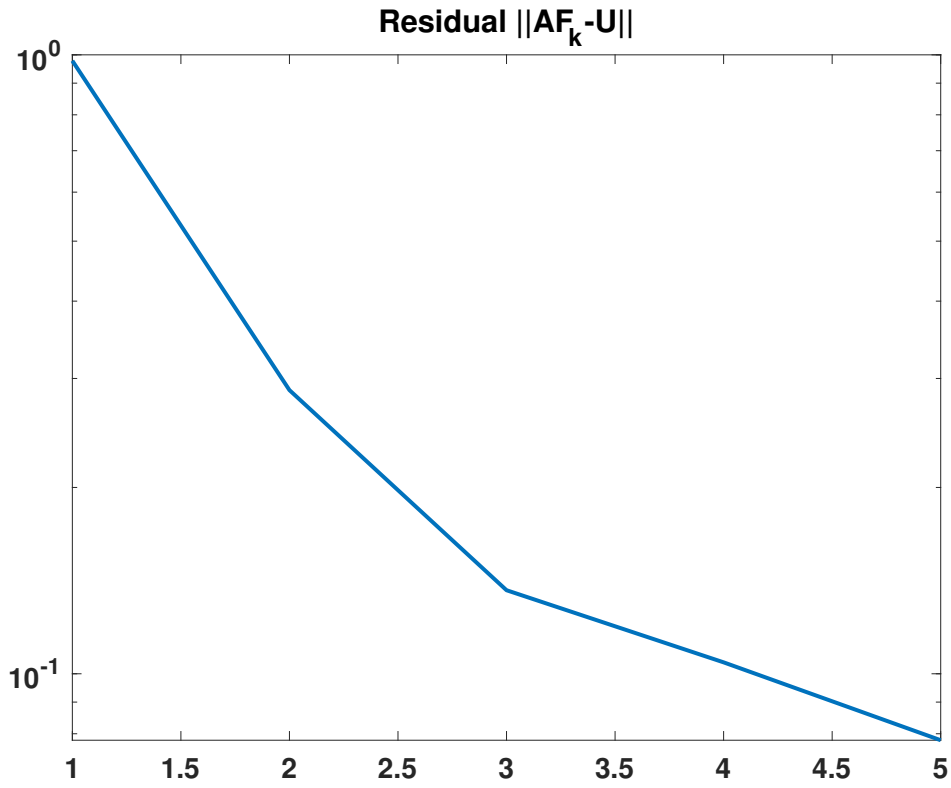
$$A^{\mathrm{T}} A F = A^{\mathrm{T}} U,$$

and terminate the algorithm for the first CG iterate F_k such that

$$\|AF_k - U\| \leq \varepsilon.$$


Evolution of CG iterates for A^TAF = A^TU
k = 1
k = 2
k = 3
k = 4
k = 5
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5


CG reconstruction for A^TAF = A^TU (k = 5 chosen using Morozov)
CG reconstruction
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5


Residual ||AF_k-U||
10^0
10^-1
1
1.5
2
2.5
3
3.5
4
4.5
5

Although we have simply scratched the surface by covering some of the basic ideas surrounding the conjugate gradient scheme and demonstrating how an "early stopping rule" can provide reasonable solutions for inverse problems, the regularizing properties of the conjugate gradient method have been analyzed more explicitly in the literature. A classic textbook specifically about this subject is:

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

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Seventh lecture, May 30, 2023

Total variation regularization for X-ray tomography

Some helpful resources on the Chambolle–Pock algorithm:


A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40**:120-145, 2011.


L. Condat. A generic proximal algorithm for convex optimization – application to total variation minimization. *IEEE Signal Proc. Letters* **21**(8):985–989, 2014.


E. Y. Sidky, J. H. Jørgensen, and X. Pan. Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle-Pock algorithm. *Phys. Med. Biol.* **57**:3065–3091, 2012.


Operator Discretization Library. `https://odl.readthedocs.io/math/solvers/nonsmooth/chambolle_pock.html`, 2017.


PORTAL. portal.readthedocs.io/en/latest/chambollepock.html, written by P. Paleo, 2015.

Additional resources on total variation regularization for X-ray tomography:


J. L. Mueller and S. Siltanen. Linear and Nonlinear Inverse Problems with Practical Applications. 2012.


S. Siltanen. Total variation regularization for X-ray tomography. FIPS Computational Blog, https://blog.fips.fi/tomography/x-ray/total-variation-regularization-for-x-ray-tomography/, 2017.

Recall that the discrete measurement model for X-ray tomography can be expressed as

$$y = Ax.$$

This time, we consider solving the inverse problem of recovering x based on noisy measurements y.

We are interested in *anisotropic total variation regularization*

$$\arg\min_{x \geq 0} \left\{ \tfrac{1}{2} \|y - Ax\|^2 + \lambda \|Dx\|_1 \right\}, \quad \lambda > 0,$$

where $\|x\|_1 = \sum_i |x_i|$, $D = \begin{bmatrix} L_H \\ L_V \end{bmatrix}$ is the discretized (image) gradient operator,

$$\|Dx\|_1 = \sum_j |(Dx)_j| = \sum_j |(L_H x)_j| + \sum_j |(L_V x)_j|,$$

and L_H and L_V denote the horizontal and vertical (image) finite difference matrices, respectively.

Special feature: TV regularization preserves sharp edges.

x_{90}	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	x_{99}
x_{80}	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	x_{88}	x_{89}
x_{70}	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	x_{77}	x_{78}	x_{79}
x_{60}	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}	x_{67}	x_{68}	x_{69}
x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

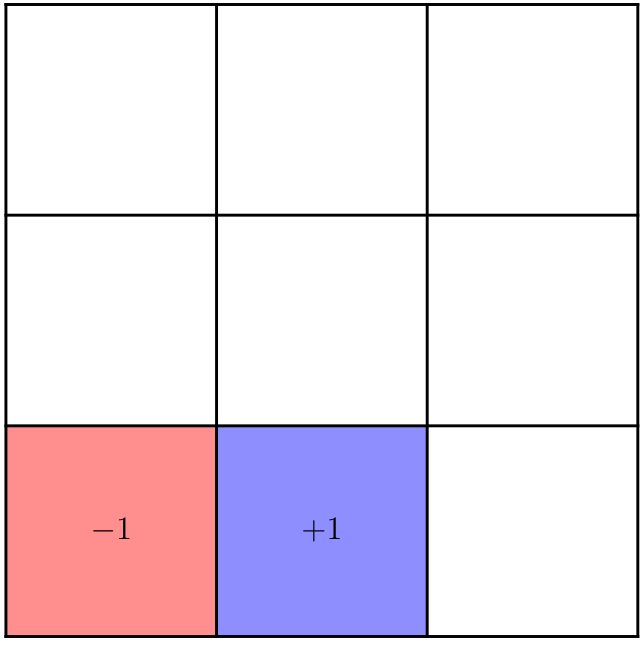
Recall that the vector x is related to the density matrix $(f_{i,j})$ of the computational domain via

$$x_{in+j} = f_{i,j}, \quad i,j \in \{0,\ldots,n-1\}.$$

`x = f.reshape((n*n,1))` and `f = x.reshape((n,n))` (Python)

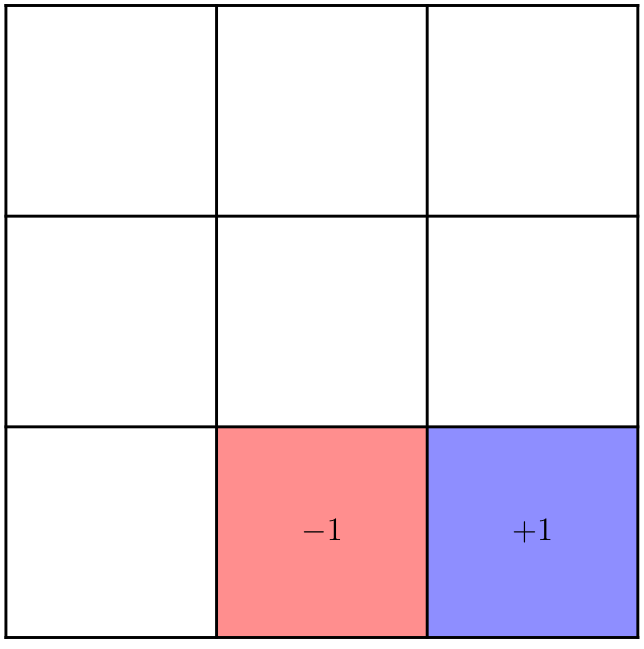
`x = f(:)` and `f = reshape(x,n,n)` (MATLAB)

Construction of L_H (periodic boundary conditions)



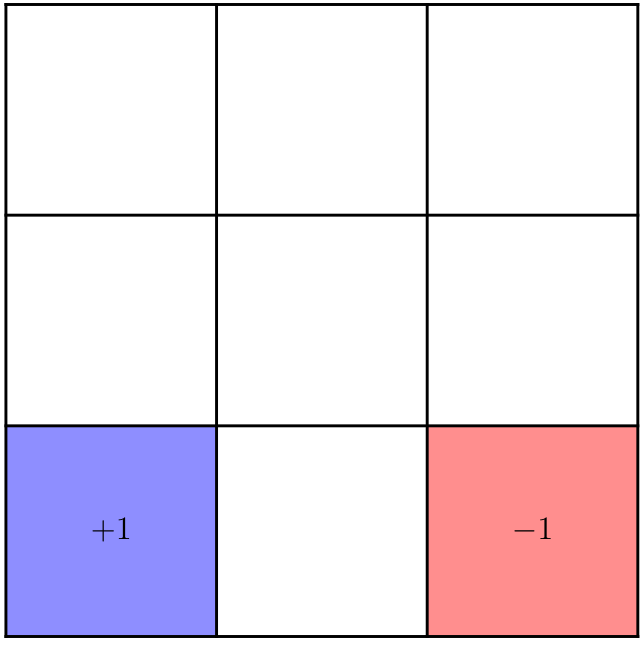
$$\begin{bmatrix} -1 & 1 & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



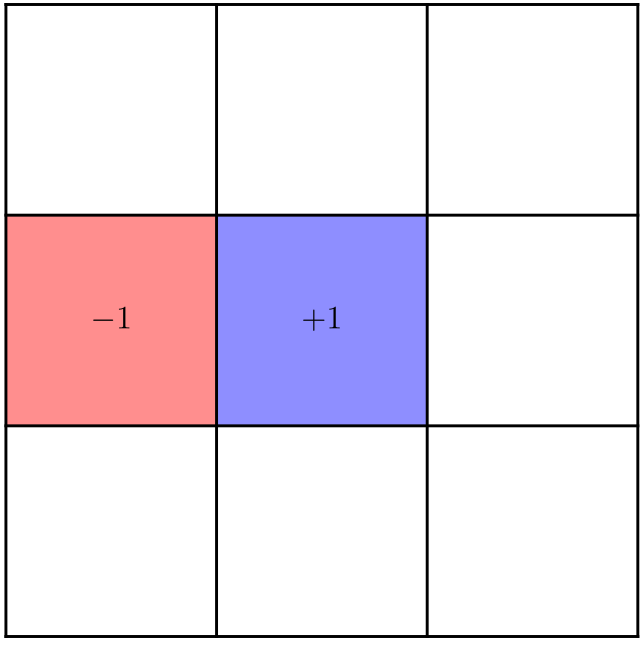
$$\begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & & \\ & & & \end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



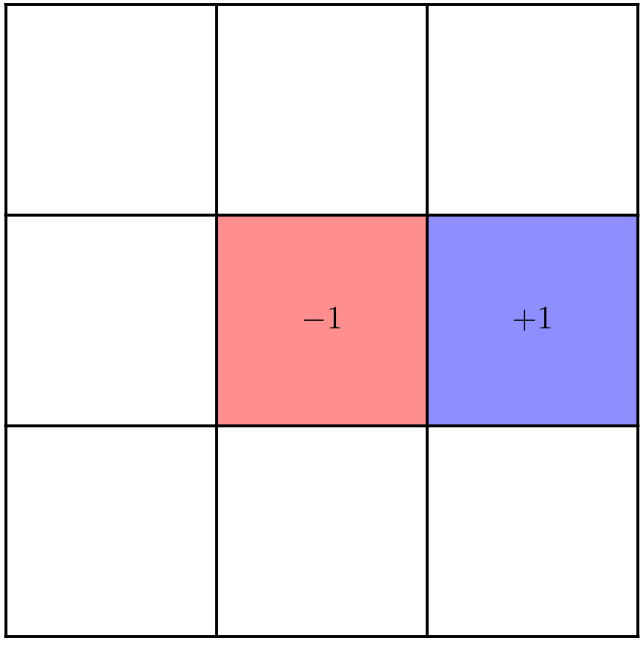
$$\begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ 1 & & -1 & \\ & & & \end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



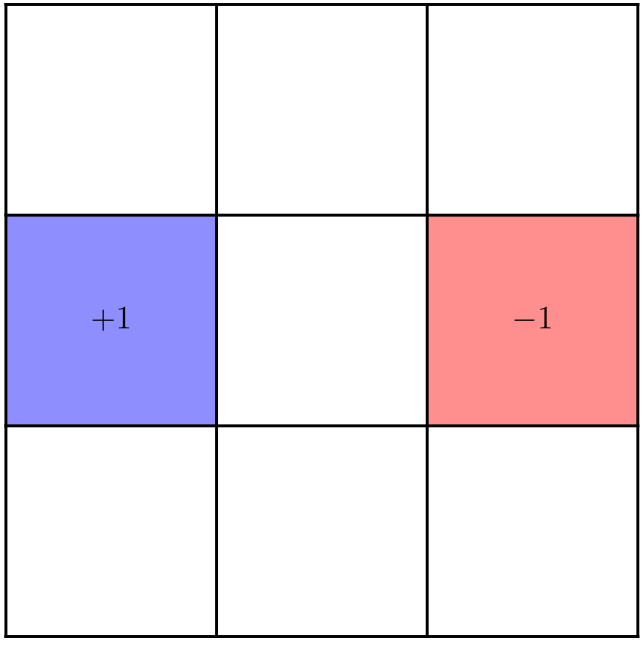
$$\begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ 1 & & -1 & & & \\ & & & -1 & 1 & \\ & & & & & \\ & & & & & \end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



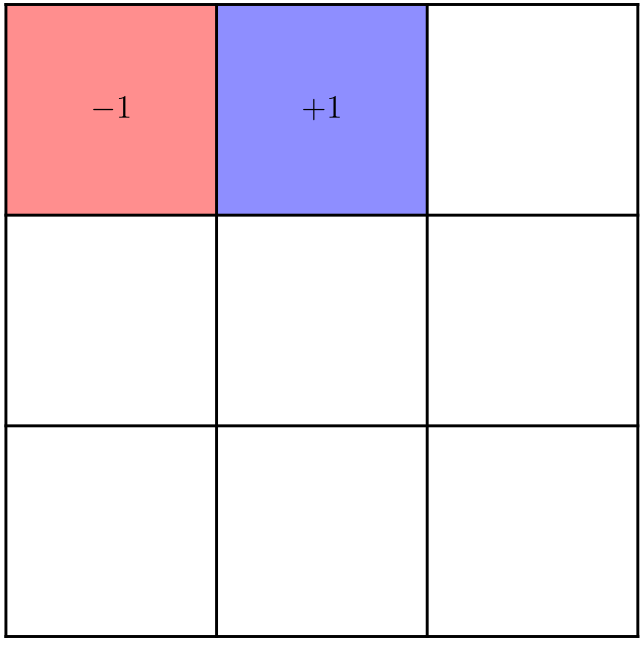
$$\begin{bmatrix} -1 & 1 & & & & & & & \\ & -1 & 1 & & & & & & \\ 1 & & -1 & & & & & & \\ & & & -1 & 1 & & & & \\ & & & & -1 & 1 & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



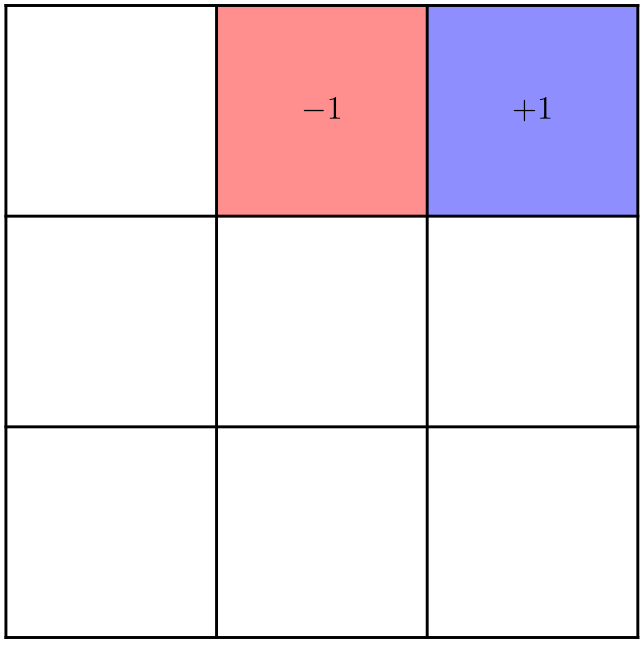
$$\begin{bmatrix}
-1 & 1 & & & & & & & \\
 & -1 & 1 & & & & & & \\
1 & & -1 & & & & & & \\
 & & & -1 & 1 & & & & \\
 & & & & -1 & 1 & & & \\
 & & & 1 & & -1 & & & \\
 & & & & & & & & \\
 & & & & & & & &
\end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



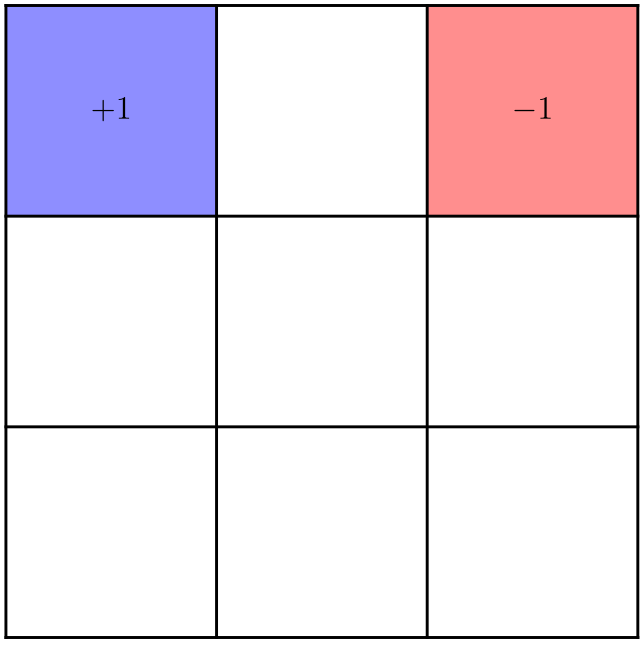
$$\begin{bmatrix}
-1 & 1 & & & & & & & \\
 & -1 & 1 & & & & & & \\
1 & & -1 & & & & & & \\
 & & & -1 & 1 & & & & \\
 & & & & -1 & 1 & & & \\
 & & & 1 & & -1 & & & \\
 & & & & & & -1 & 1 & \\
 & & & & & & & & \\
 & & & & & & & & \\
\end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



$$\begin{bmatrix}
-1 & 1 & & & & & & & \\
 & -1 & 1 & & & & & & \\
1 & & -1 & & & & & & \\
 & & & -1 & 1 & & & & \\
 & & & & -1 & 1 & & & \\
 & & & 1 & & -1 & & & \\
 & & & & & & -1 & 1 & \\
 & & & & & & & -1 & 1
\end{bmatrix}$$

Construction of L_H (periodic boundary conditions)



$$
\begin{bmatrix}
-1 & 1 & & & & & & & \\
 & -1 & 1 & & & & & & \\
1 & & -1 & & & & & & \\
 & & & -1 & 1 & & & & \\
 & & & & -1 & 1 & & & \\
 & & & 1 & & -1 & & & \\
 & & & & & & -1 & 1 & \\
 & & & & & & & -1 & 1 \\
 & & & & & & 1 & & -1
\end{bmatrix}
$$

Construction of L_H (periodic boundary conditions)

$$\begin{bmatrix}
-1 & 1 & & & & & & & \\
 & -1 & 1 & & & & & & \\
1 & & -1 & & & & & & \\
 & & & -1 & 1 & & & & \\
 & & & & -1 & 1 & & & \\
 & & & 1 & & -1 & & & \\
 & & & & & & -1 & 1 & \\
 & & & & & & & -1 & 1 \\
 & & & & & & 1 & & -1
\end{bmatrix}$$

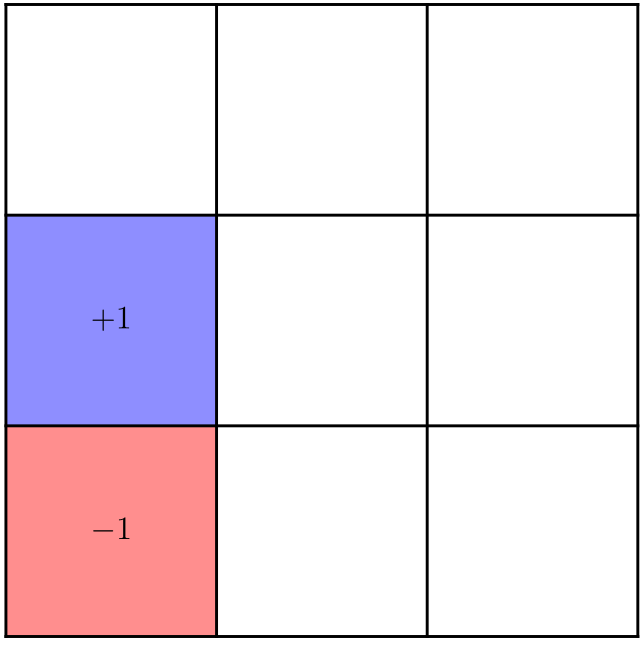
Python:

```
N = 3
block = sparse.spdiags(np.array([np.ones(N),-np.ones(N),np.ones(N)]),
                       np.array([1-N,0,1]),N,N) # form the 3x3 block
LH = sparse.block_diag([block]*N) # assemble the 9x9 block matrix
```

MATLAB:

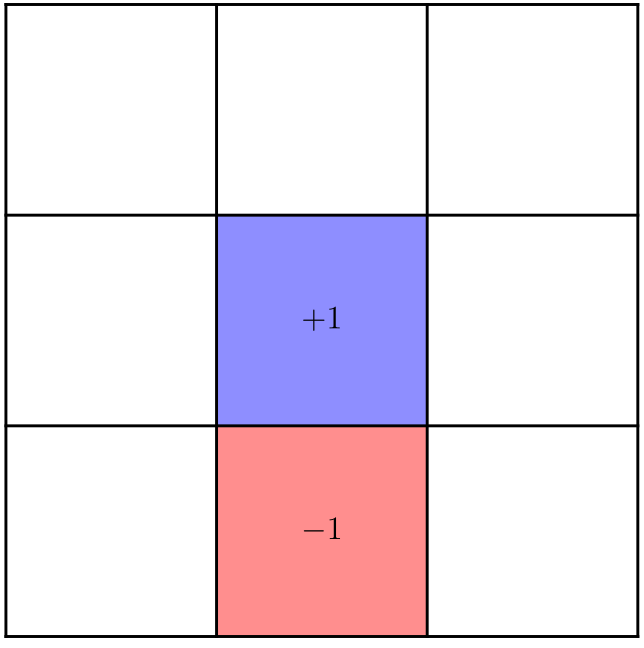
```
N = 3;
block = spdiags([1,-1,1].*ones(N,3),[1-N,0,1],N,N); % form the 3x3 block
LH = [];
for ii = 1:N
  LH = blkdiag(LH,block); % assemble the 9x9 block matrix
end
```

Construction of L_V (periodic boundary conditions)



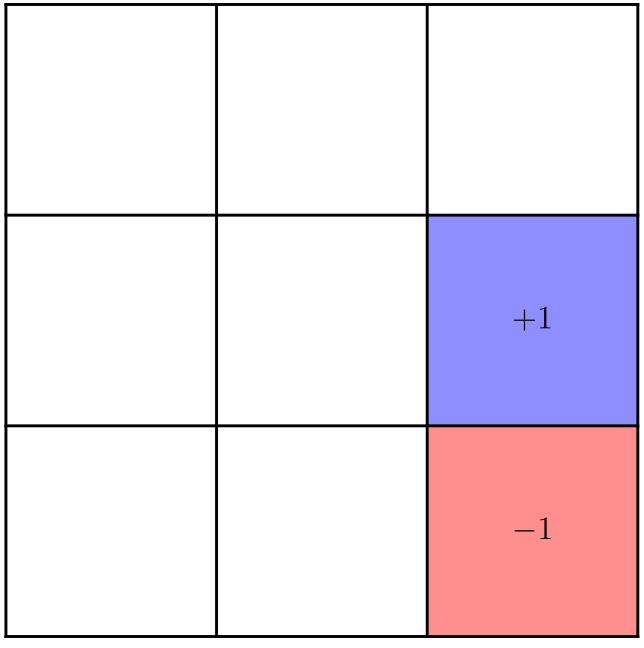
$$\begin{bmatrix} -1 & & 1 \\ & & \\ & & \\ & & \\ & & \end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



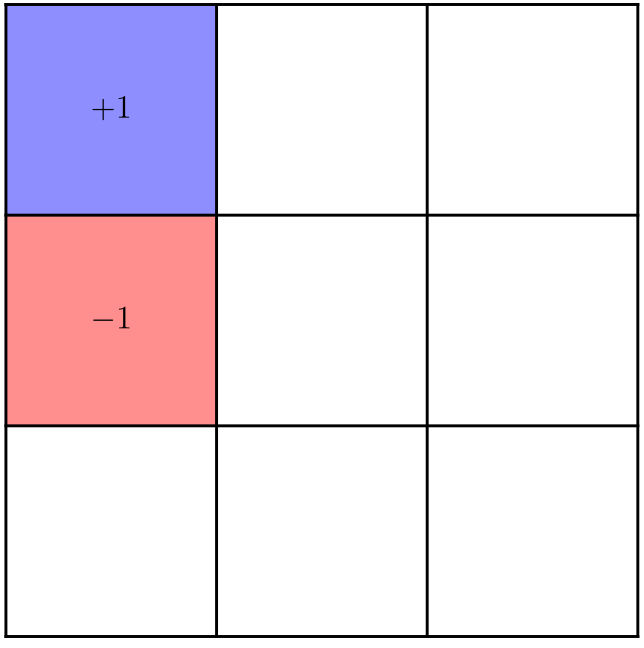
$$\begin{bmatrix} -1 & & & 1 & & \\ & -1 & & & 1 & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



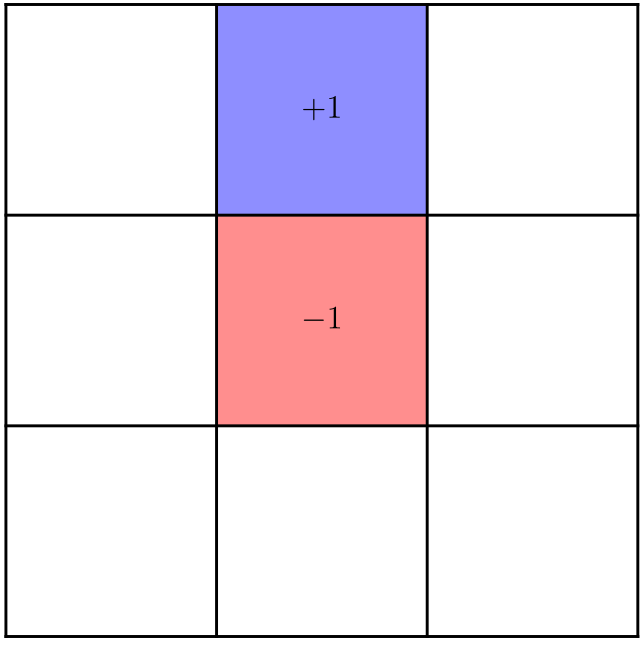
$$\begin{bmatrix} -1 & & & 1 & & & \\ & -1 & & & 1 & & \\ & & -1 & & & 1 & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



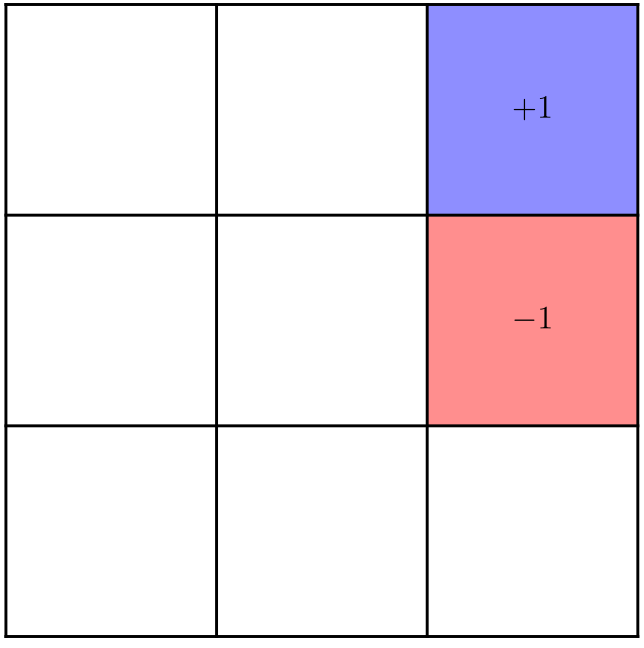
$$\begin{bmatrix} -1 & & & 1 & & & \\ & -1 & & & 1 & & \\ & & -1 & & & 1 & \\ & & & -1 & & & 1 \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



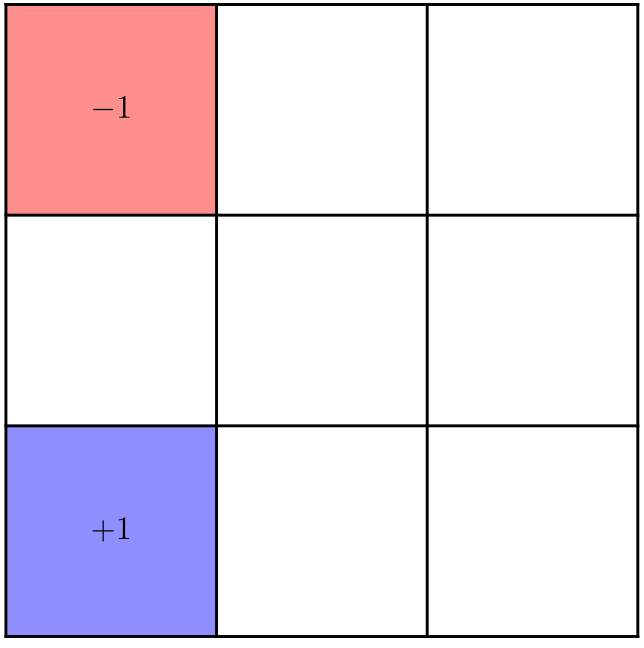
$$\begin{bmatrix} -1 & & & 1 & & & & & \\ & -1 & & & 1 & & & & \\ & & -1 & & & 1 & & & \\ & & & -1 & & & 1 & & \\ & & & & -1 & & & 1 & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



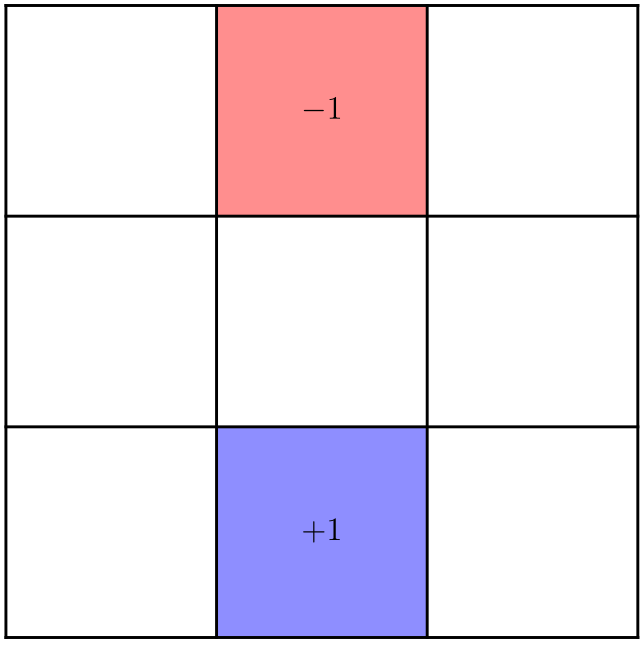
$$\begin{bmatrix}
-1 & & & 1 & & & & & \\
& -1 & & & 1 & & & & \\
& & -1 & & & 1 & & & \\
& & & -1 & & & 1 & & \\
& & & & -1 & & & 1 & \\
& & & & & -1 & & & 1 \\
& & & & & & & & \\
& & & & & & & & \\
& & & & & & & &
\end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



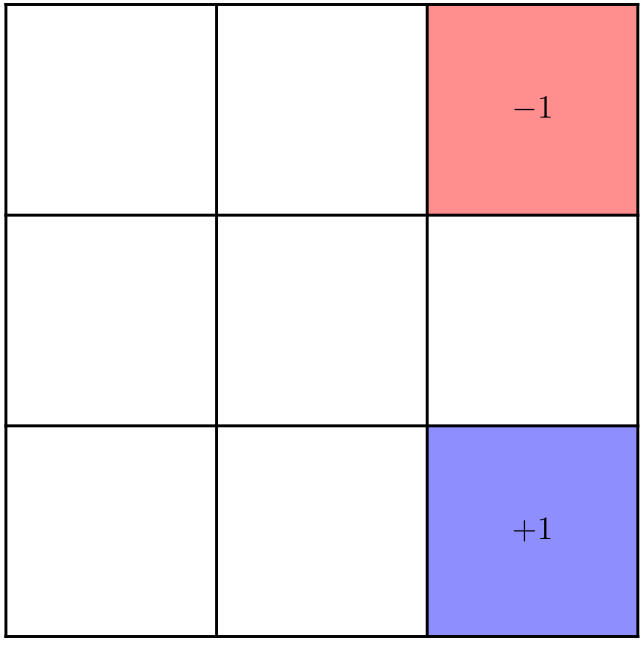
$$\begin{bmatrix}
-1 & & & 1 & & & & & \\
& -1 & & & 1 & & & & \\
& & -1 & & & 1 & & & \\
& & & -1 & & & 1 & & \\
& & & & -1 & & & 1 & \\
& & & & & -1 & & & 1 \\
1 & & & & & & -1 & & \\
& & & & & & & & \\
& & & & & & & &
\end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



$$\begin{bmatrix}
-1 & & & 1 & & & & & \\
& -1 & & & 1 & & & & \\
& & -1 & & & 1 & & & \\
& & & -1 & & & 1 & & \\
& & & & -1 & & & 1 & \\
& & & & & -1 & & & 1 \\
1 & & & & & & -1 & & \\
& 1 & & & & & & -1 & \\
& & & & & & & &
\end{bmatrix}$$

Construction of L_V (periodic boundary conditions)



$$\begin{bmatrix}
-1 & & & 1 & & & & & \\
& -1 & & & 1 & & & & \\
& & -1 & & & 1 & & & \\
& & & -1 & & & 1 & & \\
& & & & -1 & & & 1 & \\
& & & & & -1 & & & 1 \\
1 & & & & & & -1 & & \\
& 1 & & & & & & -1 & \\
& & 1 & & & & & & -1
\end{bmatrix}$$

Construction of L_V (periodic boundary conditions)

$$\begin{bmatrix} -1 & & & 1 & & & & & \\ & -1 & & & 1 & & & & \\ & & -1 & & & 1 & & & \\ & & & -1 & & & 1 & & \\ & & & & -1 & & & 1 & \\ & & & & & -1 & & & 1 \\ 1 & & & & & & -1 & & \\ & 1 & & & & & & -1 & \\ & & 1 & & & & & & -1 \end{bmatrix}$$

Python:

```
N = 3
LV = sparse.spdiags(np.array([np.ones(N**2),-np.ones(N**2),np.ones(N**2)]),
                    np.array([-N**2+N,0,N]),N**2,N**2)
```

MATLAB:

```
N = 3;
LV = spdiags([1,-1,1].*ones(N^2,3),[-N^2+N,0,N],N^2,N^2);
```

Let $F^*: \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ and $G: \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be convex lower semicontinuous functions and $K \in \mathbb{R}^{M \times N}$. Consider the abstract problem

$$\min_{x \in \mathbb{R}^N} \max_{\eta \in \mathbb{R}^M} \{\langle Kx, \eta \rangle + G(x) - F^*(\eta)\}.$$

The general form of the Chambolle–Pock algorithm can be written as

$$\begin{aligned}
\eta_{k+1} &= \operatorname{prox}_{\sigma F^*}(\eta_k + \sigma K \widetilde{x}_k), && \text{(update dual variable)} \\
x_{k+1} &= \operatorname{prox}_{\tau G}(x_k - \tau K^{\mathrm{T}} \eta_{k+1}), && \text{(update primal variable)} \\
\widetilde{x}_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k), && \text{(extrapolation)}
\end{aligned}$$

where $\tau > 0$ is the primal step size, $\sigma > 0$ is the dual step size, $\theta > 0$ is an extrapolation parameter, and the *proximal operator* of a function f is defined as

$$\operatorname{prox}_f(\eta) := \arg\min_x \left\{ f(x) + \tfrac{1}{2}\|x - \eta\|^2 \right\}.$$

If $\sigma\tau \leq 1/L^2$, $L = \|K\|_2$ (operator norm), and $\theta = 1$, then the algorithm can be shown to converge at linear rate $\mathcal{O}(k^{-1})$ [Chambolle and Pock 2011].

Let us recast the TV regularization problem

$$\min_{x\geq 0}\big\{\tfrac{1}{2}\|y-Ax\|^2+\lambda\|Dx\|_1\big\},\quad \lambda>0, \tag{1}$$

in the above framework.

- Note that

$$\tfrac{1}{2}\|Ax-y\|^2=\max_q\big\{\langle Ax-y,q\rangle-\tfrac{1}{2}\|q\|^2\big\},$$

since $0=\nabla_q(\langle Ax-y,q\rangle-\frac{1}{2}\|q\|^2)=Ax-y-q$ iff $q=Ax-y$.

- Since $\|x\|_1=\sum_i|x_i|=\langle|x|,1\rangle=\langle x,\operatorname{sign}(x)\rangle$,

$$\lambda\|Dx\|_1=\max_{\|z\|_\infty\leq 1}\langle Dx,\lambda z\rangle=\max_{\|z\|_\infty\leq\lambda}\langle Dx,z\rangle=\max_z\big\{\langle Dx,z\rangle-\iota_\lambda(z)\big\},$$

where $\iota_\lambda(z)=0$ if $\|z\|_\infty\leq\lambda$ and $\iota_\lambda(z)=+\infty$ otherwise.

Then (1) is equivalent to

$$\min_x\max_{q,z}\big\{\langle Ax-y,q\rangle+\langle Dx,z\rangle-\tfrac{1}{2}\|q\|^2-\iota_\lambda(z)+\iota_+(x)\big\},$$

where $\iota_+(x)=0$ if $x\geq 0$ and $\iota_+(x)=+\infty$ otherwise.

It is easy to see that

$$\min_{x} \max_{q,z} \big\{ \langle Ax - y, q \rangle + \langle Dx, z \rangle - \tfrac{1}{2}\|q\|^2 - \iota_\lambda(z) + \iota_+(x) \big\}$$

is tantamount to

$$\min_{x} \max_{q,z} \left\{ \left\langle Kx, \begin{bmatrix} q \\ z \end{bmatrix} \right\rangle + G(x) - F^*(q,z) \right\},$$

where

$$\begin{aligned} G(x) &= \iota_+(x), \\ F^*(q,z) &= \langle y, q \rangle + \tfrac{1}{2}\|q\|^2 + \iota_\lambda(z), \\ K &= \begin{bmatrix} A \\ D \end{bmatrix}. \end{aligned}$$

Note that if $A \in \mathbb{R}^{Q \times N}$ and $D \in \mathbb{R}^{L \times N}$, then $K \in \mathbb{R}^{(Q+L) \times N}$ and we identify the dual variable as the pair $\eta = (q, z) \in \mathbb{R}^M$, where $q \in \mathbb{R}^Q$, $z \in \mathbb{R}^L$, and $M = Q + L$.

The proximal mapping corresponding to G is simply the projection onto $\{x \geq 0 \mid x \in \mathbb{R}^N\}$:

$$\operatorname{prox}_{\tau G}(x) = (\max(x_i, 0))_i = \mathtt{max}(\mathtt{x}, 0).$$

On the other hand,

$$\operatorname{prox}_{\sigma F^*}(q, z) = \left(\frac{q - \sigma y}{1 + \sigma}, \frac{\lambda z}{\max\bigl(\lambda, |z|\bigr)} \right). \qquad \text{(N.B. } \eta = (q, z)\text{)}$$

Noting that $K^{\mathrm{T}} = [A^{\mathrm{T}}, D^{\mathrm{T}}]$, the Chambolle–Pock algorithm takes the form

$$\begin{cases} \eta_{k+1} = \operatorname{prox}_{\sigma F^*}(\eta_k + \sigma K \widetilde{x}_k) \\ x_{k+1} = \operatorname{prox}_{\tau G}(x_k - \tau K^{\mathrm{T}} \eta_{k+1}) \\ \widetilde{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k) \end{cases}$$

$$\Leftrightarrow \quad \begin{cases} q_{k+1} = \frac{q_k + \sigma A \widetilde{x}_k - \sigma y}{1+\sigma} \\ z_{k+1} = \frac{\lambda(z_k + \sigma D \widetilde{x}_k)}{\max(\lambda, |z_k + \sigma D \widetilde{x}_k|)} & \text{(elementwise division)} \\ x_{k+1} = \max(x_k - \tau A^{\mathrm{T}} q_{k+1} - \tau D^{\mathrm{T}} z_{k+1}, 0) & \text{(elementwise } \max\text{)} \\ \widetilde{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k). \end{cases}$$

Pseudocode for the Chambolle–Pock algorithm

Given: projection matrix A, data y, regularization parameter λ.

1. Form the difference matrices $\mathtt{L_H}$ and $\mathtt{L_V}$. Set $\mathtt{D = [L_H;L_V];}$
2. `L = svds([A;D],1);`
3. `tau = 1/L, sigma = 1/L, theta = 1;`
4. `x = zeros(size(A,2),1), q = zeros(size(A,1),1);`
5. `z = zeros(size(D,1),1), xhat = x;`

   Repeat

   6. `q = (q+sigma*(A*xhat-y))/(1+sigma);`
   7. `z = lambda *`
      `(z+sigma*D*xhat)./max(lambda,abs(z+sigma*D*xhat));`
   8. `xold = x;`
   9. `x = max(x-tau*A'*q-tau*D'*z,0);`
   10. `xhat = x+theta*(x-xold);`

   until convergence.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eighth lecture, June 5, 2023

Bayesian inverse problems

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

We will mainly follow

- D. Sanz-Alonso, A. M. Stuart, and A. Taeb (2018). Inverse Problems and Data Assimilation. https://arxiv.org/abs/1810.06191

Other helpful texts are

- J. Kaipio and E. Somersalo (2005). Statistical and Computational Inverse Problems. Springer, New York, NY.
- D. Calvetti and E. Somersalo (2007). Introduction to Bayesian Scientific Computing – Ten Lectures on Subjective Computing. Springer, New York, NY.

The Bayesian approach

Suppose that we have a noisy measurement model

$$y = F(x) + \eta,$$

where $F : \mathbb{R}^d \to \mathbb{R}^k$ is the forward mapping, $y \in \mathbb{R}^k$ is the measurement, $\eta \in \mathbb{R}^k$ is measurement noise, and $x \in \mathbb{R}^d$ is the unknown.

In the Bayesian approach to solving inverse problems

- both the noise η and the unknown quantity x (in a statistical context usually called the *parameter*) are modelled as *random variables* with values in $\mathbb{R}^k$ and $\mathbb{R}^d$, respectively, and their probability distributions are assumed to be known.
- the quantity of interest is now the conditional distribution of x, given the measured data y, which is considered the solution to the inverse problem in the Bayesian sense.

We consider the noisy measurement model

$$y = F(x) + \eta.$$

- The distribution of the parameter x formalizes all knowledge and beliefs about x *before* the data y is taken into account. In the Bayesian context, it is called *prior distribution*.
- The conditional distribution of x, given y, takes the data y into account, which can be understood as updating our knowledge and beliefs about the parameter x. In the Bayesian context, it is called *posterior distribution*.

The posterior distribution is usually obtained using some form of Bayes' formula. It contains all knowledge about the parameter available from the prior distribution and the measured data. It can be used to obtain parameter estimates that are most likely in some sense or that represent the posterior distribution well. In addition, the spread of the posterior distribution provides information about the remaining uncertainty in the parameter reconstruction.

While this approach has the advantage of being based upon explicit assumptions on the distribution of the noise and the parameter, it is not immediately clear why or how it should help resolving the ill-posedness of a problem. We will, however, see how under certain conditions the Bayesian approach has a regularizing effect in the sense that both the posterior distribution, and estimators based upon it, are stable with respect to changes in the data. To this end, we will introduce metrics to measure the distance of probability distributions during next week's lecture.

A brief introduction to probability theory

Here, we give a brief – and somewhat informal – overview of some fundamental notions from probability theory that are needed for our purposes, such as random variables, probability distributions and densities, as well as joint, marginal, and conditional probability densities.

Probability measures

Let Ω be a set and let $\mathcal{P}(\Omega)$ denote its power set. A subset $\mathcal{F}$ of $\mathcal{P}(\Omega)$ is called *σ-algebra* (or σ-field) if

1. $\varnothing \in \mathcal{F}$,
2. $\Omega \setminus A \in \mathcal{F}$ for every $A \in \mathcal{F}$, and
3. $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$ for every countable subset $\{A_n\}_{n\in\mathbb{N}}$ of $\mathcal{F}$.

A pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

An intuitive way of thinking about σ-algebras is that they contain information. The subsets contained in a σ-algebra represent events for which we can decide, after the observation, whether they happened or not. Hence, $\mathcal{F}$ represents all the information we can get from an experiment. For a topological space Ω (e.g., $\mathbb{R}^d$), the smallest σ-algebra containing all open sets in Ω is called *Borel σ-algebra* on Ω and it is denoted by $\mathcal{B}(\Omega)$.

A function $\mu\colon \mathcal{F} \to [0,\infty) \cup \{\infty\}$ is called *probability measure* if

(i) $\mu(\varnothing) = 0$,

(ii) for every countable subset $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ of pairwise disjoint sets (i.e., $A_i \cap A_j = \varnothing$ if $i \neq j$),

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n),$$

(iii) and $\mu(\Omega) = 1$.

We call $\mu(A)$ the *probability* of an event $A \in \mathcal{F}$. If $\mu(A) = 1$, we say that the event A occurs *almost surely*. A triple $(\Omega, \mathcal{F}, \mu)$ is called *probability space*. If only properties (i) and (ii) are satisfied, μ is called a *measure*. A measure is called σ-finite if Ω is the countable union of measurable sets with finite measure.

Example

The *Dirac measure* δ_m at a point $m \in \mathbb{R}^d$ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined by

$$\delta_m(A) = \begin{cases} 1 & \text{if } m \in A, \\ 0 & \text{if } m \notin A \end{cases} \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d).$$

Example

The Lebesgue measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is σ-finite, but not a probability measure, since $\lambda(\mathbb{R}^d) = \infty$.

Let μ and ν be two measures on the same measure space. Then μ is said to be *absolutely continuous with respect to* ν (or *dominated by* ν) if $\nu(A) = 0$ implies $\mu(A) = 0$ for each $A \in \mathcal{F}$. We denote this by $\mu \ll \nu$. Measures μ and ν are called *equivalent* if $\mu \ll \nu$ and $\nu \ll \mu$. If μ and ν are supported on disjoint sets, they are called *mutually singular*.

Theorem (Radon–Nikodym)

Let μ and ν be two measures on a measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and ν is σ-finite, then there exists a unique ν-integrable function f such that

$$\mu(A) = \int_A f(\omega)\nu(\mathrm{d}\omega) \quad \textit{for all } A \in \mathcal{F}.$$

The function f is called *Radon–Nikodym derivative* (or *density*) of μ with respect to ν and it is denoted by $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

Example

If μ is a measure which is absolutely continuous with respect to the Lebesgue measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then it has a unique density $p \in L^1(\mathbb{R}^d)$ by the Radon–Nikodym theorem.

Example

Let $\mu_1 = \mathcal{U}([0,1])$ and $\mu_2 = \mathcal{U}([0,2])$ be uniform probability measures on $\mathbb{R}$. Then $\mu_1 \ll \mu_2$ with

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(t) = \begin{cases} 2 & \text{for } t \in [0,1], \\ 0 & \text{otherwise,} \end{cases}$$

but μ_2 is not absolutely continuous with respect to μ_1 because $\mu_1([1,2]) = 0$, whereas $\mu_2([1,2]) = \frac{1}{2} > 0$.

Weak convergence of probability measures

A sequence $\{\mu_n\}_{n\in\mathbb{N}}$ of probability measures is said to *converge weakly* to μ if

$$\lim_{n\to\infty} \int_\Omega f(\omega)\mu_n(\mathrm{d}\omega) = \int_\Omega f(\omega)\mu(\mathrm{d}\omega)$$

for every bounded continuous function $f \in C_b(\Omega, \mathbb{R})$. In this case, we write $\mu_n \rightharpoonup \mu$.

Random variables

A function $x\colon \Omega \to X$ between a probability space $(\Omega, \mathcal{F}, \mu)$ and a measurable space $(X, \mathcal{X})$ is now called a *random variable (with values in X)* if it is measurable, that is, if

$$x^{-1}(A) \in \mathcal{F} \quad \text{for every } A \in \mathcal{X}.$$

Here, $x^{-1}(A) = \{\omega \in \Omega : x(\omega) \in A\}$.

A random variable x induces a probability measure ν on X, defined by

$$\nu(A) := \mu(x^{-1}(A)) \quad \text{for all } A \in \mathcal{X},$$

which is called *probability distribution* (or *law*) of x. We write $x \sim \nu$ if x is distributed according to ν.

A random variable x connects an event $A \in \mathcal{X}$ with a corresponding event $x^{-1}(A) \in \mathcal{F}$ and assigns the probability of $x^{-1}(A)$ to A. This probability is denoted by

$$\mathbb{P}(x \in A) := \nu(A) = \mu(x^{-1}(A)) = \mu(\{\omega \in \Omega : x(\omega) \in A\}).$$

Now, let x be a random variable with values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and ν its distribution.

If ν is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$, then by the Radon–Nikodym theorem there exists a unique $p \in L^1(\mathbb{R}^d)$ such that

$$\nu(A) = \int_A p(u)\mathrm{d}u \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d).$$

The function p is called *probability density* of x.

Throughout, we will work with $\mathbb{R}^d$-valued random variables and assume that they have a probability density.

- The *mean* or *expected value* of an $\mathbb{R}^d$-valued random variable x with distribution ν and density p is given by

$$\mathbb{E}[x] := \int_{\mathbb{R}^d} x\nu(\mathrm{d}x) = \int_{\mathbb{R}^d} xp(x)\mathrm{d}x.$$

- A *mode* $\bar{x}$ of a random variable x is defined as a maximizer of its density p, i.e.,

$$\bar{x} \in \operatorname*{arg\,max}_{x\in\mathbb{R}^d} p(x).$$

- The *covariance* (or *covariance matrix*) of two random variables x_1 and x_2 is defined by

$$\mathrm{Cov}(x_1, x_2) = \mathbb{E}\left[(x_1 - \mathbb{E}[x_1])(x_2 - \mathbb{E}[x_2])^{\mathrm{T}}\right].$$

- The *characteristic function* φ_x of x is defined by

$$\varphi_x(h) = \int_{\mathbb{R}^d} \exp(\mathrm{i}\, h^{\mathrm{T}}x)\, \nu(\mathrm{d}x) = \int_{\mathbb{R}^d} \exp(\mathrm{i}\, h^{\mathrm{T}}x)p(x)\mathrm{d}x \quad \text{for all } h \in \mathbb{R}^d.$$

  A random variable is uniquely determined by its characteristic function.

Gaussian random variables

Gaussian random variables arise naturally in many applications.

- A Gaussian distribution is a popular choice for the prior distribution.
- By the central limit theorem, a Gaussian distribution is often a good approximation to inherently non-Gaussian distributions when the observation is based on a large number of mutually independent random events. For this reason the noise is often assumed to have a Gaussian distribution.

Let $m \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d \times d}$ be a symmetric positive semidefinite matrix ($C \succeq 0$). An $\mathbb{R}^d$-valued random variable x is said to be *Gaussian* (or *normal*) with mean m and covariance C, denoted by $x \sim \mathcal{N}(m, C)$, if its characteristic function φ_x is given by

$$\varphi_x(h) = \exp\left(\mathrm{i}\, h^{\mathrm{T}} m - \frac{1}{2} h^{\mathrm{T}} C h\right) \quad \text{for all } h \in \mathbb{R}^d.$$

A Gaussian random variable is completely determined by its mean and its covariance.

- If, in addition, C is positive definite ($C \succ 0$), $x \sim \mathcal{N}(m, C)$ has the probability density

$$\begin{aligned}\mathbb{P}(x) &= \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(x-m)^{\mathrm{T}} C^{-1}(x-m)\right) \\ &= \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}\|C^{-\frac{1}{2}}(x-m)\|^2\right).\end{aligned}$$

  Note that C is invertible and $C^{-1/2}$ exists due to our assumptions on C. Here, $\|x\|_{C^{-1}} := \|C^{-1/2}x\|$.

- The Dirac measure δ_m at a point $m \in \mathbb{R}^d$ can be understood as a Gaussian distribution with covariance $C = 0$, i.e., $\delta_m = \mathcal{N}(m, 0)$.

- If $z_1 \sim \mathcal{N}(m_1, C_1)$ and $z_2 \sim \mathcal{N}(m_2, C_2)$ are independent and $a_1, a_2 \in \mathbb{R}$, then

$$z = a_1 z_1 + a_2 z_2 \sim \mathcal{N}(a_1 m_1 + a_2 m_2, a_1^2 C_1 + a_2^2 C_2).$$

- If $z \sim \mathcal{N}(m, C)$, $L \in \mathbb{R}^{d \times k}$, and $a \in \mathbb{R}^d$, then

$$w = Lz + a \sim \mathcal{N}(Lm + a, LCL^{\mathrm{T}}).$$

- The weak convergence of Gaussian random variables is equivalent to convergence of their means and covariances. That is, a sequence $z_n \sim \mathcal{N}(m_n, C_n)$ converges weakly towards $z \sim \mathcal{N}(m, C)$ ($z_n \rightharpoonup z$), if and only if $m_n \to m$ and $C_n \to C$.

Conditional and marginal probability densities

Let x and y be random variables with values in $\mathbb{R}^d$ and $\mathbb{R}^k$, respectively. If the random variable (x, y) has a probability density $p_{x,y}$, i.e., if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}((x, y) \in A \times B) = \int_{A \times B} p_{x,y}(u, v) \mathrm{d}(u, v),$$

for all $A \in \mathcal{B}(\mathbb{R}^d)$ and $B \in \mathcal{B}(\mathbb{R}^k)$, then $p_{x,y}$ is called *joint probability density* of x and y. Here $\mathbb{P}(x \in A, y \in B) := \mathbb{P}(x \in A \text{ and } y \in B)$. To simplify notation, we will also write $\mathbb{P}(x, y) := p_{x,y}(x, y)$.

Now, the *marginal probability density* p_x of x is defined by

$$p_x(u) = \int_{\mathbb{R}^k} p_{x,y}(u, v) \mathrm{d}v \quad \text{for all } u \in \mathbb{R}^d.$$

Analogously, the marginal density of y is

$$p_y(v) = \int_{\mathbb{R}^d} p_{x,y}(u, v) \mathrm{d}u \quad \text{for all } v \in \mathbb{R}^k.$$

The marginal density of x is indeed the probability density for x in the situation that we have no information about the random variable y, because

$$\begin{aligned}\mathbb{P}(x \in A) = \mathbb{P}(x \in A, y \in \mathbb{R}^k) &= \int_{A \times \mathbb{R}^k} p_{x,y}(u,v)\mathrm{d}(u,v) \\ &= \int_A \left(\int_{\mathbb{R}^k} p_{x,y}(u,v)\mathrm{d}v \right) \mathrm{d}u = \int_A p_x(u)\mathrm{d}u\end{aligned}$$

for every $A \in \mathcal{B}(\mathbb{R}^d)$.

The random variables x and y are called *independent* (denoted by $x \perp y$) if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}(x \in A)\,\mathbb{P}(y \in B)$$

for all $A \in \mathcal{B}(\mathbb{R}^d)$, $B \in \mathcal{B}(\mathbb{R}^k)$ or, equivalently, if

$$p_{x,y}(u,v) = p_x(u)p_y(v) \quad \text{almost surely.}$$

To simplify notation, we will also write $\mathbb{P}(x) := p_x(x)$.

Next, we consider the random variable x in the opposite situation that we know everything about the random variable y: we have observed it and know what value it has taken.

We say we consider the random variable x, *given* that we know the value y_0 taken by y, and denote this by $x|y = y_0$. For $y_0 \in \mathbb{R}^k$ with $p_y(y_0) > 0$, the *conditional probability density* of $x|y = y_0$, $p_{x|y=y_0}$, is then defined by

$$p_{x|y=y_0}(u) = \frac{p_{x,y}(u, y_0)}{p_y(y_0)}.$$

If x and y are independent and $p_y(y_0) > 0$, then

$$p_{x|y=y_0}(u) = p_x(u).$$

To simplify notation, we will also write $\mathbb{P}(x|y) := p_{x|y}(x) := p_{x|y=y}(x)$.

Bayes' formula

Let (x, y) be a random variable with joint density $\mathbb{P}(x, y)$ on $\mathbb{R}^d \times \mathbb{R}^k$. If $\mathbb{P}(y) > 0$, then the conditional probability density of x, given y, equals

$$\mathbb{P}(x|y) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(y)}, \quad \mathbb{P}(y) = \int_{\mathbb{R}^d} \mathbb{P}(x, y) \mathrm{d}x.$$

On the other hand, the conditional probability density of y in case we know the value of the unknown x, is called the *likelihood function*

$$\mathbb{P}(y|x) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)}, \quad \text{if } \mathbb{P}(x) > 0.$$

The joint density of (x, y), in turn, can be expressed in terms of the likelihood of y, given x, as $\mathbb{P}(x, y) = \mathbb{P}(y|x)\,\mathbb{P}(x)$, which leads to *Bayes' formula*

$$\mathbb{P}(x|y) = \frac{\mathbb{P}(y|x)\,\mathbb{P}(x)}{\int_{\mathbb{R}^d} \mathbb{P}(y|x)\,\mathbb{P}(x) \mathrm{d}x}.$$

Bayes' formula presents a way to express the conditional probability density of x, given y, assuming that the conditional density of y, given x, and the marginal density of x are known.

Bayes' formula for inverse problems

We return to an inverse problem of estimating an unknown parameter $x \in \mathbb{R}^d$ from data $y \in \mathbb{R}^k$ that is connected to x via the model

$$y = F(x) + \eta.$$

We make the following assumptions:

- A1 The noise η has the probability density ν on $\mathbb{R}^k$.
- A2 The parameter x has the probability density π on $\mathbb{R}^d$.
- A3 The random variables x and η are independent.

The following theorem yields the probability density of the posterior distribution, i.e., the conditional density $\pi^y(x) := \mathbb{P}(x|y)$ of the parameter x, given a specific realization y of the measured data.

Lemma

Under assumptions $\mathrm{A1} - \mathrm{A3}$*, the likelihood (i.e., the conditional probability of* y*, given* x*) is*

$$\mathbb{P}(y|x) = \nu(y - F(x)).$$

Proof. The forward model $y = F(x) + \eta$ defines the conditional probability density

$$\begin{aligned}\mathbb{P}(y|x) &= p_{y|x}(y) = p_{F(x)+\eta|x}(y) \\ &= p_{\eta|x}(y - F(x)) = p_\eta(y - F(x)) = \nu(y - F(x))\end{aligned}$$

due to the assumptions $\eta \perp x$ and $\eta \sim \nu$. □

Theorem (Bayes' theorem)

If assumptions $\mathrm{A1} - \mathrm{A3}$ *hold and*

$$Z(y) := \int_{\mathbb{R}^d} \nu(y - F(x))\pi(x)\mathrm{d}x > 0,$$

then

$$\pi^y(x) = \frac{1}{Z(y)}\nu(y - F(x))\pi(x). \tag{1}$$

Proof. By the previous Lemma, the random variable (x, y) has the joint density

$$\mathbb{P}(x, y) = \mathbb{P}(y|x)\,\mathbb{P}(x) = \nu(y - F(x))\pi(x),$$

since $x \sim \pi$ by assumption. Now, the density of the posterior distribution is defined as

$$\pi^y(x) = \mathbb{P}(x|y) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(y)} = \frac{\nu(y - F(x))\pi(x)}{\mathbb{P}(y)},$$

and the marginal density of y is given by

$$\mathbb{P}(y) = \int_{\mathbb{R}^d} \mathbb{P}(x, y)\mathrm{d}x = Z(y) > 0. \quad \square$$

- The condition that the marginal density $\mathbb{P}(y)$ of the observed data y is positive means that the observed data is assumed to be consistent with the probabilistic assumptions $\mathrm{A1} - \mathrm{A3}$.
- Bayes' formula (1) implies that the posterior distribution is absolutely continuous with respect to the prior distribution, $\pi^y \ll \pi$, with the Radon–Nikodym derivative

$$\frac{\mathrm{d}\pi^y}{\mathrm{d}\pi}(x) = \frac{\nu(y - F(x))}{Z(y)}.$$

  This means that an event cannot have positive probability under the posterior distribution if it does not have positive probability under the prior distribution.
- Bayes' theorem can be generalized to infinite-dimensional spaces, cf., e.g., [Theorem 14, Dashti–Stuart 2017]. However, its formulation involves more subtlety. There is no Lebesgue measure on infinite-dimensional spaces, so the density of the posterior distribution is stated with respect to the prior distribution instead.

Case study: source localization

Suppose that a particle with unit charge is located at some (unknown) point $x^* \in (0,1)$ and our goal is to locate it based on measurements of voltage at the interval end points $x = 0$ and $x = 1$. The mathematical model for the voltage at any point $x \in [0,1]$ is given by

$$y(x) = \frac{1}{|x^* - x|}.$$

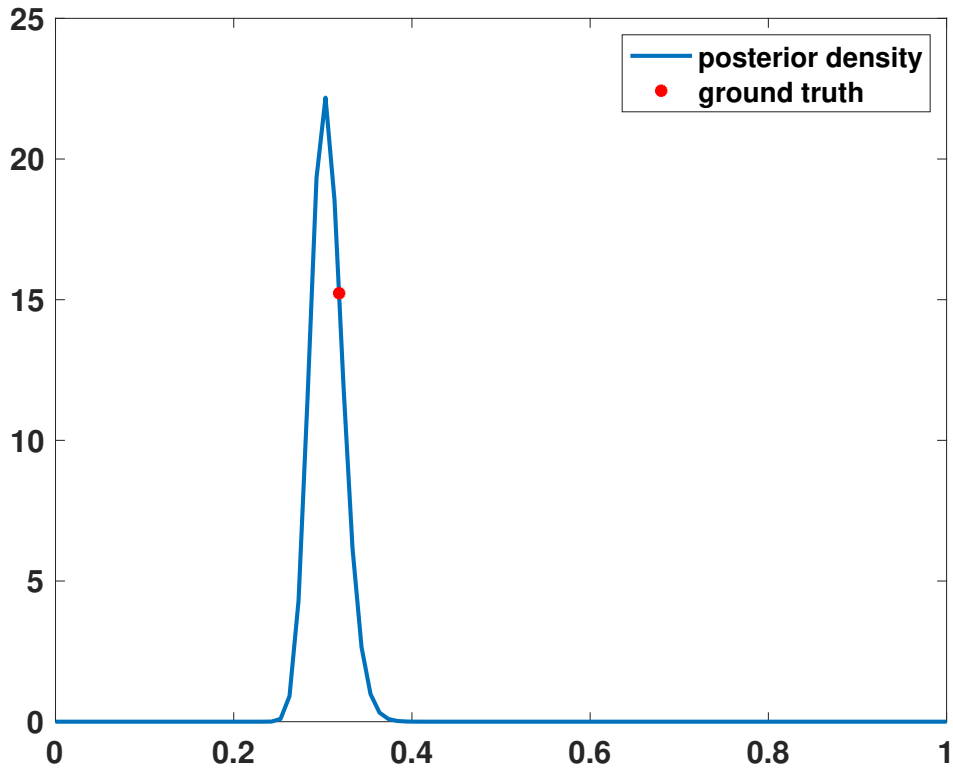
Our noisy measurements are modeled by $y_1 = \frac{1}{|x^*-0|} + \eta_1$ and $y_2 = \frac{1}{|x^*-1|} + \eta_2$, where η_1 and η_2 are i.i.d. realizations of $\mathcal{N}(0, \sigma^2)$. We take $x^* = 1/\pi$ (ground truth) and $\sigma = 0.2$ in the numerical experiments.

- The likelihood is given by $\mathbb{P}(y|x) \propto \exp(-\frac{1}{2\sigma^2} \sum_{j=0}^{1} (y_{j+1} - \frac{1}{|x-j|})^2)$.
- We consider the prior $\pi(x) = \chi_{(0,1)}(x) = \begin{cases} 1 & \text{if } x \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$

Then the posterior density is given by Bayes' formula

$$\pi^y(x) \propto \chi_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=0}^{1} \left(y_{j+1} - \frac{1}{|x-j|} \right)^2 \right).$$

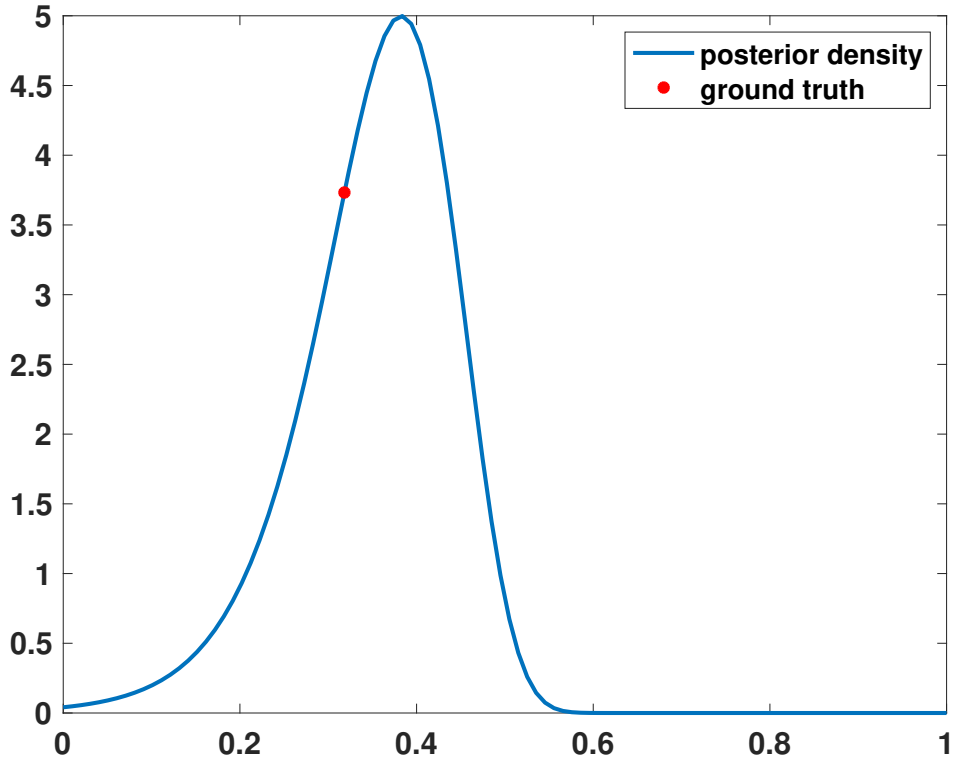
Let us visualize the posterior density against the ground truth solution. (See also files `source.py` / `source.m` on the course homepage!)



We see that the posterior is localized around the true parameter value ("ground truth"). *Note that in this case, the prior hardly plays any role.*

We could take, e.g., the mean or mode of the posterior density as a point estimate for the unknown location of the point charge. We will discuss more about Bayesian estimators next week.

What if we modify the problem so that we have access to only one boundary measurement at $x = 1$?



The resulting posterior distribution carries substantially more uncertainty since we now have less measurement data!

Note that the posterior will generally be high-dimensional, meaning that it is usually not possible to visually inspect the posterior density.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Ninth lecture, June 12, 2023

Recap: Bayes' formula for inverse problems

We are interested in the inverse problem of solving $x \in \mathbb{R}^d$ from

$$y = F(x) + \eta,$$

where $y \in \mathbb{R}^k$ is the measurement vector, $F : \mathbb{R}^d \to \mathbb{R}^k$ the forward mapping, and $\eta \in \mathbb{R}^k$ is noise. We model x, y, and η as random variables. Then we have:

Theorem (Bayes' theorem)

We assume:

- *The noise η has the probability density ν on $\mathbb{R}^k$.*
- *The parameter x has the probability density π on $\mathbb{R}^d$.*
- *The random variables x and η are independent.*

Then the likelihood is $\mathbb{P}(y|x) = \nu(y - F(x))$ and we can write

$$\pi^y(x) := \mathbb{P}(x|y) = \frac{\mathbb{P}(y|x)\mathbb{P}(x)}{\mathbb{P}(y)} =: \frac{\nu(y - F(x))\pi(x)}{Z(y)},$$

provided that $Z(y) := \int_{\mathbb{R}^d} \nu(y - F(x))\pi(x)\mathrm{d}x > 0$.

Bayes' formula:

$$\pi^y(x) = \frac{\nu(y - F(x))\pi(x)}{Z(y)}.$$

- The *prior model* $\pi(x)$ describes *a priori information*. It should assign high probability to objects x which are typical in light of *a priori information*, and low probability to unexpected x.
- The *likelihood model* $\mathbb{P}(y|x) = \nu(y - F(x))$ processes measurement information. It gives low probability to objects that produce simulated data which is very different from the measured data.
- The number $Z(y)$ can be seen as a normalization constant.
- The *posterior distribution* $\pi^y(x) = \mathbb{P}(x|y)$ represents the updated knowledge about the parameter of interest x, given the evidence y.

Since the normalization constant $Z(y)$ is often not of interest in our considerations, we frequently write the Bayes' formula as

$$\pi^y(x) \propto \nu(y - F(x))\pi(x),$$

where the symbol $\propto$ means equality up to a constant factor.

Case study: one-dimensional deconvolution

As motivation[†], suppose that we are interested in estimating a signal $f : [0, 1] \to \mathbb{R}$ from noisy, blurred observations modeled as

$$y_i = y(s_i) = \int_0^1 K(s_i, t) f(t) \, \mathrm{d}t + \eta_i, \quad i \in \{1, \ldots, k\},$$

where the blurring kernel is

$$K(s, t) = \exp\left(-\frac{1}{2\omega^2}(s - t)^2 \right), \quad \omega = 0.5,$$

and $\eta \in \mathbb{R}^k$ is measurement noise.

[†]We will consider the so-called "linear-Gaussian setting" as well as computational techniques for sampling posterior densities in more detail in a couple of weeks. Specifically, we will not consider the question of *how to draw samples from the posterior density* today. We will revisit this question in more detail at a later time.

Discrete model

Midpoint rule:

$$y_i = \int_0^1 K(s_i, t) f(t)\,\mathrm{d}t + \eta_i \approx \frac{1}{d} \sum_{j=1}^{d} K(s_i, t_j) x_j + \eta_i,$$

where $t_j = \frac{j}{d} - \frac{1}{2d}$ and $x_j = f(t_j)$ for $j \in \{1, \ldots, d\}$.

If we have $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \ldots, k\}$, then we have the discrete linear model

$$y = Ax + \eta, \quad \text{where } A_{i,j} = \frac{1}{d} K(s_i, t_j).$$

To employ the Bayesian approach, we treat y, η, and x as random variables. We assume that η is Gaussian noise with variance $\sigma^2 I$,

$$\eta \sim \mathcal{N}(0, \sigma^2 I), \qquad \nu(\eta) \propto \exp\Big(-\frac{1}{2\sigma^2} \|\eta\|^2 \Big).$$

The likelihood is then given by

$$\mathbb{P}(y|x) = \nu(y - Ax) \propto \exp\Big(-\frac{1}{2\sigma^2} \|y - Ax\|^2 \Big).$$

Next, we have to choose a prior distribution for the unknown. Assume that we know that $x(0) = x(1) = 0$ and that x is quite smooth, that is, the value of $x(t)$ in a point is more or less the same as in its neighbor. We will then model the unknown as

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + W_j, \quad j = 1, \ldots, k, \tag{1}$$

where the term W_j follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$.

The variance γ^2 determines how much the reconstructed function x departs from the smoothness model $x_j = \frac{1}{2}(x_{j-1} + x_{j+1})$. We can write (1) as

$$Lx = W, \quad \text{where} \quad L := \frac{1}{2}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & \end{bmatrix}.$$

This leads to the so-called *smoothness prior*

$$\pi(x) \propto \exp\Big(-\frac{1}{2\gamma^2}\|Lx\|^2 \Big).$$

Using Bayes' formula, we get the posterior distribution

$$\pi^y(x) \propto \exp\Big(-\frac{1}{2\sigma^2}\|y - Ax\|^2 - \frac{1}{2\gamma^2}\|Lx\|^2 \Big).$$

For the numerical experiment, we simulate measurements using the (smooth) ground truth signal

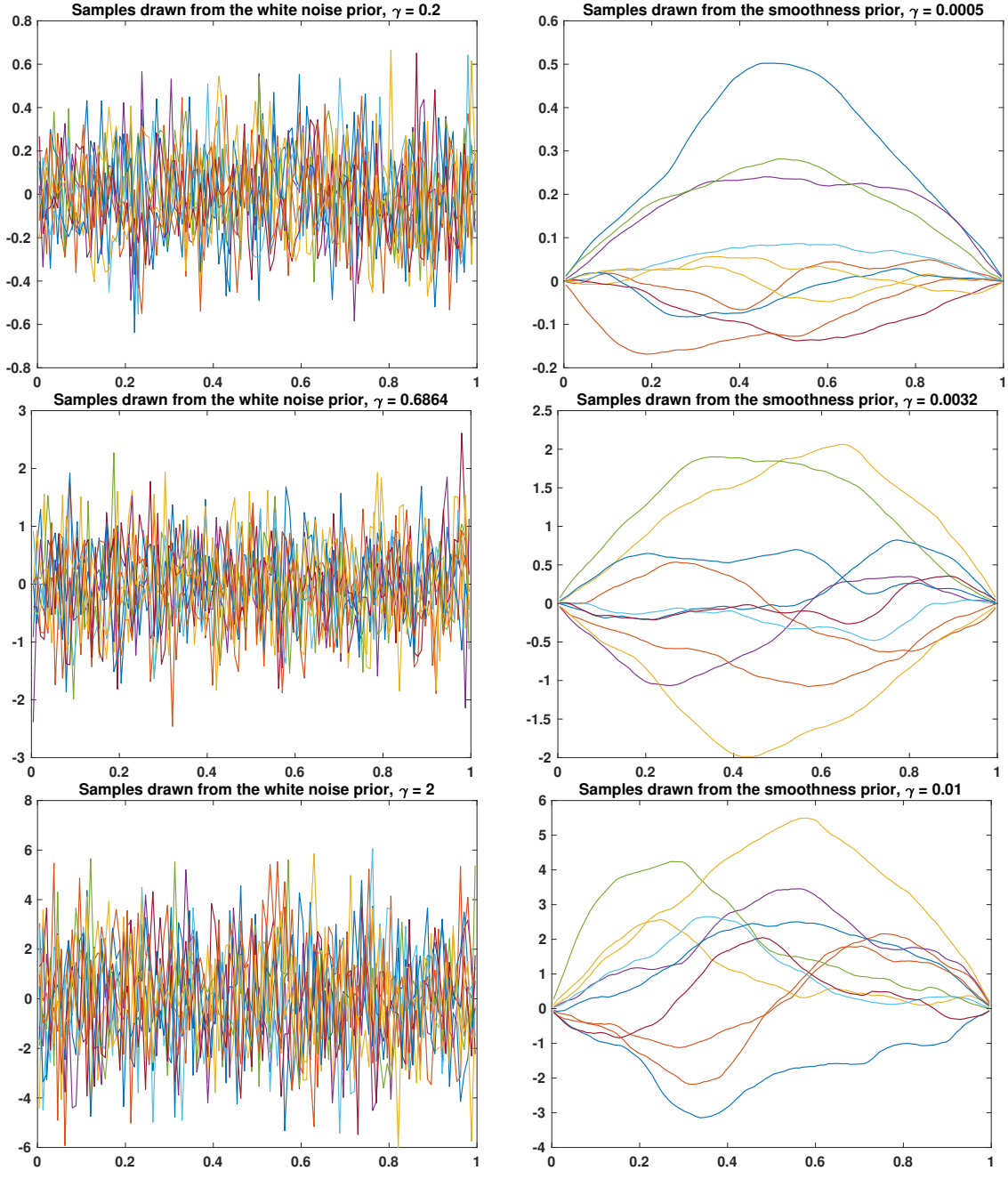
$$f(t) = 8t^3 - 16t^2 + 8t,$$

which satisfies $f(0) = f(1) = 0$. The measurements are contaminated with 10% *relative* noise ($\sigma \approx 0.0618$) and we set $d = k = 120$.

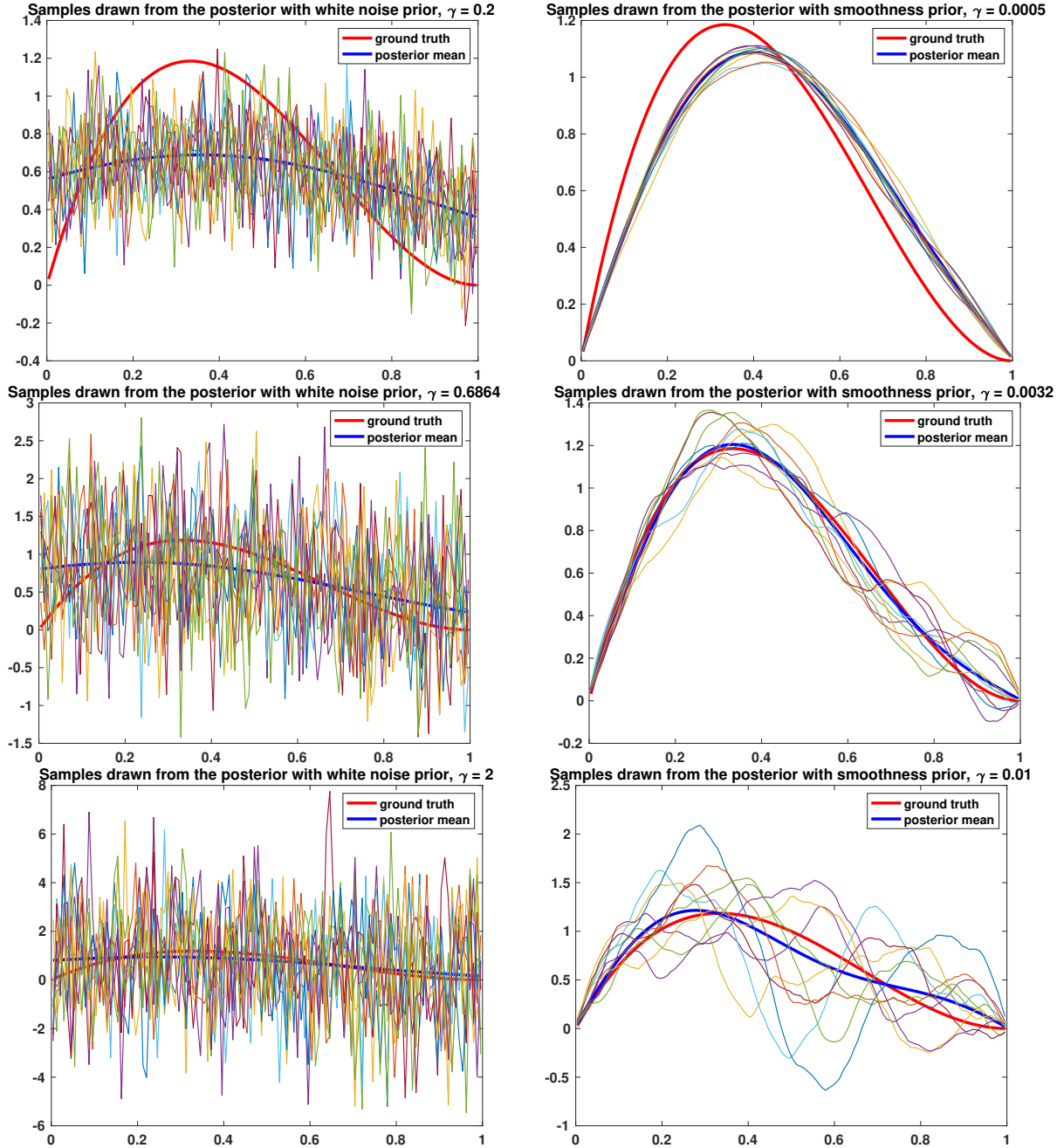
Let us draw samples from the prior and posterior. As comparison, we also consider a posterior obtained using the white noise prior, i.e.,

$$\pi_0^y(x) \propto \Big(-\frac{1}{2\sigma^2}\|y - Ax\|^2 \Big)\pi_{\mathrm{pr},0}(x), \quad \pi_{\mathrm{pr},0}(x) \propto \exp\Big(-\frac{1}{2\gamma^2}\|x\|^2 \Big).$$

Remark: Let us discuss the *implementational* details (sampling from Gaussian posterior distributions, formulae for the posterior means and variances of Gaussian posterior distributions) for this example in more detail *next week*.



Samples drawn from the white noise prior and the smoothness prior for several values of γ.



Samples drawn from the posterior corresponding to both the white noise prior and the smoothness prior for several values of γ. We also plot the ground truth solution and the posterior mean. The solutions in the middle row *roughly* satisfy the Morozov discrepancy principle.

As the previous example illustrates, many practical problems tend to be high-dimensional. The measurement model for the discretized deconvolution example

$$y = Ax + \eta,$$

with $A \in \mathbb{R}^{k \times d}$, $x \in \mathbb{R}^d$, and $y, \eta \in \mathbb{R}^k$, where k corresponds to the number of points $s_1, \ldots, s_k$ where we observe the signal and d corresponds to the number of quadrature points $t_1, \ldots, t_d$ discretizing the unknown quantity x.

A grid with only $k = d = 120$ points already corresponds to a 120-dimensional posterior, so visualization of the posterior density is highly nontrivial.

In practice, we are often interested in various point estimates, statistics, samples, or the spread of the posterior distribution.

Bayesian estimators

The posterior distribution can be used to define estimators for the conditional random variable $x|y \sim \pi^y(x)$. In general, an estimator $\hat{x}$ is any function of the data y. The estimate $\hat{x}(y)$ is itself an $\mathbb{R}^d$-valued random variable whose properties give information about the usefulness and quality of the estimator.

Bayesian estimators are those defined via the posterior distribution π^y. We present the two most prominent ones. The *conditional mean (CM) estimator*, which is defined as the mean

$$\hat{x}_{\mathrm{CM}}(y) = \mathbb{E}[x|y] = \int_{\mathbb{R}^d} u\pi^y(u)\mathrm{d}u$$

of the posterior distribution.

The *maximum a posteriori (MAP) estimator*, which is defined as the mode

$$\hat{x}_{\mathrm{MAP}}(y) = \operatorname*{arg\,max}_{u\in\mathbb{R}^d} \pi^y(u)$$

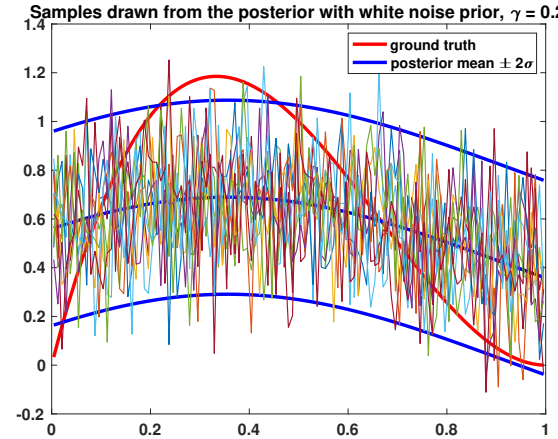
of the posterior distribution (if a unique mode exists).

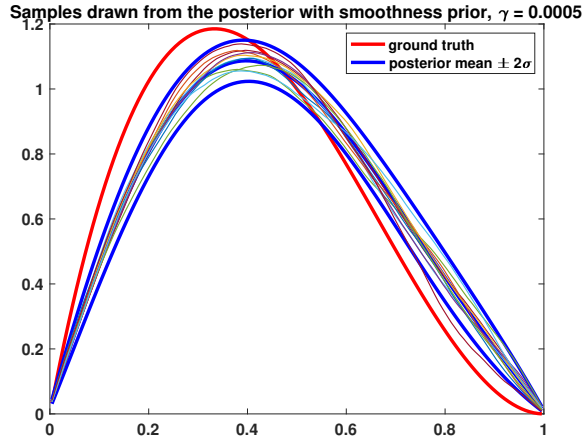
One way to estimate spread are Bayesian *credible sets*. A level $1-\alpha$ credible set $\mathcal{C}_\alpha$ with $\alpha \in (0,1)$ satisfies

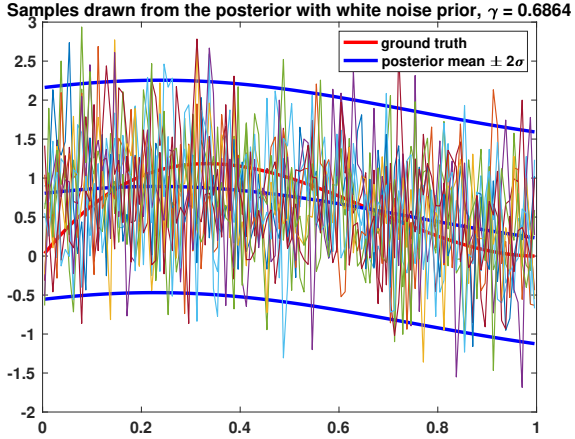
$$\mathbb{P}(x \in \mathcal{C}_\alpha | y) = \int_{\mathcal{C}_\alpha} \pi^y(u) \mathrm{d}u = 1 - \alpha.$$

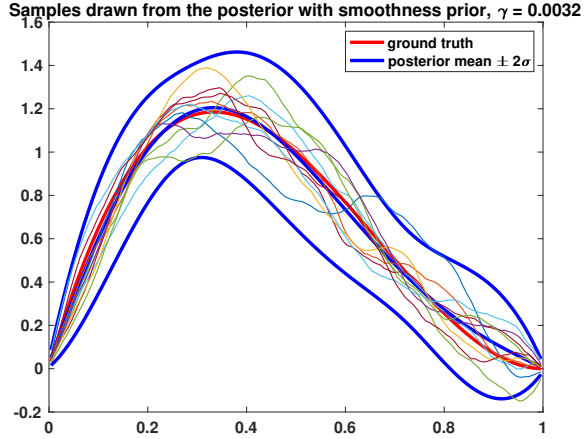
For small α, it is a region that contains a large fraction of the posterior mass.

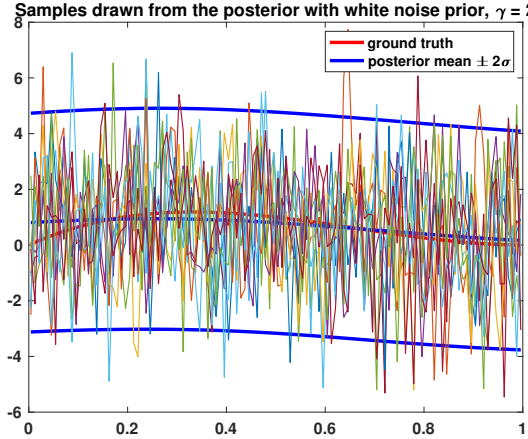
Deconvolution example: posteriors with 2σ credibility envelopes.

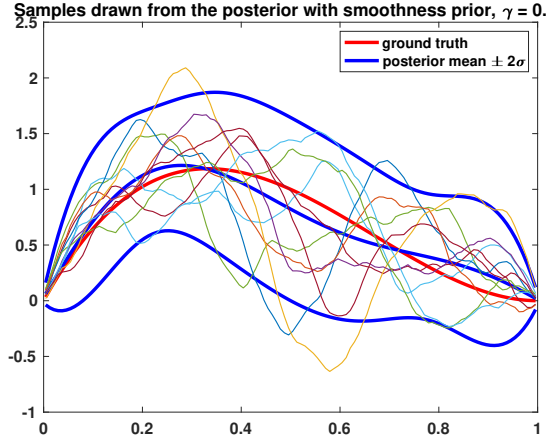












Example. Assume that $x \in \mathbb{R}$ and that the posterior density is given by

$$\pi^y(u) = \frac{c}{\sigma_1}\phi\left(\frac{u}{\sigma_1}\right) + \frac{1-c}{\sigma_2}\phi\left(\frac{u-1}{\sigma_2}\right),$$

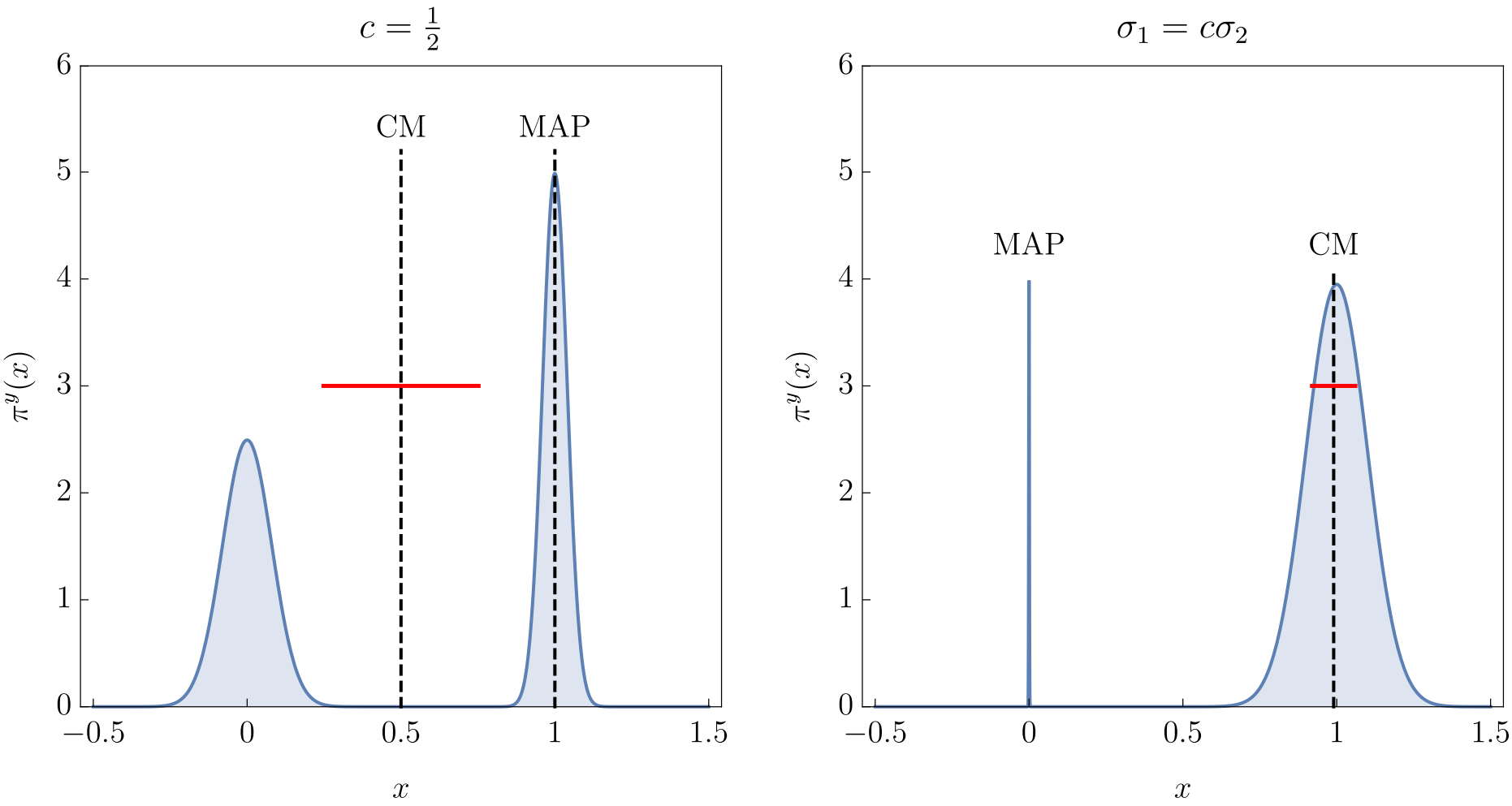
where $c \in (0,1)$, $\sigma_1, \sigma_2 > 0$, and ϕ is the density of the standard normal distribution, $\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$. In this case,

$$\hat{x}_{\mathrm{CM}} = 1 - c \quad \text{and} \quad \hat{x}_{\mathrm{MAP}} = \begin{cases} 0 & \text{if } c/\sigma_1 > (1-c)/\sigma_2, \\ 1 & \text{if } c/\sigma_1 < (1-c)/\sigma_2. \end{cases}$$

If $c = \frac{1}{2}$ and σ_1, σ_2 are small, the probability that x takes values near $\hat{x}_{\mathrm{CM}}$ is small. On the other hand, if $\sigma_1 = c\sigma_2$, then $c/\sigma_1 = 1/\sigma_2 > (1-c)/\sigma_2$, so that $\hat{x}_{\mathrm{MAP}} = 0$. If c is small, this is, however, a bad estimate for x, since the probability for x to take values near 0 is small. Last of all, we notice that when the conditional mean gives a poor estimate, this is reflected in a larger posterior variance

$$\sigma^2 = \int_{-\infty}^{\infty} (u - \hat{x}_{\mathrm{CM}})^2 \pi^y(u) \mathrm{d}u.$$

We cannot say that one estimator is better than the other in all applications.



Left: the density with $\sigma_1 = 0.08$, $\sigma = 0.04$, and $c = \frac{1}{2}$. The CM estimate represents the distribution poorly. Notice that when the CM gives a poor estimate, this is reflected in wider variance (1 standard deviation is depicted as a red line). Right: the density with $\sigma_1 = 0.001$, $\sigma_2 = 0.1$, and $c = 0.01$. The MAP gives a poor estimate since it is in an unlikely part of the computational domain.

The maximum likelihood estimate

$$\hat{x}_{\mathrm{ML}}(y) = \arg\max_{u \in \mathbb{R}^d} \mathbb{P}(y|u)$$

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. It is analogous to solving a classical inverse problem without regularization.

Well-posedness

Assume that the posterior density is given by

$$\pi^y(x) = \frac{1}{Z} g(x)\pi(x)$$

with likelihood $g(x)$ and prior density $\pi(x)$. Now consider an approximation

$$\pi^y_\delta(x) = \frac{1}{Z_\delta} g_\delta(x)\pi(x)$$

resulting from an approximated likelihood $g_\delta(x)$. Such an approximation can result, for example, from an approximation F_δ of the forward operator F or from perturbed data y_δ.

The question is therefore:

$$\text{does} \quad |g - g_\delta| = \mathcal{O}(\delta) \quad \text{imply} \quad d(\pi^y, \pi^y_\delta) = \mathcal{O}(\delta)$$

for small enough $\delta > 0$ and some metric $d(\cdot, \cdot)$ on probability densities?

- Well-posedness refers to the continuity of the method of obtaining the posterior distribution with respect to different perturbations in the parameters. In practice, this could mean for example the following: If we have two measurements close to each other, does this mean the corresponding posterior distributions are close in some metric? Recall that ill-posed problems generally are discontinuous in this regard, i.e., without regularization, small difference in measurements can induce arbitrarily large difference in reconstruction. Does the Bayesian approach then regularize the problem? The answer is yes under certain assumptions on the modeling.
- We will proceed to show that, under certain conditions, π^y and π^y_δ satisfy

$$d(\pi^y, \pi^y_\delta) \leq c\delta$$

for δ small enough, some $c > 0$, and some metric $d(\cdot, \cdot)$ on probability densities.

- To this end, we define two metrics for probability densities: the total variation distance and the Hellinger distance.

Metrics for probability densities

We introduce the total variation distance and the Hellinger distance, both of which have been used to show well-posedness results. Here, we will use the Hellinger distance to establish the well-posedness of Bayesian inverse problems.

Let π and π' be the probability densities of two random variables with values in $\mathbb{R}^d$. We define the *total variation distance* between π and π' as

$$d_{\mathrm{TV}}(\pi, \pi') = \frac{1}{2} \int_{\mathbb{R}^d} \left| \pi(x) - \pi'(x) \right| \mathrm{d}x = \frac{1}{2} \|\pi - \pi'\|_{L^1},$$

and the *Hellinger distance* between π and π' as

$$d_{\mathrm{H}}(\pi, \pi') = \left(\frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right|^2 \mathrm{d}x \right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}} \left\| \sqrt{\pi} - \sqrt{\pi'} \right\|_{L^2}.$$

The normalization constants are chosen in such a way that the largest possible distance between two densities is one, as can be seen in the following lemma.

Lemma

For any two probability densites π and π',

$$0 \leq d_{\mathsf{TV}}(\pi, \pi') \leq 1 \quad \textit{and} \quad 0 \leq d_{\mathsf{H}}(\pi, \pi') \leq 1.$$

Proof. The lower bounds follow immediately from the definition of d_{TV} and d_{H}. It remains to prove the upper bounds. To this end, we estimate

$$d_{\mathsf{TV}}(\pi, \pi') = \frac{1}{2} \int_{\mathbb{R}^d} |\pi(x) - \pi'(x)| \mathrm{d}x \leq \frac{1}{2} \int_{\mathbb{R}^d} \pi(x) \mathrm{d}x + \frac{1}{2} \int_{\mathbb{R}^d} \pi'(x) \mathrm{d}x = 1$$

and

$$\begin{aligned} d_{\mathsf{H}}(\pi, \pi')^2 &= \frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right|^2 \mathrm{d}x \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \left(\pi(x) + \pi'(x) - 2\sqrt{\pi(x)\pi'(x)} \right) \mathrm{d}x \\ &\leq \frac{1}{2} \int_{\mathbb{R}^d} \left(\pi(x) + \pi'(x) \right) \mathrm{d}x = 1. \quad \square \end{aligned}$$

In what follows, we will establish bounds between Hellinger and total variation distance and show how both distances can be used to bound the difference of expected values with respect to two different densities; these results will be useful in subsequent lectures.

Lemma

For any two probability densities π and π', the total variation and Hellinger distance are related by the inequalities

$$\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\pi, \pi') \leq d_{\mathrm{H}}(\pi, \pi') \leq \sqrt{d_{\mathrm{TV}}(\pi, \pi')}.$$

Proof. Using the Cauchy–Schwarz inequality and $(a+b)^2 \le 2a^2 + 2b^2$ leads to

$$
\begin{aligned}
d_{\mathsf{TV}}(\pi, \pi') &= \frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right| \cdot \left| \sqrt{\pi(x)} + \sqrt{\pi'(x)} \right| \mathrm{d}x \\
&\le \left(\frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right|^2 \mathrm{d}x \right)^{\frac{1}{2}} \left(\frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} + \sqrt{\pi'(x)} \right|^2 \mathrm{d}x \right)^{\frac{1}{2}} \\
&\le d_{\mathsf{H}}(\pi, \pi') \left(\frac{1}{2} \int_{\mathbb{R}^d} \left(2\pi(x) + 2\pi'(x) \right) \mathrm{d}x \right)^{\frac{1}{2}} = \sqrt{2} d_{\mathsf{H}}(\pi, \pi').
\end{aligned}
$$

Notice that $|\sqrt{\pi(x)} - \sqrt{\pi'(x)}| \le |\sqrt{\pi(x)} + \sqrt{\pi'(x)}|$, since $\sqrt{\pi(x)}, \sqrt{\pi'(x)} \ge 0$. Thus, we have

$$
\begin{aligned}
d_{\mathsf{H}}(\pi, \pi')^2 &= \frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right|^2 \mathrm{d}x \\
&\le \frac{1}{2} \int_{\mathbb{R}^d} \left| \sqrt{\pi(x)} - \sqrt{\pi'(x)} \right| \cdot \left| \sqrt{\pi(x)} + \sqrt{\pi'(x)} \right| \mathrm{d}x \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left| \pi(x) - \pi'(x) \right| \mathrm{d}x = d_{\mathsf{TV}}(\pi, \pi'). \quad \square
\end{aligned}
$$

The following lemmata show that if two densities are close in total variation or Hellinger distance, expectations computed with respect to both densities are also close.

Lemma

Let f be a real valued function on $\mathbb{R}^d$ such that $\mathbb{E}^{\pi}[f^2] + \mathbb{E}^{\pi'}[f^2] =: f_2^2 < \infty$, then

$$\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| \leq 2 f_2 d_{\mathsf{H}}(\pi, \pi'). \tag{2}$$

Proof. We estimate

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| = \left|\int_{\mathbb{R}^d} f(x)\left(\pi(x) - \pi'(x)\right)\mathrm{d}x\right| \\
&= \left|\int_{\mathbb{R}^d} f(x)\left(\sqrt{\pi(x)} - \sqrt{\pi'(x)}\right)\left(\sqrt{\pi(x)} + \sqrt{\pi'(x)}\right)\mathrm{d}x\right| \\
&\leq \left(\frac{1}{2}\int_{\mathbb{R}^d}\left|\sqrt{\pi(x)} - \sqrt{\pi'(x)}\right|^2 \mathrm{d}x\right)^{\frac{1}{2}} \left(2\int_{\mathbb{R}^d} |f(x)|^2 \left|\sqrt{\pi(x)} + \sqrt{\pi'(x)}\right|^2 \mathrm{d}x\right)^{\frac{1}{2}} \\
&\leq d_{\mathsf{H}}(\pi, \pi')\left(4\int_{\mathbb{R}^d} |f(x)|^2 \left(\pi(x) + \pi'(x)\right)\mathrm{d}x\right)^{\frac{1}{2}} = 2 f_2 d_{\mathsf{H}}(\pi, \pi'). \qquad \square
\end{aligned}$$

Lemma

Let f be a real valued function on $\mathbb{R}^d$ such that $\sup_{x\in\mathbb{R}^d} |f(x)| =: \|f\|_\infty < \infty$, *then*

$$\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| \leq 2\|f\|_\infty d_{\mathrm{TV}}(\pi, \pi').$$

Moreover, the following variational characterization of the total variation distance holds:

$$d_{\mathrm{TV}}(\pi, \pi') = \frac{1}{2} \sup_{\|f\|_\infty \leq 1} \left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right|.$$

Remark: Note that the result for the Hellinger distance only assumes that f is square integrable with respect to π and π', whereas the result for the total variation distance requires that f is bounded.

Proof. For the first part of the lemma, note that

$$\begin{aligned}\Big|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\Big| &= \Big|\int_{\mathbb{R}^d} f(x)\big(\pi(x) - \pi'(x)\big)\mathrm{d}x\Big| \\ &\leq 2\|f\|_\infty \cdot \frac{1}{2}\int_{\mathbb{R}^d} |\pi(x) - \pi'(x)|\mathrm{d}x = 2\|f\|_\infty d_{\mathsf{TV}}(\pi, \pi').\end{aligned}$$

This in particular shows that, for any f with $\|f\|_\infty = 1$,

$$d_{\mathsf{TV}}(\pi, \pi') \geq \frac{1}{2}\Big|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\Big|.$$

Our goal now is to show a choice of f with $\|f\|_\infty = 1$ that achieves equality. Define $f(x) := \mathsf{sign}\Big(\pi(x) - \pi'(x)\Big)$, so that $f(x)\Big(\pi(x) - \pi'(x)\Big) = |\pi(x) - \pi'(x)|$. Then, $\|f\|_\infty = 1$ and

$$\begin{aligned}d_{\mathsf{TV}}(\pi, \pi') &= \frac{1}{2}\int_{\mathbb{R}^d} |\pi(x) - \pi'(x)|\,\mathrm{d}x = \frac{1}{2}\int_{\mathbb{R}^d} f(x)\Big(\pi(x) - \pi'(x)\Big)\,\mathrm{d}x \\ &= \frac{1}{2}\Big|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\Big|.\end{aligned}$$

This completes the proof of the variational characterization. □

Approximation theorem

We denote by

$$g(x) = \nu(y - F(x)) \quad \text{and} \quad g_\delta(x) = \nu(y - F_\delta(x))$$

the likelihoods associated with F and F_δ, so that

$$\pi^y(x) = \frac{1}{Z} g(x)\pi(x) \quad \text{and} \quad \pi^y_\delta(x) = \frac{1}{Z_\delta} g_\delta(x)\pi(x)$$

with corresponding normalizing constants $Z, Z_\delta > 0$. We make the following assumptions on g and g_δ.

Assumption 1. There exist $\delta^+ > 0$, constants $K_1, K_2 > 0$, and a function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}^\pi[\varphi^2] \leq K_1$ and for all $\delta \in (0, \delta^+)$,

1. $\left|\sqrt{g(x)} - \sqrt{g_\delta(x)}\right| \leq \varphi(x)\delta$ for all $x \in \mathbb{R}^d$,
2. $\left|\sqrt{g(x)}\right| + \left|\sqrt{g_\delta(x)}\right| \leq K_2$ for all $x \in \mathbb{R}^d$.

Lemma

Under **Assumption 1** *there exist* $\tilde{\delta}^+ > 0$, $c_1, c_2 \in (0, +\infty)$ *such that*

$$|Z - Z_\delta| \le c_1 \delta \quad \textit{and} \quad Z, Z_\delta > c_2, \quad \textit{for } \delta \in (0, \tilde{\delta}^+).$$

Proof. Since $Z = \int_{\mathbb{R}^d} g(x)\pi(x)\mathrm{d}x$ and $Z_\delta = \int_{\mathbb{R}^d} g_\delta(x)\pi(x)\mathrm{d}x$ we have

$$\begin{aligned}
|Z - Z_\delta| &= \left| \int_{\mathbb{R}^d} \big(g(x) - g_\delta(x)\big)\pi(x)\mathrm{d}x \right| \\
&\le \Big(\int_{\mathbb{R}^d} \Big| \sqrt{g(x)} - \sqrt{g_\delta(x)} \Big|^2 \pi(x)\mathrm{d}x \Big)^{\frac{1}{2}} \Big(\int_{\mathbb{R}^d} \Big| \sqrt{g(x)} + \sqrt{g_\delta(x)} \Big|^2 \pi(x)\mathrm{d}x \Big)^{\frac{1}{2}} \\
&\le \Big(\int_{\mathbb{R}^d} \delta^2 \phi(x)^2 \pi(x)\mathrm{d}x \Big)^{\frac{1}{2}} \Big(\int_{\mathbb{R}^d} K_2^2 \pi(x)\mathrm{d}x \Big)^{\frac{1}{2}} \\
&\le \sqrt{K_1} K_2 \delta, \quad \delta \in (0, \delta^+).
\end{aligned}$$

And when $\delta \le \tilde{\delta}^+ := \min\{\frac{Z}{2\sqrt{K_1}K_2}, \delta^+\}$, we have

$$Z_\delta \ge Z - |Z - Z_\delta| \ge \frac{1}{2} Z.$$

The lemma follows by taking $c_1 = \sqrt{K_1} K_2$ and $c_2 = \frac{1}{2} Z$. □

Theorem (Well-posedness)

Under **Assumption 1**, *there exist* $\widetilde{\delta}^+ > 0$ *and* $c > 0$ *such that*

$$d_{\mathrm{H}}(\pi^y, \pi_\delta^y) \le c\delta \quad \text{for all } \delta \in (0, \widetilde{\delta}^+).$$

Proof. We break the distance into two error parts, one caused by the difference between Z and Z_δ, the other caused by the difference between g and g_δ:

$$\begin{aligned} d_{\mathrm{H}}(\pi^y, \pi_\delta^y) &= \frac{1}{\sqrt{2}} \left\| \sqrt{\pi^y} - \sqrt{\pi_\delta^y} \right\|_{L^2} \\ &= \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{g\pi}{Z}} - \sqrt{\frac{g\pi}{Z_\delta}} + \sqrt{\frac{g\pi}{Z_\delta}} - \sqrt{\frac{g_\delta \pi}{Z_\delta}} \right\|_{L^2} \\ &\le \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{g\pi}{Z}} - \sqrt{\frac{g\pi}{Z_\delta}} \right\|_{L^2} + \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{g\pi}{Z_\delta}} - \sqrt{\frac{g_\delta \pi}{Z_\delta}} \right\|_{L^2}. \end{aligned}$$

On the previous slide, we obtained

$$d_{\mathrm{H}}(\pi^y, \pi^y_\delta) \leq \frac{1}{\sqrt{2}} \Big\| \sqrt{\frac{g\pi}{Z}} - \sqrt{\frac{g\pi}{Z_\delta}} \Big\|_{L^2} + \frac{1}{\sqrt{2}} \Big\| \sqrt{\frac{g\pi}{Z_\delta}} - \sqrt{\frac{g_\delta \pi}{Z_\delta}} \Big\|_{L^2}.$$

Using the previous Lemma, for $\delta \in (0, \tilde{\delta}^+)$, we have for the first term

$$\Big\| \sqrt{\frac{g\pi}{Z}} - \sqrt{\frac{g\pi}{Z_\delta}} \Big\|_{L^2} = \Big| \frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z_\delta}} \Big| \underbrace{\Big(\int_{\mathbb{R}^d} g(x)\pi(x)\mathrm{d}x \Big)^{\frac{1}{2}}}_{=\sqrt{Z}}$$

$$= \left| 1 - \frac{\sqrt{Z}}{\sqrt{Z_\delta}} \right| = \left| \frac{\sqrt{Z_\delta} - \sqrt{Z}}{\sqrt{Z_\delta}} \right| = \frac{|Z - Z_\delta|}{(\sqrt{Z} + \sqrt{Z_\delta})\sqrt{Z_\delta}} \leq \frac{c_1}{2c_2}\delta.$$

For the second term, we obtain

$$\Big\| \sqrt{\frac{g\pi}{Z_\delta}} - \sqrt{\frac{g_\delta \pi}{Z_\delta}} \Big\|_{L^2} = \frac{1}{\sqrt{Z_\delta}} \Big(\int_{\mathbb{R}^d} \Big| \sqrt{g(x)} - \sqrt{g_\delta(x)} \Big|^2 \pi(x)\mathrm{d}x \Big)^{\frac{1}{2}} \leq \sqrt{\frac{K_1}{c_2}}\delta.$$

Therefore

$$d_{\mathrm{H}}(\pi^y, \pi^y_\delta) \leq \frac{1}{\sqrt{2}} \frac{c_1}{2c_2}\delta + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}}\delta = c\delta,$$

with $c = \frac{1}{\sqrt{2}} \frac{c_1}{2c_2} + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}}$ independent of δ. □

Notice that, together with (2), i.e., the inequality

$$\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| \leq 2 f_2 d_{\mathsf{H}}(\pi, \pi'), \quad f_2^2 := \mathbb{E}^{\pi}[f^2] + \mathbb{E}^{\pi'}[f^2],$$

this theorem guarantees that expectations computed with respect to π^y and π^y_δ are in the order of δ apart.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Tenth lecture, June 19, 2023

Today's lecture

- Sampling from multivariate Gaussian distributions, inverse transform sampling
- Prior modeling
- The linear Gaussian setting
- Numerical example

Change of variables

Consider two random variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ which are related via the formula

$$y = f(x),$$

where f is continuously differentiable and one-to-one (these conditions can be relaxed).

Then, for any $B \in \mathcal{B}(\mathbb{R}^n)$, it holds that

$$\mathbb{P}(x \in B) = \mathbb{P}(y \in f(B)) = \int_{f(B)} \pi_y(y)\,\mathrm{d}y = \int_B \pi_y(f(x))|\det Df(x)|\,\mathrm{d}x,$$

where $Df(x) \in \mathbb{R}^{n\times n}$ is the *Jacobian matrix* of f. In consequence

$$\pi_x(x) = \pi_y(f(x))|\det Df(x)|.$$

Sampling from Gaussian distributions

Suppose that we want to create a sample of realizations for a multivariate Gaussian random variable $x \sim \mathcal{N}(x_0, C)$, with the probability density

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

Since C^{-1} is (by assumption) symmetric and positive definite, it has a Cholesky decomposition

$$C^{-1} = R^{\mathrm{T}} R,$$

where R is an upper triangular matrix. The probability density of x can be alternatively written as

$$\pi_x(x) = \left(\frac{1}{(2\pi)^n \det C} \right)^{1/2} \exp\left(-\frac{1}{2}\|R(x - x_0)\|^2 \right).$$

Let us define a new random variable $w = R(x - x_0) \Leftrightarrow x = R^{-1}w + x_0$.

On the last slide, we defined $w = R(x - x_0) \Leftrightarrow x = R^{-1}w + x_0$, where $x \sim \mathcal{N}(x_0, C)$. The change of variables formula yields

$$\pi_w(w) = \pi_x(R^{-1}w + x_0)|\det R^{-1}| = \pi_x(R^{-1}w + x_0)|\det R|^{-1}.$$

Noting that

$$\frac{1}{\det C} = \det(C^{-1}) = \det R^{\mathrm{T}} \det R = \det(R)^2,$$

we obtain

$$\pi_w(w) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|w\|^2 \right).$$

In consequence, w is *Gaussian white noise*, i.e.,

$$w \sim \mathcal{N}(0, I).$$

Coloring transformation

Let $x \sim \mathcal{N}(x_0, C)$ (C p.d.) and $w \sim \mathcal{N}(0, I)$, with $C^{-1} = R^{\mathrm{T}} R$. Then it holds that

$$w = R(x - x_0) \quad \Leftrightarrow \quad x = R^{-1} w + x_0.$$

This is the basis of the *coloring transformation*:

1. Draw $w \in \mathbb{R}^n$ from $\mathcal{N}(0, I)$.
2. A realization of $x \in \mathbb{R}^n$ from $\mathcal{N}(x_0, C)$ can be obtained via

$$x = R^{-1} w + x_0.$$

Remark: The function `mvnrnd` can be used to draw from multivariate normal distributions in MATLAB (`numpy.random.multivariate_normal` in Python, respectively). However, if the Cholesky factor R (or its inverse) has been precomputed, it may be slightly more efficient to apply the coloring transformation, e.g., as

```
x = R \ randn(n,1) + x0;                                  (in MATLAB)
x = numpy.linalg.solve(R,numpy.random.normal(size=(n,1)))\
    + x0                                                  (in Python)
```

Sampling from general univariate distributions

In order to sample a real-valued random variable x directly, we can use its inverse distribution function. Let us assume that the probability density $\pi(x)$ of x is almost surely positive (this condition can be relaxed). Then, the *cumulative distribution function* $\Phi\colon \mathbb{R} \to (0,1)$ of x is defined by

$$\Phi(t) = \mathbb{P}(x < t) = \int_{-\infty}^{t} \pi(x)\,\mathrm{d}x.$$

In other words, Φ is the antiderivative of π. It follows from the fundamental theorem of calculus that Φ is strictly increasing. In particular, its inverse $\Phi^{-1}\colon (0,1) \to \mathbb{R}$ exists.

Now, we define a new random variable $u = \Phi(x)$. First, we observe that

$$\mathbb{P}(u < t) = \mathbb{P}(\Phi(x) < t) = \mathbb{P}(x < \Phi^{-1}(t))$$

for all $t \in (0, 1)$. However, by definition of the cumulative distribution function,

$$\begin{aligned}\mathbb{P}(x < \Phi^{-1}(t)) &= \int_{-\infty}^{\Phi^{-1}(t)} \pi(x)\,\mathrm{d}x = \int_{-\infty}^{\Phi^{-1}(t)} \Phi'(x)\,\mathrm{d}x \\ &= \Phi(\Phi^{-1}(t)) - \lim_{x \to -\infty} \Phi(x) = t.\end{aligned}$$

Hence $\mathbb{P}(u < t) = t$, meaning that $u \sim \mathcal{U}(0, 1)$ is distributed uniformly on the interval $[0, 1]$. On the other hand, if $u \sim \mathcal{U}(0, 1)$ is given, then we obtain a random variable x with density π by setting $x = \Phi^{-1}(u)$. This reduces drawing a sample from the distribution π to drawing a sample from a uniform distribution, which can for example be performed in MATLAB using the `rand` command (`numpy.random.uniform` in Python).

Inverse transform sampling ("Golden rule")

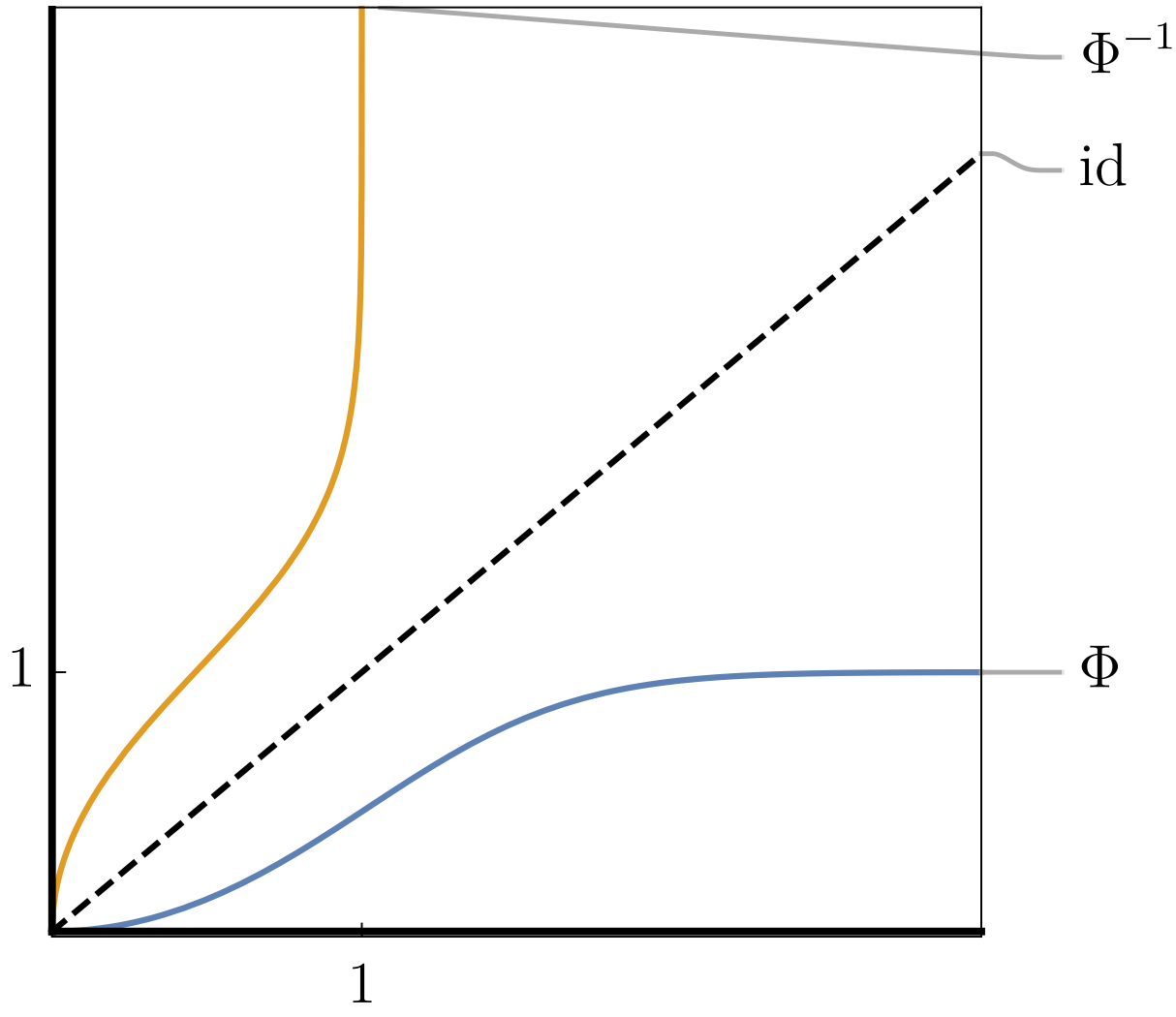
An algorithm for drawing from the density π with CDF Φ:

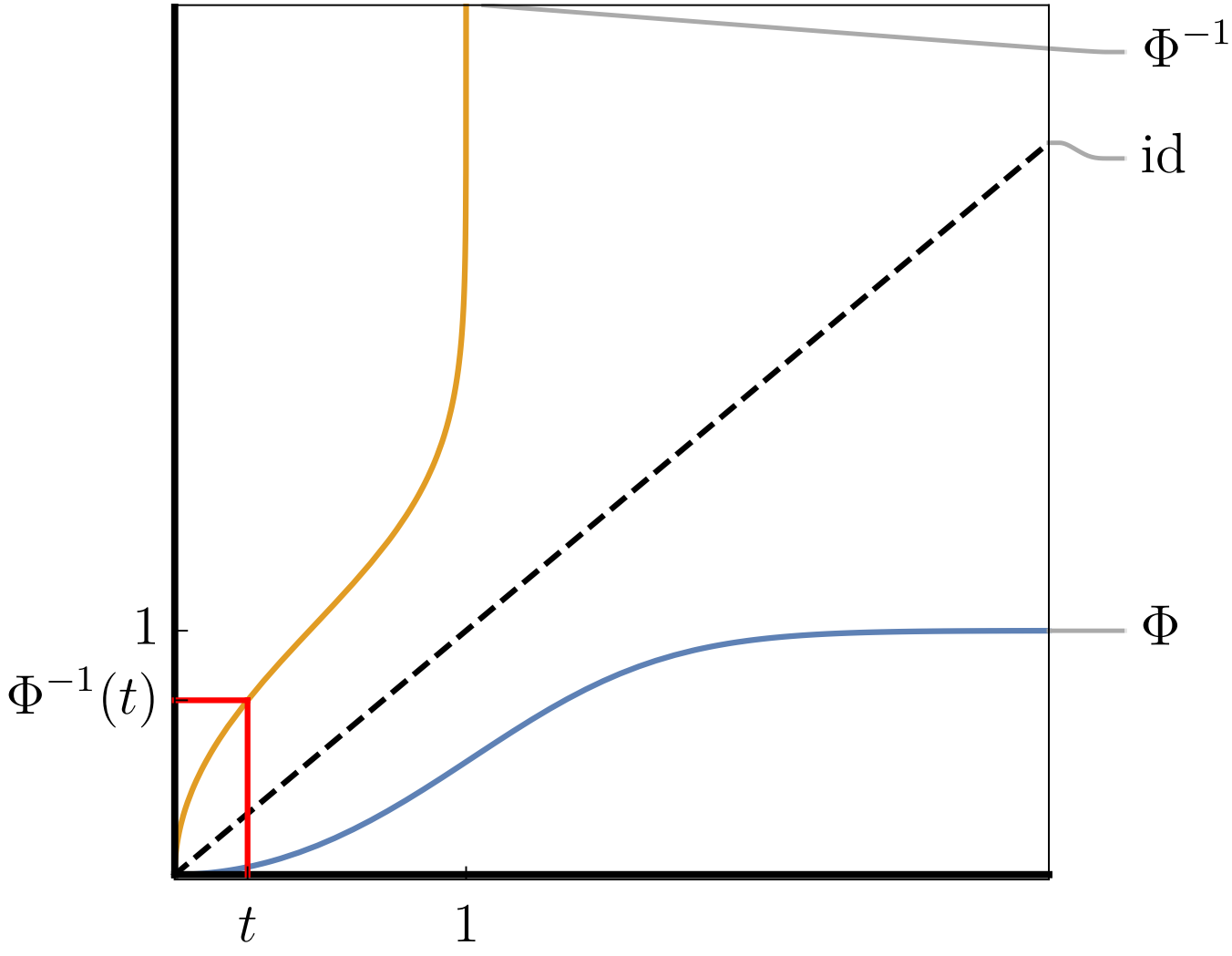
1. Draw $t \sim \mathcal{U}(0,1)$.
2. Calculate $x = \Phi^{-1}(t)$.

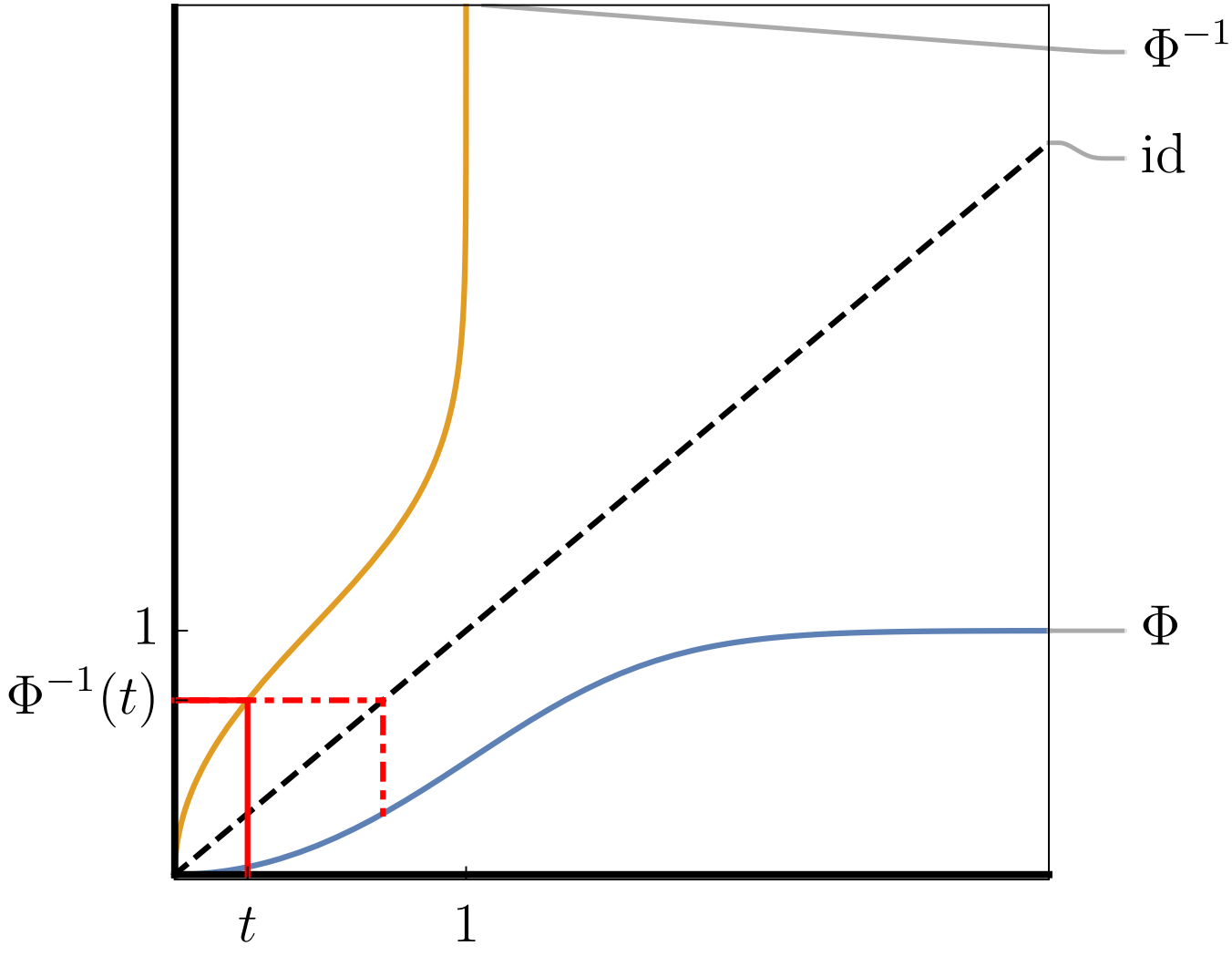
If a closed form expression for the inverse CDF is not available, then a computationally attractive formula for obtaining the value $\Phi^{-1}(t)$ at a point $t \in (0,1)$ is based on the identity

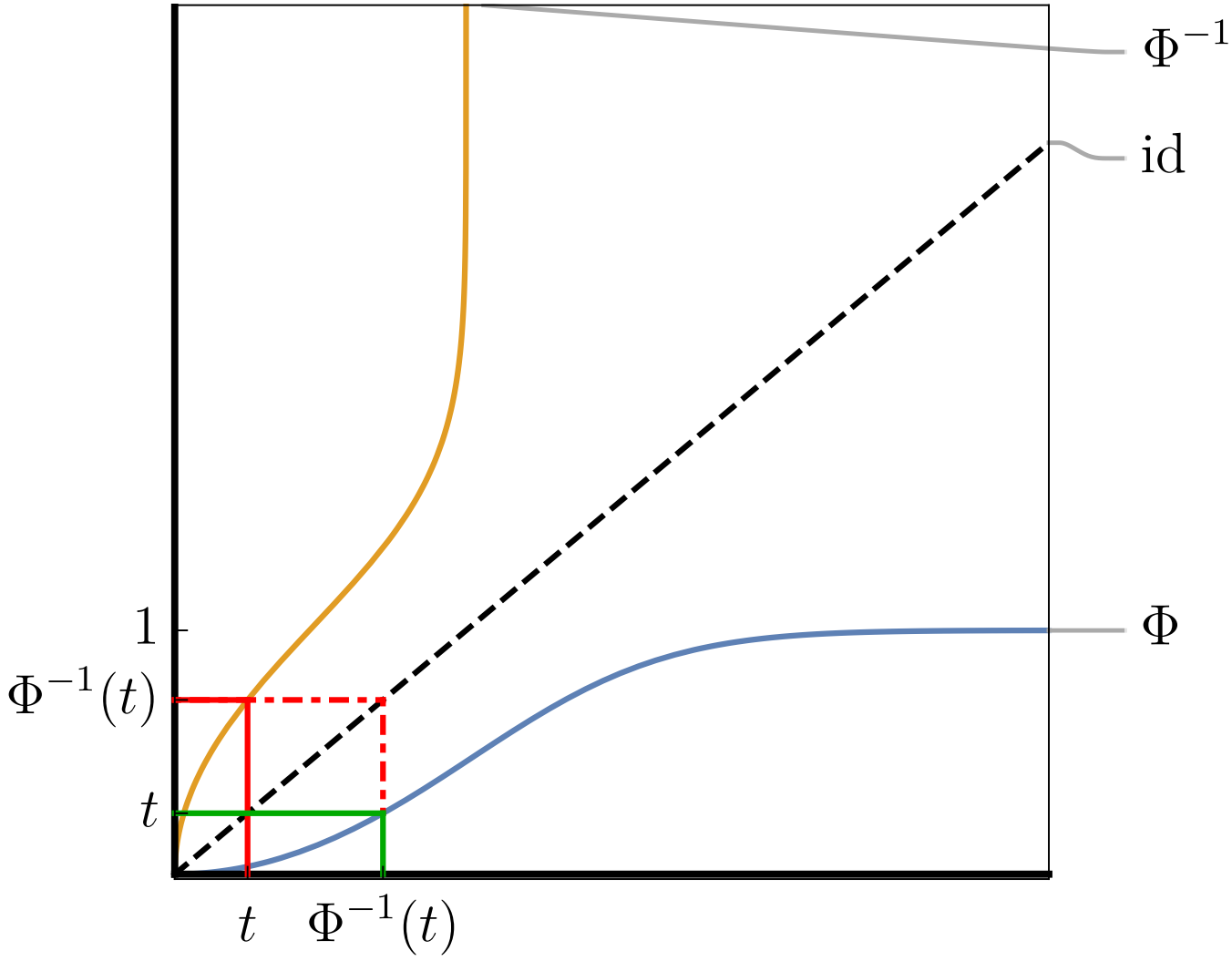
$$\Phi^{-1}(t) = \inf\{x \mid \Phi(x) \geq t\}.$$

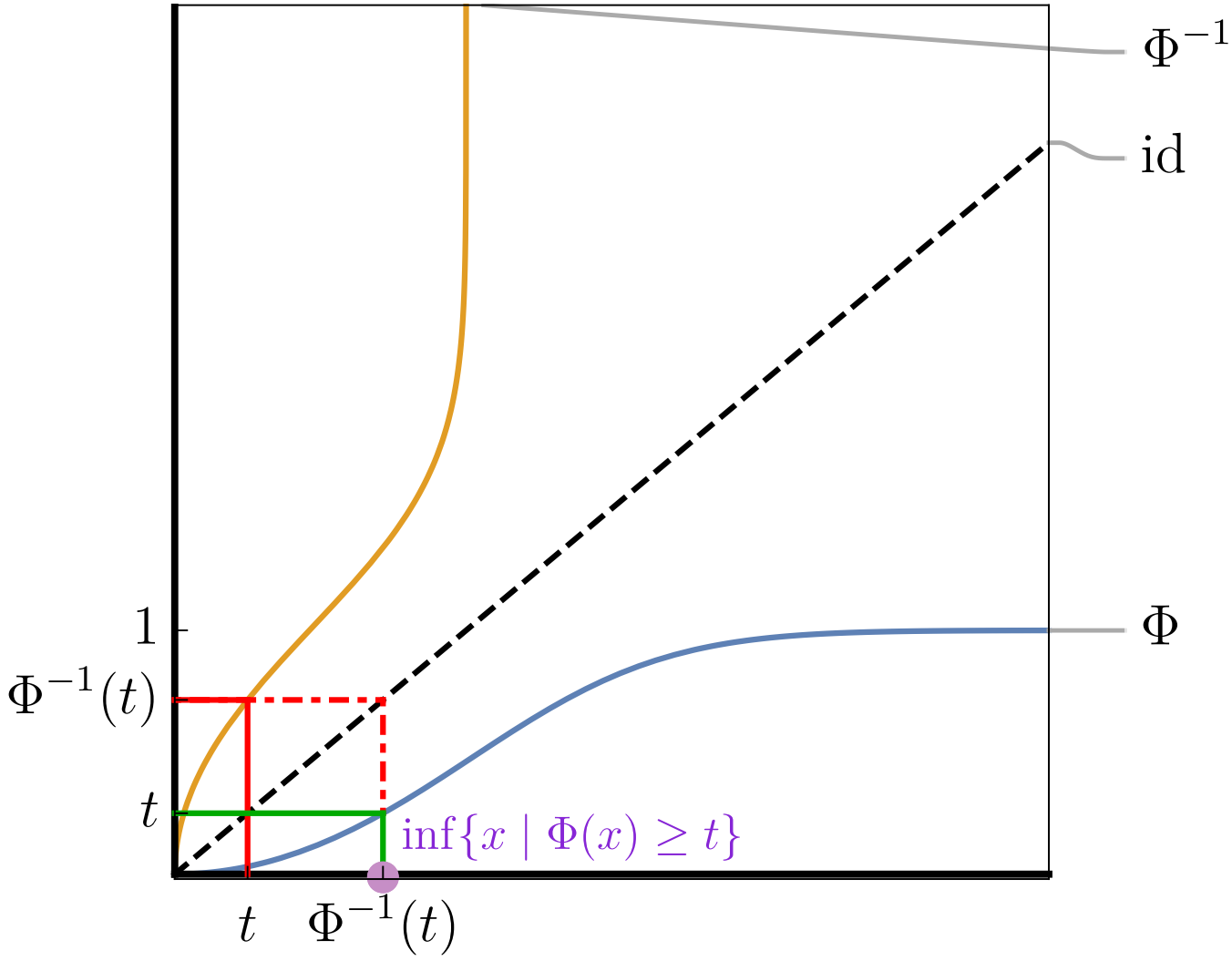
Remark: The above formula is the expression for the *generalized inverse CDF*: the formula with the infimum is valid even in the general case of weakly monotonic and right-continuous CDFs.


Φ^{-1}
id
Φ
1
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
t
$\Phi^{-1}(t)$



"Draw $t \sim \mathcal{U}(0,1)$ and find the smallest value of x such that $\Phi(x) \geq t$."

Remarks:

- The inverse transform sampling method can be used to sample univariate densities $\pi(u)$. However, if the components of a multivariate density are *mutually independent*, i.e., $\pi(u_1, \ldots, u_n) = \pi(u_1) \cdots \pi(u_n)$ holds a.e., then inverse transform sampling can be used to generate samples componentwise.
- Unfortunately, the components of multivariate posterior distributions are generally *not mutually independent*. In the next two weeks, we will discuss importance sampling and MCMC methods for sampling high-dimensional (posterior) distributions. These methods are applicable even when the components of multivariate distributions are **not** mutually independent.

Example

Suppose that we have the PDF $\pi(x) := (6x - 6x^2)\chi_{(0,1)}(x)$. We can design the following simple scheme based on inverse transform sampling to draw samples from this distribution.

MATLAB implementation:

```
n = 1e5; % sample size
x = linspace(0,1);
p = @(x) 6*x-6*x.^2; % PDF
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF of p
samples = [];
for iter = 1:n
  u = rand; % realization of U(0,1)
  ind = find(u <= P,1,'first'); % inverse CDF rule
  samples = [samples,x(ind)]; % store sample
end
histogram(samples,'Normalization','pdf'); % draw a histogram
hold on, plot(x,p(x),'LineWidth',3), legend('samples','pdf');
hold off;
```

Python implementation:

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,1,1000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
samples = []
for iter in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample
plt.hist(samples,bins='auto',
         density=True,label='samples') # draw a histogram
plt.plot(x,p(x),linewidth=2,label='pdf')
plt.legend()
plt.show()
# Thanks to Subodh Khanger for the Python implementation!
```

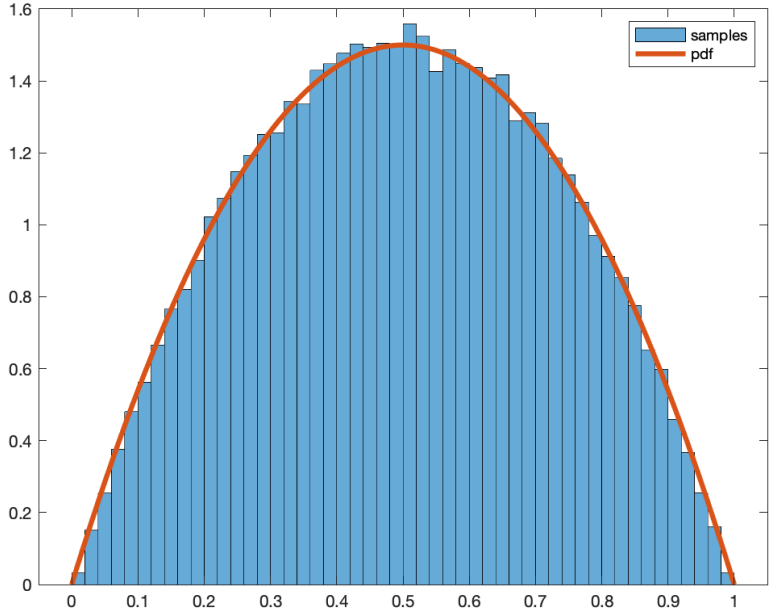


Figure: 10^5 samples drawn from the distribution given on the previous page organized as a histogram.

Prior modeling

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.

The prior probability distribution should be concentrated on those values of x we expect to see and assign a clearly higher probability to them than to the unexpected ones.

Gaussian priors

Gaussian densities

$$\pi(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}\|x-m\|^2_{C^{-1}}\right)$$

are the most used prior distribution in statistical inverse problems. They are easy to construct and form a versatile class of distributions. They also often lead to explicit estimators.

Random samples from a standard normal distribution $\mathcal{N}(0, I)$ can usually be generated directly, for example in MATLAB via `randn` or `numpy.random.normal` in Python. Samples from a general normal distribution $\mathcal{N}(m, C)$ and from a wide class of other distributions can then be derived from those, so that it is often not necessary to employ the inverse transform method.

$f_{9,0}$	$f_{9,1}$	$f_{9,2}$	$f_{9,3}$	$f_{9,4}$	$f_{9,5}$	$f_{9,6}$	$f_{9,7}$	$f_{9,8}$	$f_{9,9}$
$f_{8,0}$	$f_{8,1}$	$f_{8,2}$	$f_{8,3}$	$f_{8,4}$	$f_{8,5}$	$f_{8,6}$	$f_{8,7}$	$f_{8,8}$	$f_{8,9}$
$f_{7,0}$	$f_{7,1}$	$f_{7,2}$	$f_{7,3}$	$f_{7,4}$	$f_{7,5}$	$f_{7,6}$	$f_{7,7}$	$f_{7,8}$	$f_{7,9}$
$f_{6,0}$	$f_{6,1}$	$f_{6,2}$	$f_{6,3}$	$f_{6,4}$	$f_{6,5}$	$f_{6,6}$	$f_{6,7}$	$f_{6,8}$	$f_{6,9}$
$f_{5,0}$	$f_{5,1}$	$f_{5,2}$	$f_{5,3}$	$f_{5,4}$	$f_{5,5}$	$f_{5,6}$	$f_{5,7}$	$f_{5,8}$	$f_{5,9}$
$f_{4,0}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$	$f_{4,5}$	$f_{4,6}$	$f_{4,7}$	$f_{4,8}$	$f_{4,9}$
$f_{3,0}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{3,5}$	$f_{3,6}$	$f_{3,7}$	$f_{3,8}$	$f_{3,9}$
$f_{2,0}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{2,5}$	$f_{2,6}$	$f_{2,7}$	$f_{2,8}$	$f_{2,9}$
$f_{1,0}$	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	$f_{1,6}$	$f_{1,7}$	$f_{1,8}$	$f_{1,9}$
$f_{0,0}$	$f_{0,1}$	$f_{0,2}$	$f_{0,3}$	$f_{0,4}$	$f_{0,5}$	$f_{0,6}$	$f_{0,7}$	$f_{0,8}$	$f_{0,9}$

Let us consider an image. We divide this region into $n \times n$ pixels and label the pixels $f_{i,j}$ for $i, j \in \{0, \ldots, n-1\}$.

x_{90}	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	x_{99}
x_{80}	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	x_{88}	x_{89}
x_{70}	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	x_{77}	x_{78}	x_{79}
x_{60}	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}	x_{67}	x_{68}	x_{69}
x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

It is convenient to reshape the matrix/image $(f_{i,j})$ into a vector x of length $d = n^2$ so that

$$x_{in+j} = f_{i,j}, \quad i, j \in \{0, \ldots, n-1\}.$$

The image on the left illustrates the new numbering corresponding to the pixels.

Note that `x = f.reshape((n*n,1))` and `f = x.reshape((n,n))`.
(In MATLAB: `x = f(:)` and `f = reshape(x,n,n)`.)

As an example, consider a problem where the unknown is a two-dimensional pixel image, arranged as a vector $x \in \mathbb{R}^d$. The components x_j represent the intensity of the j^{th} pixel. Since we consider images it is natural to add a positivity constraint to our prior. Assuming that x_i and x_j are independent for $i \neq j$, the Gaussian white noise density with positivity constraint is

$$\pi(x) \propto \chi_+(x) \exp\left(-\frac{1}{2\alpha^2}\|x\|^2\right),$$

where $\chi_+(x) = 1$ if $x_j > 0$ for all j and $\chi_+(x) = 0$ otherwise.

Since we assumed that each component is independent of the others, random draws can be performed componentwise.

Impulse priors

We assume again that the unknown is a two-dimensional pixel image.

Assume that our prior information is that the image contains small and well localized objects in an almost constant background.

In such a case we could assume an impulse prior density, which means that it gives a low average amplitude but allows outliers. The tail of such a prior distribution is long, although the expected value is small.

Let $x \in \mathbb{R}^d$ represent the pixel image, where the component x_j is the intensity of the j^{th} pixel. In what follows, x_i and x_j are assumed to be independent for $i \neq j$.

One example of an impulse prior is the ℓ^1 prior. It has the density

$$\pi(x) = \left(\frac{\alpha}{2}\right)^d \exp(-\alpha \|x\|_1)$$

with $\alpha > 0$, where the ℓ^1-norm is defined as

$$\|x\|_1 = \sum_{j=1}^{d} |x_j|.$$

The impulse effect can be enhanced by choosing an even smaller power $p \in (0, 1)$ of the components of x, that is, using $\sum_{j=1}^{d} |x_j|^p$ instead of the ℓ^1-norm.

Another choice that produces images with few distinctly different pixels and a low-amplitude background is the *Cauchy density*

$$\pi(x) = \left(\frac{\alpha}{\pi}\right)^n \prod_{j=1}^{n} \frac{1}{1 + \alpha^2 x_j^2}$$

with $\alpha > 0$.

Since we consider images we add a positivity constraint to our prior. For the ℓ^1 prior, we set

$$\pi(x) = \alpha^d \chi_+(x) \exp(-\alpha \|x\|_1),$$

where $\chi_+(x) = 1$ if $x_j > 0$ for all j and $\chi_+(x) = 0$ otherwise. The components x_j are independent and each have the cumulative distribution function

$$\Phi(t) = \alpha \int_0^t \mathrm{e}^{-\alpha s} \, \mathrm{d}s = 1 - \mathrm{e}^{-\alpha t} \quad \text{for all } t \geq 0.$$

Now, we can draw samples of x_j using

$$x_j = \Phi^{-1}(u_j) = -\frac{1}{\alpha} \ln(1 - u_j),$$

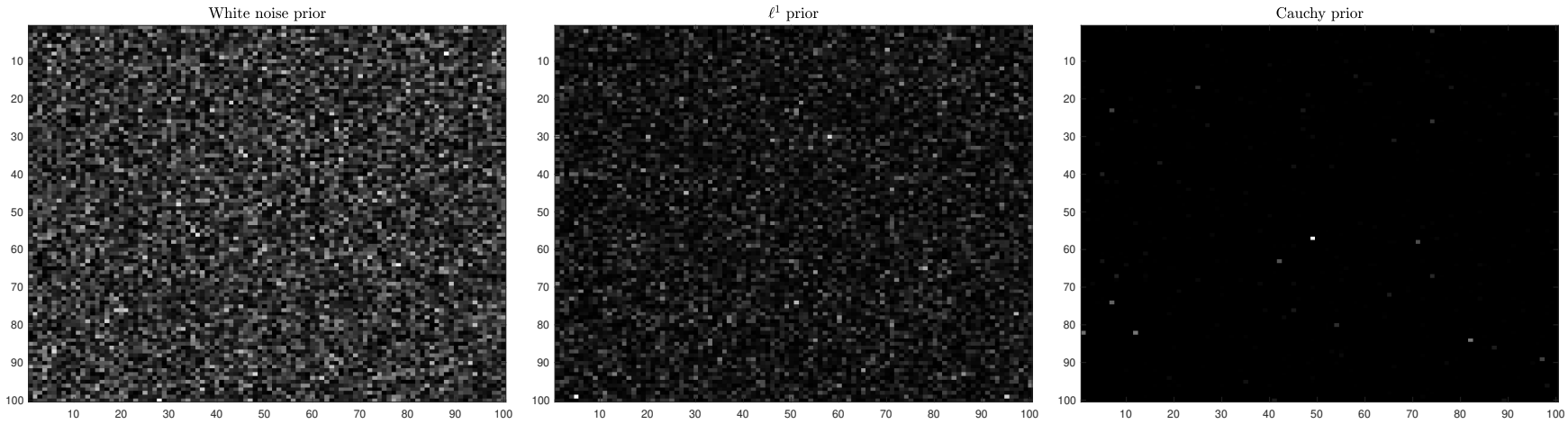
where the u_j are independent random draws from the uniform distribution $\mathcal{U}(0, 1)$.

Similarly, the components x_j of the Cauchy prior with positivity constraint are independent and have the CDF

$$\Phi(t) = \frac{2\alpha}{\pi} \int_0^t \frac{1}{1 + \alpha^2 s^2} \, \mathrm{d}s = \frac{2}{\pi} \arctan \alpha t,$$

so that the inverse cumulative distribution is $\Phi^{-1}(t) = \frac{1}{\alpha} \tan\left(\frac{\pi t}{2}\right)$.

Random draws from the white noise prior with positivity constraint, the impulse (ℓ^1) prior, and the Cauchy prior:



Note that as along as all components are independent, drawing can be done componentwise using inverse transform sampling. Here, for each pixel x_j, we draw t_j from $\mathcal{U}(0,1)$ and calculate $x_j = \Phi^{-1}(t_j)$.

Discontinuities

Assume that we want to estimate a one-dimensional signal $f\colon [0,1] \to \mathbb{R}$ with $f(0) = 0$ from indirect observations. Our prior knowledge is that the signal is usually relatively stable but can have large jumps every now and then. We may also have information on the size of the jumps or the rate of their occurrence.

We obtain one possible prior by taking the finite difference approximation of the derivative of f and assigning an impulsive noise distribution to it. Let us discretize the interval $[0,1]$ by points $t_j = j/d$ and write $x_j = f(t_j)$. Consider the density

$$\pi(x) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1 + \alpha^2 (x_j - x_{j-1})^2}.$$

To draw samples from the above distribution we define new random variables for the jumps

$$u_j = x_j - x_{j-1}, \quad j = 1, \ldots, d.$$

These each have the density

$$\pi(u) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2 u_j^2}.$$

In particular, the u_j are independent from each other, so that they can be drawn from a one-dimensional Cauchy density. Also note that $x = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ satisfies $x = Lu$, where $L \in \mathbb{R}^{d \times d}$ is a lower triangular matrix with $L_{ij} = 1$ for $i \geq j$.[†] Generalizing the idea behind the above prior leads, e.g., to total variation priors.

[†]Note that in MATLAB, it is more efficient to implement this as `x = cumsum(u)` (similarly `x = numpy.cumsum(u)` in Python).

Hierarchical models

The prior density may depend on some parameter, such as variance or mean. So far we have assumed that these parameters are known. However, we often do not know how to choose them. If a parameter is not known, it can be estimated as a part of the statistical inference problem 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 assumed to be unknown. We then write the prior as a conditional density

$$\mathbb{P}(x|\alpha).$$

We model the unknown α with a *hyperprior* $\mathbb{P}(\alpha) = \pi_{\mathsf{h}}(\alpha)$ and write the joint distribution of x and α as

$$\mathbb{P}(x, \alpha) = \mathbb{P}(x|\alpha)\,\mathbb{P}(\alpha).$$

Assuming we have a likelihood model $\mathbb{P}(y|x)$ for the measurement y, we get the posterior density for x and α, given y, using Bayes' formula

$$\mathbb{P}(x, \alpha|y) \propto \mathbb{P}(y|x, \alpha)\, \mathbb{P}(x, \alpha) = \mathbb{P}(y|x, \alpha)\, \mathbb{P}(x|\alpha)\, \mathbb{P}(\alpha).$$

The hyperprior density π_h may again depend on some hyperparameter α_0. 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.

The linear Gaussian setting

In this chapter we study the linear Gaussian setting, where the forward map F is linear and both the prior distribution and the distribution of the observational noise η are Gaussian.

For several reasons, it plays a central role in the study of inverse problems.

It arises frequently in applications, either directly or in the form of posterior distributions that are asymptotically Gaussian in the large data limit. It also allows computing explicit solutions which can be used to gain a general understanding. Apart from that, many methods employed in a nonlinear or non-Gaussian setting build on ideas from the linear Gaussian case by performing linearization or Gaussian approximation.

Let us suppose that the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

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

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \pi = \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is symmetric and positive definite.
3. The noise is Gaussian: $\eta \sim \nu = \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, where Γ_{n} is symmetric and positive definite.
4. x and η are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) is Gaussian with $x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}})$, *where we have the posterior mean*

$$\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$$

and covariance

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}.$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$= \exp\Bigg(-\frac{1}{2}\big(y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$- x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$- \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$+ x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 + x_0^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0\big)\Bigg)$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\quad -x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y$$

$$-x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$

$$+x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0^{\mathrm{T}}$$

$$+x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \qquad \Big)\Bigg)$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\quad -x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y$$

$$-x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$

$$+x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0^{\mathrm{T}}$$

$$+x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \quad \Big)\Bigg)$$

$$= \exp\left(-\frac{1}{2}(x^{\mathrm{T}}\underbrace{(\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)}_{=\Gamma_{\mathrm{post}}^{-1}}x - 2x^{\mathrm{T}}\underbrace{(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)}_{=\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}})\right).$$

On the previous slide, we arrived at

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right).$$

To finish the proof, we "complete the square" by multiplying and dividing by $\exp(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}})$. Since this term does not depend on x, we can absorb the denominator into the implied coefficient to obtain

$$\begin{aligned}
\pi^y(x) &\propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \exp\left(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right) \\
&= \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} + \mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}}) + 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}})\right)\right),
\end{aligned}$$

as desired. □

Remark: The previous proof shows that if $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$ and $\eta \sim \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, then

$$x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}}),$$

where

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1} \tag{2}$$

and

$$\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0). \tag{3}$$

One also has the following alternative representations for the posterior mean

$$\mu_{\mathrm{post}} = x_0 + \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{n}})^{-1}(y - Ax_0 - \eta_0) \tag{4}$$

and the posterior covariance

$$\Gamma_{\mathrm{post}} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{n}})^{-1}A\Gamma_{\mathrm{pr}}. \tag{5}$$

Formula (5) can be proved, e.g., by using the Sherman–Morrison–Woodbury formula on (2). Formula (4) can be proved by plugging the formula (5) into (3) and simplifying the expression (homework).

As the posterior distribution is Gaussian, its mean and its mode coincide. This means that the conditional mean estimator and the MAP estimator coincide in the linear Gaussian setting.

Corollary

The conditional mean estimator and the maximum a posteriori estimator coincide in the linear Gaussian setting and are given by
$\hat{x}_{\mathrm{CM}} = \hat{x}_{\mathrm{MAP}} = \mu_{\mathrm{post}}$.

Example

Let $\Gamma_{\mathrm{n}} = \gamma^2 I$, $\eta_0 = 0$, $\Gamma_{\mathrm{pr}} = \sigma^2 I$, $x_0 = 0$, and set $\lambda = \frac{\gamma^2}{\sigma^2}$. Then μ_{post} minimizes

$$J_\lambda(x) := \|y - Ax\|^2 + \lambda \|x\|^2.$$

and therefore satisfies

$$(A^{\mathrm{T}} A + \lambda I)\mu_{\mathrm{post}} = A^{\mathrm{T}} y. \tag{6}$$

This example provides a connection between Bayesian inference and variational regularization: J_λ can be interpreted as the objective functional in a linear regression model with a regularization term $\lambda\|x\|^2$. Equation (6) for μ_{post} is then exactly the normal equation. In the general case, equation $\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$ can thus be viewed as a generalized normal equation. This point of view helps to understand the structure of Bayesian regularization by linking it to well-understood optimization approaches for inverse problems.

Numerical example: one-dimensional deconvolution

Let us revisit the deconvolution example from last week: we are interested in estimating a signal $f : [0, 1] \to \mathbb{R}$ from noisy, blurred observations modeled as

$$y_i = y(s_i) = \int_0^1 K(s_i, t) f(t) \, \mathrm{d}t + \eta_i, \quad i \in \{1, \ldots, k\},$$

where the blurring kernel is

$$K(s, t) = \exp\left(-\frac{1}{2\omega^2}(s - t)^2 \right), \quad \omega = 0.5,$$

and we have Gaussian measurement noise $\eta \sim \mathcal{N}(\eta_0, \Gamma_{\text{noise}})$ with a symmetric, positive definite covariance matrix Γ_{noise}.

If $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \ldots, k\}$ and we discretize the integral using the midpoint rule with $t_j = \frac{j}{d} - \frac{1}{2d}$ and $x_j = f(t_j)$ for $j \in \{1, \ldots, d\}$, then we have the discrete linear model

$$y = Ax + \eta, \quad \text{where } A_{i,j} = \frac{1}{d} K(s_i, t_j).$$

Linear Gaussian setting

Suppose that we set a Gaussian prior for the unknown $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is a symmetric, positive definite covariance matrix.

Now the posterior probability density of x given the measurement y is

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(x-\overline{x})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\overline{x})\right),$$

where we have the posterior mean

$$\overline{x} = x_0 + \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}})^{-1}(y - Ax_0 - \eta_0)$$

and posterior covariance

$$\Gamma_{\mathrm{post}} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}})^{-1}A\Gamma_{\mathrm{pr}}.$$

With additive noise $\eta \sim \nu(\eta) = \mathcal{N}(\eta_0, \sigma^2 I)$, we have the likelihood

$$\mathbb{P}(y|x) = \nu(y - Ax) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax - \eta_0\|^2 \right).$$

Let $L = \mathrm{tridiag}(-1, 2, -1)$ and consider the following priors

$$\pi_{\mathrm{pr},1}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|x - x_0\|^2 \right) \quad \text{with covariance } \Gamma_{\mathrm{pr},1} = \gamma^2 I;$$

$$\pi_{\mathrm{pr},2}(F) \propto \exp\left(-\frac{1}{2\gamma^2}\|L(x - x_0)\|^2 \right)$$

$$= \exp\left(-\frac{1}{2\gamma^2}(x - x_0)^{\mathrm{T}}(L^{\mathrm{T}}L)(x - x_0) \right) \quad \text{with covariance } \Gamma_{\mathrm{pr},2} = \gamma^2(L^{\mathrm{T}}L)^{-1},$$

where $x_0 \in \mathbb{R}^d$ is the prior mean (assumed to be the same in both cases). Hence (from the previous page)

$$\overline{x_j} = x_0 + \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1}(y - Ax_0 - \eta_0),$$

$$\Gamma_{\mathrm{post},j} = \Gamma_{\mathrm{pr},j} - \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1} A \Gamma_{\mathrm{pr},j},$$

where $G_j = A\Gamma_{\mathrm{pr},j}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}}$ and $\Gamma_{\mathrm{noise}} = \sigma^2 I$.

For the numerical experiment, we simulate measurements using the (smooth) ground truth signal

$$f(t) = 8t^3 - 16t^2 + 8t,$$

which satisfies $f(0) = f(1) = 0$. The measurements are contaminated with zero-mean 10% *relative* noise ($\sigma \approx 0.0618$) and we set $d = k = 120$.

Remark: When we simulate the measurement data, it is important to avoid the *inverse crime*. One way to do this is to generate the measurement data using a denser grid and then interpolate the forward solution onto a coarser computational grid, which is actually used to compute the reconstruction.

Since both the prior and the posterior are now Gaussian, we can use the coloring transformation to draw samples from the prior and posterior.

See the scripts `deconv.m` / `deconv.py` on the course webpage!

A note on marginal Gaussian distributions

Let

$$\pi(x) \propto \exp(-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Gamma^{-1}(x-\mu))$$

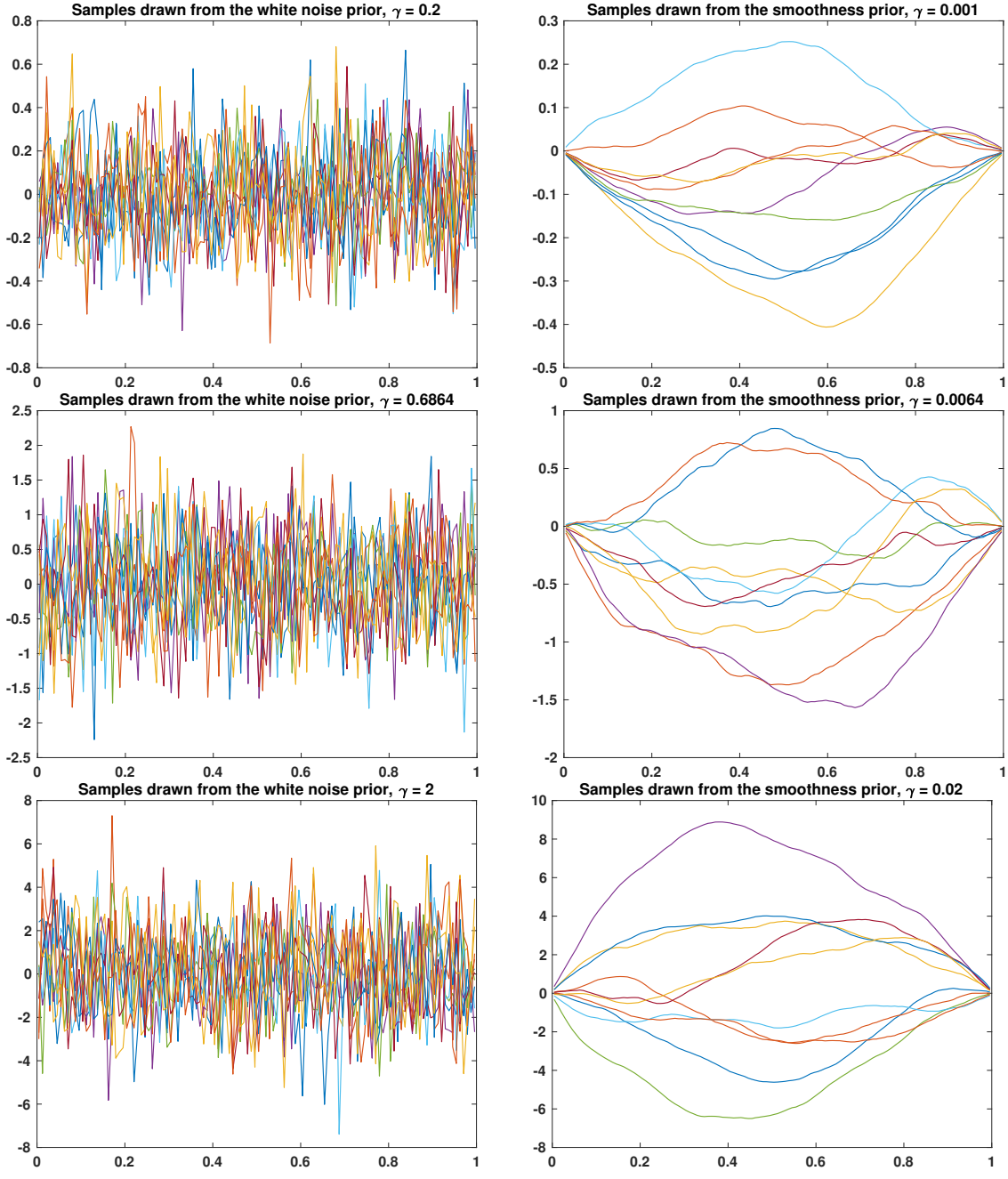
be a multivariate Gaussian PDF with mean μ and positive definite and symmetric covariance matrix Γ.

Q: What is Γ_{ii}?
A: $\sigma_i^2 := \Gamma_{ii}$ is the variance of the marginal distribution with PDF

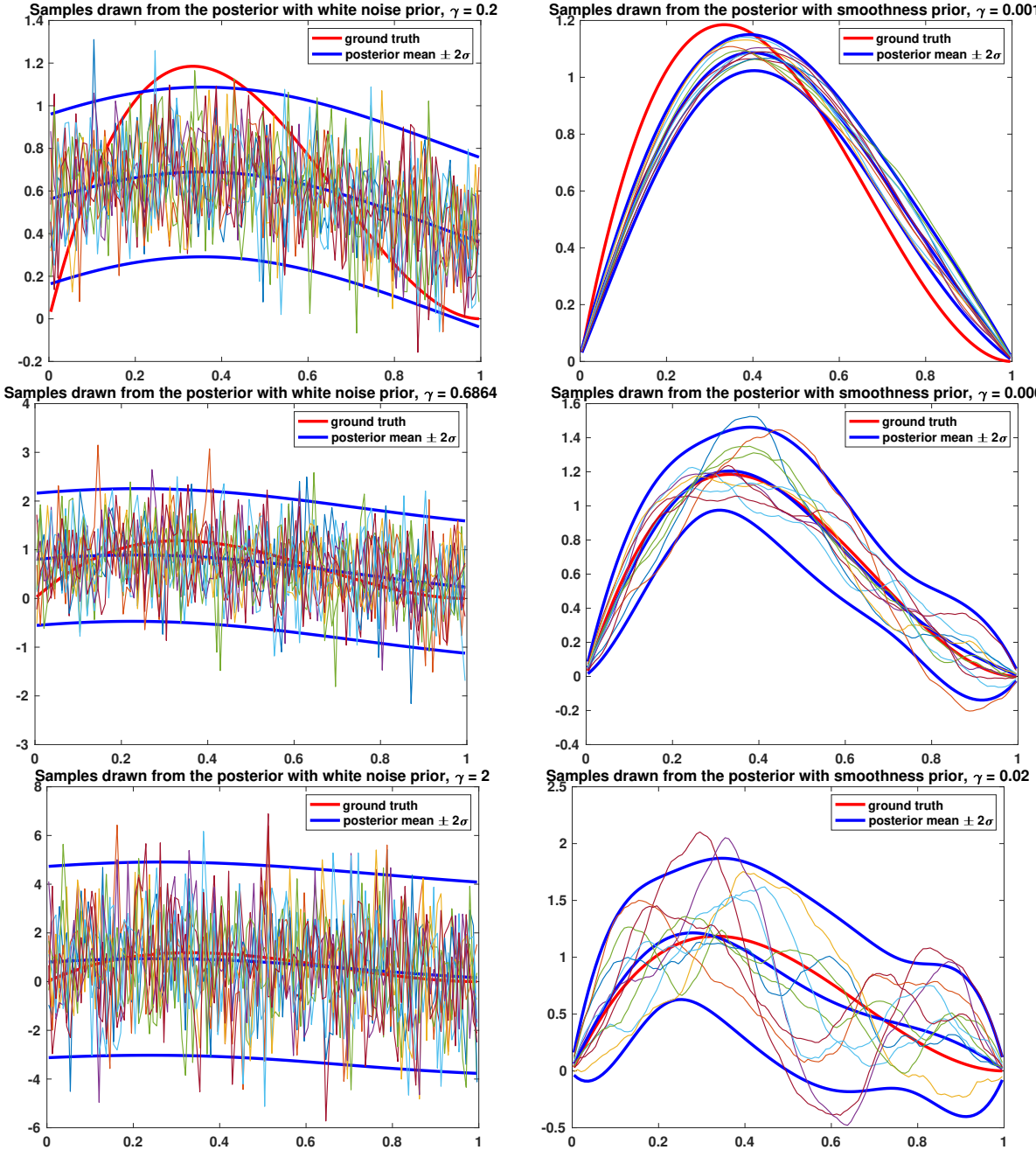
$$\pi(x_i) = \int_{\mathbb{R}^{n-1}} \pi(x_1, \ldots, x_i, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n,$$

which is itself a (univariate) Gaussian PDF with mean μ_i.

This is why we can obtain the credibility envelopes by taking the square roots of the diagonal values of $\Gamma_{\mathrm{post},j}$.



Samples drawn from the white noise prior and the smoothness prior for several values of γ.



Samples drawn from the posterior corresponding to both the white noise prior and the smoothness prior for several values of γ. We also plot the ground truth solution and the posterior mean.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eleventh lecture, June 26, 2023

Recap: the linear Gaussian setting

Let the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

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

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \pi = \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is symmetric and positive definite.
3. The noise is Gaussian: $\eta \sim \nu = \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, where Γ_{n} is symmetric and positive definite.
4. x and η are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) *is Gaussian with* $x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}})$, *where we have*

$$\mu_{\mathrm{post}} = (A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A + \Gamma_{\mathrm{pr}}^{-1})^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0),$$
$$\Gamma_{\mathrm{post}} = (A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A + \Gamma_{\mathrm{pr}}^{-1})^{-1}.$$

Small noise limit of the posterior distribution

Now we assume that the observational noise has the distribution $\eta \sim \mathcal{N}(0, \gamma^2 \Gamma_0)$ with $\gamma > 0$ and Γ_0 is a fixed symmetric and positive definite matrix, and consider the limiting behavior of the posterior mean and covariance as $\gamma \to 0$.

Substituting $\Gamma_{\mathrm{n}} = \gamma^2 \Gamma_0$ in the expressions for the posterior mean and covariance yield

$$m(\gamma) := \left(A^{\mathrm{T}} \Gamma_0^{-1} A + \gamma^2 \Gamma_{\mathrm{pr}}^{-1}\right)^{-1} \left(A^{\mathrm{T}} \Gamma_0^{-1} y + \gamma^2 \Gamma_{\mathrm{pr}}^{-1} x_0\right), \tag{2}$$

$$C(\gamma) := \gamma^2 \left(A^{\mathrm{T}} \Gamma_0^{-1} A + \gamma^2 \Gamma_{\mathrm{pr}}^{-1}\right)^{-1}. \tag{3}$$

We distinguish between overdetermined, determined, and underdetermined problems.

Overdetermined and determined case

Recall that $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$.

Theorem (Overdetermined and determined case)

Suppose in the linear Gaussian setting that $\Gamma_{\mathrm{n}} = \gamma^2 \Gamma_0$ *with* $\gamma > 0$, *and that* $\mathrm{Ker}(A) = \{0\}$.

1. *If* $d < k$, *then the posterior distribution* π^y *satisfies*

$$\pi^y \rightharpoonup \delta_{m^\dagger} \quad \text{as } \gamma \to 0,$$

where $m^\dagger$ *is the solution to the least squares problem*

$$m^\dagger = \arg\min_{u \in \mathbb{R}^d} \left\| \Gamma_0^{-\frac{1}{2}} (Au - y) \right\|^2.$$

2. *If* $d = k$, *then we have*

$$\pi^y \rightharpoonup \delta_{A^{-1}y} \quad \text{as } \gamma \to 0.$$

Proof. **(i):** As A has a trivial null space, $Au \neq 0$, and thus

$$(u, A^{\mathrm{T}}\Gamma_0^{-1}Au) = (Au, \Gamma_0^{-1}Au) > 0$$

for all $u \in \mathbb{R}^d \setminus \{0\}$. Therefore, the matrix $A^{\mathrm{T}}\Gamma_0^{-1}A$ is invertible. Now we can take γ to zero in (2) and (3) and get

$$m(\gamma) = \left(A^{\mathrm{T}}\Gamma_0^{-1}A + \gamma^2\Gamma_{\mathrm{pr}}^{-1}\right)^{-1}\left(A^{\mathrm{T}}\Gamma_0^{-1}y + \gamma^2\Gamma_{\mathrm{pr}}^{-1}m_0\right) \overset{\gamma\to 0}{\to} (A^{\mathrm{T}}\Gamma_0^{-1}A)^{-1}A^{\mathrm{T}}\Gamma_0^{-1}y =: m^*$$

as well as $C(\gamma) = \gamma^2\left(A^{\mathrm{T}}\Gamma_0^{-1}A + \gamma^2\Gamma_{\mathrm{pr}}^{-1}\right)^{-1} \overset{\gamma\to 0}{\to} 0$. This shows that $\pi^y = \mathcal{N}(m, C) \rightharpoonup \mathcal{N}(m^*, 0) = \delta_{m^*}$.

Due to the trivial null space of A, the minimizer $m^\dagger$ of

$$\left\|\Gamma_0^{-\frac{1}{2}}(Au - y)\right\|^2$$

is the unique solution to the normal equation

$$A^{\mathrm{T}}\Gamma_0^{-1}Am^\dagger = A^{\mathrm{T}}\Gamma_0^{-1}y,$$

which shows that $m^* = m^\dagger$.

(ii): As in part (i), we have $m(\gamma) \to m^*$ and $C(\gamma) \to 0$. Since A is now invertible, we obtain

$$m^* = \left(A^{-1}\Gamma_0(A^{\mathrm{T}})^{-1}\right)A^{\mathrm{T}}\Gamma_0^{-1}y = A^{-1}y. \quad \square$$

Reminder: singular value decomposition (SVD)

Let $A \in \mathbb{R}^{k \times d}$ be *any* matrix. Then we can *always* write

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

where $U \in \mathbb{R}^{k \times k}$, $\Lambda \in \mathbb{R}^{k \times d}$, and $V \in \mathbb{R}^{d \times d}$ are matrices such that

$$UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_k \quad \text{and} \quad VV^{\mathrm{T}} = V^{\mathrm{T}}V = I_d \qquad (U \text{ and } V \text{ are } \textit{orthogonal} \text{ matrices})$$

and

$$\Lambda = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & O_{k \times (d-k)} \\ & & \sigma_k & \end{array}\right) \quad \text{if } k < d,$$

$$\Lambda = \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \\ \hline \multicolumn{3}{c}{O_{(k-d) \times d}} \end{array}\right) \quad \text{if } k > d,$$

and $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ if $k = d$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{k,d\}} \geq 0$ are called the *singular values* of matrix A.

Underdetermined case

Both in the overdetermined and the determined case, the small noise limit of the posterior distribution is a Dirac distribution. Note that the prior plays no role in the limit.

This case is of particular relevance because practical inverse problems are usually underdetermined. Here, we assume that the matrix $A \in \mathbb{R}^{k\times d}$ has $\mathrm{Rank}(A) = k < d$ and write

$$A \overset{(*)}{=} \begin{pmatrix} A_1 & 0 \end{pmatrix} Q^{\mathrm{T}} = \begin{pmatrix} A_1 & 0 \end{pmatrix} \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}^{\mathrm{T}} = A_1 Q_1^{\mathrm{T}} \tag{4}$$

with an invertible matrix $A_1 \in \mathbb{R}^{k\times k}$ and an orthogonal matrix $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \in \mathbb{R}^{d\times d}$ (i.e., $Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I_d$).

To get an idea of what is going on in the underdetermined case, we first consider a basic example.

$^{(*)}$ To see this, consider the SVD $A = U\Lambda V^{\mathrm{T}}$. Since $k < d$, we have $\Lambda =: \begin{pmatrix} \Lambda_1 & 0 \end{pmatrix}$ with $\Lambda_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$; thus $A = U\Lambda V^{\mathrm{T}} = U \begin{pmatrix} \Lambda_1 & 0 \end{pmatrix} V^{\mathrm{T}} = \begin{pmatrix} U\Lambda_1 & 0 \end{pmatrix} V^{\mathrm{T}}$. Finally, define $A_1 := U\Lambda_1$ (invertible) and $Q := V$ (orthogonal).

Example. Assume that $A = \begin{pmatrix} A_1 & 0 \end{pmatrix}$, $\eta \sim \mathcal{N}(0, \gamma^2 I_k)$, and $x \sim \mathcal{N}(0, I_d)$. Let

$$x =: \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

with $x_1 \in \mathbb{R}^k$ and $x_2 \in \mathbb{R}^{d-k}$. Then, the data satisfies

$$y = Ax + \eta = A_1 x_1 + \eta.$$

The posterior density is given by $\pi^y(x) = \frac{1}{Z} \exp(-J(x))$, where

$$\begin{aligned} J(x) &= \frac{1}{2\gamma^2} \|y - A_1 x_1\|^2 + \frac{1}{2} \|x\|^2 \\ &= \left(\frac{1}{2\gamma^2} \|y - A_1 x_1\|^2 + \frac{1}{2} \|x_1\|^2 \right) + \frac{1}{2} \|x_2\|^2, \end{aligned}$$

and Z is a normalization constant.

We can write it as a product

$$\pi^y(x_1, x_2) = \frac{1}{\tilde{Z}}\nu(y - A_1x_1)\pi_1(x_1) \cdot \pi_2(x_2) =: \pi_1^y(x_1)\pi_2(x_2)$$

where $\pi_1(x_1) = \mathcal{N}(0, I_k)$ and $\pi_2(x_2) = \mathcal{N}(0, I_{d-k})$ are Gaussian densities. We can interpret the factor $\frac{1}{\tilde{Z}}\nu(y - A_1x_1)\pi_1(x_1)$ as posterior density π_1^y resulting from the determined problem $y = A_1x_1 + \eta$ with prior density $x_1 \sim \pi_1$. By the small noise limit in the determined case, we know that $\pi_1^y \rightharpoonup \delta_{A_1^{-1}y}$ as $\gamma \to 0$, whereas π_2 remains constant. Since x_1 and x_2 are independent, we would expect the posterior distribution to converge weakly towards

$$\pi^y(x_1, x_2) \rightharpoonup \delta_{A_1^{-1}y}(x_1)\pi_2(x_2).$$

This means that in the limit, the data determines the posterior distribution on a subspace of dimension k, whereas uncertainty remains in a subspace of dimension $d - k$.

In order to generalize these observations, we need the following decomposition of the identity.

Lemma

Let $\Gamma_{\mathrm{pr}} \in \mathbb{R}^{d\times d}$ *be symmetric and positive definite and* $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ *an orthogonal matrix with* $Q_1 \in \mathbb{R}^{d\times k}$, $Q_2 \in \mathbb{R}^{d\times(d-k)}$. *Then we have*

$$I_d = \Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1} Q_1^{\mathrm{T}} + Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1}. \tag{5}$$

Proof. Let R denote the right-hand side of (5). Since Q is orthogonal, we have $Q_1^{\mathrm{T}} Q_2 = Q_2^{\mathrm{T}} Q_1 = 0$, and thus

$$Q_1^{\mathrm{T}}(R - I_d) = 0, \quad Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1}(R - I_d) = 0.$$

If $B := \begin{pmatrix} Q_1 & \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix}$ has full rank, then the above identities, written as $B^{\mathrm{T}}(R - I_d) = 0$, imply $R = I$. B in turn is invertible, since

$$Q^{\mathrm{T}} B = \begin{pmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} Q_1 & \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix} = \begin{pmatrix} I_k & Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2 \\ 0 & Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix}$$

is invertible and Q is orthogonal. □

Theorem (Underdetermined case)

Suppose in the linear Gaussian setting that $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, $\eta \sim \mathcal{N}(0, \gamma^2 \Gamma_0)$ *with* $\gamma > 0$, *and that* $\mathrm{Rank}(A) = k < d$. *Then*

$$\pi^y \rightharpoonup \mathcal{N}(m^*, C^*),$$

where

$$\begin{aligned} m^* &= \Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1} A_1^{-1} y + Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x_0, \\ C^* &= Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}}. \end{aligned}$$

Proof. Using the previous lemma, we can decompose x into

$$x = \underbrace{\Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1}}_{=:\, S} \underbrace{Q_1^{\mathrm{T}} x}_{=:\, x_1} + \underbrace{Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1}}_{=:\, T} \underbrace{Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x}_{=:\, x_2} = S x_1 + T x_2.$$

This way, $x_1 = Q_1^{\mathrm{T}} x$ and $x_2 = Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x$ are Gaussian, and[†]

$$x_2 \sim \mathcal{N}(Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x_0,\, Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2).$$

Now x_1 and x_2 are independent, since

$$\begin{aligned} \mathrm{Cov}(x_1, x_2) &= \mathbb{E}[(x_1 - \mathbb{E}\, x_1)(x_2 - \mathbb{E}\, x_2)^{\mathrm{T}}] \\ &= Q_1^{\mathrm{T}}\, \mathbb{E}[(x - \mathbb{E}\, x)(x - \mathbb{E}\, x)^{\mathrm{T}}] \Gamma_{\mathrm{pr}}^{-1} Q_2 \\ &= Q_1^{\mathrm{T}} Q_2 = 0, \end{aligned}$$

where we used $\mathrm{Cov}(x, x) = \mathbb{E}[(x - \mathbb{E}\, x)(x - \mathbb{E}\, x)^{\mathrm{T}}] = \Gamma_{\mathrm{pr}}$.[(*)]

[(*)] Note that, in general, *uncorrelated random variables are not necessarily independent*. However, this assertion is true for jointly Gaussian random variables.

[†] Recall task 4 of exercise 6: if $z \sim \mathcal{N}(m, C)$, then $Lz + a \sim \mathcal{N}(Lm + a, LCL^{\mathrm{T}})$.

By (4), we have

$$y = Ax + \eta = A_1 Q_1^{\mathrm{T}} x + \eta = A_1 x_1 + \eta. \tag{6}$$

As $\eta \perp x$, this implies $x_2 \perp y, x_1$ and hence $\mathbb{P}(x_1, x_2|y) = \mathbb{P}(x_1|y)\,\mathbb{P}(x_2)$. The random variable x_1 is Gaussian, so problem (6) satisfies the assumptions of the linear Gaussian setting, and thus the posterior distribution $\mathbb{P}(x_1|y)$ is Gaussian. The small noise limit in the determined case in turn shows that $\mathbb{P}(x_1|y) \rightharpoonup \delta_{A_1^{-1}y}(x_1)$ as $\gamma \to 0$. As a consequence, the limiting posterior distribution of $(x_1, x_2)|y$ is

$$\mathbb{P}(x_1, x_2|y) \rightharpoonup \delta_{A_1^{-1}y}(x_1)\,\mathbb{P}(x_2).$$

Now, the mean and covariance of the limiting posterior distribution of $x|y$ are given by

$$\begin{aligned}
m^* &= \mathbb{E}[Sx_1 + Tx_2|y] = SA_1^{-1}y + T\,\mathbb{E}[x_2] \\
&= SA_1^{-1}y + TQ_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0, \\
C^* &= \mathrm{Var}(Sx_1 + Tx_2|y) = \mathrm{Var}(Sx_1|y) + \mathrm{Var}(Tx_2) \\
&= TQ_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}Q_2\,T^{\mathrm{T}} = Q_2(Q_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}Q_2)^{-1}Q_2^{\mathrm{T}}. \quad \square
\end{aligned}$$

Q: How to interpret the limiting distribution in the underdetermined case?

A: Uncertainty remains in the subspace $\mathrm{Ker}(A) = \mathrm{Ran}(Q_2)$ of dimension $d - k$, where the posterior is fully described by the prior.

Monte Carlo and Importance Sampling

Suppose that we are interested in estimating the integral

$$\pi(f) := \mathbb{E}^{\pi}[f(x)] := \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x, \tag{7}$$

where π is a probability density function and $f : \mathbb{R}^d \to \mathbb{R}$ is a quantity of interest.

In the Bayesian framework, we have $\pi(x) = \frac{1}{Z} g(x)\rho(x)$, where Z is a normalization constant, π is the posterior, $g(x) := \nu(y - F(x))$ is the likelihood, and ρ is the prior. Note that here we change the notations slightly to improve readability.

In a non-Gaussian setting, we usually have to resort to approximating the integral (7) by means of sampling. To this end, we will consider the following techniques:

- The Monte Carlo method (today's lecture)
- Importance sampling (today's lecture)
- Markov Chain Monte Carlo (MCMC) methods (next week's lecture)

The Monte Carlo method

A simple technique to approximate the integral

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x, \quad d \in \mathbb{Z}_+,$$

is to use a sample average. If we are able to draw the i.i.d. samples $x_1, \ldots, x_n$ from the probability distribution corresponding to π, then one can consider the Monte Carlo estimate

$$\pi_n^{\mathrm{MC}}(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

Generally speaking, the Law of Large Numbers and the Central Limit Theorem imply that

$$\lim_{n\to\infty} \pi_n^{\mathrm{MC}}(f) = \pi(f) \quad \text{and} \quad \mathrm{Var}(\pi_n^{\mathrm{MC}}(f) - \pi(f)) \approx \frac{\mathrm{Var}(f(X))}{n},$$

provided that $f(X)$ has finite mean and variance with X distributed according to the probability distribution that corresponds to π.

Some properties of the Monte Carlo estimator

If we have the i.i.d. random samples $x_1, \ldots, x_n$ distributed according to π, then π can be estimated by

$$\pi_n^{\mathrm{MC}} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Theorem ([Theorem 5.1, Sanz-Alonso, Stuart, and Taeb 2018])

For $f : \mathbb{R}^d \to \mathbb{R}$, denote $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$. Then

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[\pi(f) - \pi_n^{\mathrm{MC}}(f) \right] \right| = 0 \quad \textit{and} \quad \sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[(\pi(f) - \pi_n^{\mathrm{MC}}(f))^2 \right] \right| \le \frac{1}{n}.$$

This shows that the Monte Carlo estimator π_n^{MC} is an unbiased estimator of π. While the convergence rate is slow with respect to n, the error is independent of the dimension d or the properties of f, its supremum notwithstanding.

Proof. Let $x_1, \ldots, x_n$ be i.i.d. according to π. Define

$$\bar{f}(x) = f(x) - \pi(f).$$

To prove the first result, namely that the estimator is unbiased, note that

$$\mathbb{E}\left[\pi_n^{\mathrm{MC}}(f) - \pi(f)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(x_i) - \pi(f)] = \frac{1}{n}\sum_{i=1}^{n}\left(\pi(f) - \pi(f)\right)$$
$$= \frac{1}{n} \cdot 0 = 0.$$

Therefore the supremum of its absolute value is also zero. For the second result, which bounds the variance of the estimator, we observe that $\mathbb{E}[\bar{f}] = 0$ and, then,

$$\mathbb{E}\left[\left(\pi_n^{\mathrm{MC}}(f) - \pi(f)\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\bar{f}(x_i)\,\bar{f}(x_j)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\bar{f}(x_i)^2\right] = \frac{1}{n}\mathbb{E}\left[\bar{f}(x_1)^2\right] = \frac{1}{n}\mathrm{Var}_\pi[f]$$

since x_i are i.i.d.

In particular we have

$$\mathbb{E}\left[\left(\pi_n^{\mathrm{MC}}(f) - \pi(f)\right)^2\right] = \frac{1}{n}\mathrm{Var}_\pi[f] \leq \frac{1}{n}\pi(f^2) \tag{8}$$

since

$$\mathrm{Var}_\pi[f] = \pi(f^2) - \pi(f)^2 \leq \pi(f^2).$$

Therefore

$$\sup_{\|f\|_\infty \leq 1} \left|\mathbb{E}\left[\left(\pi_n^{\mathrm{MC}}(f) - \pi(f)\right)^2\right]\right| = \sup_{\|f\|_\infty \leq 1} \left|\frac{1}{n}\mathrm{Var}_\pi[f]\right| \leq \frac{1}{n}. \quad \square$$

Example

Suppose that we have the PDF $\pi(x) := (6x - 6x^2)\chi_{(0,1)}(x)$ and $f(x) = x$. We can design the following simple scheme based on inverse transform sampling to draw samples from this distribution.

MATLAB implementation:

```
n = 1e5; % sample size
x = linspace(0,1);
p = @(x) 6*x-6*x.^2; % PDF
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF of p
samples = [];
for iter = 1:n
  u = rand; % realization of U(0,1)
  ind = find(u <= P,1,'first'); % inverse CDF rule
  samples = [samples,x(ind)]; % store sample
end
histogram(samples,'Normalization','pdf'); % draw a histogram
hold on, plot(x,p(x),'LineWidth',3), legend('samples','pdf');
hold off;
```

Python implementation:

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,1,1000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
samples = []
for iter in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample
plt.hist(samples,bins='auto',
         density=True,label='samples') # draw a histogram
plt.plot(x,p(x),linewidth=2,label='pdf')
plt.legend()
plt.show()
# Thanks to Subodh Khanger for the Python implementation!
```

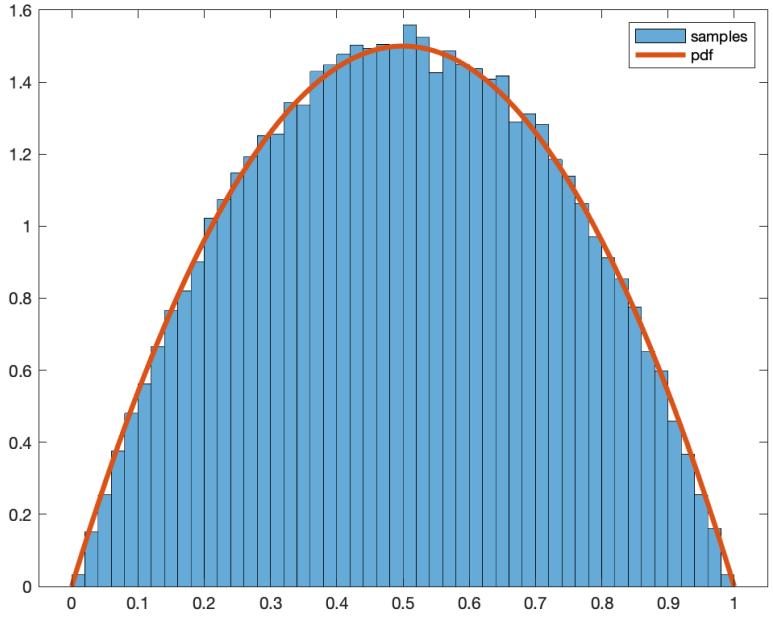


Figure: 10^5 samples drawn from the distribution given on the previous page organized as a histogram.

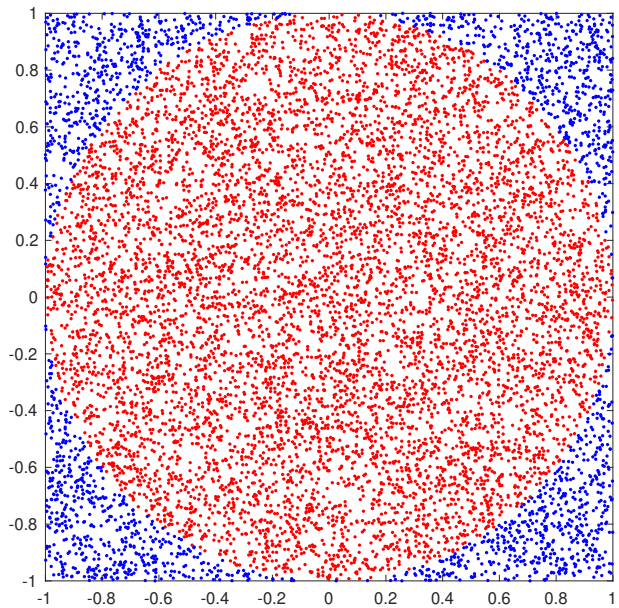
MATLAB:

```
>> mean(samples) % Monte Carlo estimate of the mean

ans =

    0.5001
```

Python: `np.mean(samples) # Monte Carlo estimate of the mean`

Example

Use Monte Carlo to estimate the value of $\int_{\mathbb{R}^2} \chi_{\{x^2+y^2<1\}}(x,y)\,\mathrm{d}x\,\mathrm{d}y$.



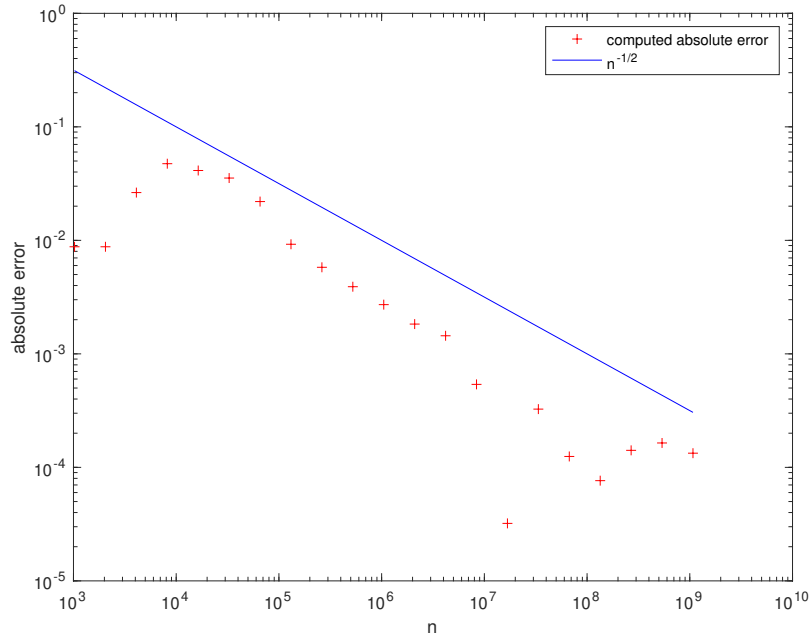


Figure: Left: 2^{13} samples drawn from $U((-1,1)^2)$. We calculate the value of the integral as $4 \cdot \frac{\#\text{samples inside unit disk}}{\#\text{samples inside unit disk}+\#\text{samples outside unit disk}}$. Right: the absolute integration error for $n = 2^k$, $k \in \{10, \ldots, 30\}$.

Sample average at $n = 2^{30}$: 3.141725998371840.

Importance sampling

Let us focus on the setting

$$\pi(x) = \frac{1}{Z} g(x)\rho(x), \tag{9}$$

where Z is a normalization constant. Unless π is some well-understood distribution (e.g., Gaussian), the basic Monte Carlo method is generally infeasible due to the difficulties associated with drawing samples from π directly in the high-dimensional setting.

An alternative tactic is to use ρ as a *proposal density*, drawing samples from it instead of π. By substituting the identity (9) into $\pi(f)$, we obtain

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x = \frac{\int_{\mathbb{R}^d} (f(x)g(x))\,\rho(x)\,\mathrm{d}x}{\int_{\mathbb{R}^d} g(x)\,\rho(x)\,\mathrm{d}x}.$$

If the samples $x_1, \ldots, x_n$ are now distributed i.i.d. according to ρ, we can replace the numerator and denominator by their respective Monte Carlo estimates:

$$\pi_n^{\mathrm{IS}}(f) := \sum_{i=1}^{n} w_i f(x_i), \quad w_i := \frac{g(x_i)}{\sum_{j=1}^{n} g(x_j)} \quad (\text{“importance weights”}).$$

Similarly to the Monte Carlo estimator, we can define the *particle approximation measure*

$$\pi_n^{\mathrm{IS}} := \sum_{i=1}^{n} w_i \delta_{x_i}, \quad w_i := \frac{g(x_i)}{\sum_{j=1}^{n} g(x_j)}.$$

Theorem ([Theorem 5.4, Sanz-Alonso, Stuart, and Taeb 2018])

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[\pi_n^{\mathrm{IS}}(f) - \pi(f)\right] \right| \le 2 \, \frac{1 + d_{\chi^2}(\pi \| \rho)}{n},$$

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[(\pi_n^{\mathrm{IS}}(f) - \pi(f))^2\right] \right| \le 4 \, \frac{1 + d_{\chi^2}(\pi \| \rho)}{n},$$

where the χ^2 *divergence of two probability distributions* $\pi, \pi' > 0$ *is defined as*

$$d_{\chi^2}(\pi \| \pi') := \int_{\mathbb{R}^d} \left(\frac{\pi(x)}{\pi'(x)} - 1 \right)^2 \pi'(x) \, \mathrm{d}x.$$

Unlike Monte Carlo, π_n^{IS} is biased for π. The χ^2 divergence between π and ρ should not be too large for importance sampling to be accurate.

Proof. Let $x_1, \ldots, x_n$ be i.i.d. according to ρ. Given

$$\pi(x) = \frac{1}{Z} g(x)\rho(x) = \frac{1}{\rho(g)} g(x)\rho(x),$$

we obtain

$$d_{\chi^2}(\pi \| \rho) = \int_{\mathbb{R}^d} \left(\frac{\pi(x)}{\rho(x)} - 1 \right)^2 \rho(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \left(\frac{g(x)}{Z} - 1 \right)^2 \rho(x)\,\mathrm{d}x$$

$$= \underbrace{\int_{\mathbb{R}^d} \frac{g(x)^2 \rho(x)}{Z^2}\,\mathrm{d}x}_{=\frac{\rho(g^2)}{\rho(g)^2}} - 2\frac{1}{Z} \underbrace{\int_{\mathbb{R}^d} g(x)\rho(x)\,\mathrm{d}x}_{=Z} + \underbrace{\int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x}_{=1} = \frac{\rho(g^2)}{\rho(g)^2} - 1.$$

Let $\zeta := \frac{\rho(g^2)}{\rho(g)^2}$. Noting that

$$\pi(f) = \frac{\rho(gf)}{\rho(g)} \approx \frac{\rho_n^{\mathrm{MC}}(gf)}{\rho_n^{\mathrm{MC}}(g)} = \pi_n^{\mathrm{IS}}(f),$$

it follows that

$$\begin{aligned} \pi_n^{\mathrm{IS}}(f) - \pi(f) &= \pi_n^{\mathrm{IS}}(f) - \frac{\rho(gf)}{\rho(g)} \\ &= \frac{\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)}{\rho(g)} - \frac{\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)}{\rho(g)}. \end{aligned} \tag{10}$$

Let us prove the second inequality first. We use the splitting of $\pi_n^{\mathrm{IS}}(f) - \pi(f)$ into the sum of two terms from the previous slide together with $\mathbb{E}[(\rho(f) - \rho_n^{\mathrm{MC}}(f))^2] \le \frac{1}{n}\rho(f^2)$ (see (8)) and the inequality $(a-b)^2 \le 2(a^2+b^2)$ such that for all $\|f\|_\infty \le 1$ we have $|\pi_n^{\mathrm{IS}}(f)| \le 1$ and

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right| \\
&\le \frac{2}{\rho(g)^2}\left(\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f)\right)^2\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)^2\right]\right) \\
&\le \frac{2}{\rho(g)^2}\left(\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)^2\right]\right) \\
&= \frac{2}{\rho(g)^2 n}\left(\mathrm{Var}_\rho[g] + \mathrm{Var}_\rho[gf]\right) \\
&\le \frac{2}{\rho(g)^2 n}\left(\rho(g^2) + \rho(g^2 f^2)\right) \le \frac{4}{n}\frac{\rho(g^2)}{\rho(g)^2} = \frac{4\zeta}{n}.
\end{aligned}
$$

Therefore, since $\zeta = d_{\chi^2}(\pi\|\rho) + 1$, we obtain

$$
\sup_{|f|_\infty \le 1}\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right| \le 4\,\frac{1 + d_{\chi^2}(\pi\|\rho)}{n}.
$$

To prove the first inequality, we start again with the splitting (10), i.e.,

$$\pi_n^{\mathrm{IS}}(f) - \pi(f) = \frac{\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)}{\rho(g)} - \frac{\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)}{\rho(g)}.$$

The expectation of the second term vanishes since

$$\left|\mathbb{E}\left[\frac{\rho(gf) - \rho_n^{\mathrm{MC}}(gf)}{\rho(g)}\right]\right| = \frac{1}{\rho(g)}\left|\mathbb{E}\left[\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right]\right| = 0.$$

The Cauchy–Schwarz inequality together with $\mathbb{E}[(\rho(g) - \rho_n^{\mathrm{MC}}(g))^2] \leq \frac{1}{n}\rho(g^2)$ (see (8)) and the previous result yield that

$$\begin{aligned}
\left|\mathbb{E}\left[\pi_n^{\mathrm{IS}}(f) - \pi(f)\right]\right| &= \frac{1}{\rho(g)}\left|\mathbb{E}\left[\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right]\right| \\
&\leq \frac{1}{\rho(g)}\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right] + \pi(f)\underbrace{\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right]}_{=0}\right| \\
&\leq \frac{1}{\rho(g)}\left(\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right]\right)^{1/2} \\
&\leq \frac{1}{\rho(g)}\left(\frac{4\zeta}{n}\right)^{1/2}\left(\frac{\rho(g^2)}{n}\right)^{1/2} = \frac{2\zeta}{n} = 2\,\frac{d_{\chi^2}(\pi\|\rho) + 1}{n}. \quad \square
\end{aligned}$$

Case study: source localization

Suppose that a particle with unit charge is located at some (unknown) point $x^* \in (0,1)$ and our goal is to locate it based on measurements of voltage at the interval end points $x = 0$ and $x = 1$. The mathematical model for the voltage at any point $x \in [0,1]$ is given by

$$y(x) = \frac{1}{|x^* - x|}.$$

Our noisy measurements are modeled by $y_1 = \frac{1}{|x^*-0|} + \eta_1$ and $y_2 = \frac{1}{|x^*-1|} + \eta_2$, where η_1 and η_2 are i.i.d. realizations of $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.2$.

- The likelihood is given by $\mathbb{P}(y|x) \propto \exp\big(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}\big(y_j - \frac{1}{|x-j|}\big)^2\big)$.
- We consider the prior $\pi(x) = \chi_{(0,1)}(x) = \begin{cases} 1 & \text{if } x \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$

Then the posterior density is given by Bayes' formula

$$\pi^y(x) \propto \chi_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}\left(y_j - \frac{1}{|x-j|}\right)^2\right).$$

Computation of the CM estimate (MATLAB)

First, let us generate the measurements.

MATLAB:

```
format long
x_ast = 1/pi; % Fix "ground truth", i.e., particle location
sigma = .2; % Std for noise
v = 1./abs(x_ast-[0,1]); % Measurements at end points
v = v+sigma*randn(1,2); % Add noise
x = linspace(0,1); % Discretize the unit interval
% Define the (unnormalized) posterior density
p = @(x) exp(-1/(2*sigma^2)*((v(1)-1./abs(x-0)).^2+ ...
         (v(2)-1./abs(x-1)).^2));
```

```
%% Monte Carlo
n = 1e5;
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF
% For the Monte Carlo method, we need to sample the posterior.
% We do this using inverse transform sampling.
samples = [];
for ii = 1:n
   u = rand; % realization of U(0,1)
   ind = find(u <= P,1,'first'); % inverse CDF rule
   samples = [samples,x(ind)]; % store sample
end
% Sanity check: plot samples in histogram.
histogram(samples,'Normalization','probability', ...
                  'BinWidth',.01), axis([0,1,0,.25]);
hold on;
plot(x,p(x)/sum(p(x)),'LineWidth',2), hold off;
title([num2str(n),' samples from the posterior density']);
mean(samples) % Monte Carlo estimate
```

```
%% Importance sampling
n = 1e5;
samples = rand(1,n); % Sample our prior, i.e., U(0,1)
weights = p(samples); % Compute the importance weights
weights = weights/sum(weights); % Normalize the weights

% Compute the IS estimate
dot(weights,samples)
```

Computation of the CM estimate (Python)

First, let us generate the measurements.

Python:

```
import numpy as np
x_ast = 1/np.pi # Fix "ground truth", i.e., particle location
sigma = .2 # Std for noise
v = 1/np.abs(x_ast-np.array([0,1])) # Measurements at
                                    # end points
v = v+sigma*np.random.normal(size=v.shape) # Add noise
x = np.linspace(0,1) # Discretize the unit interval
x = x[1:-1] # Drop end points to avoid numerical issues...
# Define the (unnormalized) posterior density
p = lambda x: (x > 0) * (x < 1) *\
              np.exp(-1/(2*sigma**2)*((v[0]-1/np.abs(x-0))**2\
              +(v[1]-1/np.abs(x-1))**2))
```

```
## Monte Carlo
n = int(1e5)
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF
# For the Monte Carlo method, we need to sample the posterior.
# We do this using inverse transform sampling.
samples = []
for ii in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample

# Compute the Monte Carlo estimate
print(np.mean(samples))
```

```
## Importance sampling
n = int(1e5)
samples = np.random.uniform(size=(1,n)) # Sample our prior,
                                        # i.e., U(0,1)
weights = p(samples) # Compute the importance weights
weights = weights/np.sum(weights) # Normalize the weights

# Compute the IS estimate
print(np.sum(weights*samples))
```

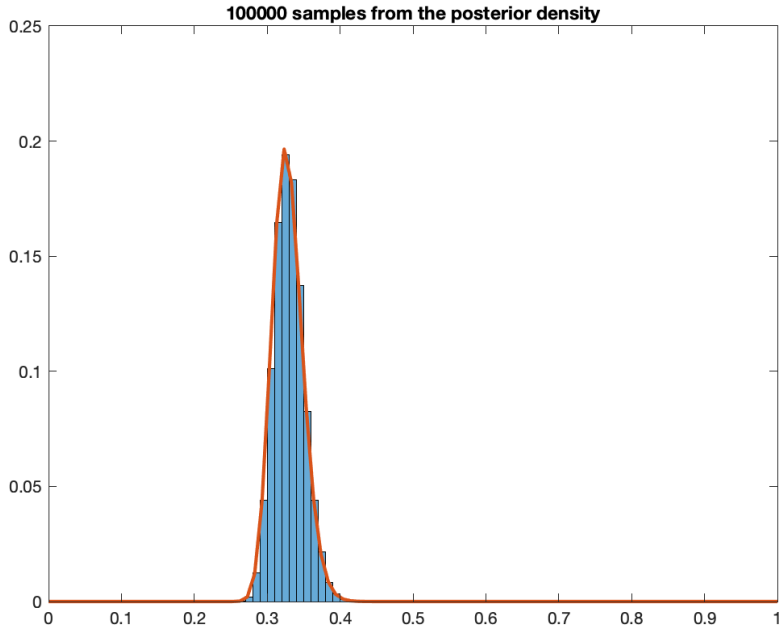


Figure: Histogram of the samples drawn from the posterior density.

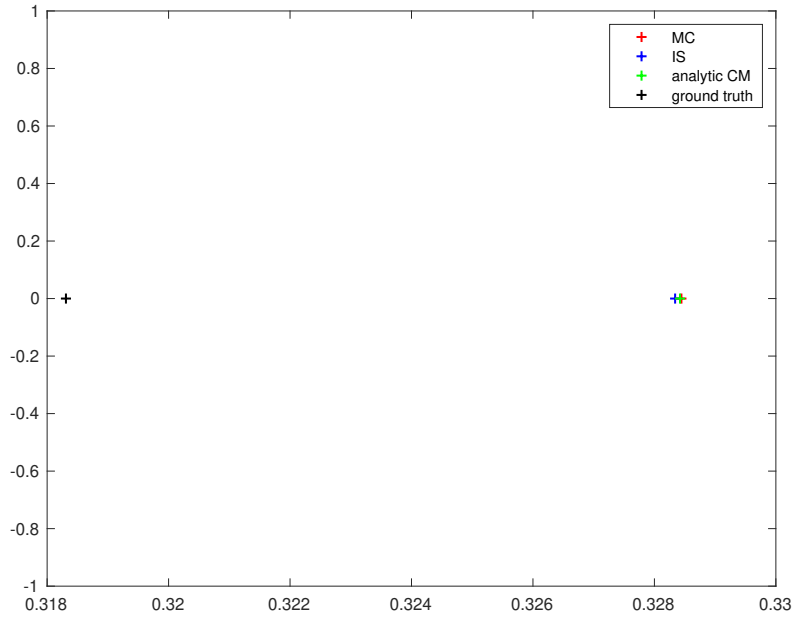
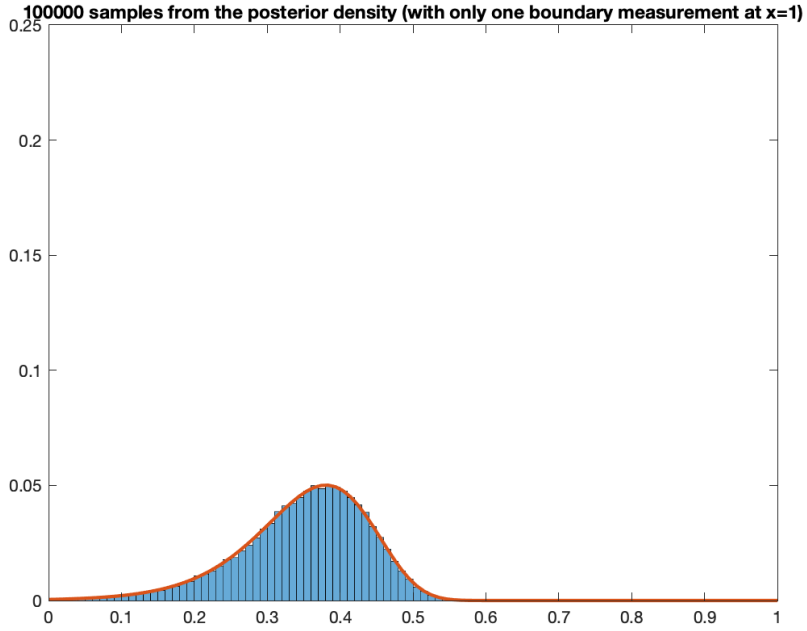


Figure: Comparison of MC and IS estimates vs. the analytic CM estimate and ground truth.

Monte Carlo estimate
0.328444646464649
Importance sampling estimate
0.328340981036045

Ground truth
0.318309886183791
Analytic CM estimate
0.328421554655529

What if we modify the problem so that we have access to only one boundary measurement at $x = 1$?



Monte Carlo estimate
0.349233333333324
Importance sampling estimate
0.349743141888635
Analytic CM estimate
0.349675613936670
Ground truth
0.318309886183791

The problem becomes substantially more ill-posed!

N.B. In the implementation above, a discretized version of the inverse transform sampling rule was used to obtain the MC estimate. The repeating digits are an artifact of the relatively coarse discretization used in the actual implementation.

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Twelfth lecture, July 3, 2023

Course schedule

Monday	*Tuesday*	*Wednesday*
July 3 (today)	**July 4 (tomorrow)**	
Lecture	Deadline for 9^{th} exercise sheet	
July 10	**July 11**	
Lecture	Deadline for 10^{th} (final) exercise sheet	
July 17	**July 18**	
Summary	**no exercise session!**	
		July 26
		Oral exam

Make-up oral exam in early October (date TBA).

The content of these slides follows roughly the material presented in the following monographs.


D. Calvetti and E. Somersalo. Introduction to Bayesian Scientific Computing – Ten Lectures on Subjective Computing. 2007.


J. Kaipio and E. Somersalo. Statistical and Computational Inverse Problems. 2005.

Discrete time Markov chains

A sequence $\{X_k\}_{k=0}^{\infty}$ of random variables is called a *discrete time Markov chain* if the probability distribution of any X_{k+1} depends only on the previous state X_k:

$$\pi(x_{k+1} \mid x_0, \ldots, x_k) = \pi(x_{k+1} \mid x_k).$$

Suppose in addition that there exists a *probability transition kernel* $q(x, y)$ such that

$$\pi(x_{k+1} \mid x_k) = q(x_k, x_{k+1}).$$

Then the Markov chain is called *time invariant* (or *time homogeneous*) since the kernel q is independent of the time k.

Remark. We assume here and in the sequel that transition kernels satisfy

- $x \mapsto \int_B q(x, y)\,\mathrm{d}y$ is measurable for all $x \in \mathbb{R}^d$ and $B \in \mathcal{B}(\mathbb{R}^d)$;
- $B \mapsto \int_B q(x, y)\,\mathrm{d}y$ is a probability distribution for all $x \in \mathbb{R}^d$, $B \in \mathcal{B}(\mathbb{R}^d)$. In particular, $\mathbb{P}(Y \in B \mid X = x) = \int_B q(x, y)\,\mathrm{d}y$ and $\int_{\mathbb{R}^d} q(x, y)\,\mathrm{d}y = 1$.

Let X be a random variable with probability density $p(x)$.

Let $q(x, y)$ be an arbitrary transition kernel used to generate a new random variable Y given $X = x$, i.e.,

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

The probability density of Y can be found through marginalization:

$$\pi(y) = \int_{\mathbb{R}^d} \pi(y \mid x) p(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} q(x, y) p(x) \, \mathrm{d}x.$$

If the probability density of Y is equal to the probability density of X,

$$\int_{\mathbb{R}^d} q(x, y) p(x) \, \mathrm{d}x = p(y),$$

then we call p an *invariant density* of the transition kernel q.

Definition (Irreducible transition kernel)

The transition kernel q is *irreducible* if, regardless of the starting point, the Markov chain generated by q can visit any set of positive measure with positive probability.

Definition (Periodic transition kernel)

The transition kernel q is *periodic* if, for some integer $m \geq 2$, there is a set of disjoint nonempty sets $\{E_1, \ldots, E_m\} \subset \mathbb{R}^d$ such that for all $j \in \{1, \ldots, m\}$ and for all $x \in E_j$:

$$\mathbb{P}(Y \in E_{\mathrm{mod}(j,m)+1} | X = x) = \int_{E_{\mathrm{mod}(j,m)+1}} q(x, y)\, \mathrm{d}y = 1.$$

That is, the Markov chain generated by q remains in a periodic loop forever.

Definition (Aperiodic transition kernel)

The transition kernel q is *aperiodic* if it is not periodic.

Theorem

Let $\{X_k\}_{k=0}^{\infty}$ be a time invariant Markov chain with the transition kernel q, i.e.,

$$\pi(x_{k+1} \mid x_k) = q(x_k, x_{k+1}).$$

Assume that p is an invariant density of q and the following technical conditions hold:

- *q is* irreducible*;*
- *q is* aperiodic*.*

Then for all $x_0 \in \mathbb{R}^d$ and any $B \in \mathcal{B}(\mathbb{R}^d)$, it holds that

$$\lim_{N\to\infty} \mathbb{P}(X_N \in B \mid X_0 = x_0) = \int_B p(x)\,\mathrm{d}x.$$

Moreover, for any $f \in L_p^1(\mathbb{R}^d)$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} f(X_j) = \int_{\mathbb{R}^d} f(x)p(x)\,\mathrm{d}x \quad \text{a.s.}$$

Suppose we want to sample some probability density p and we know that it is invariant with respect to transition kernel q. Then we can proceed as follows:

1. Select 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.

The previous theorem implies that the sample $\{x_k\}_{k=0}^{N}$ is asymptotically distributed according to p as $N \to \infty$.

This raises the question: *given a probability density p, how do you find a kernel q such that p is its invariant density?*

The *Metropolis–Hastings algorithm* is a method to construct such a kernel!

Derivation of the Metropolis–Hastings algorithm

We are interested in obtaining samples from the probability density p. Consider the following Markov process: if you are currently situated at some $x \in \mathbb{R}^d$, 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.

Here, both $R(x, y)$ and $r(x)$ are as yet undetermined—the trick will be to calibrate these in order to find a kernel such that p is its invariant density as discussed on the previous slide.

Since R is a transition kernel, $y \mapsto R(x, y)$ is a probability density and hence

$$\int_{\mathbb{R}^d} R(x, y) \, \mathrm{d}y = 1 \quad \text{for all } x \in \mathbb{R}^d.$$

Denote by $\mathcal{A}$ the event of moving away from x and by $\neg\mathcal{A}$ the event of not moving. Clearly

$$\mathbb{P}(\mathcal{A}) = 1 - r(x) \quad \text{and} \quad \mathbb{P}(\neg\mathcal{A}) = r(x).$$

Given a current state $X = x$, we want to know what is the probability density of Y generated by the aforementioned strategy. Let $B \in \mathcal{B}(\mathbb{R}^d)$ and consider the probability of the event $Y \in B$. Then

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) = \mathbb{P}(Y \in B \mid X = x, \mathcal{A})\mathbb{P}(\mathcal{A}) \quad &\text{(move away from } x\text{)} \\ + \mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A})\mathbb{P}(\neg\mathcal{A}). \quad &\text{(stay put at } x\text{)}\end{aligned}$$

The probability of arriving in B through a move is

$$\mathbb{P}(Y \in B \mid X = x, \mathcal{A}) = \int_B R(x, y)\,\mathrm{d}y.$$

The only way to arrive in B without moving is if x is already in B:

$$\mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A}) = \chi_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{cases}$$

Hence

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) &= \int_B \overbrace{(1 - r(x))R(x, y)}^{=:K(x,y)}\,\mathrm{d}y + r(x)\chi_B(x) \\ &= \int_B K(x, y)\,\mathrm{d}y + r(x)\chi_B(x).\end{aligned}$$

The probability of $Y \in B$ can be obtained by marginalizing over x:

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \int_{\mathbb{R}^d} \mathbb{P}(Y \in B \mid X = x) p(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(\int_B K(x,y)\,\mathrm{d}y \right) p(x)\,\mathrm{d}x + \int_{\mathbb{R}^d} r(x)\chi_B(x)p(x)\,\mathrm{d}x \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y)p(x)\,\mathrm{d}x \right) \mathrm{d}y + \int_B r(x)p(x)\,\mathrm{d}x \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y)p(x)\,\mathrm{d}x + r(y)p(y) \right) \mathrm{d}y \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y)p(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} K(y,x)p(y)\,\mathrm{d}x + p(y) \right) \mathrm{d}y,
\end{aligned}$$

where we used $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = (1 - r(y)) \int_{\mathbb{R}^d} R(y,x)\,\mathrm{d}x = 1 - r(y)$.
If the *balance equation*

$$\int_{\mathbb{R}^d} p(y)K(y,x)\,\mathrm{d}x = \int_{\mathbb{R}^d} p(x)K(x,y)\,\mathrm{d}x \tag{1}$$

holds, then

$$\mathbb{P}(Y \in B) = \int_B p(y)\,\mathrm{d}y \quad \text{as desired.}$$

The Metropolis–Hastings algorithm is a technique for finding a kernel K that satisfies the *detailed balance equation*

$$p(y)K(y,x) = p(x)K(x,y),$$

which implies (1). Let us start with a *proposal density* $q(x,y)$, chosen so that generating a Markov chain with it is easy. (For this reason, a Gaussian kernel is a very popular choice.) There are three separate cases:

1. If $p(y)q(y,x) = p(x)q(x,y)$, then set $r(x) = 0$, $R(x,y) = K(x,y) = q(x,y)$ and the previous analysis ensures that p is an invariant density for kernel q.
2. If $p(y)q(y,x) < p(x)q(x,y)$, then define the kernel K to be

   $$K(x,y) = \alpha(x,y)q(x,y),$$

   where α is chosen s.t. $p(y)\alpha(y,x)q(y,x) = p(x)\alpha(x,y)q(x,y)$. We can make the selection

   $$\alpha(y,x) = 1 \quad \text{and} \quad \alpha(x,y) = \frac{p(y)q(y,x)}{p(x)q(x,y)} < 1.$$
3. If $p(y)q(y,x) > p(x)q(x,y)$, then in complete analogy to the above:

   $$\alpha(x,y) = 1 \quad \text{and} \quad \alpha(y,x) = \frac{p(x)q(x,y)}{p(y)q(y,x)} < 1.$$

In summary, we define K as

$$K(x, y) = \alpha(x, y)q(x, y), \quad \alpha(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\}.$$

Even though the expression for K seems complicated, it turns out that the drawing can be performed according to the following procedure.

Metropolis–Hastings algorithm

1. Choose $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
2. Given $x = x^{(k)}$, draw y using the transition kernel $q(x, y)$ of your choosing.
3. Calculate the acceptance ratio

$$\alpha(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\}.$$

4. Flip the α-coin: draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, set $x^{(k+1)} = y$, otherwise stay put at x and set $x^{(k+1)} = x^{(k)}$.
5. Set $k \leftarrow k + 1$ and return to step 2.

Remark. Note that due to the form of α, both the target p and the proposal density q can be *unnormalized* within the Metropolis–Hastings algorithm.

Why does this work?

Let us focus on the main loop of the Metropolis–Hastings algorithm:

- Given x, draw y using the transition kernel $q(x, y)$.
- Calculate the acceptance ratio $\alpha(x, y) = \min\left\{1, \frac{p(y)q(x,y)}{p(x)q(x,y)}\right\}$.
- Draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, accept y, otherwise stay put at x.

Recall that $\mathcal{A}$ was the event of moving in the Markov chain. Then

$$\mathbb{P}(\mathcal{A}|y, x) = \text{“probability of accepting transition”} = \alpha(x, y),$$
$$\mathbb{P}(y|x) = \text{“probability of drawing } y\text{”} = q(x, y).$$

Then

$$\begin{aligned}\text{“probability of accepted } y\text{”} &= \mathbb{P}(\mathcal{A}, y|x) \\ &= \mathbb{P}(\mathcal{A}|y, x)\mathbb{P}(y|x) \\ &= \alpha(x, y)q(x, y) = K(x, y),\end{aligned}$$

as desired.

Example

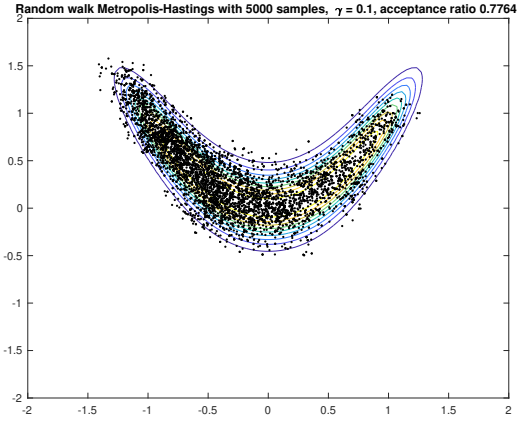
Let us consider sampling from the density

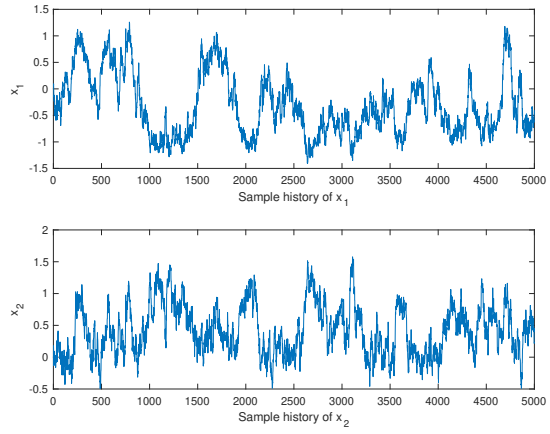
$$p(x_1, x_2) \propto \exp(-10(x_1^2 - x_2)^2 - (x_2 - \tfrac{1}{4})^4).$$

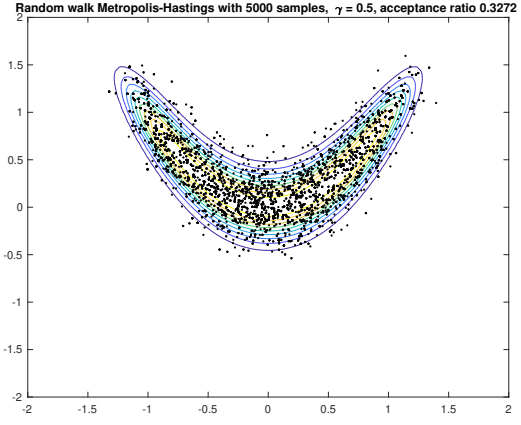
As the proposal distribution, we use the random walk model $Y = X + W$, $W \sim \mathcal{N}(0, \gamma^2 I)$, with the kernel

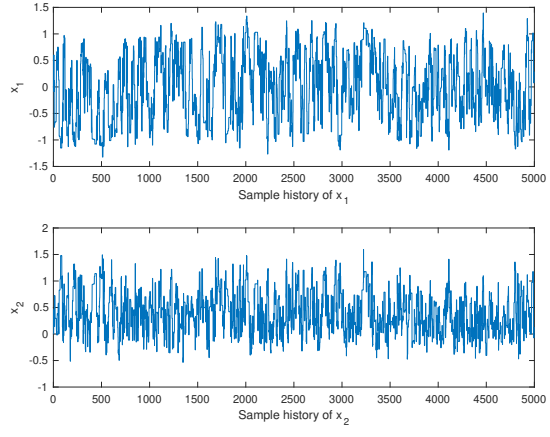
$$q(x, y) \propto \exp\left(-\frac{1}{2\gamma^2} \|x - y\|^2 \right).$$

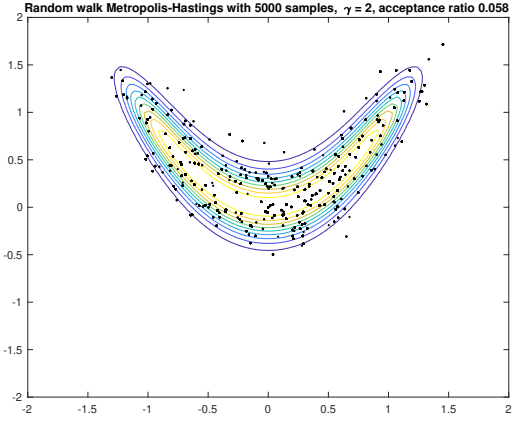
We draw 5000 samples from the probability distribution with density p using three different step sizes: $\gamma = 0.1$, $\gamma = 0.5$, and $\gamma = 2$.

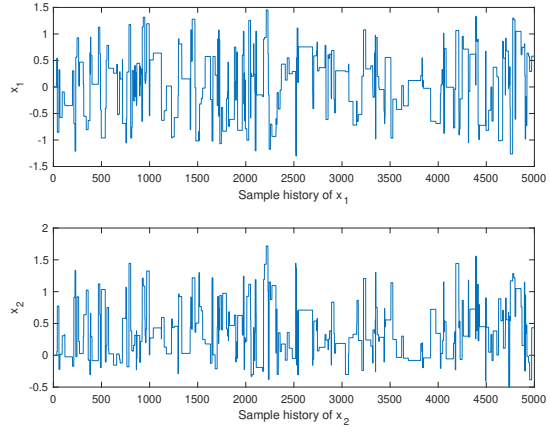

Random walk Metropolis-Hastings with 5000 samples, γ = 0.1, acceptance ratio 0.7764


x1
Sample history of x1
x2
Sample history of x2


Random walk Metropolis-Hastings with 5000 samples, γ = 0.5, acceptance ratio 0.3272


x1
Sample history of x1
x2
Sample history of x2


Random walk Metropolis-Hastings with 5000 samples, γ = 2, acceptance ratio 0.058


x1
Sample history of x1
x2
Sample history of x2

Derivation of the single component Gibbs sampler

We continue to be interested in sampling the distribution with density $p(x)$. The single component Gibbs sampler is based on the same Markov process that was introduced in the derivation of Metropolis–Hastings: if you are currently situated at some $x \in \mathbb{R}^d$, 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 the definition we made in the Metropolis–Hastings derivation:

$$K(x, y) = (1 - r(x))R(x, y).$$

Suppose that x is a d-variate random variable. For the single component Gibbs sampler, we set $r(x) = 0$ (moving is obligatory) and define the transition kernel

$$K(x, y) = R(x, y) = \prod_{i=1}^{d} p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d),$$

where $p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) = \dfrac{p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)}{\int_{\mathbb{R}} p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)\, \mathrm{d}y_i}$.

This transition kernel K does not in general satisfy the detailed balance equation, but it does satisfy the standard balance equation, which is sufficient to ensure that p is the invariant density of the Markov chain (see derivation of the Metropolis–Hastings method).

Theorem

The transition kernel

$$K(x, y) = \prod_{i=1}^{d} p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d),$$

where $p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) = \dfrac{p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)}{\int_{\mathbb{R}} p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)\,\mathrm{d}y_i}$,

satisfies

$$\int_{\mathbb{R}^d} p(y)K(y, x)\,\mathrm{d}x = \int_{\mathbb{R}^d} p(x)K(x, y)\,\mathrm{d}x.$$

Remark. We only consider the *single component* Gibbs sampler here. The Gibbs sampler can be written in slightly more general form; see, e.g., [chapter 3.6.3, Kaipio and Somersalo 2005].

Proof. We begin with the left-hand side of the balance equation and consider $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x$. We integrate inductively over the variables in the order $x_d, x_{d-1}, \ldots, x_1$:

$$\begin{aligned}
\int_{\mathbb{R}} K(y,x)\,\mathrm{d}x_d &= \int_{\mathbb{R}} \left(\prod_{i=1}^{d} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \mathrm{d}x_d \\
&= \left(\prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_d|x_1,\ldots,x_{d-1})\,\mathrm{d}x_d}_{=1} \\
&= \prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \\
\Rightarrow \int_{\mathbb{R}} \int_{\mathbb{R}} K(y,x)\,\mathrm{d}x_d\,\mathrm{d}x_{d-1} &= \int_{\mathbb{R}} \left(\prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \mathrm{d}x_{d-1} \\
&= \left(\prod_{i=1}^{d-2} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_{d-1}|x_1,\ldots,x_{d-1},y_d)\,\mathrm{d}x_{d-1}}_{=1} \\
&= \prod_{i=1}^{d-2} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \quad \Rightarrow \quad \ldots
\end{aligned}$$

Proceeding by inductively integrating over $x_{d-2}, x_{d-3}, \ldots, x_1$, we obtain $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = 1$ and thus $\int_{\mathbb{R}^d} p(y)K(y,x)\,\mathrm{d}x = p(y)\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = p(y)$.

Next we consider the right-hand side of the balance equation. Recall that $K(x,y) = \prod_{i=1}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1}\ldots,x_d)$. We integrate inductively over the variables, this time in the order $x_1,\ldots,x_d$:

$$\int_{\mathbb{R}} p(x)K(x,y)\,\mathrm{d}x_1 = K(x,y)\int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1 \qquad (K \text{ is independent of } x_1)$$

$$= \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) \underbrace{p(y_1|x_2,\ldots,x_d)}_{=\frac{p(y_1,x_2,\ldots,x_d)}{\int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1}} \int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1$$

$$= \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,x_2,\ldots,x_d)$$

$$\Rightarrow \int_{\mathbb{R}}\int_{\mathbb{R}} p(x)K(x,y)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{\mathbb{R}} \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,x_2,\ldots,x_d)\,\mathrm{d}x_2$$

$$= \left(\prod_{i=3}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) \underbrace{p(y_2|y_1,x_3,\ldots,x_d)}_{=\frac{p(y_1,y_2,x_3,\ldots,x_d)}{\int_{\mathbb{R}} p(y_1,x_2,x_3,\ldots,x_d)\,\mathrm{d}x_2}} \int_{\mathbb{R}} p(y_1,x_2,\ldots,x_d)\,\mathrm{d}x_2$$

$$= \left(\prod_{i=3}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,y_2,x_3,\ldots,x_d) \quad \Rightarrow \quad \ldots$$

Proceeding by inductively integrating over $x_3,\ldots,x_d$, we eventually obtain $\int_{\mathbb{R}^d} p(x)K(x,y)\,\mathrm{d}x = p(y)$. Therefore the balance equation holds. □

Single component Gibbs sampler

1. Choose the initial value $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
2. Draw the next sample as follows:
   (i) Set $x = x^{(k)}$ and $j = 1$.
   (ii) Draw $t \in \mathbb{R}$ from the one-dimensional distribution

$$p(t \mid y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_d) \propto p(y_1, \ldots, y_{j-1}, t, x_{j+1}, \ldots, x_d)$$

   and set $y_j = t$.
   (iii) If $j = d$, set $y = (y_1, \ldots, y_d)$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (ii).
3. Set $x^{(k+1)} = y$, increase $k \leftarrow k + 1$ and return to step 2.

Example

Let us consider the density from before

$$p(x_1, x_2) = \frac{1}{Z} \exp(-10(x_1^2 - x_2)^2 - (x_2 - \tfrac{1}{4})^4),$$

where the normalizing constant is $Z = 1.1813\ldots$

This time we use the Gibbs sampler. To sample the univariate densities that arise in the process, we use inverse transform sampling. In this case, the explicit algorithm we use is written below.

Fix $x^{(0)} \in \mathbb{R}^2$ and set $x = x^{(0)}$;

For $k = 1, \ldots, N$, do

Calculate $\Phi_1(t) = \int_{-\infty}^{t} p(x_1, x_2)\, \mathrm{d}x_1$;

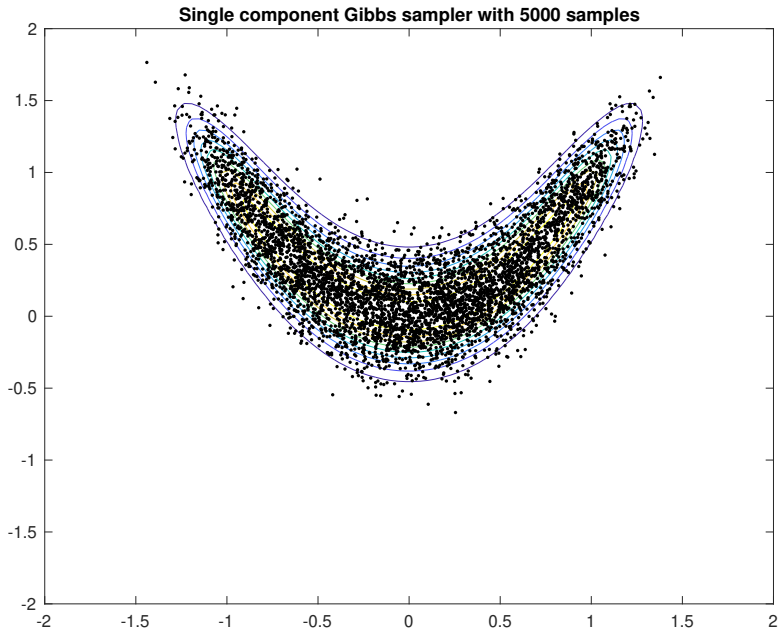
Draw $u \sim \mathcal{U}([0,1])$, set $x_1 = \Phi_1^{-1}(u)$;

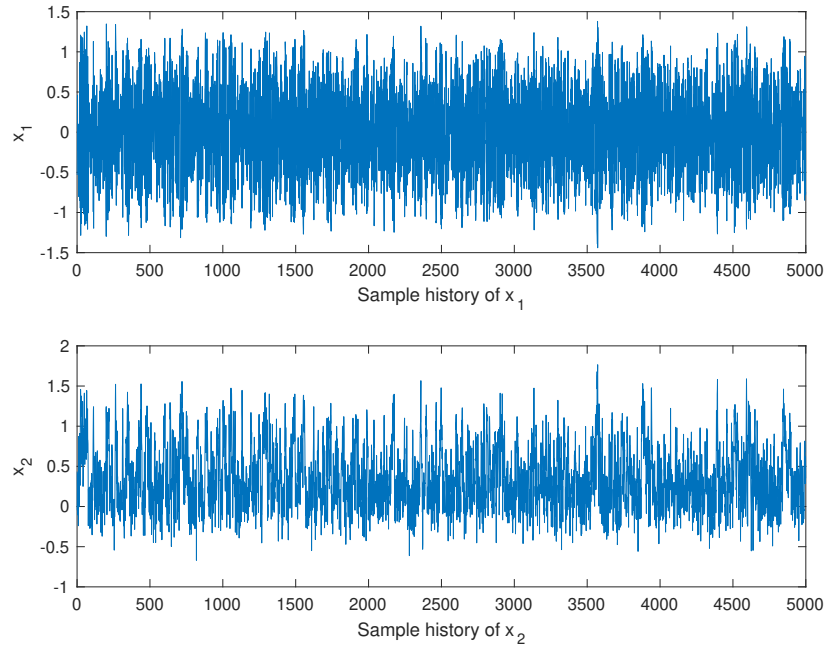
Calculate $\Phi_2(t) = \int_{-\infty}^{t} p(x_1, x_2)\, \mathrm{d}x_2$;

Draw $u \sim \mathcal{U}([0,1])$, set $x_2 = \Phi_2^{-1}(u)$;

Set $x^{(k)} = x$.

End


Single component Gibbs sampler with 5000 samples
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

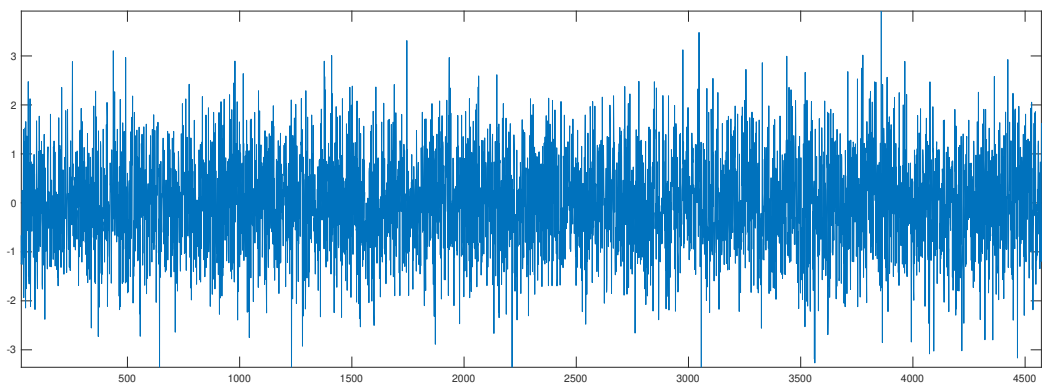

1.5
1
0.5
0
-0.5
-1
-1.5
x1
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample history of x 1
2
1.5
1
0.5
0
-0.5
-1
x2
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample history of x 2

Computational remarks about MCMC

- As a general rule of thumb, one should aim at roughly 30% acceptance rates when using Gaussian (or close to Gaussian) proposal and target densities with MH.
- It usually takes the Markov chain a number of iterations to reach the steady state. To this end, it is usually advisable to discard the first N_0 obtained samples since they may not be representative of the target distribution—this is the so-called "burn-in" period. The length of the burn-in period varies depending on the application, but one might consider throwing away the first $\sim 5 - 10\%$ steps for a sufficiently large sample size as an example.
- In MH, using a Gaussian kernel (e.g., random walk Metropolis–Hastings) is a popular choice due to the ease of implementation. While it is a safe choice, it does not take into account the form of the posterior density. To increase efficiency, it is advisable to take the shape of the density into account when designing the proposal density. In the high-dimensional setting, this is especially useful if the posterior density is *anisotropic* (stretched in some directions).

Computational remarks about MCMC

- The proposal distribution in MH can also be updated while the sampling algorithm moves around the posterior density. This process is called *adaptation*.
- Visual assessment: we are aiming for independent sample points, where the sample histories look like a "fuzzy worm". One could aim at something like the Gaussian white noise signal below:



- More quantitatively, the independence of consecutive draws can be estimated from the sample itself by computing its (sample-based) autocovariance.

A note on convergence

The success of the Metropolis–Hastings and Gibbs sampler algorithms depends largely on whether they satisfy the ergodicity conditions from before. There are known sufficient conditions concerning the density p that guarantee the ergodicity of these methods. For example, the following proposition gives some relatively general conditions.

Proposition (Proposition 3.12. in Kaipio and Somersalo 2005)

(a) Let $p\colon \mathbb{R}^d \to \mathbb{R}_+$ and let $q\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a candidate-generating kernel. If the Markov chain corresponding to q is aperiodic, then the Metropolis–Hastings chain is also aperiodic. Further, if the Markov chain corresponding to q is irreducible and $\alpha(x, y) > 0$ for all $(x, y) \in E_+ \times E_+$, where $E_+ := \{x \in \mathbb{R}^d \mid p(x) > 0\}$, then the Metropolis–Hastings chain is irreducible.

(b) Let p be a lower semicontinuous density and E_+ as above. The Gibbs sampler defines an irreducible and aperiodic transition kernel if E_+ is connected and each $(d-1)$-dimensional marginal $p(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) = \int_{\mathbb{R}} p(x)\,\mathrm{d}x_j$ is locally bounded.

Let us consider

$$\int_{\mathbb{R}^d} f(x)p(x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j).$$

Assume the variables $Y_j = f(x_j)$ are i.i.d. with $\mathbb{E}[Y_j] = \overline{y}$ and $\mathrm{Var}(Y_j) = \sigma^2$. Define

$$\widetilde{Y}_N = \frac{1}{N}\sum_{j=1}^{N} Y_j \quad \text{and} \quad Z_N = \frac{\sqrt{N}(\widetilde{Y}_N - \overline{y})}{\sigma}.$$

Then $\widetilde{Y}_N \to \mathbb{E}[Y]$ a.s. (law of large numbers) and, asymptotically, Z_N is (standard) normally distributed according to $\mathcal{N}(0, I)$ (central limit theorem).

Loosely speaking, this means that

$$\sqrt{\left|\frac{1}{N}\sum_{j=1}^{N} f(x_j) - \int_{\mathbb{R}^d} f(x)p(x)\,\mathrm{d}x\right|^2} \approx \frac{\sigma}{\sqrt{N}} \quad \text{for } N \gg 1$$

provided that x_j are independent and f has finite mean and variance.

Autocovariance and correlation length

The independence of consecutive draws can be estimated from the sample itself. Suppose that we are interested in the convergence of the integral of $f(x)$ with respect to the probability density $p(x)$. Let us denote $z_j = f(x_j)$, where $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ is a MCMC sample and let $\overline{z} = N^{-1} \sum_{j=1}^{N} z_j$. Then we define the normalized autocovariance of the sample as

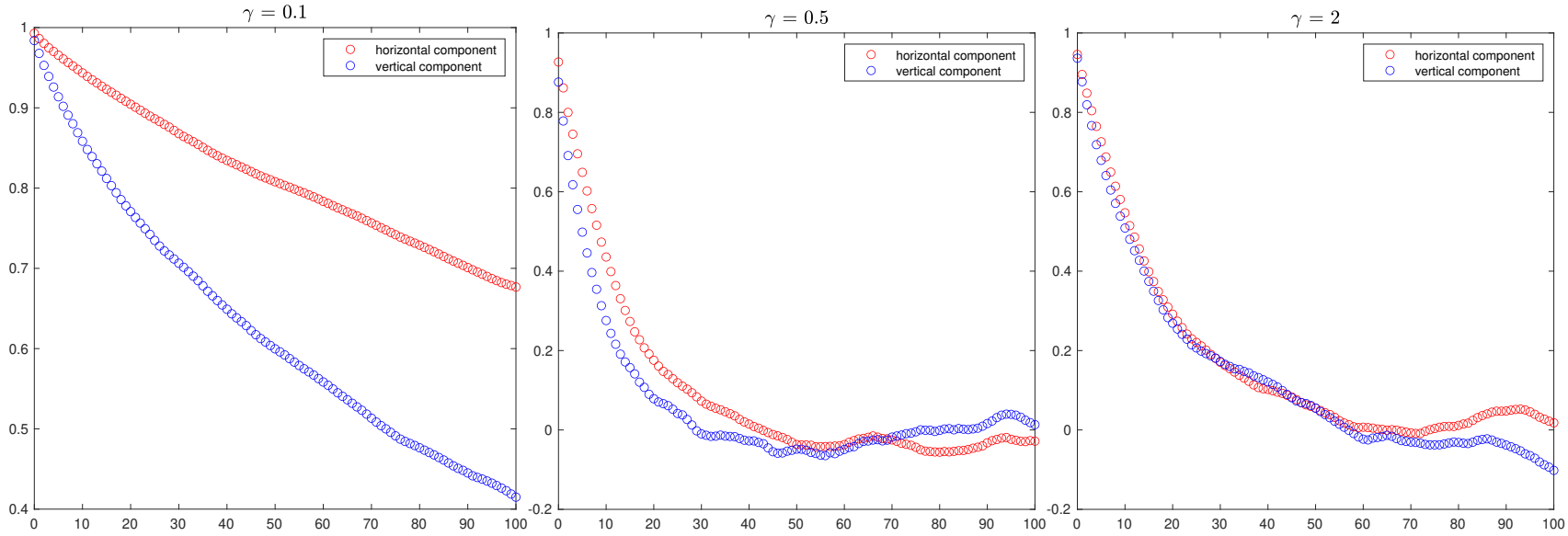
$$\gamma_k = \frac{1}{(N-k)\gamma_0} \sum_{j=1}^{N-k} (z_j - \overline{z})(z_{j+k} - \overline{z}), \quad k \geq 1,$$

where $\gamma_0 = N^{-1} \sum_{j=1}^{N} z_j^2$.

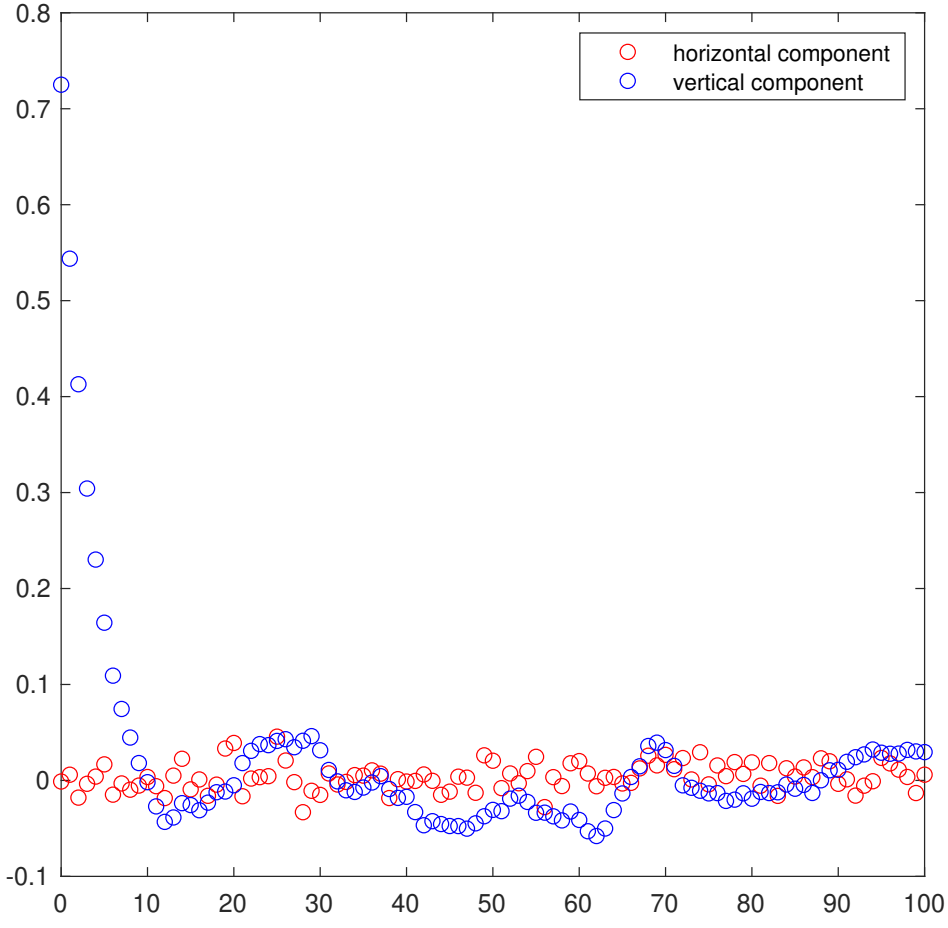
The correlation length of the sample $\{z_j\}_{j=1}^{N}$ can be estimated based on the decay of the normalized autocovariance sequence of the sample.

If every k^{th} sample point is independent, one might expect the discrepancy to behave as $1/\sqrt{N/k} = \sqrt{k/N}$ instead of $1/\sqrt{N}$. In consequence, one should try to choose the proposal distribution so that the *correlation length* is as small as possible.

Normalized autocovariance sequences for the Metropolis–Hastings example



Normalized autocovariance sequences for the Gibbs example



Preconditioned Crank–Nicolson algorithm

- The preconditioned Crank–Nicolson (pCN) algorithm is an instance of the Metropolis–Hastings algorithm with a specially chosen proposal density.
- The proposal is drawn using the model $Y = \sqrt{1-\beta^2}X + \beta W$, where $W \sim \mathcal{N}(0, C_0)$, C_0 is a symmetric and positive definite matrix, with the (*non-symmetric!*) kernel

$$q(x, y) \propto \exp\left(-\frac{1}{2\beta^2}(y - \sqrt{1-\beta^2}x)^{\mathrm{T}} C_0^{-1}(y - \sqrt{1-\beta^2}x) \right).$$

  Here, the step size $0 < \beta < 1$ is a free parameter (which can be optimized for statistical efficiency).
- The pCN method is *dimension robust*: the acceptance probability does not degenerate to zero as the dimension $d \to \infty$. Contrast this with, e.g., random walk Metropolis, whose acceptance probability degenerates to zero as the dimension $d \to \infty$.

Further variations of MCMC

We have only scratched the surface of some basic ideas surrounding MCMC methods. In the literature and practical applications, one can find many variations of these ideas to boost the performance of MCMC for "difficult" / "high-dimensional" problems. To list a couple of notable ones:

- Adaptive Metropolis: as the proposal density $q(x, y)$, use a random walk model $Y = X + W$ with $W \sim \mathcal{N}(0, \Gamma)$, where the covariance Γ is replaced by the *sample covariance* (plus some small perturbation of identity) computed using the sample history. The updating can happen either at every step or after every M steps of the Metropolis iteration. The main theoretical challenge is proving the ergodicity of the chain—this was proved by Haario, Saksman, and Tamminen (2001). Computationally, stable updating formulae for the sample means and covariances are needed in practice.
- Independence Metropolis: as the proposal density $q(x, y)$, use a probability density that is independent of the previous sample x, i.e., $q(x, y) = q(y)$. The proposal density should be similar to the target density.
- Metropolis-within-Gibbs, Delayed rejection adaptive Metropolis, ...

Software: https://mjlaine.github.io/mcmcstat/
https://mc-stan.org/

Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Thirteenth lecture, July 10, 2023

The setting

We work in the inverse problem setting of finding $x \in \mathbb{R}^d$ from $y \in \mathbb{R}^k$ given by

$$y = F(x) + \eta$$

with noise $\eta \sim \nu$ and prior $x \sim \pi$ such that $\eta \perp x$. The posterior density π^y of $x|y$ is given by Bayes' theorem

$$\pi^y(x) = \frac{1}{Z}\nu(y - F(x))\pi(x).$$

We have the negative log-likelihood:

$$\mathrm{L}(x) = -\log \nu\big(y - F(x)\big),$$

and a regularizer

$$\mathrm{R}(x) = -\log \pi(x).$$

So far we have mainly discussed point estimators: the MAP estimate

$$\hat{x}_{\mathsf{MAP}} = \arg\max_{x \in \mathbb{R}^d} \pi^y(x) = \arg\min_{x \in \mathbb{R}^d} (\mathsf{L}(x) + \mathsf{R}(x))$$

requires solving an optimization problem, and the CM estimate

$$\hat{x}_{\mathsf{CM}} = \int_{\mathbb{R}^d} x \pi^y(x) \, \mathrm{d}x$$

requires solving a high-dimensional integral. Recall that the latter can be achieved, e.g, by using MCMC to draw a sufficiently large sample from the posterior and computing the sample average. If we have a sample drawn from the posterior, we can use the sample to estimate other statistics such as the variance or credibility regions as well. Some alternatives to MCMC include importance sampling, high-dimensional cubature rules, etc.

Using point estimators reduces the complexity of Bayesian inference from determination of an entire distribution to determination of a single point. However, the approach has some limitations, in particular for noisy, multi-peaked or high-dimensional posterior distributions, where a point estimator may not capture enough information about the density.

Unimodal distributions

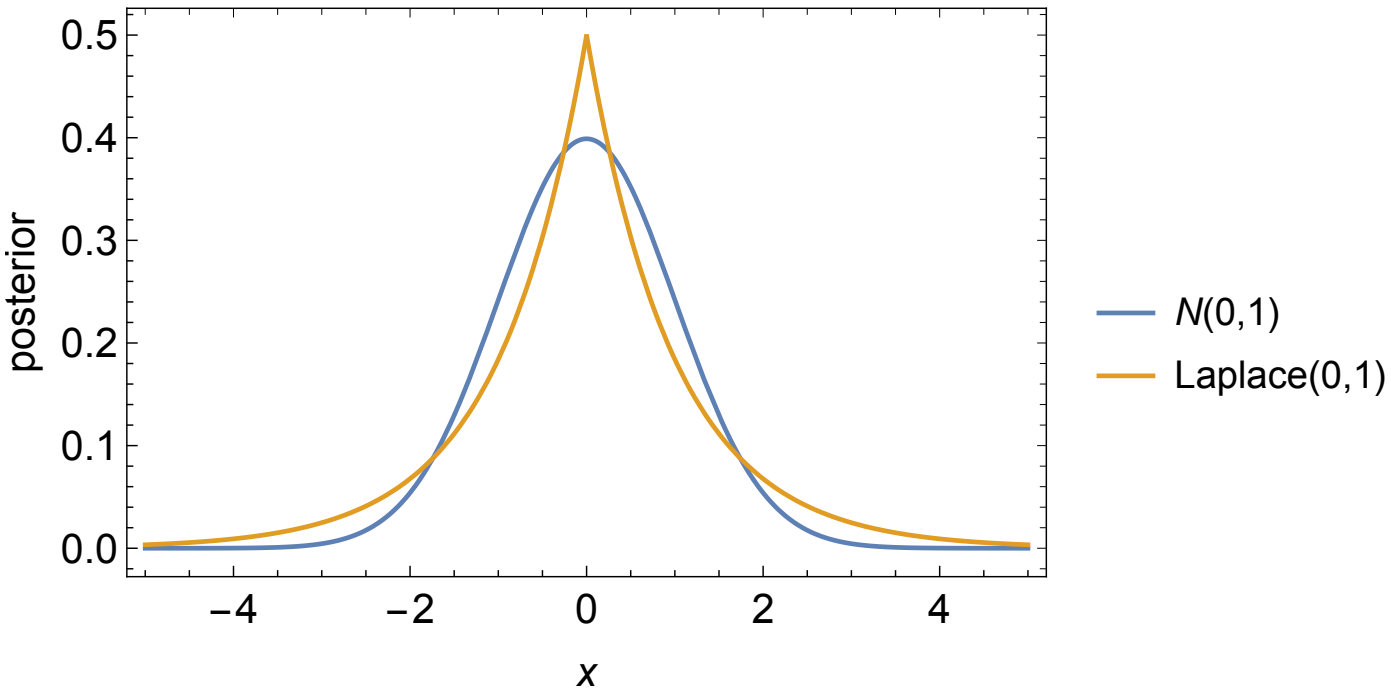


Figure: If the posterior is single-peaked, the MAP estimator reasonably summarizes the most likely value of the unknown parameter.

Problems with uneven distributions

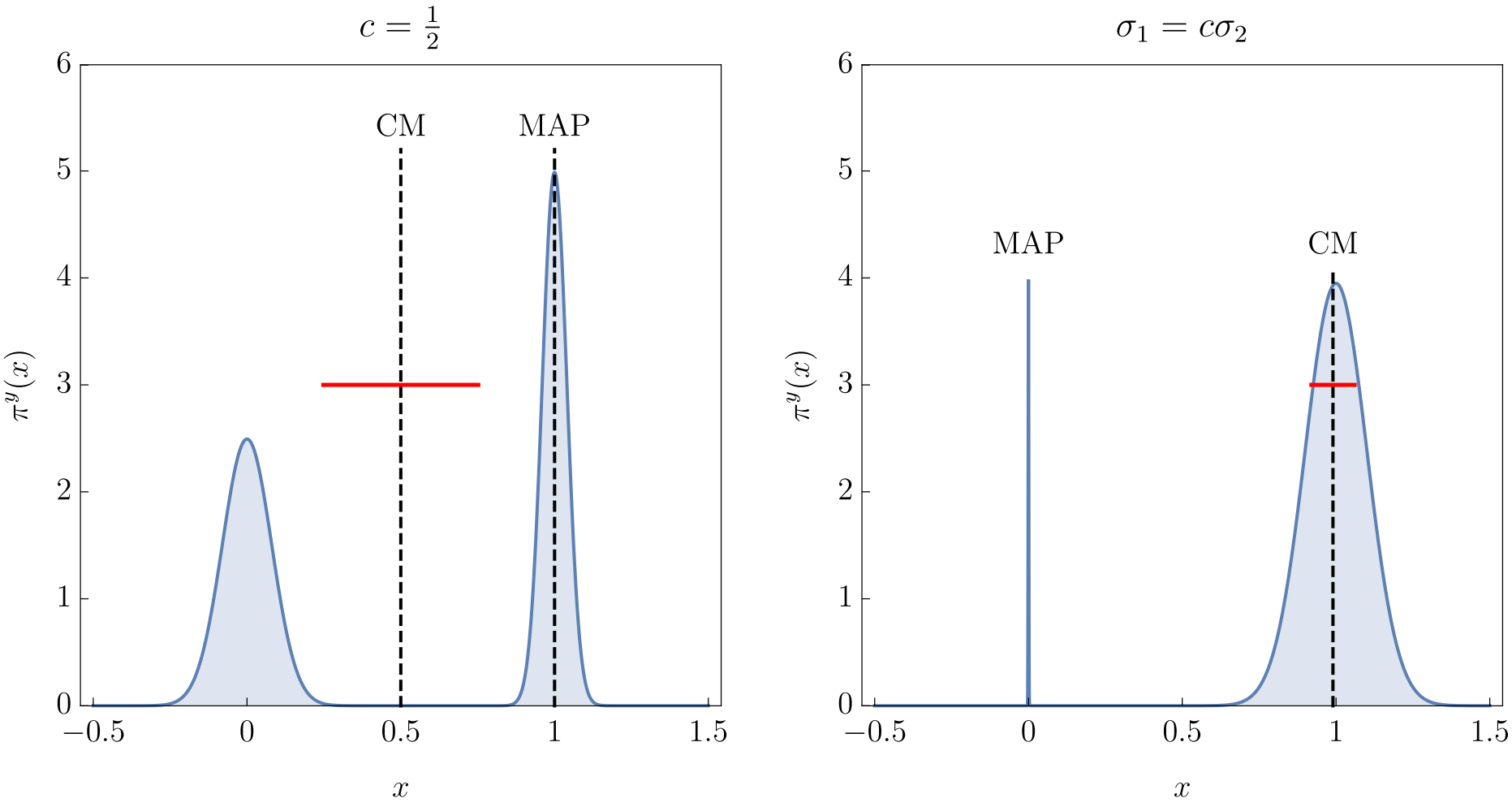


Figure: If the posterior is unevenly distributed, then it is less clear that the MAP or CM estimators usefully summarize the posterior.

Problems with rough distributions

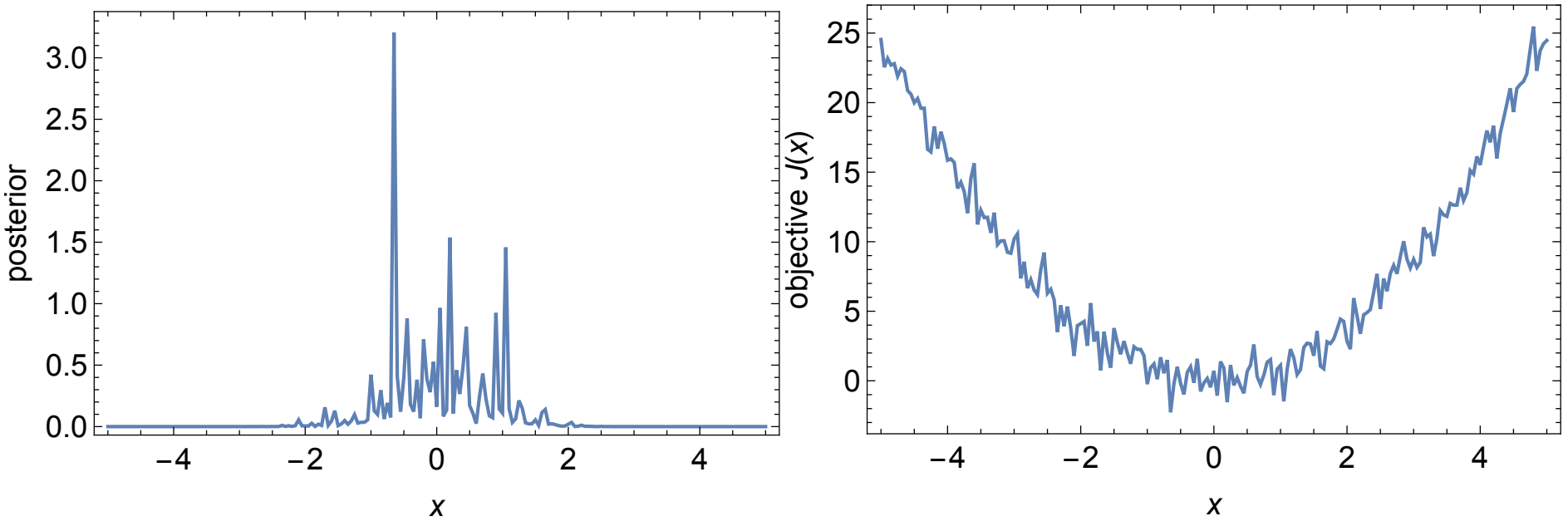


Figure: If the objective function J(x) is very rough (here it is a quadratic function contaminated with white noise), then the resulting posterior density is very rough.

The objective function has small-scale roughness, but it has a larger pattern. The MAP estimator cannot capture this larger pattern as it is found by minimizing the objective function. Arguably, $x = 0$ might be a better point estimate.

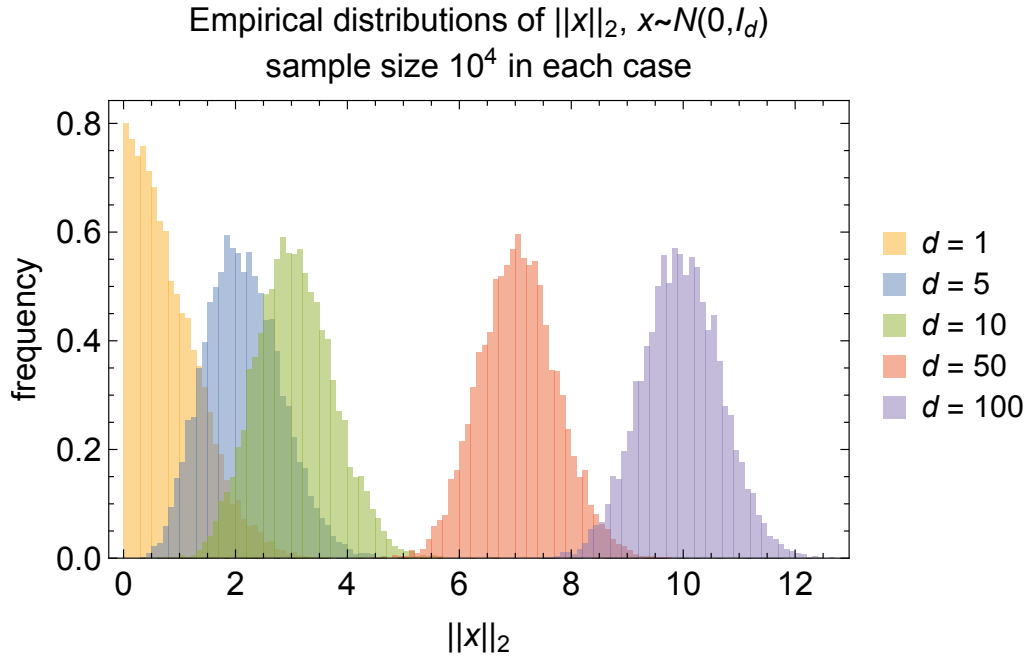
Problems with high dimension

Gaussian Annulus Theorem: Nearly all the probability of a d-dimensional spherical Gaussian distribution with unit variance is concentrated in a thin annulus of width $\mathcal{O}(1)$ at radius $\sqrt{d}$.

For example, if $x \sim \mathcal{N}(0, I_d)$, then

d	$\mathbb{P}(\|x\| < 5)$
10	0.99465
50	0.00119
100	1.135e-15

A point estimator may not capture enough information about the density.



Gaussian approximation

Instead of seeking a point estimator, we can try seeking a Gaussian distribution $p = \mathcal{N}(\mu, \Sigma)$ that minimizes the Kullback–Leibler divergence from the posterior $\pi^y(x)$. Since the Kullback–Leibler divergence is not symmetric this leads to two distinct problems, which we will consider separately.

The Kullback–Leibler divergence

Definition

Let $\pi, \pi' > 0$ be two probability distributions on $\mathbb{R}^d$. The *Kullback–Leibler (KL) divergence*, or *relative entropy*, of π with respect to π' is defined by

$$\begin{aligned} d_{\mathrm{KL}}(\pi \| \pi') &:= \int_{\mathbb{R}^d} \log\left(\frac{\pi(x)}{\pi'(x)}\right) \pi(x)\,\mathrm{d}x \\ &= \mathbb{E}^{\pi}\left[\log\left(\frac{\pi}{\pi'}\right)\right] \\ &= \mathbb{E}^{\pi'}\left[\log\left(\frac{\pi}{\pi'}\right)\frac{\pi}{\pi'}\right]. \end{aligned}$$

Kullback–Leibler is a divergence in that $d_{\mathrm{KL}}(\pi \| \pi') \geq 0$, with equality if and only if $\pi = \pi'$ a.e. However, unlike Hellinger and total variation, it is not a distance. In particular, the KL divergence is not symmetric: in general

$$d_{\mathrm{KL}}(\pi \| \pi') \neq d_{\mathrm{KL}}(\pi' \| \pi).$$

The KL divergence is useful for at least the following reasons:

- it provides an upper bound for many distances;
- its logarithmic structure allows explicit computations that are difficult using actual distances;
- it satisfies many convenient analytical properties such as being convex in both arguments and lower-semicontinuous in the topology of weak convergence;
- it has an information-theoretic and physical interpretation.

Lemma

The KL divergence provides the following upper bounds for Hellinger and total variation distance:

$$d_{\mathrm{H}}(\pi,\pi')^2 \leq \frac{1}{2} d_{\mathrm{KL}}(\pi\|\pi'), \quad d_{\mathrm{TV}}(\pi,\pi')^2 \leq d_{\mathrm{KL}}(\pi\|\pi').$$

Proof. Recall from Week 9 that $\frac{1}{\sqrt{2}} d_{\mathrm{TV}}(\pi,\pi') \leq d_{\mathrm{H}}(\pi,\pi')$ $\Leftrightarrow d_{\mathrm{TV}}(\pi,\pi')^2 \leq 2 d_{\mathrm{H}}(\pi,\pi')^2$. Thus the second inequality follows from the first one. We prove only the first inequality.

Consider the function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$ defined by

$$\phi(x) = x - 1 - \log x.$$

Note that

$$\begin{aligned}
\phi'(x) &= 1 - \frac{1}{x}, \\
\phi''(x) &= \frac{1}{x^2}, \\
\lim_{x\to+\infty} \phi(x) &= \infty = \lim_{x\to 0+} \phi(x).
\end{aligned}$$

Thus the function is convex on its domain. As the minimum of ϕ is attained at $x = 1$, and as $\phi(1) = 0$, we deduce that $\phi(x) \geq 0$ for all $x \in (0, \infty)$. Hence,

$$\begin{aligned}
x - 1 &\geq \log x \qquad \text{for all } x > 0, \\
\sqrt{x} - 1 &\geq \frac{1}{2} \log x \qquad \text{for all } x > 0.
\end{aligned}$$

We can use this last inequality to bound the Hellinger distance:

$$
\begin{aligned}
d_{\mathsf{H}}(\pi, \pi')^2 &= \frac{1}{2} \int_{\mathbb{R}^d} \left(1 - \sqrt{\frac{\pi'}{\pi}}\right)^2 \pi \, \mathrm{d}x \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left(1 + \frac{\pi'}{\pi} - 2\sqrt{\frac{\pi'}{\pi}}\right) \pi \, \mathrm{d}x \\
&= 1 - \int_{\mathbb{R}^d} \sqrt{\frac{\pi'}{\pi}} \pi \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(1 - \sqrt{\frac{\pi'}{\pi}}\right) \pi \, \mathrm{d}x \\
&\leq -\frac{1}{2} \int_{\mathbb{R}^d} \log\left(\frac{\pi'}{\pi}\right) \pi \, \mathrm{d}x = \frac{1}{2} \int_{\mathbb{R}^d} \log\left(\frac{\pi}{\pi'}\right) \pi \, \mathrm{d}x = \frac{1}{2} d_{\mathsf{KL}}(\pi \| \pi').
\end{aligned}
$$

Lemma

$d_{\mathrm{KL}}(\pi\|\pi') = 0$ if and only if $\pi = \pi'$ a.e.

Proof. The sufficient direction is trivial. For the necessary direction, suppose that $d_{\mathrm{KL}}(\pi\|\pi') = 0$. From the previous lemma, we deduce that

$$0 \leq d_{\mathrm{TV}}(\pi, \pi')^2 \leq d_{\mathrm{KL}}(\pi\|\pi') = 0$$

and therefore

$$d_{\mathrm{TV}}(\pi, \pi') = \frac{1}{2} \int_{\mathbb{R}^d} |\pi(x) - \pi'(x)| \, \mathrm{d}x = 0,$$

which can only hold if $\pi = \pi'$ a.e. □

Best Gaussian approximation

Let π be the target distribution, e.g., the posterior. We consider two different minimization problems, both leading to a "best Gaussian":

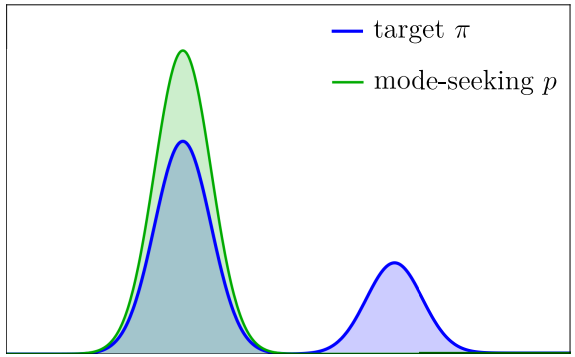
$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(p \| \pi) \qquad \text{("Mode-seeking Gaussian approximation")}$$

and

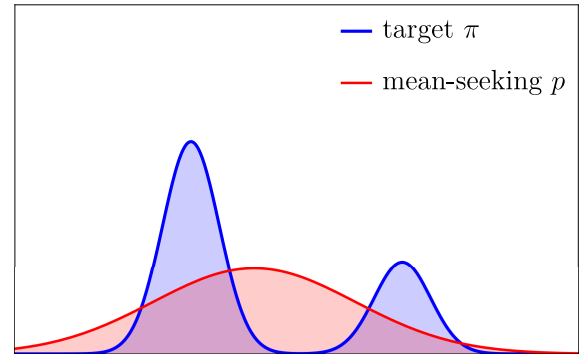
$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(\pi \| p), \qquad \text{("Mean-seeking Gaussian approximation")}$$

where the minimization is performed over the set of Gaussian distributions on $\mathbb{R}^d$ with positive definite covariance, i.e.,

$$\mathcal{A} := \{\mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d \times d} \text{ positive definite}\}.$$



(a) Minimizing $d_{\mathrm{KL}}(p\|\pi)$



(b) Minimizing $d_{\mathrm{KL}}(\pi\|p)$

- Fig. (a): Minimizing $d_{\mathrm{KL}}(p\|\pi)$ may miss out components of π – we want $\log\left(\frac{p}{\pi}\right)p$ to be small, which can happen when $p \approx \pi$ or $p \ll \pi$. Minimizing $d_{\mathrm{KL}}(p\|\pi)$ over Gaussians p can only give a single mode approximation which is achieved by matching one of the modes; we may think of this as "mode-seeking".
- Fig. (b): Minimizing $d_{\mathrm{KL}}(\pi\|p)$ over Gaussians p we want $\log\frac{\pi}{p}$ to be small where p appears as the denominator. Wherever π has some mass we must let p also have some mass there in order to keep $\frac{\pi}{p}$ as close as possible to one. The mass of p is allocated in a way such that on average the divergence between p and π attains its minimum; hence, it can be thought of as "mean-seeking".

Different applications will favor different choices between the mean and mode seeking approaches to Gaussian approximation.

Best Gaussian fit by minimizing $d_{\mathrm{KL}}(p\|\pi)$ ("mode-seeking")

Theorem (Best Gaussian approximation / "mode-seeking")

Suppose that the loss function $\mathsf{L}(x) := -\log \nu\big(y - F(x)\big)$ *is non-negative and bounded above and that the prior* $\pi \sim \mathcal{N}(0, \lambda^{-1} I)$. *Then there exists at least one probability distribution* $p \in \mathcal{A}$ *at which the infimum*

$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(p\|\pi^y)$$

is attained.

Proof. Let $p(x) = \frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}$, $\pi^y(x) = \frac{1}{Z}\mathrm{e}^{-\mathsf{L}(x)-\frac{\lambda}{2}\|x\|^2}$. Then

$$d_{\mathrm{KL}}(p\|\pi^y) = \mathbb{E}^p\left[\log\left(\frac{1}{(2\pi)^{d/2}(\det \Sigma)^{1/2}} \mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}\right) - \log\left(\frac{1}{Z}\mathrm{e}^{-\mathsf{L}(x)-\frac{\lambda}{2}\|x\|^2}\right)\right]$$

$$= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \log Z + \mathbb{E}^p\left[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}} + \mathsf{L}(x) + \frac{\lambda}{2}\|x\|^2\right]$$

$$d_{\mathsf{KL}}(p\|\pi^y) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \log Z + \mathbb{E}^p\bigg[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}} + \mathsf{L}(x) + \frac{\lambda}{2}\|x\|^2\bigg].$$

Note that Z is the normalization constant for π and is independent of p and hence of μ and Σ. We can represent a given random variable $x \sim p$ by writing $x = \mu + \Sigma^{1/2}\xi$, where $\xi \sim \mathcal{N}(0, I)$, and hence

$$\|x-\mu\|^2_{\Sigma^{-1}} = \|\Sigma^{1/2}\xi\|^2_{\Sigma^{-1}} = \|\xi\|^2 \quad \Rightarrow \quad \mathbb{E}^p\bigg[-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}\bigg] = -\frac{d}{2}.$$

Moreover,

$$\begin{aligned}
\mathbb{E}^p[\|x\|^2] &= \int_{\mathbb{R}^d} \|x-\mu+\mu\|^2 p(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \|x-\mu\|^2 p(x)\,\mathrm{d}x + 2\langle\mu, \int_{\mathbb{R}^d} x p(x)\,\mathrm{d}x\rangle - 2\langle\mu, \int_{\mathbb{R}^d} \mu p(x)\,\mathrm{d}x\rangle + \int_{\mathbb{R}^d} \|\mu\|^2 p(x)\,\mathrm{d}x \\
&= \mathsf{tr}(\Sigma) + 2\langle\mu,\mu\rangle - 2\langle\mu,\mu\rangle + \|\mu\|^2 = \mathsf{tr}(\Sigma) + \|\mu\|^2.
\end{aligned}$$

We obtain

$$d_{\mathrm{KL}}(p\|\pi^y) = -\frac{d}{2} - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma + \mathbb{E}^p \mathsf{L}(x) + \frac{\lambda}{2}\|\mu\|^2 + \frac{\lambda}{2}\mathrm{tr}(\Sigma) + \log Z.$$

Define $\mathcal{I}(\mu,\Sigma) = \mathbb{E}^p \mathsf{L}(x) + \frac{\lambda}{2}\|\mu\|^2 + \frac{\lambda}{2}\mathrm{tr}(\Sigma) - \frac{1}{2}\log\det\Sigma$. Note that there is a correspondence between minimizing $d_{\mathrm{KL}}(p\|\pi^y)$ over $p \in \mathcal{A}$ and minimizing $\mathcal{I}(\mu,\Sigma)$ over $\mu \in \mathbb{R}^d$ and positive definite Σ. Moreover:

- $\mathcal{I}(0, I) < \infty$.
- For any Σ, $\mathcal{I}(\mu,\Sigma) \to \infty$ as $\|\mu\| \to \infty$.
- For any μ, $\mathcal{I}(\mu,\Sigma) \to \infty$ as $\mathrm{tr}(\Sigma) \to 0$ or $\mathrm{tr}(\Sigma) \to \infty$.

Therefore, there are $M, r, R > 0$ such that the infimum of $\mathcal{I}(\mu,\Sigma)$ over $\mu \in \mathbb{R}^d$ and positive definite Σ is equal to the infimum of $\mathcal{I}(\mu,\Sigma)$ over

$$\tilde{\mathcal{A}} := \{(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d} \text{ positive-definite symmetric}, \|\mu\| \le M, r \le \mathrm{tr}(\Sigma) \le R\}.$$

Since $\mathcal{I}$ is continuous in $\tilde{\mathcal{A}}$ it achieves its infimum and the proof is complete. □

We remark that the theorem establishes the existence of a best Gaussian approximation. However, minimizers need not be unique.

Best Gaussian fit by minimizing $d_{\mathrm{KL}}(\pi\|p)$ ("mean-seeking")

The best Gaussian approximation in Kullback–Leibler with respect to its second argument is unique and given by moment matching.

Theorem (Best Gaussian by moment matching / "mean-seeking")

Assume that $\bar{\mu} := \mathbb{E}^{\pi}[x]$ *is finite and that* $\bar{\Sigma} := \mathbb{E}^{\pi}[(x-\bar{\mu})(x-\bar{\mu})^{\mathrm{T}}]$ *is positive definite. (Here,* π *denotes the target distribution, e.g., the posterior.) Then the infimum*

$$\inf_{p \in \mathcal{A}} d_{\mathrm{KL}}(\pi\|p)$$

is attained by $p = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$.

Proof. Note that $d_{\mathrm{KL}}(\pi\|p) = -\mathbb{E}^{\pi}[\log p] + \overbrace{\mathbb{E}^{\pi}[\log \pi]}^{\text{independent of } p}$. Since we want a Gaussian minimizer, write $p(x) = ((2\pi)^d |\det \Sigma|)^{-1/2} \exp\big(-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}\big)$

$$\begin{aligned}
\Rightarrow -\mathbb{E}^{\pi}[\log p] &= -\mathbb{E}^{\pi}[\log\big((2\pi)^{-d/2}(\det\Sigma)^{-1/2}\mathrm{e}^{-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}}\big)] \\
&= \tfrac{1}{2}\mathbb{E}^{\pi}[\|x-\mu\|^2_{\Sigma^{-1}}] + \tfrac{1}{2}\log\det\Sigma + \tfrac{d}{2}\log(2\pi).
\end{aligned}$$

Note that the final term is irrelevant for the optimization problem.

Let $\Lambda := \Sigma^{-1}$. Our task is equivalent to finding the minimizer of

$$I(\mu, \Lambda) := \frac{1}{2}\mathbb{E}^{\pi}[(x-\mu)\Lambda(x-\mu)^{\mathrm{T}}] - \frac{1}{2}\log\det\Lambda.$$

Let $\Lambda = (\Lambda_{ij})_{i,j=1}^{d}$. We can view the above functional as the $d + d^2$ variate function $I(\mu_1, \ldots, \mu_d, \Lambda_{11}, \Lambda_{12}, \ldots, \Lambda_{dd})$. Thus, we only need to show that

$$\nabla I(\bar{\mu}, \bar{\Sigma}^{-1}) = 0 \quad \text{and} \quad \nabla^2 I(\mu, \Sigma^{-1}) > 0 \quad \text{for all } \mu, \Sigma.$$

($(\bar{\mu}, \bar{\Sigma}^{-1})$ is the critical point and the objective function is convex.)

By defining the notations $\partial_\mu f := \left(\frac{\partial f}{\partial \mu_i}\right)_{i=1}^{d}$ (gradient w.r.t. vector μ) and $\partial_\Lambda f := \left(\frac{\partial f}{\partial \Lambda_{ji}}\right)_{i,j=1}^{d}$ (gradient w.r.t. vector $(\Lambda_{11}, \Lambda_{12}, \ldots, \Lambda_{dd})$, reshaped into a $d \times d$ matrix), we easily see that $\nabla I = 0$ can be expressed as the pair

$$\begin{cases} 0 = \partial_\mu I & = -\mathbb{E}^{\pi}[\Lambda(x-\mu)] = 0 \\ 0 = \partial_\Lambda I & = \frac{1}{2}\partial_\Lambda(\mathbb{E}^{\pi}[(x-\mu)\Lambda(x-\mu)^{\mathrm{T}}]) - \frac{1}{2\det\Lambda}\partial_\Lambda \det\Lambda \\ & = \frac{1}{2}\mathbb{E}^{\pi}[(x-\mu)(x-\mu)^{\mathrm{T}}] - \frac{1}{2}\Lambda^{-1}, \end{cases}$$

where we used a special case of Jacobi's formula $\partial_\Lambda \det\Lambda = \det\Lambda \cdot \Lambda^{-1}$. Clearly, $(x, \Lambda) = (\bar{\mu}, \bar{\Sigma}^{-1})$ is the critical point satisfying the above condition.

Finally, we need to show that $\nabla^2 I(\mu, \Sigma^{-1})$ is positive definite. To this end, we note that

$$p(x) = \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Lambda(x-\mu)} = \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x - \frac{1}{2}\mu^{\mathrm{T}}\Lambda\mu}$$
$$= \sqrt{\frac{\det \Lambda}{(2\pi)^d}} \mathrm{e}^{-\frac{1}{2}\mu^{\mathrm{T}}\Lambda\mu} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x} = \frac{\mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x}}{\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x}\,\mathrm{d}x}.$$

Noting that $x^{\mathrm{T}}\Lambda x = \sum_{i,j=1}^d \Lambda_{ij} x_i x_j = \sum_{i,j=1}^d \Lambda_{ij}(xx^{\mathrm{T}})_{ij}$, we can write $x^{\mathrm{T}}\Lambda x = \mathrm{vec}(\Lambda) \cdot \mathrm{vec}(xx^{\mathrm{T}})$, where we define

$$\mathrm{vec}(M) := (M_{11}, M_{12}, \ldots, M_{dd})^{\mathrm{T}} \quad \text{for } M \in \mathbb{R}^{d\times d}.$$

In particular,

$$-\frac{1}{2}x^{\mathrm{T}}\Lambda x + \mu^{\mathrm{T}}\Lambda x = \underbrace{\begin{bmatrix} \Lambda\mu \\ -\frac{1}{2}\mathrm{vec}(\Lambda) \end{bmatrix}}_{=:\theta}^{\mathrm{T}} \underbrace{\begin{bmatrix} x \\ \mathrm{vec}(xx^{\mathrm{T}}) \end{bmatrix}}_{=:T(x)}$$

and we can write $p_\theta(x) := p(x) = \frac{1}{Z(\theta)} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}$, $Z(\theta) := \int_{\mathbb{R}^d} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}\,\mathrm{d}x$.

The importance of the characterization

$$p_\theta(x) = \frac{1}{Z(\theta)} \mathrm{e}^{\theta^{\mathrm{T}} T(x)}, \quad Z(\theta) := \int_{\mathbb{R}^d} \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x,$$

lies in the fact that *every possible Gaussian PDF* can be parameterized by the vector $\theta = (\theta_1, \ldots, \theta_{d+d^2})^{\mathrm{T}}$. Thus, the KL divergence $d_{\mathsf{KL}}(\pi \| p_\theta)$ that we are interested in can be recast as

$$\begin{aligned} H(\theta) :=& d_{\mathsf{KL}}(\pi \| p_\theta) = -\mathbb{E}^\pi[\log p_\theta] + \mathbb{E}^\pi[\log \pi] \\ =& -\theta^{\mathrm{T}} \mathbb{E}^\pi[T(x)] + \log Z(\theta) + \mathbb{E}^\pi[\log \pi]. \end{aligned}$$

Noting that $\nabla_\theta^2(\theta^{\mathrm{T}} \mathbb{E}^\pi[T(x)]) = 0$ and $\frac{\partial \log Z(\theta)}{\partial \theta_i} = \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} \frac{\partial}{\partial \theta_i} \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x$ $= \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x$, we compute

$$\begin{aligned} &[\nabla_\theta^2 H(\theta)]_{ij} = \frac{\partial^2 \log Z(\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(\frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) \\ &= -\frac{1}{Z(\theta)^2} \left(\int_{\mathbb{R}^d} T_i(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) \left(\int_{\mathbb{R}^d} T_j(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \right) + \frac{1}{Z(\theta)} \int_{\mathbb{R}^d} T_i(x) T_j(x) \mathrm{e}^{\theta^{\mathrm{T}} T(x)} \, \mathrm{d}x \\ &= \mathbb{E}^{p_\theta}[T_i T_j] - \mathbb{E}^{p_\theta}[T_i] \mathbb{E}^{p_\theta}[T_j] = [\mathrm{Cov}^{p_\theta}(T)]_{ij}, \end{aligned}$$

which is positive definite. □

Remark. Notice that the preceding proof of convexity holds for any distribution p that can be parameterized by the following more general expression:

$$p_\theta(x) = h(x)\exp\left(\theta^T T(x) - A(\theta)\right) \tag{1}$$

$$\text{with} \qquad A(\theta) = \log\left[\int_{\mathbb{R}^d} h(x)\exp\left(\theta^T T(x)\right) dx\right].$$

Since $h(x)$ is independent of θ, the conclusion of the previous theorem carries over to distributions with the form of (1). Such distributions belong to the *exponential family* in the statistics literature. Here, θ is called the natural parameter, $T(x)$ the sufficient statistic, $h(x)$ the base measure, and $A(\theta)$ the log-partition.

The Gaussian distribution is a special case in which $h(x)$ is constant with respect to x.

Variational formulation of Bayes' theorem

We have been concerned with finding the best Gaussian approximations to a measure with respect to KL divergences. Bayes' theorem itself can be formulated through a closely related minimization principle. Consider a posterior $\pi^y(x)$ in the following form:

$$\pi^y(x) = \frac{1}{Z}\exp\big(-\mathsf{L}(x)\big)\pi(x),$$

where $\pi(x)$ is the prior, $\mathsf{L}(x)$ is the negative log-likelihood, and Z the normalization constant. We assume here for exposition that all densities are positive. Let p be an *arbitrary* PDF. Then we can express $d_{\mathsf{KL}}(p\|\pi^y)$ as

$$\begin{aligned} d_{\mathsf{KL}}(p\|\pi^y) &= \int_{\mathbb{R}^d} \log\left(\frac{p}{\pi^y}\right) p\,\mathrm{d}x = \int_{\mathbb{R}^d} \log\left(\frac{p}{\pi}\frac{\pi}{\pi^y}\right) p\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \log\left(\frac{p}{\pi}\exp\big(\mathsf{L}(x)\big)Z\right) p\,\mathrm{d}x \\ &= d_{\mathsf{KL}}(p\|\pi) + \mathbb{E}^p[\mathsf{L}(x)] + \log Z. \end{aligned}$$

If we define

$$\mathcal{J}(p) = d_{\mathsf{KL}}(p\|\pi) + \mathbb{E}^p[\mathsf{L}(x)]$$

then we have the following:

Theorem (Bayes' theorem as an optimization principle)

The posterior distribution π^y is given by the following minimization principle:

$$\pi^y = \arg\min_{p\in\mathscr{P}} \mathcal{J}(p),$$

where $\mathscr{P}$ contains all probability densities on $\mathbb{R}^d$.

Proof.
Since Z is the normalization constant for π^y and is independent of p, the minimizer of $d_{\mathsf{KL}}(p\|\pi^y)$ will also be the minimizer of $\mathcal{J}(p)$. Since the global minimizer of $d_{\mathsf{KL}}(p\|\pi^y)$ is attained at $p = \pi^y$ the result follows. □

Why is it useful to view the posterior as the minimizer of an energy?

- The variational formulation provides a natural way to approximate the posterior by restricting the minimization problem to distributions satisfying some computationally desirable property.
    - For instance, variational Bayes methods often restrict the minimization to densities with product structure and in this chapter we have studied restriction to the class of Gaussian distributions.
- Variational formulations provide natural paths, defined by a gradient flow, towards the posterior. Understanding these flows and their rates of convergence is helpful in the choice of sampling algorithms.

Appendix

The material on slides 28–33 was **not** considered during the 2023 course and it is **not** part of the course exam.

Consider still the problem of finding $x \in \mathbb{R}^d$ from $y \in \mathbb{R}^k$ given by

$$y = F(x) + \eta$$

with noise $\eta \sim \nu$ and prior $x \sim \pi$ such that $\eta \perp x$. The posterior density π^y of $x|y$ is given by Bayes' theorem

$$\pi^y(x) = \frac{1}{Z}\nu(y - F(x))\pi(x).$$

We have the negative log-likelihood:

$$\mathsf{L}(x) = -\log \nu\big(y - F(x)\big),$$

and a regularizer

$$\mathsf{R}(x) = -\log \pi(x).$$

When added together these two functions of x comprise an objective function of the form

$$\mathsf{J}(x) = \mathsf{L}(x) + \mathsf{R}(x).$$

Furthermore

$$\pi^y(x) = \frac{1}{Z}\nu\big(y - F(x)\big)\pi(x) \propto e^{-\mathsf{J}(x)}.$$

We see that minimizing the objective function J is equivalent to maximizing the posterior π^y. Therefore, the MAP estimator can be rewritten in terms of J as follows:

$$\hat{x}_{\mathsf{MAP}} = \arg\max_{x\in\mathbb{R}^d} \pi^y(x) = \arg\min_{x\in\mathbb{R}^d} \mathsf{J}(x).$$

Let us consider conditions under which the MAP estimator is attained, and characterize the MAP estimator in terms of small ball probabilities – this interpretation generalizes the definition of MAP estimators to measures that do not possess a Lebesgue density.

For any optimization problem for an objective function with a finite infimum, it is of interest to determine whether the infimum is attained.

Theorem (Attainable MAP estimator)

Assume that J *is non-negative, continuous and that* $\mathrm{J}(x) \to \infty$ *as* $|x| \to \infty$. *Then* J *attains its infimum. Therefore, the MAP estimator of* x *based on the posterior* $\pi^y(x) \propto \exp\big(-\mathrm{J}(x)\big)$ *is attained.*

Proof.
By the assumed growth and non-negativity of J, there is R such that $\inf_{x\in\mathbb{R}^d} \mathrm{J}(x) = \inf_{x\in\bar{B}(0,R)} \mathrm{J}(x)$ where $\bar{B}(0,R)$ denotes the closed ball of radius R around the origin. Since J is assumed to be continuous, its infimum over $\bar{B}(0,R)$ is attained and the proof is complete. □

Remark. The assumption that $\mathrm{J}(x) \to \infty$ is not restrictive: this condition needs to hold in order to be able to normalize $\pi^y(x) \propto \exp\big(-\mathrm{J}(x)\big)$ into a PDF, which is implicitly assumed in the second part of the theorem statement.

Example. Suppose that

1. $F : \mathbb{R}^d \to \mathbb{R}^k$ is continuous and $\eta \sim \mathcal{N}(0, \Gamma)$;
2. the objective function $\mathrm{J}(x) = \mathrm{L}(x) + \mathrm{R}(x)$ has Γ-weighted L^2 loss

$$\mathrm{L}(x) = \frac{1}{2}\|y - F(x)\|^2_{\Gamma^{-1}}$$

and L^p regularizer

$$\mathrm{R}(x) = \frac{\lambda}{p}\|u\|_p^p, \quad p \in (0, \infty).$$

Then the assumptions on J in the previous theorem are satisfied, and the infimum of J is attained at the MAP estimator of the corresponding Bayesian problem with posterior PDF proportional to $\exp\big(-\mathrm{J}(u)\big)$.

Intuitively the MAP estimator maximizes posterior probability. We make this precise in the following theorem which links the objective function J to small ball probabilities.

Theorem (Objective function and posterior probability)

Assume that J *is non-negative, continuous and that* $\mathsf{J}(x) \to \infty$ *as* $|x| \to \infty$. *Let*

$$\alpha(x, \delta) := \int_{B(x,\delta)} \pi^y(v)\,\mathrm{d}v = \mathbb{P}^{\pi^y}\big(B(x, \delta)\big),$$

be the posterior probability of a ball with radius δ *centered at* x. *Then, for all* $x_1, x_2 \in \mathbb{R}^d$, *we have*

$$\lim_{\delta \to 0} \frac{\alpha(x_1, \delta)}{\alpha(x_2, \delta)} = e^{\mathsf{J}(x_2) - \mathsf{J}(x_1)}.$$

Remark: For fixed x_2, the right-hand side is maximized at point x_1 that minimizes J. Independently of the choice of any fixed x_2, the above result shows that the probability of a small ball of radius δ centered at x_1 is, approximately, maximized by choosing the centre at a minimizer of J.

This result essentially characterizes the MAP estimate and, since it makes no reference to Lebesgue density, it can be generalized to infinite dimensions.

Proof. Let $x_1, x_2 \in \mathbb{R}^d$, $\varepsilon > 0$. By continuity of J, for all sufficiently small δ:

$$x \in \bar{B}(x_j, \delta) \quad \Rightarrow \quad |\mathsf{J}(x) - \mathsf{J}(x_j)| \le \varepsilon, \quad j \in \{1, 2\},$$

and therefore

$$\begin{aligned} e^{-\mathsf{J}(x_1)-\varepsilon} \le e^{-\mathsf{J}(v)} \le e^{-\mathsf{J}(x_1)+\varepsilon} \quad & \text{for all } v \in B(x_1, \delta), \\ e^{-\mathsf{J}(x_2)-\varepsilon} \le e^{-\mathsf{J}(v)} \le e^{-\mathsf{J}(x_2)+\varepsilon} \quad & \text{for all } v \in B(x_2, \delta). \end{aligned}$$

It follows, for all δ sufficiently small, that

$$\begin{aligned} B_\delta e^{-\mathsf{J}(x_1)-\varepsilon} &\le \int_{B(x_1,\delta)} e^{-\mathsf{J}(v)} \, \mathrm{d}v \le B_\delta e^{-\mathsf{J}(x_1)+\varepsilon}, \\ B_\delta e^{-\mathsf{J}(x_2)-\varepsilon} &\le \int_{B(x_2,\delta)} e^{-\mathsf{J}(v)} \, \mathrm{d}v \le B_\delta e^{-\mathsf{J}(x_2)+\varepsilon}, \end{aligned}$$

where B_δ is the Lebesgue measure of a ball with radius δ. Taking the ratio of α's and using the above bounds we obtain that, for all δ sufficiently small,

$$e^{\mathsf{J}(x_2)-\mathsf{J}(x_1)-2\varepsilon} \le \frac{\alpha(x_1, \delta)}{\alpha(x_2, \delta)} \le e^{\mathsf{J}(x_2)-\mathsf{J}(x_1)+2\varepsilon}.$$

Since ε was arbitrary, the desired result follows. □