Bayesian inversion: theoretical perspective

Lecture notes, spring 2016 course

Tapio Helin

February 29, 2016

Abstract

These lecture notes are written for the theoretical part of the course "Bayesian inversion" given at University of Helsinki during Spring 2016. We rely on several references, but most important are the extensive review paper [2] by Andrew Stuart and the lecture notes [1] by Dashti and Stuart. The notes are updated (nonlinearly) as the course progresses.

Contents

1 Short motivation

Consider an indirect physical measurement, which can be approximatively modelled by a linear equation

$$m = Ax. \tag{1}$$

Above, $x, m \in \mathbb{R}^n$ describe the unknown and the measurement, respectively, and matrix $A \in \mathbb{R}^{n\times n}$ models how these two quantity are related via physics. Among inverse problems

research community, we are in the business of solving x given the measurement data ideally modelled by m. This task is made non-trivial by considering problems where the underlying mathematical model (approximated by (1)) is ill-posed. The classical definition of a well-posed problem by Hadamard states that (a) *a solution must exist* and that it is (b) *unique*. Moreover, the (c) *solution has to depend continuously on the data*. An ill-posed problem violates at least one of these conditions.

The violation of the stability condition (c) typically leads to numerical challenges in inverse problems that for problem (1) appear as a high condition number of matrix A. Recall that the condition number of A is defined by

$$\text{cond}(A) = \frac{\lambda_{max}}{\lambda_{min}}.$$

For example, let us assume that $\lambda_{max} = 1$ and $\lambda_{min} = \epsilon$, where λ_{max} and λ_{min} correspond the largest and smallest eigenvalue, respectively, and $\epsilon > 0$ is very small. Any real-life measurement is contaminated by some noise. Hence, it is reasonable to assume that our measurement is obtained as

$$m^\delta = Ax_0 + \delta,$$

where x_0 describes the 'true' value and δ describes the measurement noise. Notice that we do not know δ exactly and in the best case scenario we might have some estimate concerning its size/norm. Even if A is invertible, a naive reconstruction by

$$A^{-1}m^\delta = x_0 + A^{-1}\delta =: x_0 + \widetilde{\delta}$$

easily leads to useless approximation since in the worst case the error

$$\|\widetilde{\delta}\|_2 \approx \frac{\|\delta\|_2}{\epsilon}$$

can be arbitrarily large. This illustrates one key perspective of inverse problem theory: how to stabilize the reconstruction process while maintaining acceptable accuracy.

The theory related to deterministic problems like (1) is called *regularization theory* and is discussed in more detail in the usual *Inverse problems* course. One of the fundamental ideas of regularization theory is to approximate the problem (1) by a stable one. In the classical Tikhonov regularization (1) is replaced by a variational problem

$$\min_{x\in\mathbb{R}^n} \left(\left\| Ax - m^\delta \right\|_2^2 + \alpha \|x\|_2^2 \right). \tag{2}$$

The solution to (2) is given by

$$x_\alpha^\delta = (A^\top A + \alpha I)^{-1} A^\top m^\delta =: R_\alpha m^\delta,$$

where the reconstruction matrix R_α has a modified eigenvalue structure compared to A^{-1}. Most importantly, we have

$$\lambda_{min}(R_\alpha) = \frac{\lambda_{min}}{\lambda_{min}^2 + \alpha}$$

and hence the reconstruction is more stable (reconstruction error is comparable to $\frac{1}{\lambda_{min}(R_\alpha)}$). However, we have induced new kind of reconstruction error. Namely, we have that

$$R_\alpha m^\delta = x_0 + (R_\alpha A - I)x_0 + R_\alpha \delta$$

where the two error terms on the right hand side are approximately of size

$$\|(R_\alpha A - I)x_0\|_2 \approx \frac{\alpha}{\lambda_{min}^2 + \alpha} \|x_0\|_2 \tag{3}$$

and

$$\|R_\alpha \delta\|_2 \approx \frac{\lambda_{min}}{\lambda_{min}^2 + \alpha} \|\delta\|_2 . \tag{4}$$

Important effect is this: if α is increased, the first error term (3) increases (becoming comparable to $\|x_0\|_2$!). Meanwhile, if α is decreased, the second error term explodes. The optimal strategy is a balance between the two errors. The key observation is that the more accurate information you have related to the unknown x_0, the noise δ and structure of the problem (here: eigenvalue structure), the more effective you can make your regularization strategy.

The topic of this course, Bayesian inversion, rephrases the problem (1) as a question of statistical inference: consider a problem

$$M = AX + \mathcal{E}, \tag{5}$$

where the quantities describing our measurement, unknown and noise are replaced by random variables. Here, $X : \Omega \to \mathbb{R}^n$ and $M, \mathcal{E} : \Omega \to \mathbb{R}^d$, where Ω is our probability space. Randomness in this framework describes our lack of knowledge related to their exact values. The degree of our information is encoded into their probability distributions. The solution to (5) is so-called *posterior distribution*, i.e., the conditional probability of X given measurement $M = m^\delta$.

The randomness (or uncertainty) can appear due to several effects in a practical measurement setting. It can appear via some statistical information which is available about the unknown or the model. Randomness can also reflect the lack of information about correct parameter values in the model. Ultimately, the noise in any practical measurement is always random.

In practise, the posterior distribution is obtained via the Bayes formula which states, using probability densities, that

$$\pi_{post}(x \mid m) = \frac{\pi_{like}(m \mid x)\pi_X(x)}{\pi_M(m)}, \tag{6}$$

where π_{post}, π_X and π_M are the posterior, prior and marginal probability densities (we of course need to assume they exist). The likelihood density $\pi_{like}(m|x)$ expresses the likelihood of measurement outcome m given $X = x$. We will come back to these objects later, but let us now jump a little bit ahead of ourselves and illustrate how the stabilization discussed above plays out here.

In the Bayesian scheme the ill-posedness of the model is stabilized (mainly) by our *a priori* information regarding X. Suppose that $\mathcal{E}$ is random vector in $\mathbb{R}^d$ with normally distributed independent components. Similarly, let us assume that X has normally independent components but with variance $\frac{1}{\alpha}$. It turns out that the respective probability densities are of the form

$$\pi_X(x) \simeq \exp(-\alpha \|x\|_2^2) \quad \text{and} \quad \pi_{\mathcal{E}}(e) \simeq \exp(-\|e\|_2^2).$$

Above and throughout these notes, the notation $f \simeq g$ means that functions f and g coincide up to a constant, i.e., there is some $c > 0$ such that $f = cg$. Now since $\pi_{like}(m|x) =$

$\pi_{\mathcal{E}}(m - Ax)$, considering the posterior density in (6) as a function of x we have that

$$\pi_{post}(x \mid m) \simeq \exp\left(-\frac{\alpha}{2}\|x\|_2^2 - \frac{1}{2}\|Ax - m\|_2^2\right).$$

In consequence, the most probable solution with respect to the posterior (maximizing $\pi_{post}(\cdot|m)$) is actually the minimizer of problem (2). Although this is a very rudimentary example, it gives intuition how well-designed prior can affect the problem so that the posterior gives high probability to stable solution candidates. Similarly, a well-designed prior can overcome existence or uniqueness issues if present.

Later on, we aim to quantify and understand abstract effects like stability in a broader sense - also for problems where the unknown is function valued, i.e., the realizations of random variable X belong to some infinite-dimensional space. The computational part of this course concerns the following question: how to extract information of the possibly high-dimensional probability distribution π_{post} once it is solved. Also, the practical effects related to different prior and noise models are considered there. The computational part of this course has independent lecture notes/material.

2 A brief dive into probability theory

2.1 Preliminaries

As discussed above, our task is to understand probability of X being something given measurement data M. From basic probability theory we know that

$$\mathbb{P}(X \in E \mid M \in F) = \frac{\mathbb{P}(X \in E, M \in F)}{\mathbb{P}(M \in F)},$$

where E and F are some measurable sets. However, we would like to condition the probability of $X \in E$ with respect to a single realization of M. If M has a nice probability density, it is easy to realize that probability of single value vanishes, i.e. $\mathbb{P}(M = m) = 0$. Hence we need to do a little work-out in the modern probability theory.

A triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called probability space, if

(1) $\Omega \neq \emptyset$ is a set,

(2) $\mathcal{F}$ is a σ-algebra, i.e.

   (a) $\Omega \in \mathcal{F}$,

   (b) If $E \in \mathcal{F}$, then $\Omega \setminus E \in \mathcal{F}$ and

   (c) If $E_j \in \mathcal{F}$, $j \in \mathbb{N}$, then $\bigcup_{j=1}^{\infty} E_j \in \mathcal{F}$.

(3) $\mathbb{P}$ is a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies

   (a) $\mathbb{P}(\Omega) = 1$ and

   (b) If measurable sets $E_j \in \mathcal{F}$, $j \in \mathbb{N}$, are disjoint (i.e. $E_j \cap E_k = \emptyset$ if $j \neq k$), then

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(E_j).$$

The property 3b) of measure $\mathbb{P}$ is called σ-additivity. A (general) measure is called σ-finite if Ω is the countable union of measurable sets with finite measure. Consider Lebesgue measure on $\mathbb{R}^n$ as an example.

For a while, we consider random variable in Euclidian spaces equipped with the standard Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$. Recall that a Borel σ-algebra is the smallest σ-algebra containing the open sets.

A random variable X is a measurable mapping

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)),$$

i.e., $X^{-1}(E) \in \mathcal{F}$ whenever $E \in \mathcal{B}(\mathbb{R}^n)$. Now X induces a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ by

$$\mu(E) := \mathbb{P}(X^{-1}(E)) = \text{probability that } X \in E.$$

The measure μ is called the probability distribution of X. We will often use notation $X \sim \mu$ to underline this.

Suppose μ and ν are two measures on the same measure space. Then μ is *absolutely continuous* with respect to ν, if $\nu(E) = 0$ implies $\mu(E) = 0$. In such a case, we write $\mu \ll \nu$. Measures μ and ν are *equivalent* if $\mu \ll \nu$ and $\nu \ll \mu$. If μ and ν are supported on disjoint sets, they are called *mutually singular*.

Theorem 1. *Let μ and ν be two measures on the same measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and ν is σ-finite then there exists $f \in L^1(\Omega, \mathcal{F}, \nu)$ such that*

$$\mu(E) = \int_E f(x) d\nu(x)$$

for all $E \in \mathcal{F}$.

Theorem 1 is called Radon–Nikodym theorem and the function f is known as the Radon–Nikodym derivative of μ with respect to ν. In the following, we write

$$\frac{d\mu}{d\nu}(x) = f(x) \in L^1(\nu)$$

The proof of Theorem 1 is omitted (will add reference later).

Example 2.1. *Suppose μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}^n))$ and $\mu \ll \mathcal{L}_n$, where $\mathcal{L}_n$ is a Lebesgue measure. By Theorem 1 there exists $\pi \in L^1(\mathbb{R}^n)$ such that*

$$\mu(E) = \int_E \pi(x) dx$$

for any $E \in \mathcal{B}(\mathbb{R})$. The function π is called probability density *is X.*

Let us also define the joint distribution of random variables X and Y by

$$\mu_{X,Y}(E \times F) = \mathbb{P}(X^{-1}(E) \cap Y^{-1}(F))$$

for any measurable sets E and F (the range of X and Y can differ and thus E and F can be subsets of different spaces). Suppose $Y : \Omega \to \mathbb{R}^n$. The marginal distribution of X is (similarly for Y) is obtained by

$$\mu_X(E) = \mu_{X,Y}(E \times \mathbb{R}^n).$$

Notice that the marginal distribution of M in (6) appears frequently throughout these notes.

The random variables X and Y called *independent* if

$$\mu_{X,Y}(E \times F) = \mu_X(E)\mu_Y(F)$$

for any measurable sets E and F. It is one of the fundamental assumptions of Bayesian inference that X and M in (5) are independent.

2.2 Conditional expectation and probability

In probability theory, σ-algebras represent information. One way to think about it is that 'knowing a σ-algebra $\mathcal{G}$' means knowing for each event $E \in \mathcal{G}$ whether E happened or not. Hence, $\mathcal{F}$ represents all the information about the experiment in $(\Omega, \mathcal{F}, \mathbb{P})$ while sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ represents partial information.

A common way for σ-algebras to arise is to have them generated by random variables. For examples, if $X : \Omega \to \mathbb{R}$ then $\sigma(X)$ denotes the smallest σ-algebra containing preimages of measurerable sets, i.e., sets $X^{-1}(E)$ where $E \in \mathcal{B}(\mathbb{R})$. Knowing the actual value of X corresponds to knowing whether $X \in E$ happened for each $E \in \mathcal{B}(\mathbb{R})$. However, many sample points might produce the same realization $X(\omega)$. In this sense $\sigma(X)$ provides only partial information.

Suppose that $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-algebra. Notice carefully that measurability with respect to $\mathcal{G}$ is a stronger requirement than measurability with respect to $\mathcal{F}$ since there are fewer choices for the preimages of X.

Definition 2.2. *Any random variable $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}^n)$ is called the conditional expectation of $X \in L^1(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^n)$ with respect to $\mathcal{G}$ if*

$$\int_G X(\omega) d\mathbb{P}(\omega) = \int_G Y(\omega) d\mathbb{P}(\omega) \tag{7}$$

for all $G \in \mathcal{G}$. We write $\mathbb{E}(X|\mathcal{G}) := Y$.

Proof. To be included. □

Example 2.3. *Let $E \subset \Omega$ such that $0 < \mathbb{P}(E) < 1$ and $\mathcal{G} = \{\emptyset, E, \Omega \setminus E, \Omega\}$. Then it holds that*

$$\mathbb{E}(X|\mathcal{G})(\omega) = \frac{\mathbb{E}(X\mathbf{1}_E)}{\mathbb{P}(E)}\mathbf{1}_E(\omega) + \frac{\mathbb{E}(X\mathbf{1}_{\Omega\setminus E})}{\mathbb{P}(\Omega \setminus E)}\mathbf{1}_{\Omega\setminus E}(\omega).$$

To convince us that this is indeed the case, we have to check whether the condition (7) holds for each set in $\mathcal{G}$. For example, we have

$$\int_E \mathbb{E}(X|\mathcal{G})(\omega) d\mathbb{P}(\omega) = \frac{\mathbb{E}(X\mathbf{1}_E)}{\mathbb{P}(E)} \int_E \mathbf{1}_E(\omega) d\mathbb{P}(\omega) = \mathbb{E}(X\mathbf{1}_E) = \int_E X(\omega) d\mathbb{P}(\omega).$$

Similarly, one can check the case for $\Omega \setminus E$.

It also possible to consider conditional expectations of type $\mathbb{E}(\phi(X)|\mathcal{F})$. This leads us to conditional probability. Namely, conditional probability of an event $\{\omega \mid X(\omega) \in E\}$ with respect to $\mathcal{G}$ is defined by

$$Q(E, \omega) = \mathbb{E}(\mathbf{1}_E(X)|\mathcal{G}).$$

Let us now study the mapping $Q : \mathcal{G} \times \Omega \to [0,1]$. In our search for conditioning with respect to a single realization (see beginning of Section 2.1) it would be crucial to know that $Q(\cdot, \omega)$ defines a probability measure on $\mathcal{G}$ for all (or at least almost all) $\omega \in \Omega$. Recall that by definition

$$\int_G Q(E,\omega) d\mathbb{P}(\omega) = \int_G \mathbf{1}_E(X) d\mathbb{P}(\omega) = \mathbb{P}(G \cap \{X \in E\})$$

for all $G \in \mathcal{G}$. We find out that $Q(E, \cdot)$ is defined up to $\mathbb{P}$-almost everywhere. However, since there may be uncountably many sets in $\mathcal{G}$, it is not trivial that we find a suitable version of Q.

Definition 2.4. *A family of probability distributions $(\mu(\cdot,\omega))_{\omega\in\Omega}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called a* regular conditional distribution *of X given $\mathcal{G} \subset \mathcal{F}$ if for each $E \in \mathcal{B}(\mathbb{R}^n)$ we have*

$$\mu(E,\cdot) = \mathbb{E}(\mathbf{1}_E(X) \mid \mathcal{G}) \quad \textit{almost surely.}$$

When $(\Omega, \mathcal{F})$ is identified with $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $X(\omega) = \omega$, $(\mu(\cdot,\omega))_{\omega\in\mathbb{R}^n}$ is called a regular conditional probability *on $\mathcal{F}$ with respect to $\mathcal{G}$.*

A classical result in probability theory is the following.

Theorem 2. *Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ be a random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Then there exists a regular conditional distribution $(\mu(\cdot,\omega))_{\omega\in\Omega}$ for X with respect to $\mathcal{G}$.*

We omit the proof (will add a reference!). In fact, the space $\mathbb{R}^n$ plays here no important role. Instead, Theorem 2 can be generalized to e.g. complete separable metric spaces.

The rigorous meaning of $\mu(E, x)$ for $x \in \mathbb{R}^n$ is important for us. The idea is now to use the regular conditional probability measure

$$\mu_{post}(E, M(\omega)) = \mathbb{E}(\mathbf{1}_E(X)|\sigma(M))(\omega), \tag{8}$$

where $\sigma(M) \subset \mathcal{F}$ is the σ-algebra generated by M and identify this object with $\mu_{post}(E, m)$.

3 Playing with the Bayes formula

3.1 What is the Bayes formula?

Bayesian statistics usually begins by the notion that the joint distribution of (X, M) is given and the posterior measure is a regular conditional distribution. Notice that, in general, measurement model like (6) may not be available but the necessary information (like likelihood distribution) is given via other means. In any case, important factor is that the marginal distribution of (X, M) with respect to X is assumed to be our prior distribution. Further, according to the Bayesian 'philosophy', the prior should be independent of the measurement setup.

Interesting phenomena would appear if we would allow the (rather natural) possibility of the unknown X being generated by some different distribution than the prior. After all, prior only models our beliefs and the partial information we have. However, in the context of inverse problems we could easily end up in a situation where the posterior is not well-defined (we couldn't talk about it at all) and hence during these notes we keep the purely Bayesian setup where X is generated by the prior.

Now, let us assume that in problem (6) our prior satisfies $X \sim \mu_X$. Due to Theorem 2 and identification of type (8) we are now able to talk about regular conditional probabilities $\mu_{like}(\cdot|x)$ and $\mu_{post}(\cdot|m)$ with respect to a single realization. We are ready to state the fundamental identity of Bayesian statistics, namely, the Bayes theorem.

Theorem 3 (Bayes)**.** *Suppose $X : \Omega \to \mathbb{R}^n$ and $M : \Omega \to \mathbb{R}^d$ satisfy equation (5). Assume $\mu_{like}(\cdot|x) \ll \nu$ for μ_X-almost every $x \in \mathbb{R}^n$, where ν is a σ-finite measure. Moreover, we write*

$$\Gamma_{like}(\cdot|x) := \frac{d\mu_{like}}{d\nu}(\cdot|x) \in L^1(\mathbb{R}^d, \nu).$$

Then we have $\mu_{post}(\cdot|m) \ll \mu_X$ for μ_M-almost every $m \in \mathbb{R}^d$ and

$$\frac{d\mu_{post}}{d\mu_X}(x|m) = \frac{1}{Z(m)}\Gamma_{like}(m|x),$$

where $Z(m) = \int_{\mathbb{R}^n} \Gamma_{like}(m|x)d\mu_X(x)$.

Proof. Our first concern is what is the probability of $Z(m) = 0$ or $Z(m) = \infty$. Let us denote these events by

$$E_0 = \{m|Z(m) = 0\} \quad \text{and} \quad E_\infty = \{m|Z(m) = \infty\}.$$

We know that the marginal distribution of M satisfies

$$\mu_M(E) = \int_E \int \int_{\mathbb{R}^n} \Gamma_{like}(m|x)d\mu_X(x)d\nu(m) = \int_E Z(m)d\nu(m).$$

It directly follows that $\mu_M(E_0) = 0$. Moreover, suppose $\nu(E_\infty) > 0$. Then we have

$$\mu_M(E_\infty) = \int_{E_\infty} \infty d\nu(x) = \infty,$$

which yields a contradiction since μ_M is a probability measure. Moreover, since $\mu_M \ll \nu$, it must hold that also $\mu_M(E_\infty) = 0$.

Next, the regularity of the posterior measure guarantees that we can write

$$\begin{aligned}\mathbb{P}(X \in E, M \in F) &= \int_F \mu_{post}(E|m)d\mu_M(m) \\ &= \int_F \mu_{post}(E|m)\left(\int_{\mathbb{R}^n} \Gamma_{like}(m|x)d\mu_X(x)\right)d\nu(m).\end{aligned}$$

for any measurable sets $E \in \mathbb{R}^n$ and $F \in \mathbb{R}^d$. Similarly, by writing the joint probability via the regular likelihood yields

$$\begin{aligned}\mathbb{P}(X \in E, M \in F) &= \int_E \int_F \Gamma_{like}(m|x)d\nu(m)d\mu_X(x) \\ &= \int_F \int_E \Gamma_{like}(m|x)d\mu_X(x)d\nu(m),\end{aligned}$$

where we have applied the Fubini theorem. Since E and F are arbitrary, we obtain

$$\mu_{post}(E|m) = \frac{\int_E \Gamma_{like}(m|x)d\mu_X(x)}{\int_{\mathbb{R}^n} \Gamma_{like}(m|x)d\mu_X(x)}$$

and we are done. □

Now suppose we take $\nu = \mathcal{L}_d$ and $\mu_{like}(\cdot|x) \ll \mathcal{L}_d$, where $\mathcal{L}_d$ is the Lebesgue measure on $\mathbb{R}^d$ and denote

$$\pi_{like}(m|x) := \frac{d\mu_{like}}{d\mathcal{L}_d}(m|x).$$

Moreover, assume $\mu_X \ll \mathcal{L}_n$ and

$$\pi_X(x) := \frac{d\mu_X}{d\mathcal{L}_n}(x).$$

Then we have

$$\begin{aligned} \mu_{post}(E|m) &= \frac{1}{Z(m)} \int_E \pi_{like}(m|x) d\mu_X(x) \\ &= \frac{1}{Z(m)} \int_E \pi_{like}(m|x)\pi_X(x) d\mathcal{L}_n(x). \end{aligned}$$

Now we see that $\mu_{post}(\cdot|m) \ll \mathcal{L}_n$ and

$$\begin{aligned} \pi_{post}(x|m) &= \frac{d\mu_{post}}{d\mathcal{L}_n}(x|m) \\ &= \frac{\pi_{like}(x|m)\pi_X(x)}{Z(m)} \end{aligned}$$

Since we have

$$\begin{aligned} \int_F \pi_M(m) d\mathcal{L}_d(m) &= \mathbb{P}(M \in F) \\ &= \mathbb{P}(X \in \mathbb{R}^n, M \in F) \\ &= \int_{\mathbb{R}^n} \mu_{like}(F|x) d\mu_X(x) \\ &= \int_F \int_{\mathbb{R}^n} \pi_{like}(m|x) d\mu_X(x) d\mathcal{L}_d(m) \\ &= \int_F Z(m) d(m), \end{aligned}$$

it follows that $Z(m) = \pi_M(m)$ for μ_M-almost every $m \in \mathbb{R}^d$.

Corollary 3.1. *Suppose all probability distributions related to problem* (5) *have well-defined probability densities. Then the density function representation of the Bayes formula*

$$\pi_{post}(x \mid m) = \frac{\pi_{like}(m \mid x)\pi_X(x)}{\pi_M(m)}, \tag{9}$$

holds, where π_{post}, π_{like} and π_X represent the posterior, likelihood and prior density, respectively. Moreover, π_M is the marginal distribution of the measurement M.

3.2 Example: Gaussian posterior

Let us next move to studying how the posterior density looks like in the canonical example when the prior and likelihood have Gaussian statistics. Before proceeding, we record what is a Gaussian random variable on $\mathbb{R}^n$.

Definition 3.2. *Let $x_0 \in \mathbb{R}^n$ and $C \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. A Gaussian n-variate random variable X with mean x_0 and covariance C is a random variable with the probability density*

$$\pi_X(x) = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\left(-\frac{1}{2}(x-x_0)^\top C^{-1}(x-x_0)\right).$$

We denote the Gaussian distribution by $X \sim \mathcal{N}(x_0, C_0)$.

Let us recall that covariance matrix of (any) random variable X is defined by

$$C = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top.$$

A Gaussian distribution is completely characterized by its mean and covariance.

Notice that the expression $(x-x_0)^\top C^{-1}(x-x_0)$ can also be written in form $\left\|C^{-1/2}x\right\|_2^2$ since due to our assumptions on C the inverse square root $C^{-1/2}$ is well-defined. Sometimes, when the posteriori distribution is of the form $const \cdot \exp(-F(x))$, one can try to rewrite F as a sum of a quadratic form and constant term in order to show that the posterior is Gaussian (and to solve what is mean and covariance). This method is called *completing the square* and it is what we essentially do in the following.

Since research on inverse problems most often is based on some model equation (5), we have a connection between the likelihood and noise distributions.

Remark 3.3 (Likelihood)**.** *Suppose $\mathcal{E} \sim \mu_\mathcal{E} \ll \mathcal{L}_d$ and $\pi_{noise}(e) = \frac{d\mu_\mathcal{E}}{d\mathcal{L}_d}(e)$. The regular conditional probability satisfies*

$$\mathbb{P}(M \in E | X = x) = \mathbb{P}(Ax + \mathcal{E} \in E) = \mathbb{P}(\mathcal{E} \in \{e - Ax \mid e \in E\}).$$

Therefore, it must hold that

$$\pi_{like}(m|x) = \pi_{noise}(m - Ax).$$

In order to analyse the Gaussian posterior further, we need some machinery from linear algebra.

Definition 3.4. *Let*

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

be a positive definite symmetric matrix. We define the Schur complements $\widetilde{C}_{jj}$ *of* C_{jj}, $j = 1, 2$, *by*

$$\begin{aligned} \widetilde{C}_{22} &:= C_{11} - C_{12}C_{22}^{-1}C_{21} \quad \text{and} \\ \widetilde{C}_{11} &:= C_{22} - C_{21}C_{11}^{-1}C_{12}. \end{aligned}$$

Lemma 3.5. *The Schur complements $\widetilde{C}_{jj}$ are invertible and*

$$C^{-1} = \begin{pmatrix} \widetilde{C}_{22}^{-1} & -\widetilde{C}_{22}^{-1}C_{12}C_{22}^{-1} \\ -\widetilde{C}_{11}^{-1}C_{21}C_{11}^{-1} & \widetilde{C}_{11}^{-1} \end{pmatrix}$$

Proof. Left for exercise. □

For the following, let $X \sim \mathcal{N}(x_0, C_0)$ and $\mathcal{E} \sim \mathcal{N}(0, \Gamma)$. Recall that X and $\mathcal{E}$ are assumed to be independent. Next, consider the distribution of the measurement M. The equality (5) implies that we have $m_0 := \mathbb{E}M = Ax_0$ and

$$\mathbb{E}(M - m_0)(M - m_0)^\top = \mathbb{E}(A(X - x_0) + \mathcal{E})(A(X - x_0) + \mathcal{E})^\top = AC_0A^\top + \Gamma.$$

Moreover, we have

$$\mathbb{E}(X - x_0)(M - m_0)^\top = \mathbb{E}(X - x_0)(A(X - x_0) + \mathcal{E})^\top = C_0A^\top$$

The joint distribution of X and M then has a covariance

$$\mathrm{Cov}\begin{pmatrix} X \\ M \end{pmatrix} = \mathbb{E}\left(\begin{pmatrix} X - x_0 \\ M - m_0 \end{pmatrix}\begin{pmatrix} X - x_0 \\ M - m_0 \end{pmatrix}^\top\right) = \begin{pmatrix} C_0 & C_0A^\top \\ AC_0 & AC_0A^\top + \Gamma \end{pmatrix}.$$

Therefore, it follows that

$$\pi(x, m) \simeq \exp\left\{-\frac{1}{2}\begin{pmatrix} X - x_0 \\ M - m_0 \end{pmatrix}\begin{pmatrix} C_0 & C_0A^\top \\ AC_0 & AC_0A^\top + \Gamma \end{pmatrix}\begin{pmatrix} X - x_0 \\ M - m_0 \end{pmatrix}^\top\right\}$$

In order to ease our notations, let C_{ij}, $i, j = 1, 2$, denote the components of $\mathrm{Cov}\begin{pmatrix} X \\ M \end{pmatrix}$. In addition, we make a simplification of assuming $x_0 = 0$ (and thus also $m_0 = 0$). This can be considered as a translation of the coordinates, and the equations below can be adjusted for general case simply by replacing x with $x - x_0$ and m with $m - m_0$.

The crucial idea now is that we are interested in the posterior distribution only through behaviour of x whereas m plays the role of constant for us. Therefore, we have by the Bayes formula that

$$\pi_{post}(x|m) \simeq \pi(x, m).$$

In consequence, the joint density of the form

$$\begin{aligned} \pi(x, m) &\simeq \exp\left(-\frac{1}{2}(x^\top \widetilde{C}_{22}^{-1}x - 2x^\top \widetilde{C}_{22}^{-1}C_{12}C_{22}^{-1}m + m^\top \widetilde{C}_{11}^{-1}m)\right) \\ &\simeq \exp\left(-\frac{1}{2}(x - C_{12}C_{22}^{-1}m)^\top \widetilde{C}_{22}^{-1}(x - C_{12}C_{22}^{-1}m) + \mathrm{const}\right) \end{aligned}$$

Now we obtain the mean and covariance of the posterior $\mu_{post}(\cdot|m) \sim \mathcal{N}(\bar{x}, C_{post})$ as

$$C_{post} = \widetilde{C}_{22} = C_0 - C_0A^\top(AC_0A^\top + \Gamma)^{-1}AC_0 \tag{10}$$

and

$$x_{post} = C_{12}C_{22}^{-1}m = C_0A^\top(AC_0A^\top + \Gamma)^{-1}m.$$

Notice carefully that the covariance C_{post} is independent of the mean x_0 (also mean of the noise if that would be non-zero). Luckily, we have a more compact expression for these objects.

Theorem 4. *Let $X \sim \mathcal{N}(x_0, C_0)$ and $\mathcal{E} \sim \mathcal{N}(0, \Gamma)$, and assume that equation (5) holds. Then we have*

$$\pi_{post}(x|m) \simeq \exp\left(-\frac{1}{2}(x-\bar{x})^\top C_{post}^{-1}(x-\bar{x})\right),$$

where

$$C_{post} = (A^\top \Gamma^{-1} A + C_0^{-1})^{-1} \tag{11}$$

and

$$x_{post} = C_{post}(A^\top \Gamma^{-1} m + C_0^{-1} x_0). \tag{12}$$

Proof. To show that the two expressions coincide we simply multiply (10) with the inverse of (10) and hope to obtain an identity matrix:

$$\begin{aligned}
(C_0 - C_0 A^\top (AC_0A^\top + \Gamma)^{-1} AC_0)(A^\top \Gamma^{-1} A + C_0^{-1}) \\
&= (I - C_0 A^\top (AC_0A^\top + \Gamma)^{-1} A)(I + C_0 A^\top \Gamma^{-1} A) \\
&= I + C_0 A^\top \Gamma^{-1} A - C_0 A^\top (AC_0A^\top + \Gamma)^{-1} A(I + C_0 A^\top \Gamma^{-1} A) \\
&= I + C_0 A^\top \Gamma^{-1} A - C_0 A^\top \Gamma^{-1} A = I,
\end{aligned}$$

where on the third line we used the identity $A(I + C_0 A^\top \Gamma^{-1} A) = (\Gamma + AC_0A^\top)\Gamma^{-1} A$. Similar deduction can be made if the order of matrices is reversed (i.e. $AB = BA = I$).

For the mean value notice that we can write

$$\Gamma^{-1} - (\Gamma + AC_0A^\top)^{-1} AC_0A^\top \Gamma^{-1} = (\Gamma + AC_0A^\top)^{-1}. \tag{13}$$

From equation (12) we obtain

$$\begin{aligned}
x_{post} &= C_{post}((C_{post}^{-1} - A^\top \Gamma^{-1} A)x_0 + A^\top \Gamma^{-1} m) \\
&= x_0 + C_{post} A^\top \Gamma^{-1}(m - Ax_0).
\end{aligned}$$

A combination of (10) and (13) yields

$$\begin{aligned}
C_{post} A^\top \Gamma^{-1} &= C_0 A^\top \Gamma^{-1} - C_0 A^\top (\Gamma + AC_0A^\top)^{-1} AC_0A^\top \Gamma^{-1} \\
&= C_0 A^\top (\Gamma + AC_0A^\top)^{-1}
\end{aligned}$$

and we are done. □

4 Posterior contraction with Gaussian distributions

Recall from introduction that the reconstruction error in inverse problems is a balance between two factors: one generated by the modelling error of replacing (1) with a stable one and second generated by the noise. In consequence, an optimal solution strategy (choosing α) is based on an accurate estimate of the noise level. However, as we are talking about ill-posed problems, it is important to analyse how the solution strategy works as the noise level is reduced. Does the regularized solution converge to the 'true' solution?

In Bayesian inversion the analysis of vanishing measurement noise is often called *posterior contraction* or *consistency* related to individual estimators. In the following we consider this problem for over- and underdetermined systems. The take-away message is that in overdetermined systems, the prior plays no role in the limit. However, practical inverse problems are usually underdetermined. We will notice that in this case a well-designed prior distribution is essential in the limit.

Example 4.1 (Overdetermined system)**.** *Let us consider a measurement*

$$M = gX + \mathcal{E}$$

where $g \in \mathbb{R}^d \setminus \{0\}$ for $d \geq 2$ and the unknown X is one-dimensional. Suppose that $X \sim \mathcal{N}(0,1)$ and $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I)$. Then by Theorem 4 we have

$$\pi_{post}(x|m) \simeq \exp\left(-\frac{1}{2\sigma^2}|m - gX|_2^2 - \frac{1}{2}x^2\right).$$

and moreover

$$x_{post} = \frac{g^\top m}{\sigma^2 + |g|^2} \quad \text{and} \quad \sigma^2_{post} = \frac{\sigma^2}{\sigma^2 + |g|^2}.$$

In consequence, in the limit one obtains

$$x^+_{post} = \lim_{\sigma \to 0} x_{post}(\sigma) = \frac{g^\top m}{|g|^2} = \arg\min_{x \in \mathbb{R}} |m - gx|^2$$

and $(\sigma^2_{post})^+ = \lim_{\sigma \to 0} \sigma^2_{post} = 0$. We notice that for overdetermined problems the prior plays no role in the limit of zero measurement noise.

Example 4.2 (Underdetermined system)**.** *Suppose $d = 1$, $n \geq 2$ and*

$$M = g^\top x + \mathcal{E},$$

where $g \in \mathbb{R}^n \setminus \{0\}$ and both the noise $\mathcal{E}$ and the measurement M are one-dimensional. Let us assume $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$ and $X \sim \mathcal{N}(0, C_0)$. By Theorem 4 we get

$$\pi_{post}(x|m) \simeq \exp\left(-\frac{1}{2\sigma^2}|m - g^\top x|^2 - \frac{1}{2}x^\top C_0^{-1} x\right)$$

and

$$x_{post} = \frac{m}{\sigma^2 + g^\top C_0 g} \cdot C_0 g$$

together with

$$C_{post} = C_0 - \frac{(C_0 g)(C_0 g)^\top}{\sigma^2 + g^\top C_0 g}.$$

Again, going to the limit of zero measurement noise yields

$$x^+_{post} = \lim_{\sigma \to 0} x_{post}(\sigma) = \frac{m}{g^\top C_0 g} C_0 g$$

and

$$C^+_{post} = \lim_{\sigma \to 0} C_{post}(\sigma) = C_0 - \frac{(C_0 g)(C_0 g)^\top}{g^\top C_0 g}.$$

How to interpret this? On one hand, since $C^+_{post} g = 0$ and $(x^+_{post})^\top g = m$, we see that the posterior predicts the 'true' solution in the subspace G spanned by the vector g. On the other hand, we have no information from the complement subspace, and consequently the posterior is fully described by the prior in $G^\perp$.

Lemma 4.3. *Let $\mu_n \sim \mathcal{N}(x_n, C_n)$ and $\mu \sim \mathcal{N}(x, C)$ on $\mathbb{R}^n$. Suppose $x_n \to x$ and $C_n \to C$ in the usual two-norm as $n \to \infty$. It follows that the measures converge weakly, i.e., $\mu_n \rightharpoonup \mu$.*

Proof. This follows directly from Lemma B.3. Also, using the Definition B.1 we could prove the claim by using the Lebesgue dominated convergence. □

Theorem 5. *Let $X \sim \mathcal{N}(x_0, C_0)$ and $\mathcal{E} \sim \mathcal{N}(0, \Gamma(\sigma))$. If $\mathrm{Null}(A) = \{0\}$ and $\Gamma(\sigma) = \sigma^2\Gamma_0$, $\gamma > 0$, it follows that*

$$\mu_{post}(\cdot|m) \rightharpoonup \delta_{x^+_{post}}$$

when $\sigma \to 0$ and where

$$x^+_{post} = \arg\min_x \left\| \Gamma_0^{-1/2}(Ax - m) \right\|^2.$$

Proof. From equations (12) and (11) we see that

$$x_{post} = (A^\top \Gamma_0^{-1} A + \sigma^2 C_0^{-1})^{-1}(A\Gamma_0^{-1} m + \sigma^2 C_0^{-1} x_0) \quad \text{and} \tag{14}$$

$$C_{post} = \sigma^2 (A^\top \Gamma_0^{-1} A + \sigma^2 C_0^{-1})^{-1}. \tag{15}$$

Since A has a trivial null space, there exists $\alpha > 0$ such that

$$\langle \xi, A^\top \Gamma_0^{-1} A \xi \rangle = |\Gamma_0^{-1/2} A \xi|^2 \geq \alpha |\xi|^2$$

for all $\xi \in \mathbb{R}^n$. Therefore, the matrix $A^\top \Gamma_0^{-1} A$ is invertible. Now we can take σ to zero in (14) and get

$$x^+_{post} = \lim_{\sigma \to 0} x_{post}(\sigma) = (A^\top \Gamma_0^{-1} A)^{-1} A \Gamma_0^{-1} m$$

and $C^+_{post} \to 0$. By Lemma 4.3 we have the weak convergence.

Due to the trivial null space of A, the minimizer of

$$\frac{1}{2} \left\| \Gamma_0^{-1/2}(Ax - m) \right\|^2$$

is unique and satisfies

$$A^\top \Gamma_0^{-1} A x^+_{post} = A^\top \Gamma_0^{-1} m.$$

This yields the claim. □

In the underdetermined setting $A \in \mathbb{R}^{d \times n}$, where $d < n$. Below, we list some notations that help us to formulate and prove the contraction result efficiently. Assume $\mathrm{rank}(A) = d$ so that we can write

$$A = (A_0 \; \mathbf{0}) Q^\top,$$

where $Q \in \mathbb{R}^{n \times n}$ is orthogonal ($Q^\top Q = QQ^\top = I$), $A_0 \in \mathbb{R}^{d \times d}$ is invertible and $\mathbf{0} \in \mathbb{R}^{d \times (n-d)}$ a zero matrix. Also, denote by $L_0 = \Gamma_0^{-1}$ the *precision matrix* and

$$Q^\top L_0 Q = \begin{pmatrix} L_{11} & L_{12} \\ L_{12}^\top & L_{22} \end{pmatrix},$$

where $L_{11} \in \mathbb{R}^{d \times d}$, $L_{12} \in \mathbb{R}^{d \times (n-d)}$ and $L_{22} \in \mathbb{R}^{(n-d) \times (n-d)}$. Both L_{11} and L_{22} inherit the symmetricity and positive definiteness of L_0. Also, we write $Q = (Q_1 Q_2)$ with $Q_1 \in \mathbb{R}^{n \times d}$ and $Q_2 \in \mathbb{R}^{n \times (n-d)}$.

Next we define vectors $z \in \mathbb{R}^d$ and $z' \in \mathbb{R}^{n-d}$, which we use to define the limiting posterior mean. First, we set z to be the unique solution of

$$A_0 z = m. \tag{16}$$

We notice that if $Ax = m$ for some $x \in \mathbb{R}^n$, then $z = Q_1^\top x$. On the other hand, $Q_2^\top x$ is not determined by the identity of x. Hence, z represents the information provided by the measurement. Similarly, $Q_2^\top x$ is not determined by the measurement.

Let $w \in \mathbb{R}^d$ and $w' \in \mathbb{R}^{n-d}$ be defined via

$$L_0 x_0 = Q \begin{pmatrix} w \\ w' \end{pmatrix}$$

and set

$$z' = -L_{22}^{-1} L_{12}^\top z + L_{22}^{-1} w' \in \mathbb{R}^{n-d}. \tag{17}$$

Theorem 6. *Let $X \sim \mathcal{N}(x_0, C_0)$ and $\mathcal{E} \sim \mathcal{N}(0, \Gamma(\sigma))$. If $\Gamma(\sigma) = \sigma^2 \Gamma_0$, $\gamma > 0$, and (z, z') is defined by* (16) *and* (17)*, it follows that*

$$\mu_{post}(\cdot|m) \rightharpoonup \mathcal{N}(x_{post}^+, C_{post}^+),$$

where $x_{post}^+ = Q \begin{pmatrix} z \\ z' \end{pmatrix}$ and $C_{post}^+ = Q_2 L_{22}^{-1} Q_2^\top$.

Proof. We have

$$A^\top \Gamma_0^{-1} A = Q \begin{pmatrix} A_0^\top \Gamma_0^{-1} A_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} Q^\top$$

and consequently

$$C_{post}^{-1} = Q \begin{pmatrix} \frac{1}{\sigma^2} A^\top \Gamma_0^{-1} A + L_{11} & L_{12} \\ L_{12}^\top & L_{22} \end{pmatrix} Q^\top.$$

By utilizing Schur complements we can prove

$$C_{post} = Q \begin{pmatrix} \sigma^2 (A_0^\top \Gamma_0^{-1} A_0)^{-1} & \mathbf{0} \\ -\sigma^2 L_{22}^{-1} L_{12}^\top (A_0^\top \Gamma_0^{-1} A_0)^{-1} & L_{22}^{-1} \end{pmatrix} Q^\top + \Delta, \tag{18}$$

where $\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$ satisfies

$$\frac{1}{\sigma^2}(|\Delta_{11}| + |\Delta_{21}|) \to 0$$

as $\sigma \to 0$ and $|\Delta_{12}| + |\Delta_{22}| \leq C\sigma^2$ for some fixed constant $C > 0$ (see Exercise 3.3). Now we obtain

$$C_{post}^+ = \lim_{\sigma \to 0} C_{post}(\sigma) = Q \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & L_{22}^{-1} \end{pmatrix} Q^\top$$

as required.

Let us then consider the posterior mean. We see that

$$x_{post} = C_{post} \left(\frac{1}{\sigma^2} Q \begin{pmatrix} A_0^\top \Gamma_0^{-1} \\ \mathbf{0} \end{pmatrix} m + C_0^{-1} x_0 \right)$$

and by the definition of w and w' we deduce that

$$x_{post} = C_{post} Q \begin{pmatrix} \frac{1}{\sigma^2} A_0^\top \Gamma_0^{-1} m + w \\ w' \end{pmatrix}.$$

By equation (18) we have

$$\lim_{\sigma \to 0} x_{post}(\sigma) = Q \begin{pmatrix} z \\ -L_{22}^{-1} L_{12}^\top z + L_{22}^{-1} w' \end{pmatrix} = Q \begin{pmatrix} z \\ z' \end{pmatrix},$$

which proves the claim. □

5 Well-posedness of Bayesian inversion

By well-posedness we refer to the continuity of the method of obtaining the posterior probability distribution with respect to different perturbations in the parameters. In practise, 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 (see Section B.2)? 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 reconstructions. Does the Bayesian approach then regularize the problem? The answer is yes under certain assumptions on the modelling. In the following we also consider how the modelling error in prior is propagated to the posterior.

In the following we frequently write

$$f \lesssim g \tag{19}$$

for two functions f and g, if there is a constant $c > 0$ such that

$$f \leq cg$$

almost everywhere.

5.1 Distance of posteriors in $\mathbb{R}^n$

Let us first consider the distance between Gaussian distributions to set the scene.

Example 5.1. *Let $\mu_1 \sim \mathcal{N}(x_1, \sigma_1^2)$ and $\mu_2 \sim \mathcal{N}(x_2, \sigma_2^2)$ be two Gaussian probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. As the reference measure ν in the Hellinger distance we use the Lebesgue measure. We have*

$$d_{Hell}(\mu_1, \mu_2)^2 = 1 - \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \int_{\mathbb{R}} \exp(-Q(x)) dx,$$

where

$$Q(x) = \frac{1}{4\sigma_1^2}(x - x_1)^2 + \frac{1}{4\sigma_2^2}(x - x_2)^2.$$

Define γ by

$$\frac{1}{2\gamma^2} = \frac{1}{4\sigma_1^2} + \frac{1}{4\sigma_2^2}.$$

A change of variable $y = x - \frac{x_1+x_2}{2}$ *yields for* $r = \frac{x_1-x_2}{2}$ *that*

$$\begin{aligned} Q(y) &= \frac{1}{4\sigma_1^2}(y-r)^2 + \frac{1}{4\sigma_2^2}(y+r)^2 \\ &= \frac{y^2}{2\gamma^2} - \left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) yr + \frac{r^2}{2\gamma^2} \\ &= \frac{1}{2\gamma^2}(y - \frac{z}{2})^2 - \frac{z^2}{8\gamma^2} + \frac{r^2}{2\gamma^2}, \end{aligned}$$

where $z = \gamma^2 r \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)$. *Since*

$$-\frac{z^2}{8\gamma^2} + \frac{r^2}{2\gamma^2} = \frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)}$$

we find that

$$\begin{aligned} d_{Hell}(\mu_1,\mu_2)^2 &= 1 - \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2\gamma^2}(y-\frac{z}{2})^2\right) dy \cdot \exp\left(-\frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right) \\ &= 1 - \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} \sqrt{2\pi\gamma^2} \exp\left(-\frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right) \\ &= 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\left(-\frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right). \end{aligned}$$

This can be further approximated by

$$\begin{aligned} d_{Hell}(\mu_1,\mu_2)^2 &\leq 1 - \sqrt{1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}} \left(1 - \frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right) \\ &\leq 1 - \sqrt{1 - \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2}} + \frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)} \\ &\leq \frac{(\sigma_1-\sigma_2)^2}{\sigma_1^2+\sigma_2^2} + \frac{(x_1-x_2)^2}{4(\sigma_1^2+\sigma_2^2)} \\ &\lesssim (\sigma_1-\sigma_2)^2 + (x_1-x_2)^2, \end{aligned}$$

where the implicit constant of course depends on σ_1 *and* σ_2. *This elegantly illustrates how the study of the distance can be split into separation of standard deviations (or variances since the difference is equivalent) and mean values. For notation* $\lesssim$ *see* (19).

Theorem 7. *Let* $\mu_1 \sim \mathcal{N}(x_1, C_1)$ *and* $\mu_2 \sim \mathcal{N}(x_2, C_2)$ *be two probability measures on* $\mathbb{R}^n$. *Then we have*

$$d_{Hell}(\mu_1,\mu_2)^2 = 1 - \frac{(\det C_1)^{1/4} (\det C_2)^{1/4}}{\left(\det\left(\frac{C_1+C_2}{2}\right)\right)^{1/2}} \exp\left(-\frac{1}{8}\Delta x^\top \left(\frac{C_1+C_2}{2}\right)^{-1} \Delta x\right), \tag{20}$$

where $\Delta x = x_1 - x_2$.

Proof. Left for exercise. □

Let us next consider what implications Theorem 7 has for the posterior under perturbations of the measurement or prior.

Theorem 8. *Suppose $m_1, m_2 \in \mathbb{R}^d$ are the measurements obtained for problem (5). It follows that*

$$d_{Hell}(\mu_{post}(\cdot|m_1), \mu_{post}(\cdot|m_2)) \lesssim \|m_1 - m_2\|_2$$

Proof. According to Theorem 4 we have

$$\Delta x_{post} = x^1_{post} - x^2_{post} = C_{post} A^\top \Gamma^{-1}(m_1 - m_2),$$

where the posterior covariance

$$C_{post} = (A^\top \Gamma^{-1} A + C_0^{-1})^{-1}$$

is independent of measurements. We deduce that

$$\begin{aligned} \Delta x_{post}^\top C_{post}^{-1} \Delta x_{post} &= (m_1 - m_2)^\top \Gamma^{-1} A C_{post} A^\top \Gamma^{-1}(m_1 - m_2) \\ &= \left\| C_{post}^{1/2} A^\top \Gamma^{-1}(m_1 - m_2) \right\|_2^2 \\ &\leq c \|m_1 - m_2\|_2^2 \end{aligned}$$

for some constant $c > 0$ depending on the (bounded) norms of $C_{post}^{1/2}$, $A^\top$ and Γ^{-1}. Our result follows from

$$d_{Hell}(\mu_{post}(\cdot|m_1), \mu_{post}(\cdot|m_2)) \leq 1 - \exp(-c \|m_1 - m_2\|_2^2) \leq \tilde{c} \|m_1 - m_2\|_2^2, \tag{21}$$

where $\tilde{c} > 0$ is a constant. Above, we used the fact $\exp(-t) \geq 1 - t$ for $t > 0$. □

Next we compare Bayesian inference on the problem (5) based on one measurement $m \in \mathbb{R}^d$ but two different prior distributions μ_1 and μ_2. We denote the corresponding posterior distributions by μ^1_{post} and μ^2_{post}, respectively.

Lemma 5.2. *Let us assume the prior distributions are Gaussian with only difference in mean values, i.e. $\mu_1 \sim \mathcal{N}(x_1, C_0)$ and $\mu_2 \sim \mathcal{N}(x_2, C_0)$. We have then*

$$d_{Hell}(\mu^1_{post}, \mu^2_{post}) \lesssim \|x_1 - x_2\|_2 .$$

Proof. The proof is similar to Theorem 8. We observe that

$$\Delta x_{post} = C_{post} C_0^{-1}(x_1 - x_2)$$

and consequently

$$\Delta x_{post}^\top C_{post}^{-1} \Delta x_{post} \lesssim \|x_1 - x_2\|_2^2 .$$

An inequality similar to (21) finishes the proof. □

Lemma 5.3. *Let us assume the prior distributions are Gaussian with only difference in covariance, i.e. $\mu_1 \sim \mathcal{N}(x_0, C_1)$ and $\mu_2 \sim \mathcal{N}(x_0, C_2)$. We have then*

$$d_{Hell}(\mu^1_{post}, \mu^2_{post}) \lesssim \|C_1 - C_2\|_2 .$$

Proof. Denote the posterior covariances by $C_{post,1}$ and $C_{post,2}$, respectively. Recall that

$$C_{post,j} = \left(A^\top \Gamma^{-1} A + C_j^{-1}\right)^{-1}$$

and consider first the expressions appearing in the exponent in (20). The difference between posterior means is given by

$$\Delta x_{post} = (C_{post,1} - C_{post,2}) A^\top \Gamma^{-1} m + (C_{post,1} C_1^{-1} - C_{post,2} C_2^{-1}) x_0 \tag{22}$$

The following line of argument is rather technical (I should check if there's an easier way!). However, the two main points are that for invertible matrices T_1 and T_2 (of same size) we have

$$T_1^{-1} - T_2^{-1} = T_1^{-1}(T_2 - T_1)T_2^{-1}$$

and we are basically only interested in the difference $C_1 - C_2$. Norm of anything else is constant (bigger than zero) to us. Hence, let us denote

$$B_1 := C_{post,1} - C_{post,2} = C_{post,1}(C_{post,2}^{-1} - C_{post,1}^{-1})C_{post,2} = C_{post,1}C_2^{-1}(C_1 - C_2)C_1^{-1}C_{post,2}$$

and consequently $\|B_1\|_2 \lesssim \|C_1 - C_2\|$. Similarly, let

$$B_2 := C_{post,1}C_1^{-1} - C_{post,2}C_2^{-1} = C_{post,1}C_1^{-1}(C_2 - C_1)A^\top \Gamma^{-1} A C_{post,2} C_2^{-1}$$

which yields $\|B_2\|_2 \lesssim \|C_1 - C_2\|$. Consider writing out the term $2x_{post}^\top (C_{post,1} + C_{post,2})^{-1} x_{post}^\top$ with notations B_1 and B_2. Direct bound obtained by factoring out norms of each involved matrix yields

$$\begin{aligned}
&\left\| \Delta x_{post}^\top \left(\frac{C_{post,1} + C_{post,2}}{2}\right)^{-1} \Delta x_{post}^\top \right\| \\
&\quad \lesssim \left\|B_1^\top\right\|_2 \|B_1\|_2 + \left\|B_1^\top\right\|_2 \|B_2\|_2 + \left\|B_2^\top\right\|_2 \|B_1\|_2 + \left\|B_2^\top\right\|_2 \|B_2\|_2 \lesssim \|C_1 - C_2\|_2^2 .
\end{aligned} \tag{23}$$

Now let us turn our attention to the determinants in (20). First notice the equality

$$\frac{(\det C_{post,1})^{1/4} (\det C_{post,2})^{1/4}}{\left(\det\left(\frac{C_{post,1}+C_{post,2}}{2}\right)\right)^{1/2}} = \frac{\left(\det(C_{post,1}^{-1} C_{post,2})\right)^{1/4}}{\left(\det\left(\frac{\mathrm{I}+C_{post,1}^{-1}C_{post,2}}{2}\right)\right)^{1/2}}. \tag{24}$$

Moreover, for positive definite matrix $\widetilde{C} \in \mathbb{R}^{n\times n}$ we know that

$$\exp(\mathrm{Tr}(\mathrm{I} - \widetilde{C}^{-1})) \leq \det(\widetilde{C}) \leq \exp(\mathrm{Tr}(\widetilde{C} - \mathrm{I})).$$

The idea is now that we take $\widetilde{C} = C_{post,1}C_{post,2}^{-1}$, which is rather close to identity. In fact, the distance to identity is of order $\|C_1 - C_2\|_2$ as we will see next. Since $C_{post,1}^{-1}C_{post,2}$ is symmetric and positive definite (particularly invertible), we have a lower bound to the right hand side of (24) by

$$\begin{aligned}
&\frac{\left(\exp\left(\mathrm{Tr}(\mathrm{I} - C_{post,2}^{-1}C_{post,1})\right)\right)^{1/4}}{\left(\exp\left(\mathrm{Tr}(\frac{\mathrm{I}+C_{post,1}^{-1}C_{post,2}}{2} - \mathrm{I})\right)\right)^{1/2}} \\
&\qquad = \exp\left(-\frac{1}{4}\mathrm{Tr}(C_{post,2}^{-1}C_{post,1} + C_{post,1}^{-1}C_{post,2} - 2\mathrm{I})\right) \geq \exp\left(b\,\|C_1 - C_2\|_2^2\right)
\end{aligned} \tag{25}$$

for some constant $b > 0$, since $\widetilde{C} + \widetilde{C}^{-1} - 2I$ is positive definite (implying we can bound the trace by n-times the norm). The bound is obtained by computing

$$\begin{aligned} C_{post,2}^{-1} C_{post,1} - \mathrm{I} &= (A^\top \Gamma^{-1} A + C_2^{-1})(A^\top \Gamma^{-1} A + C_1^{-1})^{-1} - \mathrm{I} \\ &= (C_2^{-1} - C_1^{-1})(A^\top \Gamma^{-1} A + C_1^{-1})^{-1} \end{aligned}$$

(similarly to the other term) and consequently

$$C_{post,2}^{-1} C_{post,1} + C_{post,1}^{-1} C_{post,2} - 2\mathrm{I} = (C_2^{-1} - C_1^{-1}) C_{post,1} (C_2^{-1} - C_1^{-1}) C_{post,2}.$$

What we get is

$$\left\| C_{post,2}^{-1} C_{post,1} + C_{post,1}^{-1} C_{post,2} - 2\mathrm{I} \right\|_2 \lesssim \left\| C_1^{-1} - C_2^{-1} \right\|_2^2 \lesssim \| C_1 - C_2 \|_2^2$$

and hence (25). Finally, putting together (23) and (25) we obtain

$$d_{Hell}(\mu_{post}^1, \mu_{post}^2)^2 \leq 1 - \exp(c \| C_1 - C_2 \|_2^2) \leq c \| C_1 - C_2 \|_2^2$$

for some constant $c > 0$. □

Theorem 9. *Suppose our two priors are $\mu_1 \sim \mathcal{N}(x_1, C_1)$ and $\mu_2 \sim \mathcal{N}(x_2, C_2)$. It follows that*

$$d_{Hell}(\mu_{post}^1, \mu_{post}^2) \lesssim \| x_1 - x_2 \|_2 + \| C_1 - C_2 \|_2 .$$

Proof. Let us define an auxiliary prior $\mu_3 \sim \mathcal{N}(x_3, C_3)$ and the corresponding posterior μ_{post}^3. By triangle inequality and Lemmas 5.2 and 5.3 we have

$$d_{Hell}(\mu_{post}^1, \mu_{post}^2) \leq d_{Hell}(\mu_{post}^1, \mu_{post}^3) + d_{Hell}(\mu_{post}^3, \mu_{post}^2) \lesssim \| x_1 - x_2 \|_2 + \| C_1 - C_2 \|_2 ,$$

□

A Crash course on probability in Banach spaces

TO BE CONTINUED.

B Convergence of probability measures

The results in this appendix are written for probability measures on a separable Banach space $(B, \mathcal{B}(B))$ with its Borel σ-algebra. However, the structure of the space is not essential and so just take $B = \mathbb{R}^n$ if you will and keep in mind that $B^* = B = \mathbb{R}^n$ in that case.

B.1 Weak convergence

Definition B.1. *Let μ_n, $n \in \mathbb{N}$ and μ be probability measures on $(B, \mathcal{B}(B))$. Then we say that μ_n converges weakly to μ if for all $f \in C_b(B, \mathbb{R})$ it holds that*

$$\lim_{n \to \infty} \int_B f(x) d\mu_n(x) = \int_B f(x) d\mu(x).$$

If this is the case, we write $\mu_n \rightharpoonup \mu$.

Definition B.2. *Let μ be a probability measure on $(B, \mathcal{B}(B))$. The characteristic function $\psi_\mu : B^* \to \mathbb{C}$ is defined by*

$$\psi_\mu(x^*) := \int_B \exp(i\langle x^*, x\rangle_{B^*\times B}) d\mu(x).$$

Lemma B.3. *Let μ_n, $n \in \mathbb{N}$, be probability measures on $(B, \mathcal{B}(B))$. If for all $x^* \in B^*$ it holds that*

$$\psi_{\mu_n}(x^*) \to \exp\left(i\langle x^*, x_0\rangle_{B^*\times B} - \frac{1}{2}\langle x^*, C_0 x^*\rangle_{B^*\times B}\right),$$

where $x_0 \in B$ and $C_0 : B^* \to B$ is symmetric positive-definite operator then

$$\mu_n \rightharpoonup \mathcal{N}(x_0, C_0),$$

i.e., the measures converge to a Gaussian distribution with mean $x_0 \in B$ and covariance $C_0 : B^* \to B$.

B.2 Metrics on probability measures

Suppose μ_1 and μ_2 are probability measures and ν a σ-finite measure on $(B, \mathcal{B}(B))$. In this subsection we assume that ν satisfies $\mu_1 \ll \nu$ and $\mu_2 \ll \nu$ simultaneously. Such a measure surely exists since one can take $\nu = \frac{1}{2}(\mu_1 + \mu_2)$.

Definition B.4. *Total variation distance between μ_1 and μ_2 is defined by*

$$d_{TV}(\mu_1, \mu_2) = \frac{1}{2}\int_B \left|\frac{d\mu_1}{d\nu} - \frac{d\mu_2}{d\nu}\right| d\nu.$$

In particular, if $\mu_1 \ll \mu_2$, we have

$$d_{TV}(\mu, \mu') = \frac{1}{2}\int_B \left|1 - \frac{d\mu_1}{d\mu_2}\right| d\mu.$$

Definition B.5. *The Hellinger distance between μ_1 and μ_2 is defined by*

$$d_{Hell}(\mu_1, \mu_2) = \sqrt{\frac{1}{2}\int_B \left(\sqrt{\frac{d\mu_1}{d\nu}} - \sqrt{\frac{d\mu_2}{d\nu}}\right)^2 d\nu}.$$

Again, if it holds that $\mu_1 \ll \mu_2$, then

$$d_{Hell}(\mu_1, \mu_2) = \sqrt{\frac{1}{2}\int_B \left(1 - \sqrt{\frac{d\mu_1}{d\mu_2}}\right)^2 d\nu} = \sqrt{1 - \int_B \sqrt{\frac{d\mu_1}{d\mu_2}} d\mu_2}.$$

Notice that the constant $\frac{1}{2}$ guarantees that

$$0 \le d_{TV}(\mu_1, \mu_2), d_{Hell}(\mu_1, \mu_2) \le 1.$$

Lemma B.6. *We have*

$$\frac{1}{\sqrt{2}} d_{TV}(\mu_1, \mu_2) \le d_{Hell}(\mu_1, \mu_2) \le d_{TV}(\mu_1, \mu_2)^{\frac{1}{2}}.$$

Proof. A lengthy calculation yields

$$\begin{aligned}
d_{TV}(\mu_1,\mu_2) &= \frac{1}{2}\int_B \left|\sqrt{\frac{d\mu_1}{d\nu}}-\sqrt{\frac{d\mu_2}{d\nu}}\right|\left|\sqrt{\frac{d\mu_1}{d\nu}}+\sqrt{\frac{d\mu_2}{d\nu}}\right| d\nu \\
&\leq \sqrt{\frac{1}{2}\int_B \left(\sqrt{\frac{d\mu_1}{d\nu}}-\sqrt{\frac{d\mu_2}{d\nu}}\right)^2 d\nu}\sqrt{\frac{1}{2}\int_B \left(\sqrt{\frac{d\mu_1}{d\nu}}+\sqrt{\frac{d\mu_2}{d\nu}}\right)^2 d\nu} \\
&\leq \sqrt{\frac{1}{2}\int_B \left(\sqrt{\frac{d\mu_1}{d\nu}}-\sqrt{\frac{d\mu_2}{d\nu}}\right)^2 d\nu}\sqrt{\int_B \left(\frac{d\mu_1}{d\nu}+\frac{d\mu_2}{d\nu}\right) d\nu} \\
&= \sqrt{\int_B \left(\sqrt{\frac{d\mu_1}{d\nu}}-\sqrt{\frac{d\mu_2}{d\nu}}\right)^2 d\nu} \\
&= \sqrt{2}d_{Hell}(\mu_1,\mu_2),
\end{aligned}$$

where we applied Cauchy–Schwarz inequality. The second inequality follows by using $|\sqrt{a}-\sqrt{b}| \leq \sqrt{a}+\sqrt{b}$, since

$$\begin{aligned}
d_{Hell}(\mu_1,\mu_2)^2 &= \frac{1}{2}\int_B \left(1-\sqrt{\frac{d\mu_1}{d\mu_2}}\right)^2 d\nu = \sqrt{1-\int_B \sqrt{\frac{d\mu_1}{d\mu_2}} d\mu_2} \\
&\leq \frac{1}{2}\int_B \left|\sqrt{\frac{d\mu_1}{d\nu}}-\sqrt{\frac{d\mu_2}{d\nu}}\right|\left|\sqrt{\frac{d\mu_1}{d\nu}}+\sqrt{\frac{d\mu_2}{d\nu}}\right| d\nu \\
&= d_{TV}(\mu_1,\mu_2).
\end{aligned}$$

□

References

[1] M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2013.

[2] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.