Inverse Problems
Sommersemester 2022

Vesa Kaarnioja

vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fifth lecture, May 23, 2022

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 $F \colon \mathbb{R}^d \to \mathbb{R}^k$ 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 $i \neq j$),

$$\mu\bigg(\bigcup_{k=1}^{\infty} A_k \bigg) = \sum_{k=1}^{\infty} \mu(A_n),$$

(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

$$\delta_m(A) = \begin{cases} 1 & \text{if } m \in A, \\ 0 & \text{if } m \notin A \end{cases} \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d).$$

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

$$\mu(A) = \int_A f(\omega)\nu(\mathrm{d}\omega) \quad \text{for all } A \in \mathcal{F}.$$

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.,

$$\bar{x} \in \arg\max_{x \in \mathbb{R}^d} p(x).$$

- 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

$$\varphi_x(h) = \exp\left(i\, h^{\mathrm{T}} m - \frac{1}{2} h^{\mathrm{T}} C h \right) \quad \text{for all } h \in \mathbb{R}^d.$$

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

$$\mathbb{P}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(x-m)^{\mathrm{T}} C^{-1}(x-m)\right)$$

$$= \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}\|C^{-\frac{1}{2}}(x-m)\|^2\right).$$

  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

$$\mathbb{P}(x \in A) = \mathbb{P}(x \in A, y \in \mathbb{R}^k) = \int_{A \times \mathbb{R}^k} p_{x,y}(u, v) \mathrm{d}(u, v)$$
$$= \int_A \left(\int_{\mathbb{R}^k} p_{x,y}(u, v) \mathrm{d}v \right) \mathrm{d}u = \int_A p_x(u) \mathrm{d}u$$

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 realisation y of the measured data.

Lemma

Under assumptions $A1 - A3$, 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 $\eta \sim \nu$. $\qquad\qquad\square$

Theorem (Bayes' theorem)

If assumptions $A1 - A3$ *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 $A1 - A3$.

- 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 $\mathbb{P}(y|x) \propto \exp(-\frac{1}{2\sigma^2} \sum_{j=0}^{1} (y_j - \frac{1}{|x-j|})^2)$.

- 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

$$\pi^y(x) \propto \chi_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=0}^{1} \left(y_j - \frac{1}{|x - j|} \right)^2 \right).$$

Let us visualize the posterior density against the ground truth solution.
(See also file week5.m on the course homepage!)
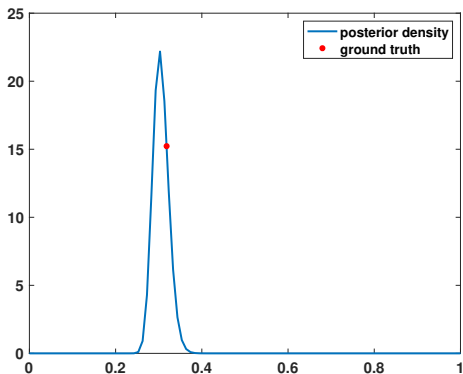
We see that the posterior is localized around the true parameter value
("ground truth"). *Note that in this case, the prior hardly plays any role.*

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 $x = 1$?
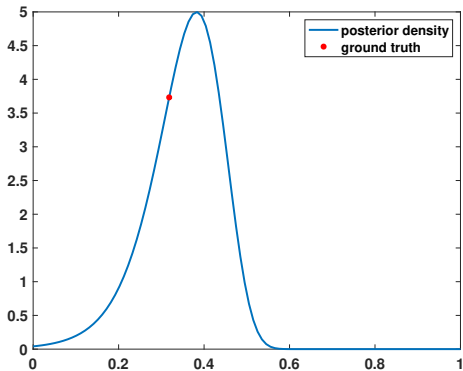
The resulting posterior distribution carries substantially more uncertainty since we now have less measurement data!

Note that the posterior will generally be high-dimensional, meaning that it is usually not possible to visually inspect the posterior density.