Bayesian inversion: theoretical perspective

Lecture notes, spring 2016 course

Tapio Helin

February 5, 2016

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

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

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

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

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 we notice that the reconstruction matrix has a modified eigenvalue structure. Namely, we have

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

and hence the problem is stabilized. Moreover, we find 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

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

We immediately notice the effect that if α is increased, the first error term (3) increases (becoming comparable to $\|x_0\|_2$!). Meanwhile, if α decreases, the second error term explodes. The optimal strategy is a balance between the two errors. As illustrated here: the more accurate information you have related to the unknown x_0, the noise δ and structure of the problem (here: eigenvalue structure), the better choices you can make with 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

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

Therefore, it follows that

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

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

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

and

$$\bar{x} = C_{12}C_{22}^{-1} m = C_0 A^\top (AC_0A^\top + \Gamma)^{-1} m.$$

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

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

and

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

Proof. Will follow on the next lecture. □