Numerical Methods for Bayesian Inverse Problems

Robert Scheichl and Jakob Zech

Universität Heidelberg

October 6, 2021

[Draft of October 6, 2021. Not for dissemination.]

Acknowledgment

These lecture notes have been prepared for the lecture "Numerical Methods for Bayesian Inverse Problems" held in the winter semester 2020/21 at Heidelberg University. They are based on various different sources, but mainly on the lecture notes "Inverse Probleme" by B. Sprungk (Freiberg) and C. Schillings (Mannheim). These are themselves based on other sources including the notes by C. Clason (Duisburg-Essen), B. von Harrach (Stuttgart) and M. Burger (Erlangen-Nürnberg).

More specifically, Chapters 1 and 2 are up to minor modifications directly translated from the notes of Sprungk and Schillings. In Chapter 3, Sections 3.1-3.2 are based on the lecture notes "Stochastic evoluation equations" of J. van Neerven (Delft), and Sections 3.3-3.5 on the book "Wahrscheinlichkeitstheorie" by A. Klenke (Mainz). Section 3.6 was up to few modifications taken from lecture notes on "Numerical Methods for Stochastic Modeling and Inference" by Y. Marzouk (MIT) and T. Cui (Monash), and Section 3.7 is based on the lecture notes "A Bayesian Approach To Inverse Problems" by M. Dashti (Sussex) and A. Stuart (Caltech). Chapter 4 follows again loosely Chapter 3 of the lecture notes by Sprungk and Schillings, but with significant changes both in presentation and the shown results. Chapter 5 has been prepared using several resources. The most important ones are always listed at the beginning of the relevant section. In addition to the listed references, Section 5.1 is based on Chapter 6 of [Kaipio, Somersalo, 2004]; Section 5.4 is based on the Lecture Notes of "High Dimensional Approximation and Applications in Uncertainty Quantification" held by A. Gilbert and R. Scheichl in the summer semester 2020 at Heidelberg University; Section 5.5 is based on Lecture Notes by Art Owen from Stanford University available at ⟨https://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf⟩, but includes also some material from other sources; Section 5.6 are again up to minor modifications directly translated from Section 3.6 of the notes of Sprungk and Schillings; Section 5.7 is based on the Article "Variational Inference: A Review for Statisticians" by D. M. Blei, A. Kucukelbir and J. D. McAuliffe, and on Chapter 2 of the book "Optimal Transport for Applied Mathematicians" by F. Santambrogio. Finally, Section 5.8 was contributed by Simon Weissmann (Heidelberg), and the relevant references can be found at the beginning of that section.

Main References

- B. Sprungk and C. Schillings, Inverse Probleme, lecture notes, Uni Mannheim, 2017.
- J. van Neerven, Stochastic evolution equations, ISEM lecture notes, 2007.
- T. Cui, Y. Marzouk, Numerical Methods for Stochastic Modeling and Inference, lecture, MIT, 2018.
- M. Dashti and A. Stuart, The Bayesian Approach To Inverse Problems, *Handbook of Uncertainty Quantification* (R. Ghanem, D. Higdon, H. Owhadi, Eds.), Springer, 2015.

- D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational Inference: A Review for Statisticians, 2016.
- F. Santambrogio, Optimal Transport for Applied Mathematicians, Springer, 2015.
- H.W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, 2000.
- L.C. Evans, *An Introduction to Stochastic Differential Equations*, American Mathematical Society, Providence, RI, 2013.
- T. Hytönen, J. van Neerven, M. Veraar, L. Weis, *Analysis in Banach spaces. Vol. I. Martingales and Littlewood-Paley theory.*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, Springer, Cham, 2016.
- J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, 2004.
- A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, 2nd edition, Springer, 2011.
- A. Klenke, *Wahrscheinlichkeitstheorie*, 4th edition, Springer, 2020.
- J. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2001.
- A. Rieder, *Keine Probleme mit Inversen Problemen*, Vieweg, 2003.
- C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd edition, Springer, 2004.
- J. Rosenthal, *A First Look at Rigorous Probability Theory*, World Scientific Publishing Co. Pte. Ltd., 2006.
- A. M. Stuart, *Inverse problems: A Bayesian perspective.* Acta Numerica **19**:451-559, 2010.
- L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004.

Contents

Chapter 1

Introduction

Definition 1.0.1. According to **Hadamard** (1865-1963) a problem is called **well-posed**, if

(i) a solution exists **(existence)**,

(ii) the solution is unique **(uniqueness)**,

(iii) the solution depends continuously on the input data **(stability)**.

If any of these properties is violated, we speak of an **ill-posed** problem.

Let X, Y be Hilbert spaces and $A : X \to Y$ be linear and bounded (we write $A \in \mathcal{L}(X, Y)$). Then the (forward) problem to compute $y = Ax$ for a given $x \in X$, is clearly well-posed. For the corresponding **inverse problem**, to solve the linear equation $Ax = y$ for a given $y \in Y$, the conditions of Hadamard are:

(i) Existence: $y \in \mathcal{R}(A)$, i.e. A is surjective.

(ii) Uniqueness: A is injective

(iii) Stability: A^{-1} is bounded.

For finite dimensional problems, these conditions may be satisfied for a bounded linear operator A, although the problem typically gets more and more **ill-conditioned** as the dimension increases. In infinite dimensions on the other hand, it is in general impossible. In particular, for compact operators A the singular values have to accumulate at 0, which implies that A^{-1} is unbounded. Thus, fundamentally the inverse problem is ill-posed if the forward problem is well-posed.

In this course, we will study how ill-posed problems can be solved in a numerically stable manner. Our particular focus will lie on the Bayesian approach and Bayesian techniques, but before we get there, we will first study classical approaches.

1.1 A motivating example

Many processes in science and engineering can be modelled via differential equations. Assuming complete knowledge of all the necessary parameters, initial and boundary conditions, the solution of such a differential equation allows in principle to fully predict the process.

Consider for example a rod of length 1 with thermal diffusivity coefficient α. The temperature at the two ends of the rod is assumed to be 0. Then the temperature distribution $u(x,t)$ satisfies

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}, \qquad \text{for } \ 0 < x < 1, \ t > 0, \tag{1.1}$$

with boundary conditions

$$u(0,t) = u(1,t) = 0, \qquad \text{for } \ t > 0, \tag{1.2}$$

and initial condition

$$u(x,0) = u_0(x), \qquad \text{for } \ 0 < x < 1. \tag{1.3}$$

We can now consider the following inverse problem in this simple setting: Given the temperature distribution at some time $T > 0$, can we recover the initial temperature profile u_0 at time $t = 0$?

Using the Laplace transform method, the solution to the heat equation (1.1)-(1.3) has the general form

$$u(x,t) = \sum_{n=1}^{\infty} \theta_n e^{-(n\pi)^2 \alpha t} \sin(n\pi x),$$

where θ_n are the Fourier-sine-coefficients of the initial condition u_0, i.e.,

$$u_0(x) = \sum_{n=1}^{\infty} \theta_n \sin(n\pi x).$$

Thus, in principle the coefficients θ_n of the initial condition u_0 can be estimated from measurements of $u(x,T)$ at time $T > 0$. However, consider two initial conditions $u_0^{(1)}$ and $u_0^{(2)}$ with $\theta_1^{(1)} = \theta_1^{(2)} = 1$ that differ only in one single frequency component, i.e.,

$$u_0^{(1)}(x) - u_0^{(2)}(x) = \theta_N \sin(N\pi x), \quad \text{for some } \ N > 1.$$

At time $T > 0$ the two solutions will differ by

$$u^{(1)}(x,T) - u^{(2)}(x,T) = \theta_N e^{-(N\pi)^2 \alpha T} \sin(N\pi x),$$

which is exponentially small. Therefore, any information about this difference will be lost due to measurement noise for T or N sufficiently large, even if the noise is extremely small.

To demonstrate this consider the case of $\alpha = 0.01$ (which after nondimensionalisation roughly corresponds to a copper rod of length 10cm in dimensionless quantities with time measured in seconds) and let

$$\theta_1^{(1)} = \theta_1^{(2)} = 1, \quad \theta_5^{(2)} = 0.5 \quad \text{and} \quad \theta_i^{(j)} = 0 \quad \text{otherwise.}$$

In Figure 1.1, the solution is plotted at the initial time $t = 0$ and at $t = 1$ and $t = 4$. Even though the two initial conditions clearly differ significantly and the difference is not even particularly oscillatory, it is already very difficult to distinguish the two solutions at $t = 1$. At $t = 4$, it will be impossible to say whether the observed temperature profile came from $u_0^{(1)}$ or from $u_0^{(2)}$.

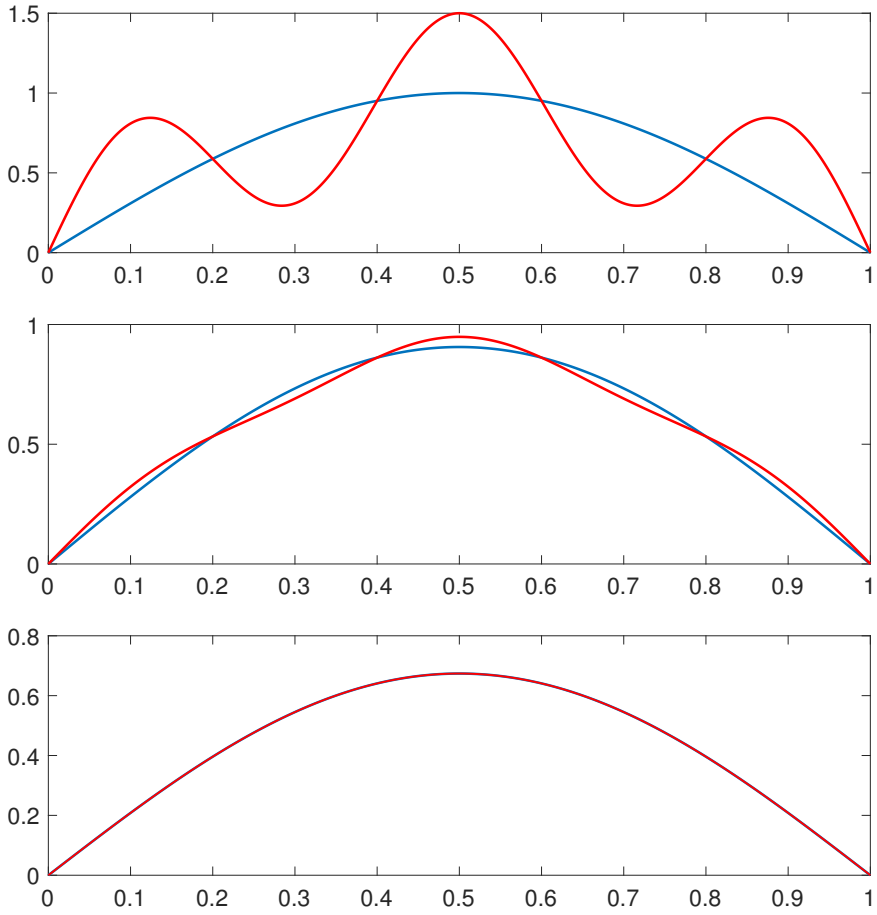


Figure 1.1: Inverse heat equation example with two initial conditions that differ only in one frequency: $u^{(1)}(x,t)$ (blue curve) and $u^{(2)}(x,t)$ (red curve) at times $t = 0$ (top), $t = 1$ (middle) and $t = 4$ (bottom).

Chapter 2

Linear Inverse Problems and Regularisation

To motivate the remainder of this chapter and to relate to things that you have already come across in earlier courses (e.g., in an introductory numerical analysis course, such as Numerik 0 in Heidelberg), we will first consider the finite dimensional setting. However, the most relevant issue in the numerical treatment of ill-posed problems, namely the lack of continuous dependence on the data, only emerges in infinite dimensions. Thus, in the remainder of this chapter we analyse infinite dimensional linear inverse problems and introduce regularisation techniques to solve them approximatively in a numerically stable way.

2.1 Finite dimensional ill-posed problems (matrix equations)

It suffices to consider matrix equations. Every finite dimensional vector space X over $\mathbb{R}$ is isomorphic to $\mathbb{R}^n$ and every linear operator on $\mathbb{R}^n$ has a matrix representation.

The regular case. Thus, to begin with consider a linear equation system of the form

$$Ax = y \tag{2.1.1}$$

with a symmetric, positive definite (SPD) $n \times n$ square matrix $A \in \mathbb{R}^{n\times n}$. Recall that such a matrix A has n positive, real eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n > 0$ with corresponding eigenvectors $u_i \in \mathbb{R}^n$, $i = 1, \ldots, n$, with $\|u_i\| = 1$. Furthermore, A has the spectral decomposition

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \left(= U\Lambda U^\top \right), \tag{2.1.2}$$

where the ith column of U is u_i and Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$. W.l.o.g. assume that $\lambda_1 = \mathcal{O}(1)$, in particular independent of n, otherwise rescale A and y.

As studied in detail in Numerik 0 (or an equivalent course), the condition number of A provides a measure for how accurate and stable the system (2.1.1) can be solved. It is given by the ratio of the largest and the smallest eigenvalue of A, i.e., $\kappa(A) = \lambda_1/\lambda_n$.

Consider that the data, namely the right hand side y, is only available in only a perturbed (or noisy) form as y^δ, such that

$$\|y^\delta - y\| \leq \delta \tag{2.1.3}$$

for some $\delta > 0$ in the Euclidean norm on $\mathbb{R}^n$, and denote by x^δ the solution of the perturbed system with right hand side y^δ. Using the decomposition (2.1.2) of A, we get

$$x^\delta - x = \sum_{i=1}^{n} \frac{u_i^\top (y^\delta - y)}{\lambda_i} \, u_i \, .$$

Since the eigenvectors of A can be chosen to be orthonormal (Numerik 0), we can apply the Bessel inequality (A.1) to obtain the bound

$$\|x^\delta - x\|^2 = \sum_{i=1}^{n} \lambda_i^{-2} |u_i^\top (y^\delta - y)|^2 \leq \lambda_n^{-2} \|y^\delta - y\|^2 \leq \lambda_n^{-2} \delta^2 \, .$$

for the error in the solution. Using the condition number and our assumption on the scaling of λ_1 this can also be expressed as

$$\|x^\delta - x\| \leq \kappa \lambda_1^{-1} \delta = \mathcal{O}(\kappa \delta) \, .$$

The bound is sharp, which can be seen easily by choosing $y^\delta - y = \delta u_n$. Thus, any growth in the condition number of A directly leads to an amplification of noise in the data in the solution.

Thus, for large condition numbers we say that the problem (2.1.1) is ill-posed – recall for example that the condition number of the stiffness matrix A in finite element discretisations of elliptic PDEs typically grows like $\mathcal{O}(h^{-2})$, where h is the mesh width. Note however that for finite dimensional problems Hadamard's third condition is not strictly speaking violated and so (2.1.1) is not ill-posed in the sense of Hadamard, it is only **ill-conditioned**, but it is **asymptotically ill-posed** for $\kappa \to \infty$ (e.g. as $h \to 0$ in the FE problem).

On the positive side, we also note that the above expansion shows clearly that errors in the low frequency components $i \ll n$, i.e., the components in the direction of eigenvectors corresponding to the larger eigenvalues, are not amplified as much. This is a typical situation in inverse problems (recall the introductory example in Section 1.1).

The singular case. Let us now consider the case that A in (2.1.1) is positive semi-definite, i.e. it has a nontrivial kernel. Since $A^* = A^T = A$, we can decompose the vector space in

$$\mathbb{R}^n = \mathcal{N}(A) + \mathcal{R}(A) \, ,$$

where $\mathcal{R}$ is the range and $\mathcal{N}$ is the kernel (cf. Appendix A). Let λ_m be the smallest nonzero eigenvalue and let $\kappa_{\text{eff}} = \lambda_1 / \lambda_m$ be the **effective condition number**. Then

$$x = \sum_{i=1}^{m} \lambda_i^{-1} u_i u_i^\top y$$

and the problem is solvable (Hadamard's first condition) iff $u_i^\top y = 0$ for $i > m$.

In the general noisy case, this will usually not be satisfied, but we can for example project the noisy data y^δ into the range of A via a projection $P : \mathbb{R}^n \to \mathcal{R}(A)$. Now the problem is solvable and the solution x_P^δ with data Py^δ satisfies

$$x_P^\delta - x = \sum_{i=1}^{m} \lambda_i^{-1} u_i u_i^\top (Py^\delta - y) \, .$$

Since by construction $u_i^\top P y^\delta = u_i^\top y^\delta$, for $i \le m$, we have

$$\|x_P^\delta - x\| \le \lambda_m^{-1}\delta = \mathcal{O}(\kappa_{\text{eff}}\delta)\,.$$

No (arbitrary) contributions in the kernel components are included and the error amplification is again determined by the smallest nonzero eigenvalue (or equivalently by the effective condition number). This is typical for finite dimensional operators, i.e. matrices. However, in practice it may be difficult to find P without first performing a spectral decomposition of A.[1]

Outlook to infinite dimensions. In the general case of a linear operator A between two infinite dimensional Hilbert spaces X and Y, the range of A and A^* are not necessarily closed. In that case we have

$$X = \mathcal{N}(A) + \overline{\mathcal{R}(A^*)} \quad \text{and} \quad Y = \mathcal{N}(A^*) + \overline{\mathcal{R}(A)}$$

(cf. Appendix A). If the range of A is not closed, i.e., $\overline{\mathcal{R}(A)} \neq \mathcal{R}(A)$, then the projection P is not bounded, which leads again to instabilities. Any operator A with eigenvalues arbitrarily close to 0 will have this behaviour, in particular every compact operator (see below).

Regularisation. Let us now discuss ideas for numerically stable ways to solve such ill-posed problems and introduce **regularisation methods** for matrix equations.

We saw above that small eigenvalues of A are causing instabilities. A natural approach would thus be to approximate the matrix A with a family of matrices with eigenvalues bounded away from zero. One such family is

$$A_\alpha := A + \alpha I\,, \qquad \alpha > 0\,.$$

The eigenvalues of A_α are $\lambda_i + \alpha$, $i = 1, \dots, n$ and the eigenvectors remain unchanged.

To estimate the **regularisation error** consider again the regular (SPD) case, i.e. $\lambda_n > 0$ and let $x = A^{-1}y$ and $x_\alpha = A_\alpha^{-1}y$. (The singular case can be handled similarly.) Then

$$x - x_\alpha = \sum_{i=1}^n \left(\frac{1}{\lambda_i} - \frac{1}{\lambda_i + \alpha}\right) u_i u_i^\top y = \sum_{i=1}^n \frac{\alpha}{\lambda_i(\lambda_i + \alpha)} (u_i^\top y)\, u_i$$

and using again the Bessel inequality we can estimate the regularisation error by

$$E_\alpha(\alpha) := \|x - x_\alpha\| \le \frac{\alpha}{\lambda_n(\lambda_n + \alpha)} \|y\|\,.$$

In particular, we have $E_\alpha \to 0$ as $\alpha \to 0$. In the case of a noisy data y^δ, with x_α^δ the solution of $A_\alpha x_\alpha^\delta = y^\delta$, the spectral decomposition gives

$$x_\alpha^\delta - x_\alpha = \sum_{i=1}^n (\lambda_i + \alpha)^{-1} u_i u_i^\top (y^\delta - y)\,.$$

and thus the perturbation error can be estimated by

$$E_\delta(\alpha, \delta) := \|x_\alpha^\delta - x_\alpha\| \le \frac{\delta}{\lambda_n + \alpha}\,.$$

[1] Another way to deal with the singularity of A is to multiply (2.1.1) with A^T and to form the normal equations (see below), but $\kappa(A^T A) = \kappa(A)^2$ and so the ill-conditioning gets even worse!

Finally, using the triangle inequality the total error between the exact solution and the solution of the regularised problem with noisy data can be bounded by

$$\|x - x_\alpha^\delta\| \leq E_\alpha(\alpha) + E_\delta(\alpha, \delta) \leq \left(\frac{\alpha}{\lambda_n(\lambda_n + \alpha)} \|y\| + \frac{\delta}{\lambda_n + \alpha} \right) .$$

In practice, the exact data is not known, but we can bound $\|y\| \leq \|y^\delta\| + \delta$ using (2.1.3) and thus obtain

$$\|x - x_\alpha^\delta\| \leq \left(\frac{\alpha}{\lambda_n(\lambda_n + \alpha)} (1 + \delta_{\text{rel}}) + \frac{\delta_{\text{rel}}}{\lambda_n + \alpha} \right) \|y^\delta\| ,$$

where $\delta_{\text{rel}} = \delta / \|y^\delta\|$ is the relative noise level (or the inverse signal-to-noise ratio).

For fixed δ_{rel} the two terms in the error bound behave very differently with respect to α. The first term decreases monotonically as $\alpha \to 0$ while the second one grows monotonically.

The main task in regularisation is thus to determine the optimal α that minimises the total error, either through an a priori choice $\alpha = \alpha(\delta)$ or through an a posteriori choice $\alpha = \alpha(\delta_{\text{rel}})$ that takes into account the size of the data $\|y^\delta\|$. Clearly the optimal α will always depend on δ, but any regularisation strategy needs to at least satisfy the requirement that $\alpha(\delta) \to 0$ as $\delta \to 0$, so that in the noise-free case the exact solution is recovered.

The discussion can easily be generalised also to arbitrary rectangular linear equation systems with $A \in \mathbb{R}^{n \times m}$ (and thus also to arbitrary linear operators between finite dimensional vector spaces of possibly different dimension) by considering the normal equations

$$A^\top A x = A^\top y .$$

However, the ill-conditioning is significantly worse since $\kappa(A^T A) = \kappa(A)^2$.

In the next section we will go one step further and look at general linear inverse problems on arbitrary, infinite dimensional Hilbert spaces.

2.2 Generalised inverse - the infinite dimensional setting

In this section, throughout $A \in \mathcal{L}(X, Y)$ is a linear bounded operator between the Hilbert spaces X and Y, and we are interested in solutions of the linear operator equation

$$Ax = y \tag{2.2.1}$$

for possibly non-injective and/or non-surjective A. For $y \notin \mathcal{R}(A)$, (2.2.1) has no solution (Hadamard 1). In this case, a sensible thing to do – as we also did in Numerik 0 in finite dimensions – is to find $x \in X$ that minimises $\|Ax - y\|_Y$. On the other hand, for $\mathcal{N}(A) \neq \{0\}$ there are infinitely many solutions (Hadamard 2). In that case, we choose the one that minimises $\|x\|_X$. This leads to the following definition.

Definition 2.2.1. An element $x \in X$ is called

- **least-squares solution** of $Ax = y$ (more precisely the Y**-best approximate solution**), if

$$\|Ax - y\|_Y = \min_{z \in X} \|Az - y\|_Y ,$$

[Draft of October 6, 2021. Not for dissemination.]

- **minimum-norm (or (X,Y)-best approximate) solution** of $Ax = y$, if x is least-squares solution and
$$\|x\|_X = \min\{\|z\|_X : z \text{ is least squares solution of } Az = y\}\,.$$

For A bijective, $x = A^{-1}y$ is the only minimum-norm solution. However, a minimum-norm solution does not have to exist if $\mathcal{R}(A)$ is not closed. To study which $y \in Y$ admit a minimum-norm solution, we introduce an operator that maps y to the minimum-norm solution; it is called **generalised inverse** or **pseudoinverse**.

To do this we first restrict its domain to the range of A to guarantee invertibility before extending the domain as much as possible.

Definition 2.2.2. Let $A \in \mathcal{L}(X,Y)$ and define
$$\tilde{A} := A\big|_{\mathcal{N}(A)^\perp} : \mathcal{N}(A)^\perp \to \mathcal{R}(A)\,. \tag{2.2.2}$$

The **Moore-Penrose** (or **generalised**) **inverse** $A^\dagger$ is the unique, linear extension of $\tilde{A}^{-1}$ with
$$\begin{aligned} \mathcal{D}(A^\dagger) &:= \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp\,, \quad \text{and} \\ \mathcal{N}(A^\dagger) &= \mathcal{R}(A)^\perp\,. \end{aligned} \tag{2.2.3}$$

Due to the restriction to $\mathcal{N}(A)^\perp$ and $\mathcal{R}(A)$ the operator $\tilde{A}$ in (2.2.2) is bijective (cf. Appendix A). Thus, $A^\dagger$ is well-defined on $\mathcal{R}(A)$. For arbitrary $y \in \mathcal{D}(A^\dagger)$, an orthogonal decomposition guarantees the existence of $y_1 \in \mathcal{R}(A)$ and $y_2 \in \mathcal{R}(A)^\perp$ such that $y = y_1 + y_2$. Finally, due to $\mathcal{N}(A^\dagger) = \mathcal{R}(A)^\perp$ we have
$$A^\dagger y = A^\dagger y_1 + A^\dagger y_2 = A^\dagger y_1 = \tilde{A}^{-1} y_1\,, \tag{2.2.4}$$
and thus $A^\dagger$ is well-defined on all of $\mathcal{D}(A^\dagger)$, defined in (2.2.3).

Theorem 2.2.3. *The Moore-Penrose inverse $A^\dagger$ satisfies $\mathcal{R}(A^\dagger) = \mathcal{N}(A)^\perp$, as well as the* **Moore-Penrose equations**

(i) $AA^\dagger A = A$

(ii) $A^\dagger A A^\dagger = A^\dagger$

(iii) $A^\dagger A = Id_X - P_{\mathcal{N}}$

(iv) $AA^\dagger = (P_{\overline{\mathcal{R}}})\big|_{\mathcal{D}(A^\dagger)}$

where $P_{\mathcal{N}}$ and $P_{\overline{\mathcal{R}}}$ are the orthogonal projections to $\mathcal{N}(A)$ and $\overline{\mathcal{R}(A)}$, respectively. (The Moore-Penrose equations characterise $A^\dagger$ uniquely.)

Proof. As shown in (2.2.4), for all $y \in \mathcal{D}(A^\dagger)$, it follows that $A^\dagger y \in \mathcal{R}(\tilde{A}^{-1}) = \mathcal{N}(A)^\perp$, i.e.. $\mathcal{R}(A^\dagger) \subset \mathcal{N}(A)^\perp$. Conversely, it follows from the definition of $\tilde{A}$ that for all $x \in \mathcal{N}(A)^\perp$
$$A^\dagger A x = A^\dagger \tilde{A} x = \tilde{A}^{-1}\tilde{A} x = x,$$
i.e., $x \in \mathcal{R}(A^\dagger)$. Thus, $\mathcal{R}(A^\dagger) = \mathcal{N}(A)^\perp$.

Since orthogonal projections are always closed (cf. Appendix A), $\mathcal{R}(A)^\perp$ is closed and thus $\mathcal{R}(P_{\overline{\mathcal{R}}}) \cap \mathcal{D}(A^\dagger) = \mathcal{R}(A)$. Thus, for all $y \in \mathcal{D}(A^\dagger)$

$$A^\dagger y = \tilde{A}^{-1} P_{\overline{\mathcal{R}}} y \tag{2.2.5}$$

which due to $\tilde{A}^{-1} P_{\overline{\mathcal{R}}} y \in \mathcal{N}(A)^\perp$ implies $AA^\dagger y = P_{\overline{\mathcal{R}}} y$ and thus equation (iv).

The proof of the other three Moore-Penrose equations is left as an exercise. □

We will now show that the Moore-Penrose inverse provides the minimum-norm solution.

Theorem 2.2.4. *Let $y \in \mathcal{D}(A^\dagger)$. Then $Ax = y$ has a unique minimum-norm solution $x^\dagger \in X$, which is given by*

$$x^\dagger = A^\dagger y \,.$$

The set of all least-squares solution is $x^\dagger + \mathcal{N}(A)$.

Proof. Let $y \in \mathcal{D}(A^\dagger)$. To show existence of least-squares solutions consider the set

$$S := \{z \in X : Az = P_{\overline{\mathcal{R}}} y\} \,,$$

which is non-empty, since $P_{\overline{\mathcal{R}}}$ maps $\mathcal{D}(A^\dagger)$ to $\mathcal{R}(A)$ (cf. the discussion before (2.2.5)). Let $z \in S$. Then, due to the optimality of the orthogonal projection

$$\|Az - y\|_Y = \|P_{\overline{\mathcal{R}}} y - y\|_Y = \min_{w \in \overline{\mathcal{R}(A)}} \|w - y\|_Y \leq \|Ax - y\| \quad \text{for all } x \in X \,,$$

i.e., z is least-squares solution of $Ax = y$. Conversely, let $z \in X$ be a least squares solution. Then it follows again from $P_{\overline{\mathcal{R}}} y \in \mathcal{R}(A)$ that

$$\|P_{\overline{\mathcal{R}}} y - y\|_Y \leq \|Az - y\| = \min_{x \in X} \|Ax - y\|_Y = \min_{w \in \mathcal{R}(A)} \|w - y\|_Y \leq \|P_{\overline{\mathcal{R}}} y - y\|_Y \,,$$

i.e., Az is the orthogonal projection of y onto $\overline{\mathcal{R}(A)}$. In summary,

$$\{x \in X : x \text{ is least squares solution of } Ax = y\} = S \neq \emptyset \,.$$

Each element $z \in S$ can be decomposed uniquely into $x = \tilde{x} + x_0$ with $\tilde{x} \in \mathcal{N}(A)^\perp$ and $x_0 \in \mathcal{N}(A)$, but we have already seen in (2.2.5) that the unique solution to $Az = P_{\overline{\mathcal{R}}} y$ in $N(A)^\perp$ is

$$\tilde{x} = \tilde{A}^{-1} P_{\overline{\mathcal{R}}} y = A^\dagger y = x^\dagger.$$

Thus, the set of all least squares solution is $x^\dagger + \mathcal{N}(A)$. Finally, due to the orthogonality of $x^\dagger$ and x_0,

$$\|z\|_X^2 = \|x^\dagger + x_0\|_X^2 = \|x^\dagger\|_X^2 + \|x_0\|_X^2 \geq \|x^\dagger\|_X^2$$

so that $x^\dagger$ is also the unique minimum-norm solution. □

Theorem 2.2.5. *Let $y \in \mathcal{D}(A^\dagger)$. Then $x \in X$ is least-squares solution of $Ax = y$ iff x satisfies the* **normal equations**

$$A^* Ax = A^* y \,.$$

If in addition $x \in \mathcal{N}(A)^\perp$, then $x = x^\dagger$.

[Draft of October 6, 2021. Not for dissemination.]

Proof. As in the previous proof, $x \in X$ is least-squares solution iff $Ax = P_{\overline{\mathcal{R}}} y$, which is equivalent to $Ax \in \overline{\mathcal{R}(A)}$ and $Ax - y \in \overline{\mathcal{R}(A)}^{\perp} = \mathcal{N}(A^*)$. This in turn is equivalent to $A^*(Ax - y) = 0$. The final part was already proved in the previous theorem. □

The minimum-Norm solution $x^\dagger$ of $Ax = y$ is the solution and thus in particular the least-squares solution of the normal equations with minimum norm, i.e.,

$$x^\dagger = (A^*A)^\dagger A^* y \,.$$

Thus, $x^\dagger$ can be computed as the minimum-norm solution of the normal equations.

So far we have considered the generalised inverse on $\mathcal{D}(A^\perp) = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$ without studying this domain in detail. Since orthogonal complements are always closed,

$$\overline{\mathcal{D}(A^\dagger)} = \overline{\mathcal{R}(A)} \oplus \mathcal{R}(A)^\perp = \mathcal{N}(A^*)^\perp \oplus \mathcal{N}(A^*) = Y \,,$$

i.e., $\mathcal{D}(A^\dagger)$ is dense in Y. Thus, $\mathcal{D}(A^\dagger) = Y$ iff $\mathcal{R}(A)$ is closed. Furthermore, for any $y \in \mathcal{R}(A)^\perp = \mathcal{N}(A^\dagger)$ the minimum- norm solution is $x^\dagger = 0$.

The central question is if $R(A)$ is closed. If it is then $A^\dagger$ is even bounded. Conversely, if there exists any $y \in \overline{\mathcal{R}(A)} \setminus \mathcal{R}(A)$, then $A^\dagger$ cannot be bounded.

Theorem 2.2.6. *Let $A \in \mathcal{L}(X, Y)$. Then $A^\dagger \in \mathcal{L}(\mathcal{D}(A^\dagger), X)$ iff $\mathcal{R}(A)$ is closed.*

Proof. We apply Theorem A.1.2, the Closed Graph Theorem, and show first that $A^\dagger$ is closed.

Let $(y_n)_{n\in\mathbb{N}} \subset \mathcal{D}(A^\dagger)$ be a convergent sequence with $y_n \to y \in Y$ and $A^\dagger y_n \to x \in X$. From Moore-Penrose equation (iv) and the continuity of orthogonal projections it follows that

$$AA^\dagger y_n = P_{\overline{\mathcal{R}}} y_n \to P_{\overline{\mathcal{R}}} y \,,$$

which due to the continuity of A implies

$$P_{\overline{\mathcal{R}}} y = \lim_{n\to\infty} P_{\overline{\mathcal{R}}} y_n = \lim_{n\to\infty} AA^\dagger y_n = Ax \,,$$

i.e., x is least-squares solution. Furthermore, $A^\dagger y_n \in \mathcal{R}(A^\dagger) = \mathcal{N}(A)^\perp$ and so

$$A^\dagger y_n \to x \in \mathcal{N}(A)^\perp \,,$$

since $\mathcal{N}(A)^\perp = \overline{\mathcal{R}(A^*)}$ is closed. Hence, x is in fact the minimum-norm solution of $Ax = y$, i.e., $x = A^\dagger y$, and $A^\dagger$ is closed.

Now let $\mathcal{R}(A)$ be closed. Then $\mathcal{D}(A^\dagger) = Y$ and Theorem A.1.2 implies that $A^\dagger : Y \to X$ is bounded. Conversely, let $A^\dagger$ be bounded on $\mathcal{D}(A^\dagger)$. In that case, since $\mathcal{D}(A^\dagger)$ is dense in Y, $A^\dagger$ can be continuously extended to an operator $\overline{A^\dagger} \in \mathcal{L}(Y, X)$ by defining

$$\overline{A^\dagger} y := \lim_{n\to\infty} A^\dagger y_n \quad \text{for some sequence } (y_n)_{n\in\mathbb{N}} \subset \mathcal{D}(A^\dagger) \text{ with } y_n \to y \in Y.$$

Due to its continuity, $A^\dagger$ maps Cauchy sequences to Cauchy sequences, and thus $\overline{A^\dagger}$ is well-defined and bounded. Now let $y \in \overline{\mathcal{R}(A)}$ and $(y_n)_{n\in\mathbb{N}} \subset \mathcal{R}(A)$ with $y_n \to y$. It follows from Moore-Penrose equation (iv) and the continuity of A that

$$y = P_{\overline{\mathcal{R}}} y = \lim_{n\to\infty} P_{\overline{\mathcal{R}}} y_n = \lim_{n\to\infty} AA^\dagger y_n = A\overline{A^\dagger} y \in \mathcal{R}(A) \,,$$

and thus $\overline{\mathcal{R}(A)} = \mathcal{R}(A)$, which completes the proof. □

Unfortunately this excludes the most interesting case of a compact operator on a Hilbert space.

Lemma 2.2.7. *Let $K \in \mathcal{K}(X,Y)$, i.e., K is compact, with infinite dimensional image $\mathcal{R}(K)$. Then $K^\dagger$ is not bounded.*

Proof. Suppose that $K^\dagger$ is bounded. Then it follows from Theorem 2.2.6 that $\mathcal{R}(K)$ is closed and

$$\tilde{K} := K\big|_{\mathcal{N}(K)^\perp} : \mathcal{N}(K)^\perp \to \mathcal{R}(K)$$

is a bijective operator with bounded inverse $\tilde{K}^{-1} \in \mathcal{L}\big(\mathcal{R}(K), \mathcal{N}(K)^\perp\big)$. Since K is compact, $K \circ \tilde{K}^{-1}$ is also compact. But $K \circ \tilde{K}^{-1}$ is the identity on $\mathcal{R}(K)$, which can only be compact iff $\mathcal{R}(K)$ is finite dimensional. □

2.3 Singular value decomposition of compact operators

We now use an orthonormal system to characterise the Moore-Penrose inverse of compact operators $K \in \mathcal{K}(X,Y)$. To do this for general non-selfadjoint operators we need to generalise the Spectral Theorem (Thm. A.2.6). Because of Theorem 2.2.5 we can look at the selfadjoint operator K^*K instead. This leads to the singular value decomposition (cf. again Numerik 0 for matrices).

Theorem 2.3.1. *Let $K \in \mathcal{K}(X,Y)$. Then there exists*

(i) a (null) sequence $(\sigma_n)_{n\in\mathbb{N}}$ with $\sigma_1 \geq \sigma_2 \geq \ldots > 0$ and $\sigma_n \to 0$ as $n \to \infty$,

(ii) an orthonormal basis $(u_n)_{n\in\mathbb{N}} \subset Y$ of $\overline{\mathcal{R}(K)}$, and

(iii) an orthonormal basis $(v_n)_{n\in\mathbb{N}} \subset X$ of $\overline{\mathcal{R}(K^*)}$,

with

$$Kv_n = \sigma_n u_n \quad \text{and} \quad K^*u_n = \sigma_n v_n\,, \quad \text{for all } n \in \mathbb{N}\,, \tag{2.3.1}$$

and

$$Kx = \sum_{n\in\mathbb{N}} \sigma_n \langle x, v_n\rangle_X u_n\,, \quad \text{for all } x \in X\,. \tag{2.3.2}$$

A sequence $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$ that provides such a **singular value decomposition (SVD)** (2.3.2) *of K, is called* **singular system**.

Proof. Since $K^*K : X \to X$ is compact and selfadjoint, it follows from Theorem A.2.6 that there exists a null sequence $(\lambda_n)_{n\in\mathbb{N}} \subset \mathbb{R}\setminus\{0\}$ and an orthonormal system $(v_n)_{n\in\mathbb{N}} \subset X$ such that

$$K^*Kx = \sum_{n\in\mathbb{N}} \lambda_n \langle x, v_n\rangle_X v_n \quad \text{for all } x \in X\,.$$

Moreover, (v_n) is an orthonormal basis (ONB) of $\overline{\mathcal{R}(K^*K)}$.

Now, since

$$\lambda_n = \lambda_n \|v_n\|_X^2 = \langle \lambda_n v_n, v_n\rangle_X = \langle K^*Kv_n, v_n\rangle_X = \|Kv_n\|_X^2 > 0\,,$$

we can define for all $n \in \mathbb{N}$

$$\sigma_n := \sqrt{\lambda_n} > 0 \quad \text{and} \quad u_n := \frac{1}{\sigma_n} Kv_n \in Y$$

[Draft of October 6, 2021. Not for dissemination.]

so that (σ_n) is a strictly positive null sequence and the first equation in (2.3.1) is satisfied. Moreover,

$$\langle u_i, u_j\rangle_Y = \frac{1}{\sigma_i\sigma_j}\langle Kv_i, Kv_j\rangle_Y = \frac{1}{\sigma_i\sigma_j}\langle K^*Kv_i, v_j\rangle_X = \frac{\lambda_i}{\sigma_i\sigma_j}\langle v_i, v_j\rangle_X = \begin{cases} 1\,, & \text{if } i = j,\\ 0\,, & \text{otherwise,}\end{cases}$$

and thus $(u_n)_{n\in\mathbb{N}}$ is an orthonormal system in Y. Furthermore, for all $n \in \mathbb{N}$,

$$K^*u_n = \sigma_n^{-1}K^*Kv_n = \sigma_n^{-1}\lambda_n v_n = \sigma_n v_n\,,$$

i.e., the second equation in (2.3.1) holds.

To show that $(v_n) \subset X$ is not only an ONB of $\overline{\mathcal{R}(K^*K)}$ but also of $\overline{\mathcal{R}(K^*)}$, it suffices to show that $\overline{\mathcal{R}(K^*)} \subset \overline{\mathcal{R}(K^*K)}$. Let $x \in \overline{\mathcal{R}(K^*)}$. For any $\epsilon > 0$, there exists a

$$y \in \mathcal{N}(K^*)^\perp = \overline{\mathcal{R}(K)} \text{ with } \|K^*y - x\|_X < \frac{\epsilon}{2} \quad \text{and} \quad \tilde{x} \in X \text{ with } \|K\tilde{x} - y\| < \frac{\epsilon}{2}\|K\|^{-1}_{\mathcal{L}(X,Y)}\,,$$

such that

$$\|K^*K\tilde{x} - x\|_X \le \|K^*K\tilde{x} - K^*y\|_X + \|K^*y - x\|_X < \epsilon$$

and thus $x \in \overline{\mathcal{R}(K^*K)}$.

To prove the SVD (2.3.2) consider first an arbitrary $\tilde{x} \in \mathcal{N}(K)^\perp$ and

$$\tilde{x}_N := \sum_{j=1}^{N}\langle \tilde{x}, v_j\rangle_X v_j\,,$$

i.e., the partial basis representation of $\tilde{x}$ with respect to the ONB (v_n) of $\overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^\perp$. Clearly

$$K\tilde{x}_N = \sum_{j=1}^{N}\langle \tilde{x}, v_j\rangle_X Kv_j = \sum_{j=1}^{N}\sigma_j\langle \tilde{x}, v_j\rangle_X u_j\,.$$

Since $\tilde{x}_N \to \tilde{x}$ and K is bounded,

$$K\tilde{x} = \lim_{N\to\infty} K\tilde{x}_N = \sum_{j=1}^{\infty}\sigma_j\langle \tilde{x}, v_j\rangle_X u_j\,. \tag{2.3.3}$$

Now, let $x \in X$ be arbitrary. Then, there exist unique $\tilde{x} \in \mathcal{N}(K)^\perp$, $x_0 \in \mathcal{N}(K)$ such that $x = \tilde{x} + x_0$ and

$$\sigma_j\langle x, v_j\rangle_X = \langle x, K^*u_j\rangle_X = \langle Kx, u_j\rangle_Y = \langle K\tilde{x}, u_j\rangle_Y = \sigma_j\langle \tilde{x}, v_j\rangle_X$$

Substituting this into (2.3.3) and using the fact that $Kx = K\tilde{x}$ leads to the SVD (2.3.2).

Finally, to show that $(u_n) \subset Y$ is an ONB of $\overline{\mathcal{R}(K)}$ let $y \in \overline{\mathcal{R}(K)}$ be arbitrary. Then, there exists a sequence $(x_n) \subset X$ such that

$$y = \lim_{n\to\infty} Kx_n = \lim_{n\to\infty}\sum_{j=1}^{\infty}\langle Kx_n, u_j\rangle_Y u_j = \sum_{j=1}^{\infty}\langle y, u_j\rangle_Y u_j \qquad \text{and} \qquad \|y\|_Y^2 = \sum_{j=1}^{\infty}|\langle y, u_j\rangle_Y|^2\,.$$

This implies that $(u_n)_{n\in\mathbb{N}} \subset Y$ is an ONB of $\overline{\mathcal{R}(K)}$. □

Since eigenvalues λ_n of K^*K with eigenvector v_n are eigenvalues of KK^* with eigenvector u_n as well, (2.3.1) also provides a SVD of K^*:

$$K^*y = \sum_{n\in\mathbb{N}} \sigma_n \langle y, u_n\rangle_Y v_n \quad \text{for all } y \in Y\,.$$

We will now use the SVD of K to characterise the domain $\mathcal{D}(K^\dagger) = \mathcal{R}(K) \oplus \mathcal{R}(K)^\perp$ of the Moore-Penrose inverse $K^\dagger$. Recall that minimum-norm solution for $y \in \mathcal{R}(K)^\perp = \mathcal{N}(K^*)$ is $x^\dagger = 0$, and conversely $\mathcal{N}(K^*)^\perp = \overline{\mathcal{R}(K)}$. Thus, the crucial question is whether an element $y \in \overline{\mathcal{R}(K)}$ also lies in $\mathcal{R}(K)$.

Theorem 2.3.2. *Let $K \in \mathcal{K}(X,Y)$ with singular system $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$ and $y \in \overline{\mathcal{R}(K)}$. Then, $y \in \mathcal{R}(K)$ iff the* ***Picard-condition***

$$\sum_{n\in\mathbb{N}} \sigma_n^{-2} |\langle y, u_n\rangle_Y|^2 < \infty \tag{2.3.4}$$

is satisfied. In this case

$$K^\dagger y = \sum_{n\in\mathbb{N}} \sigma_n^{-1} \langle y, u_n\rangle_Y v_n\,. \tag{2.3.5}$$

Proof. First let $y \in \mathcal{R}(K)$, i.e. there exists a $x \in X$ with $Kx = y$. Then, using Bessel's inequality

$$\sum_{n\in\mathbb{N}} \sigma_n^{-2} |\langle y, u_n\rangle_Y|^2 = \sum_{n\in\mathbb{N}} \sigma_n^{-2} |\langle x, K^*u_n\rangle_X|^2 = \sum_{n\in\mathbb{N}} |\langle x, v_n\rangle_X|^2 \le \|x\|_X^2 < \infty\,.$$

To show the reverse implication, let $y \in \overline{\mathcal{R}(K)}$ and suppose that (2.3.4) holds. Then, the sequence $(s_N)_{N\in\mathbb{N}}$ of partial sums $s_N := \sum_{n=1}^N \sigma_n^{-2} |\langle y, u_n\rangle_Y|^2$ is a Cauchy sequence and thus

$$(x_N)_{N\in\mathbb{N}} \quad \text{with} \quad x_N := \sum_{n=1}^N \sigma_n^{-1} \langle y, u_n\rangle_Y v_n$$

is also a Cauchy sequence. In other words,

$$\|x_N - x_M\|_X^2 = \|\sum_{n=N}^M \sigma_n^{-1} \langle y, u_n\rangle_Y v_n\|_X^2 = \sum_{n=N}^M |\sigma_n^{-1} \langle y, u_n\rangle_Y|^2 \to 0\,,$$

where we used that $(v_n)_{n\in\mathbb{N}}$ is an orthonormal system in $\overline{\mathcal{R}(K^*)}$. Thus, $(x_N)_{N\in\mathbb{N}} \subset \overline{\mathcal{R}(K^*)}$ converges to

$$x := \sum_{n\in\mathbb{N}} \sigma_n^{-1} \langle y, u_n\rangle_Y v_n \;\in\; \overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^\perp$$

(since $\overline{\mathcal{R}(K^*)}$ is closed). Now,

$$Kx = \sum_{n\in\mathbb{N}} \sigma_n^{-1} \langle y, u_n\rangle_Y K v_n = \sum_{n\in\mathbb{N}} \langle y, u_n\rangle_Y u_n = P_{\overline{\mathcal{R}(K)}} y = y\,,$$

so that $y \in \mathcal{R}(K)$.

However, due to Theorem 2.2.4, $x \in \mathcal{N}(K)^\perp$ and $Kx = P_{\overline{\mathcal{R}(K)}} y$ is equivalent to $K^\dagger y = x$. □

The Picard-condition states that a minimum-Norm solution exists only if the coefficients $\langle y, u_n\rangle_Y$ of y with respect to the ONB (u_n) decay faster than the singular values σ_n. The representation shows clearly how perturbations of y will affect $x^\dagger$: In particular, if $y^\delta = y + \delta u_n$ then

$$\|K^\dagger y^\delta - K^\dagger y\|_X = \delta \|K^\dagger u_n\|_X = \sigma_n^{-1}\delta \to \infty \quad \text{for } n \to \infty\,.$$

Therefore, the faster the singular values decay the more data errors are amplified for a fixed n. We call a problem

- **moderately ill-posed**, if there are $c, r > 0$ such that $\sigma_n \geq cn^{-r}$ for all $n \in \mathbb{N}$,
- **severely ill-posed**, if this is not the case, and
- **exponentially ill-posed**. If there are $c, r > 0$ such that $\sigma_n \leq ce^{n^{-r}}$ for all $n \in \mathbb{N}$.

For exponentially ill-posed problems, such as the inverse heat equation in Section 1.1, we can typically expect only very crude estimates for the solution. However, if $\mathcal{R}(K)$ is finite dimensional, the sequence (σ_n) truncates at a finite N, i.e. $\sigma_n = 0$ for $n > N$ and the error remains bounded; in this case $K^\dagger$ is bounded.

In practice, infinite dimensional problems typically need to be discretised. In general, integral equations or differential equations can not be solved explicitly like the simple one-dimensional inverse heat equation in Section 1.1. So strictly speaking, in practice we will always solve finite dimensional inverse problems. But as highlighted already in Section 2.1, the problem will be asymptotically ill-posed as the discretisation parameter $h \to 0$, and so we need to find a way to deal with this ill-posedness in a more uniform way, independently of h. One way to achieve this is via **regularisation** which we will now discuss in the linear case in the remainder of this chapter. The other is to apply a statistical (**Bayesian**) approach, which we will return to in Chapter 4.

2.4 Regularisation

We have seen that for $y \in \mathcal{D}(A^\dagger)$ the minimum-norm solution $x^\dagger = A^\dagger y$ of the ill-posed operator equation $Ax = y$ exists. Now consider, as in Section 2.1, the situation where y is known only up to the measurement (or representation) error δ (the **noise level**), i.e. we only know y^δ with

$$\|y^\delta - y\|_Y \leq \delta\,.$$

Since $A^\dagger$ is in general not bounded, $A^\dagger y^\delta$ will normally be a bad approximation to $x^\dagger$, even if $y^\delta \in \mathcal{D}(A^\dagger)$. Thus, in a **regularisation method** we will typically aim to find an approximation x_α^δ, that depends on the one hand continuously on y^δ and thus on δ, and on the other hand can be selected as close to $x^\dagger$ as the noise level δ allows by a judicious choice of the **regularisation parameter** $\alpha > 0$. In particular, the choice of $\alpha(\delta)$ should guarantee that $x^\delta_{\alpha(\delta)} \to x^\dagger$ as $\delta \to 0$.

In the case of a linear operator on a Hilbert space this can be achieved by defining a family of regularisation operators that provide bounded replacements of the unbounded pseudoinverse $A^\dagger$.

Definition 2.4.1. Let X, Y be two Hilbert spaces and $A \in \mathcal{L}(X, Y)$ a bounded, linear operator. A family $(A_\alpha^\dagger)_{\alpha>0}$ of linear operators $A_\alpha^\dagger : Y \to X$ is called a **regularisation** of $A^\dagger$ for $\alpha > 0$ if

(i) $A_\alpha^\dagger \in \mathcal{L}(Y, X)$ for all $\alpha > 0$,

(ii) $A_\alpha^\dagger y \to A^\dagger y$ for all $y \in \mathcal{D}(A^\dagger)$, as $\alpha \to 0$.

Thus, a regularisation is a pointwise approximation of the Moore-Penrose inverse by a sequence of bounded, linear operators. Since $A^\dagger$ is in general not bounded, it follows from Theorem A.1.4 (Banach-Steinhaus) that the convergence will in general **not** be uniform.

Theorem 2.4.2. *Let $A \in \mathcal{L}(X,Y)$ and $(A_\alpha^\dagger)_{\alpha>0} \subset \mathcal{L}(Y,X)$ a regularisation. If $A^\dagger$ is unbounded then the family $(A_\alpha^\dagger)_{\alpha>0}$ is not uniformly bounded. In particular, there exists a $y \in Y$ such that $\|A_\alpha^\dagger y\|_X \to \infty$ as $\alpha \to 0$.*

In fact, by adding an additional condition we can show divergence for all $y \notin \mathcal{D}(A^\dagger)$.

Theorem 2.4.3. *Let $A \in \mathcal{L}(X,Y)$ with $A^\dagger$ unbounded and $(A_\alpha^\dagger)_{\alpha>0} \subset \mathcal{L}(Y,X)$ a regularisation. If*

$$\sup_{\alpha>0} \|AA_\alpha^\dagger\|_{\mathcal{L}(Y,Y)} < \infty\,, \tag{2.4.1}$$

then $\|A_\alpha^\dagger y\|_X \to \infty$ as $\alpha \to 0$ for all $y \notin \mathcal{D}(y^\dagger)$.

Since in general $y^\delta \notin \mathcal{D}(A^\dagger)$, to analyse the total error we decompose it as

$$\begin{aligned} \|A_\alpha^\dagger y^\delta - A^\dagger y\|_X &\leq \|A_\alpha^\dagger y^\delta - A_\alpha^\dagger y\|_X + \|A_\alpha^\dagger y - A^\dagger y\|_X \\ &\leq \delta \|A_\alpha^\dagger\|_{\mathcal{L}(Y,X)} + \|A_\alpha^\dagger y - A^\dagger y\|_X\,. \end{aligned} \tag{2.4.2}$$

This decomposition is a fundamental tool of regularisation theory that will be used throughout. The first term represents the (propagated) **data error** that remains unbounded for $\alpha \to 0$ while $\delta > 0$. The second term is the **regularisation error** that due to the pointwise convergence of $A_\alpha^\dagger$ converges to zero as $\alpha \to 0$. Thus, to obtain a meaningful approximation, the regularisation parameter α has to be chosen correctly as a function of δ, in particular such that the total error converges to zero as $\delta \to 0$.

2.4.1 Parameter choice

Definition 2.4.4. A function $\alpha : \mathbb{R}_+ \times Y \to \mathbb{R}_+, (\delta, y^\delta) \mapsto \alpha(\delta, y^\delta)$ is called **parameter choice rule**. A regularisation $(A_\alpha^\dagger)_{\alpha>0} \subset \mathcal{L}(Y,X)$ of $A^\dagger$ together with a parameter choice rule α is called a **regularisation method** of (2.2.1). The regularisation method $(A_\alpha^\dagger, \alpha)$ is called **convergent** if

$$\limsup_{\delta\to 0}\{\|A^\dagger_{\alpha(\delta,y^\delta)} y^\delta - A^\dagger y\|_X : y^\delta \in Y, \|y^\delta - y\|_Y \leq \delta\} = 0\,, \quad \text{for all } y \in \mathcal{D}(A^\dagger)\,. \tag{2.4.3}$$

We distinguish between

- **a priori parameter choice rules** that only depend on δ;
- **a posterior parameter choice rules** that depend on δ and y^δ;
- **heuristic rules** that only depend on y^δ.

It can be shown that for all regularisations there exists an a priori rule and thus a convergent regularisation method. Furthermore, the following characterisation of a priori rules leads to convergent regularisation methods.

Theorem 2.4.5. *Let $(A_\alpha^\dagger)_{\alpha>0}$ be a regularisation and $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ an a-priori rule with*

(i) $\lim_{\delta\to 0}\alpha(\delta)=0$,

(ii) $\lim_{\delta\to 0}\delta\|A^\dagger_{\alpha(\delta)}\|_{\mathcal{L}(Y,X)}=0$.

Then $(A^\dagger_\alpha,\alpha)$ *is a convergent regularisation method.*

Proof. Due to the decomposition (2.4.2) it follows that

$$\|A^\dagger_{\alpha(\delta)}y^\delta - A^\dagger y\|_X \le \delta\|A^\dagger_{\alpha(\delta)}\|_{\mathcal{L}(X,Y)} + \|A^\dagger_{\alpha(\delta)}y - A^\dagger y\|_X \to 0 \quad \text{as } \delta\to 0,$$

where we have used (ii) and the pointwise convergence of the regularisation operators under condition (i). □

Let $y\in\mathcal{D}(A^\dagger)$ and $y^\delta\in Y$ with $c\delta\le\|y^\delta-y\|_Y\le\delta$ for some $0<c\le 1$. The main idea of a-posteriori rules can be described as follows: for $x^\delta_\alpha := A^\dagger_\alpha y^\delta$ we consider the **residual**

$$\|Ax^\delta_\alpha - y^\delta\|_Y\,.$$

Even for $y\in\mathcal{R}(A)$ and the minimum-norm solution $Ax^\dagger=y$ we only have

$$\|Ax^\dagger - y^\delta\|_Y = \|y-y^\delta\|_Y \ge c\delta\,.$$

Thus, it makes no sense to expect a smaller residual for the approximation x^δ_α. This motivates:

Definition 2.4.6 (Discrepancy Principle of Morozov)**.** Given $\delta>0$ and y^δ, choose $\alpha=\alpha(\delta,y^\delta)$ such that

$$\|Ax^\delta_\alpha - y^\delta\|_Y \le \tau\delta \qquad \text{for some } \tau>1\,. \tag{2.4.4}$$

This principle does not have to be satisfied: if $y\in\mathcal{D}(A^\dagger)$ such that $y=Ax+y^\perp$ for some $x\in X$ and $0\neq y^\perp\in\mathcal{R}(A)^\perp$ and $\delta<\frac{1}{2}\|y^\perp\|_Y$, then even for exact data $y^\delta=y$,

$$\|Ax^\dagger - y\|_Y = \|AA^\dagger y - y\|_Y = \|P_{\overline{\mathcal{R}(A)}}y - y\|_Y = \|y^\perp\|_Y > 2\delta\,.$$

Thus, we have to assume that this is not possible. It suffices to assume that $\mathcal{R}(A)$ is dense in Y, since in that case $\mathcal{R}(A)^\perp=\{0\}$.

A practical approach to implement such an a posteriori rule is to choose a null sequence $(\alpha_n)_{n\in\mathbb{N}}$, to successively calculate $x^\delta_{\alpha_n}$ for $n=1,\dots$ and to terminate the iteration as soon as the discrepancy principle (2.4.4) is satisfied. The following theorem justifies this approach.

Theorem 2.4.7. *Let* $(A^\dagger_\alpha)_{\alpha>0}$ *be a regularisation of* $A\in\mathcal{L}(X,Y)$ *with* $\mathcal{R}(A)$ *dense in* Y*, and suppose that the family* $(AA^\dagger_\alpha)_{\alpha>0}$ *is uniformly bounded. Consider a strictly monotonic null sequence* $(\alpha_n)_{n\in\mathbb{N}}$ *and* $\tau>1$*. Then for all* $y\in\mathcal{D}(A^\dagger)$ *and* $y^\delta\in Y$ *with* $\|y-y^\delta\|_Y\le\delta$ *there exists* $n^*\in\mathbb{N}$ *such that*

$$\|Ax^\delta_{\alpha_{n^*}} - y^\delta\|_Y \le \tau\delta < \|Ax^\delta_{\alpha_n} - y^\delta\|_Y \qquad \textit{for all } n<n^*\,.$$

Proof. For all $y\in\mathcal{D}(A^\dagger)$, $AA^\dagger_\alpha y$ converges pointwise to $AA^\dagger y = P_{\overline{\mathcal{R}}}y$. Thus, due to the uniform boundedness of $(AA^\dagger_\alpha)_{\alpha>0}$ this convergence extends to all $y\in Y=\overline{\mathcal{D}(A^\dagger)}$. This implies for all $y\in\mathcal{D}(A^\dagger)=\mathcal{R}(A)$ and $y^\delta\in Y$ with $\|y^\delta-y\|_Y\le\delta$ that

$$\begin{aligned}\lim_{n\to\infty}\|Ax^\delta_{\alpha_n}-y^\delta\|_Y &= \lim_{n\to\infty}\|AA^\dagger_{\alpha_n}y^\delta - y^\delta\|_Y \\ &= \|P_{\overline{\mathcal{R}}}y^\delta - y^\delta\|_Y = \min_{z\in\overline{\mathcal{R}(A)}}\|z-y^\delta\|_Y \le \|y-y^\delta\|_Y \le \delta\,.\end{aligned}$$

The existence of an $n^*\in\mathbb{N}$ then follows directly, since $\tau>1$. □

[Draft of October 6, 2021. Not for dissemination.]

Heuristic rules do not even assume any knowledge of the noise level δ, which is highly relevant in practice, since often it is hard or impossible to estimate δ accurately. However, such a strategy cannot work in general, as the following important result (called the *"Bakushinskii veto"*) shows.

Theorem 2.4.8 (Bakushinskii, 1985)**.** *Let $(A_\alpha^\dagger)_{\alpha>0}$ be a regularisation. If there exists a heuristic parameter choice rule $\alpha \neq \alpha(\delta)$ such that $(A_\alpha^\dagger, \alpha)$ is a convergent regularisation method, then $A^\dagger$ is bounded.*

2.4.2 Convergence rates

We will only give a very general discussion here and refer to the references given at the start of the notes for details.[2]

A central question in the regularisation of inverse problems is the derivation of error bounds of the form

$$\|A_{\alpha(\delta,y^\delta)}^\dagger y^\delta - A^\dagger y\| \leq \phi(\delta)$$

for some function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t\to 0} \phi(t) = 0$ that is independent of y. We are interested in particular in the **worst case error**

$$e(y,\delta) := \sup\{\|A_{\alpha(\delta,y^\delta)}^\dagger y^\delta - A^\dagger y\|_X : y^\delta \in Y \text{ with } \|y - y^\delta\|_Y \leq \delta\}\,. \tag{2.4.5}$$

This would allow us to provide a priori error bounds for the regularisation method. However, without any further assumptions on y or on $x^\dagger = A^\dagger y$ this hope is futile.

Theorem 2.4.9. *Let $(A_\alpha^\dagger, \alpha)$ be a convergent regularisation method. If there exists a function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t\to 0} \phi(t) = 0$ and*

$$\sup_{y\in\mathcal{D}(A^\dagger)\ s.t.\ \|y\|_Y\leq 1} e(y,\delta) \leq \phi(\delta)\,, \tag{2.4.6}$$

then $A^\dagger$ is bounded.

Proof. Let $y \in \mathcal{D}(A^\dagger)$ with $\|y\|_Y \leq 1$ and $(y_n)_n \subset \mathcal{D}(A^\dagger)$ with $\|y_n\|_Y \leq 1$ a sequence satisfying $y_n \to y$ as $n \to \infty$. With $\delta_n := \|y - y_n\|_Y \to 0$ it then follows that

$$\begin{aligned} \|A^\dagger y_n - A^\dagger y\|_X &\leq \|A^\dagger y_n - A_{\alpha(\delta_n,y_n)}^\dagger y_n\|_X + \|A_{\alpha(\delta_n,y_n)}^\dagger y_n - A^\dagger y\|_X \\ &\leq e(y_n,\delta_n) + e(y,\delta_n) \ \leq\ 2\phi(\delta_n) \end{aligned}$$

and thus, $A^\dagger$ is bounded on $\mathcal{D}(A^\dagger) \cap B_Y$. Since $A^\dagger$ is linear the boundedness extends to all of $\mathcal{D}(A^\dagger)$. □

Thus the convergence can be arbitrarily slow without further assumptions on y or on $x^\dagger$. For compact operators $K \in \mathcal{K}(X,Y)$ (which are smoothing operators), this can be achieved via an abstract smoothness condition on $x^\dagger$, called a **source condition**, namely that

$$x^\dagger \in \mathcal{R}(|K|^\nu), \quad \text{for some } \ \nu > 0,$$

[2]Note that this is also one of the central research questions in Prof. Jan Johannes' group and the topic of specialist seminars offered by him and his group.

where $|K|^\nu x := \sum_{n\in\mathbb{N}} \sigma_n^\nu \langle x, v_n\rangle_X v_n$, using the singular system (σ_n, u_n, v_n) of K. This is in fact equivalent to a strengthened Picard-condition

$$\sum_{n\in\mathbb{N}} \sigma_n^{-2(\nu+1)} |\langle y, u_n\rangle_Y|^2 < \infty\,,$$

i.e. a faster decay of the coefficients of y (or equivalently of $Kx^\dagger$) with respect to (u_n) than necessary to purely guarantee the existence of $x^\dagger$ (as in Theorem 2.3.2).

Under this condition, it can be shown that the convergence rate as $\delta \to 0$ for the total error (with respect to δ) is bounded below by $\frac{\nu}{\nu+1}$ for any regularisation method. A regularisation method is called **order-optimal** for $\nu > 0$, if for all $x^\dagger \in R(|K|^\nu)$ there exists a constant $c = c(x^\dagger) > 0$ such that

$$e(Kx^\dagger, \delta) \le c\, \delta^{\frac{\nu}{\nu+1}}\,.$$

2.5 Construction of regularisation methods

Let us now consider the construction of regularisation methods for linear ill-posed problems. We focus on compact operators $K \in \mathcal{K}(X, Y)$ and recall that stability issues with the Moore-Penrose inverse arose from error amplification through small singular values. Therefore, we aim to construct regularisation methods in such a way that they modify the smallest singular values appropriately.

Thus, recall the SVD of $K^\dagger$ with respect to the singular system $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$ of K. It suggests to construct regularisation operators of the form

$$K_\alpha^\dagger y := \sum_{n=1}^{\infty} g_\alpha(\sigma_n) \langle y, u_n\rangle_Y v_n \qquad \text{for } y \in Y\,,$$

with a suitable function $g_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ that satisfies $g_\alpha(\sigma) \to \frac{1}{\sigma}$ for all $\sigma > 0$ as $\alpha \to 0$. We will see that $(K_\alpha^\dagger)_{\alpha\ge 0}$ is a regularisation if

$$g_\alpha(\sigma) \le C_\alpha < \infty\,, \qquad \text{for all } \sigma > 0\,. \tag{2.5.1}$$

Note that (2.5.1) implies

$$\|K_\alpha^\dagger y\|_X^2 = \sum_{n=1}^{\infty} (g_\alpha(\sigma_n))^2 |\langle y, u_n\rangle_Y|^2 \le C_\alpha^2 \sum_{n=1}^{\infty} |\langle y, u_n\rangle_Y|^2 \le C_\alpha^2 \|y\|_Y^2\,,$$

i.e. C_α is a bound for the norm of $K_\alpha^\dagger$. Moreover, the condition $\lim_{\delta\to 0} \delta \|K_{\alpha(\delta)}^\dagger\|_{\mathcal{L}(Y,X)} = 0$ can be replaced by

$$\lim_{\delta\to 0} \delta C_{\alpha(\delta)} = 0\,.$$

Before proving that such a construction leads to a convergent regularisation method, we illustrate the idea on some examples.

Example 2.5.1 (**Truncated SVD**)**.** Here, all singular values smaller than a prescribed value (controlled by α) are ignored (i.e. set to 0). We choose

$$g_\alpha(\sigma) := \begin{cases} \frac{1}{\sigma}\,, & \text{if } \sigma \geq \alpha, \\ 0\,, & \text{otherwise.} \end{cases} \tag{2.5.2}$$

Clearly, $g_\alpha(\sigma) \to \frac{1}{\sigma}$ as $\alpha \to 0$ and $C_\alpha = \frac{1}{\alpha}$. Thus, this regularisation with a-priori parameter choice rule leads to a convergent regularisation method provided $\frac{\delta}{\alpha} \to 0$. Furthermore, $\sup_{\sigma,\alpha} \sigma g_\alpha(\sigma) = 1$. The regularised solution is

$$x_\alpha^\delta := K_\alpha^\dagger y^\delta = \sum_{\sigma_n \geq \alpha} \frac{1}{\sigma_n} \langle y^\delta, u_n \rangle_Y v_n\,, \qquad \text{for } y^\delta \in Y\,,$$

motivating the name of the method. The sum in x_α^δ is always finite for $\alpha > 0$, since zero is the only accumulation point of the sequence (σ_n).

Example 2.5.2 (**Lavrentiev Regularisation**)**.** Here, all singular values are shifted away from zero by α, i.e. $g_\alpha(\sigma) = \frac{1}{\sigma+\alpha}$, and

$$x_\alpha^\delta := K_\alpha^\dagger y^\delta = \sum_{n=1}^{\infty} \frac{1}{\sigma_n + \alpha} \langle y^\delta, u_n \rangle_Y v_n\,, \qquad \text{for } y^\delta \in Y\,,$$

As in Example 2.5.1, the computation of this approximation requires an explicit knowledge of the singular system $(\sigma_n, u_n, v_n)_{n\in\mathbb{N}}$ of K, which is not very useful in practice. However, for selfadjoint, positive semidefinite operators K (i.e. $Y = X$, $\lambda_n = \sigma_n$ and $u_n = v_n$) we have

$$(K + \alpha I) x_\alpha^\delta = \sum_{n=1}^{\infty} (\sigma_n + \alpha) \langle x_\alpha^\delta, u_n \rangle_X u_n = \sum_{n=1}^{\infty} \langle y^\delta, u_n \rangle_Y u_n = y^\delta$$

and the regularised solution can be found without explicit knowledge of the singular system of K by solving

$$(K + \alpha I) x_\alpha^\delta = y^\delta\,.$$

Since $\frac{1}{\sigma+\alpha} \leq \frac{1}{\alpha}$, we have again $C_\alpha = \frac{1}{\alpha}$. Furthermore, $g_\alpha(\sigma) \to \frac{1}{\sigma}$ as $\alpha \to 0$ and $\sigma g_\alpha(\sigma) < 1$.

Example 2.5.3 (**Tikhonov Regularisation**)**.** Here

$$g_\alpha(\sigma) = \frac{\sigma}{\sigma^2 + \alpha}\,,$$

such that

$$x_\alpha^\delta := K_\alpha^\dagger y^\delta = \sum_{n=1}^{\infty} \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y^\delta, u_n \rangle_Y v_n\,, \qquad \text{for } y^\delta \in Y\,.$$

Since $\sigma^2 + \alpha \geq 2\sigma\sqrt{\alpha}$, we can choose $C_\alpha = \frac{1}{2\sqrt{\alpha}}$. Thus, a necessary condition for convergence with a-priori parameter rule is $\frac{\delta}{\sqrt{\alpha}} \to 0$. Furthermore, again $g_\alpha(\sigma) \to \frac{1}{\sigma}$ as $\alpha \to 0$ and $\sigma g_\alpha(\sigma) = \frac{\sigma^2}{\sigma^2+\alpha} < 1$.

As in Example 2.5.2, x_α^δ can be computed without explicit knowledge of the singular system, however, in this case for arbitrary $K \in \mathcal{K}(X, Y)$. In particular,

$$(K^*K + \alpha I) x_\alpha^\delta = K^* y^\delta\,, \tag{2.5.3}$$

a well-posed linear system for $\alpha > 0$. Tikhonov regularisation is in fact equivalent to Lavrentiev regularisation applied to the normal equations.

Example 2.5.4 (**Landweber Iteration**). For $\omega > 0$, consider the fixed point iteration

$$x_0 = 0 \quad \text{and} \quad x_{k+1} = x_k + \omega K^*(y^\delta - Kx_k)\,, \quad \text{for } k \geq 0,$$

to compute regularised solutions x_k of $Kx = y^\delta$. The associated family of regularisation operators $(K_k^\dagger)_{k\in\mathbb{N}}$ satisfies $K_k^\dagger y^\delta = x_k$. Using the SVD of K and K^* we get

$$\sum_{j=1}^{\infty} \langle x_{k+1}, v_j\rangle_X v_j = \sum_{j=1}^{\infty} \Big((1-\omega\sigma_j^2)\langle x_k, v_j\rangle_X + \omega\sigma_j\langle y^\delta, u_j\rangle_Y\Big)\, v_j$$

and due to orthogonality

$$\langle x_{k+1}, v_j\rangle_X = (1-\omega\sigma_j^2)\langle x_k, v_j\rangle_X + \omega\sigma_j\langle y^\delta, u_j\rangle_Y\,.$$

Since $x_0 = 0$, we get

$$\langle x_k, v_j\rangle_X = \omega\sigma_j\langle y^\delta, u_j\rangle_Y \sum_{i=1}^{k}(1-\omega\sigma_j^2)^{k-i} = \frac{1-(1-\omega\sigma_j^2)^k}{\sigma_j}\langle y^\delta, u_j\rangle_Y\,.$$

Now we interpret the iteration number as the regularisation parameter and set $\alpha := 1/k$, so that

$$x_\alpha^\delta = K_\alpha^\dagger y^\delta = \sum_{j=1}^{\infty} \frac{1-(1-\omega\sigma_j^2)^{\frac{1}{\alpha}}}{\sigma_j}\langle y^\delta, u_j\rangle_Y\,,$$

i.e. $g_\alpha(\sigma) = (1-(1-\omega\sigma^2)^{\frac{1}{\alpha}})\frac{1}{\sigma}$. This function converges to $\frac{1}{\sigma}$ as $\alpha \to 0$ provided $|1-\omega\sigma^2| < 1$. A sufficient condition for $\sigma \in \{\sigma_n\}$ is

$$0 < \omega < 2\|K\|^{-2}_{\mathcal{L}(X,Y)}\,.$$

Since g_α is continuous and $\lim_{\sigma\to 0} g_\alpha(\sigma) = 0$, it is also bounded and $\sigma g_\alpha(\sigma) < 1$ for any $\alpha > 0$.

Under the stated conditions on $g_\alpha(\sigma)$, we can now prove the convergence of all the above regularisation methods with parameter choice rule $\alpha = \alpha(\delta, y^\delta)$ satisfying $\lim_{\delta\to 0} \delta C_{\alpha(\delta,y^\delta)} = 0$.

Theorem 2.5.5. *Let $g_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be a piecewise continuous function such that $g_\alpha(\sigma) \to \frac{1}{\sigma}$ for $\sigma > 0$ as $\alpha \to 0$, and suppose that there exist a constant $C_\alpha > 0$ depending on α and a constant $\gamma > 0$ independent of α such that*

$$\sigma g_\alpha(\sigma) \leq \gamma \qquad \text{and} \qquad g_\alpha(\sigma) \leq C_\alpha < \infty \quad \text{for all } \sigma, \alpha > 0\,.$$

Consider (2.1.1) *with $A = K \in \mathcal{K}(X,Y)$, $y \in \mathcal{D}(K^\dagger)$ and perturbed data $y^\delta \in Y$ with $\|y - y^\delta\|_Y \leq \delta$. Then the regularisation method $(K_\alpha^\dagger, \alpha)$ with*

$$K_\alpha^\dagger y := \sum_{n=1}^{\infty} g_\alpha(\sigma_n)\langle y, u_n\rangle_Y v_n\,, \qquad \text{for } y \in Y\,,$$

and parameter choice rule $\alpha = \alpha(\delta, y^\delta)$ converges provided

$$C_{\alpha(\delta,y^\delta)}\delta \to 0 \quad \text{as} \quad \delta \to 0\,.$$

Proof. To show convergence we bound again the two terms in the decomposition (2.4.2).

First to show that $K_\alpha^\dagger \to K^\dagger$ on $\mathcal{D}(K^\dagger)$ let $y \in D(K^\dagger)$. Then,

$$K_\alpha^\dagger y - K^\dagger y = \sum_{n=1}^{\infty} (g_\alpha(\sigma_n) - \frac{1}{\sigma_n}) \langle y, u_n \rangle_Y v_n = \sum_{n=1}^{\infty} (\sigma_n g_\alpha(\sigma_n) - 1) \langle x^\dagger, v_n \rangle_X v_n \,.$$

Due to the assumptions on g_α, the coefficients in the above expansion satisfy

$$|(\sigma_n g_\alpha(\sigma_n) - 1) \langle x^\dagger, v_n \rangle_X| \leq (\gamma + 1) |\langle x^\dagger, v_n \rangle_X| \,,$$

i.e. the sequence is bounded, and thus

$$\begin{aligned} \limsup_{\alpha \to 0} \|K_\alpha^\dagger y - K^\dagger y\|_X^2 &\leq \limsup_{\alpha \to 0} \sum_{n=1}^{\infty} |\sigma_n g_\alpha(\sigma_n) - 1|^2 |\langle x^\dagger, v_n \rangle_X|^2 \\ &\leq \sum_{n=1}^{\infty} (\lim_{\alpha \to 0} |\sigma_n g_\alpha(\sigma_n)) - 1|^2 |\langle x^\dagger, v_n \rangle_X|^2 \;=\; 0 \,, \end{aligned}$$

since $\sigma g_\alpha(\sigma) \to 1$ pointwise. Thus, $\|K_\alpha^\dagger y - K^\dagger y\|_X \to 0$ for $\alpha \to 0$, independently of δ.

To bound the propagated data error, note that, for all $\alpha, \delta > 0$,

$$\begin{aligned} \|K_\alpha^\dagger y - K_\alpha^\dagger y^\delta\|_X^2 &\leq \sum_{n=1}^{\infty} g_\alpha(\sigma_n)^2 |\langle y - y^\delta, u_n \rangle_Y|^2 \\ &\leq C_\alpha^2 \sum_{n=1}^{\infty} |\langle y - y^\delta, u_n \rangle_Y|^2 \leq C_\alpha^2 \|y - y^\delta\|_Y^2 \leq (C_\alpha \delta)^2 \,. \end{aligned}$$

Thus, under the condition on the limit of $\delta C_{\alpha(\delta, y^\delta)}$ for the parameter choice rule $\alpha(\delta, y^\delta)$, this term also converges with $\delta \to 0$ and the proof is complete. □

Remark 2.5.6. The bound

$$\|K_\alpha^\dagger y - K_\alpha^\dagger y^\delta\|_X \leq C_\alpha \delta$$

in the proof suggests that the propagated data error is of order δ. However, this is not true since C_α depends on δ and will in general grow with $\delta \to 0$. However, since we required that $C_{\alpha(\delta, y^\delta)} \delta \to 0$ as $\delta \to 0$, C_α grows slower than δ decreases such that $C_\alpha \delta$ will be of order δ^ν for some $0 < \nu < 1$.

2.6 Variational regularisation and extensions

The solution of the Tikhonov regularised linear system (2.5.3) is equivalent to minimising the following quadratic functional

$$K_\alpha^\dagger y^\delta := \arg\min_{x \in X} \left\{ \frac{1}{2} \|Kx - y^\delta\|_Y^2 + \frac{\alpha}{2} \|x\|_X^2 \right\} . \tag{2.6.1}$$

This is how Tikhonov regularisation is typically introduced. In this variational setting, it is easier to generalise to other regularising functionals and to nonlinear inverse problems.

Indeed, if $\Phi(x) := \frac{1}{2}\|Kx - y^\delta\|_Y^2 + \frac{\alpha}{2}\|x\|_X^2$, the first-order optimality condition for a minimiser $x^* \in X$ of Φ is equivalent to setting $\frac{\mathrm{d}}{\mathrm{d}t}\Phi(x^* + th)\big|_{t=0} = 0$ for arbitrary $h \in X$ with $\|h\|_X = 1$. Expanding, we get

$$\begin{aligned}\Phi(x+th) &= \frac{1}{2}\big\langle K(x+th) - y^\delta, K(x+th) - y^\delta\big\rangle_Y + \frac{\alpha}{2}\langle x+th, x+th\rangle_X \\ &= \Phi(x) + t\Big(\langle Kx - y^\delta, Kh\rangle_Y + \alpha\langle x, h\rangle_X\Big) + \frac{t^2}{2}\Big(\|Kh\|_Y^2 + \alpha\|h\|_X^2\Big)\end{aligned}$$

and thus

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\Phi(x^* + th)|_{t=0} = \langle Kx^* - y^\delta, Kh\rangle_Y + \alpha\langle x^*, h\rangle_X = \langle K^*(Kx^* - y^\delta) + \alpha x^*, h\rangle_X\,,$$

which is equivalent to x^* being the solution of (2.5.3), since $h \in X$ with $\|h\|_X = 1$ was arbitrary.

Let $J : X \to \mathbb{R}$ be a functional on X, then **generalised Tikhonov regularisation** seeks the regularised solution as the minimum of

$$\Phi_{\alpha,y^\delta}(x) := \frac{1}{2}\|Kx - y^\delta\|_Y^2 + \frac{\alpha}{2}J(x)\,.$$

Example 2.6.1 (**Tikhonov-Philipps Regularisation**)**.** A simple generalisation of the classical Tikhonov regularisation consists in simply replacing $\frac{1}{2}\|x\|_X$ by $\frac{1}{2}\|Dx\|_Z$ for some linear (not necessarily bounded) operator $D : X \to Z$ from X to some Hilbert space Z. Then minimising

$$\Phi_{\alpha,y^\delta}(x) := \frac{1}{2}\|Kx - y^\delta\|_Y^2 + \frac{\alpha}{2}\|Dx\|_Z^2\,,$$

constitutes the so-called Tikhonov-Philipps regularisation. It allows to penalise certain properties of x through a suitable choice of D. In image processing, a typical choice for D is the gradient operator, i.e. the regularisation functional J is chosen to be the square of the H^1-seminorm. This penalises only variations in x, but not the size of x.

Example 2.6.2. (l^1**-Regularisation**) For non-injective operators $K \in \mathcal{L}(l^1, l^2)$ on sequence spaces, a popular choice for J is the l^1-norm which is suitable to enforce sparsity in the regularised solution, i.e.,

$$\Phi_{\alpha,y^\delta}(x) := \frac{1}{2}\|Kx - y^\delta\|_Y^2 + \alpha\sum_{j=1}^{\infty}|x_j|\,.$$

As mentioned above, in variational form, Tikhonov regularisation can also easily be generalised to nonlinear inverse problems

$$F(x) = y\,, \tag{2.6.2}$$

where $F : \mathcal{D}(F) \subset X \to Y$ is a nonlinear bounded operator with domain $\mathcal{D}(F)$ between two Hilbert spaces X and Y.

Definition 2.6.3. Let $x \in \mathcal{D}(F)$. The nonlinear equation (2.6.2) is called **locally ill-posed in** x, if for every $r > 0$ there exists a sequence $(x_n)_{n\in\mathbb{N}} \subset B_r(x) \subset \mathcal{D}(F)$ such that

$$F(x_n) \to F(x)\,, \quad \text{but} \quad x_n \nrightarrow x\,.$$

Otherwise (2.6.2) is called **locally well-posed in** x.

For nonlinear problems, the classical Tikhonov regularisation computes a regularised approximation x_α^δ of x by minimising the functional

$$\Phi_{\alpha,y^\delta}(x) := \frac{1}{2}\|F(x) - y^\delta\|_Y^2 + \frac{\alpha}{2}\|x\|_X^2 \,. \tag{2.6.3}$$

Under certain conditions on the nonlinear operator F it can again be shown that together with an a priori parameter choice rule $\alpha = \alpha(\delta)$ such that

$$\alpha(\delta) \to 0 \quad \text{and} \quad \frac{\delta^2}{\alpha(\delta)} \to 0 \quad \text{as } \delta \to 0\,,$$

(2.6.3) provides a convergent regularisation method for (2.6.2). As in the linear case, the regularisation functional $\|\cdot\|_X^2$ can again be replaced by another functional $J : X \to \mathbb{R}$ that penalises other features in the solution x.

For more details on regularisation for nonlinear inverse problems see, e.g., [Engl, Hanke, Neubauer, *Regularization of Inverse Problems*, Kluwer, 2000] or [Rieder, *Keine Probleme mit Inversen Problemen*, Kluwer, 2003].

Chapter 3

Basic Concepts of Probability Theory

Bertrand's paradox from Exercise sheet 0 shows that one must be careful when introducing the notion of "randomness". In this chapter, among others we formally introduce probability spaces, random variables and most importantly conditional expectations and probabilities. In uncertainty quantification and inverse problems for partial differential equations, we often deal with quantities of interest or random objects belonging to a Sobolev space. For this reason, throughout we concentrate on random variables taking values in separable Banach spaces. For such random variables we will show the existence of so-called "regular conditional distributions", which allows to consider a version of Bayes' theorem in this setting.

Proofs for the stated results on probability theory that are not given in this chapter, can be found in the book "Wahrscheinlichkeitstheorie" by Achim Klenke.

3.1 Measure spaces

3.1.1 σ-algebras

In the following Ω denotes a set, interpreted as the collection of all "elementary events". We write 2^Ω for its power set (the set of all subsets of Ω). A σ-algebra is a specific subset of 2^Ω, on which we will be able to define measures. For $A \subseteq \Omega$ we denote its complement by $A^c := \Omega \backslash A$.

Definition 3.1.1 (σ-algebra)**.** We call $\mathcal{A} \subseteq 2^\Omega$ a **σ-algebra** iff

(i) $\Omega \in \mathcal{A}$,

(ii) $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$,

(iii) $A_i \in A$ for all $i \in \mathbb{N}$ implies $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{A}$.

Definition 3.1.2. For $\Omega \neq \emptyset$ and a σ-algebra $\mathcal{A}$ on Ω, the tuple $(\Omega, \mathcal{A})$ is called a **measurable space**. A subset $A \subseteq \Omega$ is called **measurable** iff it belongs to $\mathcal{A}$.

Remark 3.1.3. Note that (i) and (ii) imply $\Omega^c = \emptyset \in \mathcal{A}$ and (iii) and (ii) imply $(\bigcup_{i\in\mathbb{N}} A_i)^c = \bigcap_{i\in\mathbb{N}} A_i^c \in \mathcal{A}$. In particular $\bigcap_{i\in\mathbb{N}} A_i \in \mathcal{A}$ whenever $A_i \in \mathcal{A}$ for all $i \in \mathbb{N}$.

Recall that $(\Omega, \mathcal{T})$ is called a **topological space**, if $\mathcal{T} \subseteq 2^\Omega$ is a topology on Ω, i.e. $\mathcal{T}$ is the collection of all "open sets" and satisfies that

(i) $\emptyset, \Omega \in \mathcal{T}$

(ii) $\bigcap_{j=1}^{N} O_j \in \mathcal{T}$,

(iii) $\bigcup_{j\in I} O_j \in \mathcal{T}$,

whenever $O_1, \dots, O_N$ and $(O_j)_{j\in I}$ belong to $\mathcal{T}$. Here I is an arbitrary index set and need not be countable. On a topological space we can define the Borel σ-algebra, which is the σ-algebra generated by the open sets. To introduce it, we need the following result:

Proposition 3.1.4. *Let $\mathcal{E} \subseteq 2^\Omega$ be nonempty. Then*

$$\sigma(\mathcal{E}) := \bigcap_{\mathcal{A} \text{ is a } \sigma\text{-algebra s.t. } \mathcal{E}\subseteq\mathcal{A}} \mathcal{A} \tag{3.1.1}$$

defines a σ-algebra called the **σ-algebra generated by $\mathcal{E}$.**

Proof. The intersection in (3.1.1) is not empty since 2^Ω is a σ-algebra containing $\mathcal{E}$. Let $(\mathcal{A}_i)_{i\in I}$ be a family of σ-algebras (I is not necessarily countable). Then it is simple to check that $\bigcap_{i\in I} \mathcal{A}_i$ (by which we mean $\{A \subseteq \Omega : A \in \mathcal{A}_i \ \forall i \in I\}$) is again a σ-algebra (by verifying each item in Def. 3.1.1 for this intersection). Hence (3.1.1) defines a σ-algebra containing $\mathcal{E}$. □

Evidently $\sigma(\mathcal{E})$ is the smallest σ-algebra containing $\mathcal{E}$.

Definition 3.1.5 (Borel σ-algebra)**.** For a topological space $(\Omega, \mathcal{T})$ we call $\sigma(\mathcal{T})$ the Borel σ-algebra and denote it by $\mathcal{B}(\Omega)$.

In case there is no confusion about the topology, we simply say that "$\mathcal{B}$ is the Borel σ-algebra on Ω". In case of Ω being an open or closed subset of $\mathbb{R}^d$, the "Borel σ-algebra on Ω" is always understood w.r.t. the Euclidean topology on $\mathbb{R}^d$.

Remark 3.1.6. There exist sets $A \subseteq \mathbb{R}^d$ which do not belong to $\mathcal{B}(\mathbb{R}^d)$, i.e. $\mathcal{B}(\mathbb{R}^d) \neq 2^{\mathbb{R}^d}$.

Since complements of open sets are closed (and closed sets are in general not open), the Euclidean topology on $\mathbb{R}^d$ is not a σ-algebra. Furthermore:

Exercise 3.1.7. Use Rmk. 3.1.6 to show that $\mathcal{B}(\mathbb{R}^d)$ is not a topology.

3.1.2 Measures

We are now in position to introduce measures. These are functions assigning nonnegative numbers to each set in $\mathcal{A}$:

Definition 3.1.8 (measure)**.** Let $\mathcal{A}$ be a σ-algebra on $\Omega \neq \emptyset$. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a **measure** iff

(i) $\mu(\emptyset) = 0$,

(ii) if $A_i \in \mathcal{A}$ for all $i \in \mathbb{N}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$ then

$$\mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

A measure is called **σ-finite** if there exist $(A_j)_{j\in\mathbb{N}} \in \mathcal{A}$ such that $\Omega = \bigcup_{j\in I} A_j$ and $\mu(A_j) < \infty$ for all $j \in \mathbb{N}$. A measure μ with $\mu(\Omega) = 1$ is called a **probability measure**.

Definition 3.1.9. For a set $\Omega \neq \emptyset$, a σ-algebra $\mathcal{A}$ on Ω and a measure μ on $\mathcal{A}$ we call the triple $(\Omega, \mathcal{A}, \mu)$ a **measure space**. If $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{A})$, we call $(\Omega, \mathcal{A}, \mathbb{P})$ a **probability space**.

Example 3.1.10. Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a finite set and let $0 \leq p_j \leq 1$ for $j = 1, \dots, N$ such that $\sum_{j=1}^N p_j = 1$. Set $\mathcal{A} := 2^\Omega$. Then $\mu(A) := \sum_{\omega_j \in A} p_j$ defines a probability measure on $(\Omega, \mathcal{A})$.

Example 3.1.11. Let $f : \mathbb{R}^n \to \mathbb{R}$ be nonnegative and integrable with $\int_{\mathbb{R}^n} f(x)\,\mathrm{d}x = 1$. Then

$$\mu(A) := \int_A f(x)\,\mathrm{d}x \tag{3.1.2}$$

defines a probability measure on $(\mathbb{R}^n, \mathcal{B})$.

The following theorem is often useful, as it allows to check for equality of two measures:

Theorem 3.1.12. *Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. Let $\mathcal{E} \subseteq 2^\Omega$ satisfy $A \cap B \in \mathcal{E}$ for all $A, B \in \mathcal{E}$ as well as $\sigma(\mathcal{E}) = \mathcal{A}$. If there exists a sequence $(E_n)_{n\in\mathbb{N}}$ with $\Omega = \bigcup_{n\in\mathbb{N}} E_n$, $E_n \subseteq E_{n+1}$ and $\mu(E_n) < \infty$ for all n, then μ is uniquely defined through $\mu(E)$ for all $E \in \mathcal{E}$.*

3.1.3 Product measures

For measurable spaces $(\Omega_j, \mathcal{A}_j)_{j=1}^n$ the σ-algebra

$$\otimes_{j=1}^n \mathcal{A}_j := \sigma(\{\times_{j=1}^n A_j \,:\, A_j \in \mathcal{A}_j \ \forall j\})$$

is called the **product σ-algebra** on the space $\times_{j=1}^n \Omega_j$.

Theorem 3.1.13. *Let $(\Omega_j, \mathcal{A}_j, \mu_j)$ for $j = 1, \dots, n$ be a family of σ-finite measure spaces. Then there exists a unique measure μ on $(\times_{j=1}^n \Omega_j, \otimes_{j=1}^n \mathcal{A}_j)$ such that*

$$\mu(\times_{j=1}^n A_j) = \prod_{j=1}^n \mu_j(A_j) \qquad \forall A_j \in \mathcal{A}_j.$$

We call μ the **product measure** *and use the notation $\mu = \otimes_{j=1}^n \mu_j$.*

The product measure can also be constructed for $n = \infty$: Consider the σ-algebra $\mathcal{A} := \sigma(\mathcal{E})$ generated by the cylindrical sets

$$\mathcal{E} := \{\times_{j\in\mathbb{N}} A_j \,:\, A_j \in \mathcal{B}(\mathbb{R})\}.$$

Suppose that $(\mu_j)_{j\in\mathbb{N}}$ is a family of probability measures on $\mathbb{R}$. Then there is a unique measure μ on $(\mathbb{R}^\mathbb{N}, \mathcal{A})$ satisfying

$$\mu(\times_{j=1}^n A_j \bigtimes \times_{i\in\mathbb{N}} \mathbb{R}) = \prod_{j=1}^n \mu_j(A_j).$$

One of the most important measures is the Lebesgue measure on the measurable space $(\mathbb{R}, \mathcal{B})$, which satisfies

$$\lambda_1((a,b]) = b - a \qquad \forall b > a. \tag{3.1.3}$$

Note that since $\mathcal{E} = \{(a,b] : a < b\}$ generates $\mathcal{B}(\mathbb{R})$, i.e. $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$, Thm. 3.1.12 implies that the Lebesgue measure is unique. Furthermore, by Thm. 3.1.13 there is a unique measure $\lambda_d = \otimes_{j=1}^d \lambda$ on $(\mathbb{R}^d, \otimes_{j=1}^d \mathcal{B}(\mathbb{R}))$ with the property

$$\lambda_d(\times_{j=1}^d (a_j, b_j]) = \prod_{j=1}^d (b_j - a_j) \qquad \forall a_j < b_j,$$

which is again called the Lebesgue measure. Whenever d is clear from the context, we drop the index and simply write λ instead of λ_d. We also mention that $\otimes_{j=1}^d \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^d)$ (exercise).

3.2 Integration in Banach spaces

Let V denote a Banach space over the field $\mathbb{R}$ (most results are easily generalized to Banach spaces over $\mathbb{C}$). In this section, we discuss integrals of the type $\int_\Omega f(\omega)\,\mathrm{d}\mu(\omega)$, where μ is a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$ and f maps from Ω to the Banach space V.

Throughout we adhere to the following notational conventions: The norm of elements $x \in V$ is denoted by $\|x\|_V$ (or simply $\|x\|$ in case there's no confusion about V) and we write $V' := \mathcal{L}(V; \mathbb{R})$ for the topological dual space of V (the space of continuous linear maps from $V \to \mathbb{R}$). For $v' \in V'$, we denote by $\langle v, v' \rangle_V$ the dual pairing (or simply $\langle v, v' \rangle$ if there's no confusion about V). We consider V as a measurable space equipped with the Borel σ-algebra $\mathcal{B}(V)$. For a function $f : \Omega \to V$ and a set $B \subseteq V$ we use the shorthand $f^{-1}(B) := \{\omega \in \Omega : f(\omega) \in B\}$.

3.2.1 Measurability

Definition 3.2.1 (measurability)**.** Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be two measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is called $\mathcal{A}_1/\mathcal{A}_2$**-measurable** iff $f^{-1}(A_2) \in \mathcal{A}_1$ for all $A_2 \in \mathcal{A}_2$. If there's no confusion about $\mathcal{A}_2$ and/or $\mathcal{A}_1$ we also say that f is $\mathcal{A}_1$**-measurable** or simply **measurable**.

Remark 3.2.2. Note that measurability of a function depends only on the σ-algebras, but no measure needs to be defined.

If $\mathcal{A}_1$ and $\mathcal{A}_2$ are both the Borel-σ-algebras, then we say that $f : \Omega_1 \to \Omega_2$ is **Borel measurable**. To check for Borel measurability it suffices to consider preimages of open sets; more generally:

Proposition 3.2.3. *Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be two measurable spaces and assume that $\mathcal{A}_2 = \sigma(\mathcal{E})$ for some $\mathcal{E} \subseteq 2^{\Omega_2}$. A function $f : \Omega_1 \to \Omega_2$ is $\mathcal{A}_1/\mathcal{A}_2$-measurable iff $f^{-1}(E) \in \mathcal{A}_1$ for all $E \in \mathcal{E}$.*

Proof. Measurability implies that $f^{-1}(E) \in \mathcal{A}_1$ for all $E \in \mathcal{E} \subseteq \sigma(\mathcal{E})$.

To show the other direction define

$$\mathcal{C} := \{B \subseteq \Omega_2 : f^{-1}(B) \in \mathcal{A}_1\}.$$

For any $B \in \mathcal{C}$

$$f^{-1}(B^c) = \{\omega \in \Omega_1 : f(\omega) \in B^c\} = \{\omega \in \Omega_1 : f(\omega) \notin B\} = \Omega_1 \backslash f^{-1}(B) = (f^{-1}(B))^c \in \mathcal{A}_1$$

[Draft of October 6, 2021. Not for dissemination.]

and thus $B^c \in \mathcal{C}$. Similarly for all $B_i \in \mathcal{C}$

$$f^{-1}\Big(\bigcup_{i\in\mathbb{N}} B_i\Big) = \bigcup_{i\in\mathbb{N}} f^{-1}\Big(B_i\Big),$$

and thus $\bigcup_{i\in\mathbb{N}} B_i \in \mathcal{C}$ whenever $B_i \in \mathcal{C}$ for all $i \in \mathbb{N}$. Hence $\mathcal{C}$ is a σ-algebra on Ω_2. By assumption every $E \in \mathcal{E}$ belongs to $\mathcal{C}$. Since $\sigma(\mathcal{E})$ is the smallest σ-algebra containing $\mathcal{E}$ it holds $\mathcal{C} \supseteq \sigma(\mathcal{E})$. Thus $f^{-1}(B) \in \mathcal{A}_1$ for all $B \in \sigma(\mathcal{E}) = \mathcal{A}_2$. □

Remark 3.2.4. The previous proposition implies in particular that continuous functions are always Borel-measurable.

To give meaning to integrals over V-valued functions, we require a stronger notion of measurability. A function $f : \Omega \to V$ is called $\mathcal{A}$-**simple** iff

$$f(\omega) = \sum_{j=1}^{N} v_j \mathbb{1}_{A_j}(\omega) \tag{3.2.1}$$

for finite $N \in \mathbb{N}$, measurable $A_j \in \mathcal{A}$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $v_j \in V$. Here $\mathbb{1}_{A_j}(\omega)$ denotes the indicator function, that is $\mathbb{1}_{A_j}(\omega) = 1$ if $\omega \in A_j$ and $\mathbb{1}_{A_j}(\omega) = 0$ otherwise.

Definition 3.2.5 (strong measurability)**.** A function $f : \Omega \to V$ is **strongly measurable** iff there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of $\mathcal{A}$-simple functions such that $\lim_{n\to\infty} f_n = f$ pointwise.

As the name suggests, strong measurability is in general indeed stronger than measurability. In case V is a separable Banach space, the two notions are in fact equivalent. This follows by Pettis measurability theorem, which we show next.

Recall that V is called **separable** if there exists a countable dense subset of V. A function $f : \Omega \to V$ is called **separably valued** if it takes values in a separable subspace $V_0 \subseteq V$. If V is separable, then any $f : \Omega \to V$ is necessarily separably valued. To show Pettis theorem, we'll need the following result:

Proposition 3.2.6. *Let $(\Omega, \mathcal{A})$ be a measurable space and let $f_n : \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be a sequence of $\mathcal{A}$-measurable functions. Then*

- *if $f(\omega) := \sup_{n\in\mathbb{N}} f_n(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, then f is $\mathcal{A}$-measurable,*
- *if $f(\omega) := \inf_{n\in\mathbb{N}} f_n(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, then f is $\mathcal{A}$-measurable,*
- *if $f(\omega) := \lim_{n\to\infty} f_n(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, then f is $\mathcal{A}$-measurable.*

The proof is left as an exercise (Hint: Use that $\mathcal{E} = \{(a, \infty) : a \in \mathbb{R}\}$ generates $\mathcal{B}(\mathbb{R})$ and write $\lim_n f_n = \sup_{n\in\mathbb{N}} \inf_{m\geq n} f_m$).

Theorem 3.2.7 (Pettis measurability theorem, first version)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space. For $f : \Omega \to V$ the following are equivalent:*

(i) f is strongly measurable,

(ii) f is separably valued and $\langle f, v' \rangle$ is $\mathcal{A}$-measurable for every $v' \in V'$.

Proof. (i) $\Rightarrow$ (ii): Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{A}$-simple functions converging pointwise to f, and let V_0 be the closed subspace spanned by the countably many values taken by the functions $(f_n)_{n\in\mathbb{N}}$. Then V_0 is separable and $f:\Omega\to V_0$. Furthermore each $\langle f,v'\rangle:\Omega\to\mathbb{R}$ is $\mathcal{A}$-measurable as the pointwise limit of the $\mathcal{A}$-measurable functions $\langle f_n,v'\rangle$ by Prop. 3.2.6.

(ii) $\Rightarrow$ (i): Let V_0 be a separable subspace of V such that $f:\Omega\to V_0$. First we show that there exists a sequence $(v'_n)_{n\in\mathbb{N}}\subseteq V'$ such that for all $v\in V_0$

$$\|v\|=\sup_{n\in\mathbb{N}}|\langle v,v'_n\rangle|. \tag{3.2.2}$$

To this end let $(v_n)_{n\in\mathbb{N}}$ be dense sequence in V_0. By the Hahn-Banach theorem there exist $v'_n\in V'$ such that $\|v'_n\|=1$ and $\|v_n\|=\langle v_n,v'_n\rangle$. Now, for every $v\in V_0$ and $\varepsilon>0$ there exists $n\in\mathbb{N}$ so large that $\|v-v_n\|<\varepsilon$. Then

$$\langle v,v'_n\rangle\geq\langle v_n,v'_n\rangle-|\langle v_n-v,v'_n\rangle|\geq\|v_n\|-\varepsilon\geq\|v\|-\|v-v_n\|-\varepsilon=\|v\|-2\varepsilon.$$

Also note that for any $n\in\mathbb{N}$ $|\langle v,v_n\rangle|\leq\|v\|\|v_n\|=\|v\|$. Since $\varepsilon>0$ was arbitrary, the claim follows. Now let $v_0\in V_0$. By the $\mathcal{A}$-measurability of $\omega\mapsto\langle f(\omega),v'_n\rangle$, for each $v_0\in V_0$

$$\omega\mapsto\|f(\omega)-v_0\|=\sup_{n\in\mathbb{N}}\langle f(\omega)-v_0,v'_n\rangle\qquad\text{is }\mathcal{A}\text{-measurable.} \tag{3.2.3}$$

Next define $s_n:V_0\to\{v_1,\dots,v_n\}$ as follows: for all $w\in V_0$ let $k(n,w)$ be the smallest integer in $\{1,\dots,n\}$ such that

$$\|w-v_k\|=\min_{1\leq j\leq n}\|w-v_j\|,$$

and set $s_n(w):=v_{k(n,w)}$. By density of $(v_n)_{n\in\mathbb{N}}$ in V_0

$$\lim_{n\to\infty}\|w-s_n(w)\|=0\qquad\forall w\in V_0.$$

Next, set

$$f_n(\omega):=s_n(f(\omega))\qquad\forall\omega\in\Omega.$$

Then for $1\leq k\leq n$

$$\begin{aligned}\{\omega\in\Omega:f_n(\omega)=v_k\}=\{\omega\in\Omega:\|f(\omega)-v_k\|=\min_{1\leq j\leq n}\|f(\omega)-v_j\|\}\\ \cap\{\omega\in\Omega:\|f(\omega)-v_l\|>\min_{1\leq j\leq n}\|f(\omega)-v_j\|\ \forall l=1,\dots,k-1\}.\end{aligned}$$

The set on the right-hand side is in $\mathcal{A}$ due to (3.2.3). Since f_n takes values in $\{v_1,\dots,v_n\}$, we conclude that f_n is $\mathcal{A}$-simple. The proof is finished since for every $\omega\in\Omega$

$$\lim_{n\to\infty}\|f_n(\omega)-f(\omega)\|=\lim_{n\to\infty}\|s_n(f(\omega))-f(\omega)\|=0.$$ □

Corollary 3.2.8. *The pointwise limit of a sequence of strongly $\mathcal{A}$-measurable functions is strongly $\mathcal{A}$-measurable.*

Proof. Let $\lim_{n\to\infty}f_n=f$ pointwise, where each f_n is strongly $\mathcal{A}$-measurable, and thus takes values in a separable subspace $V_n\subseteq V$. The closure V_0 of $\bigcup_{n\in\mathbb{N}}V_n$ is separable, and thus f is separably valued. Moreover, Pettis theorem implies $\langle f_n,v'\rangle:\Omega\to\mathbb{R}$ to be measurable for every n and every $v'\in V'$. Now, $\lim_{n\to\infty}\langle f_n,v'\rangle=\langle f,v'\rangle$ for every $v'\in V'$, and since the limit of $\mathbb{R}$-valued $\mathcal{A}$-measurable functions is $\mathcal{A}$-measurable by Prop. 3.2.6, we conclude that $\langle f,v'\rangle:\Omega\to\mathbb{R}$ is $\mathcal{A}$-measurable for every $v'\in V'$ so that by Pettis theorem f is strongly measurable. □

Corollary 3.2.9. *Let $f : \Omega \to V$ be strongly $\mathcal{A}$-measurable. Let W be another Banach space and let $\phi : V \to W$ be continuous. Then $\phi \circ f : \Omega \to W$ is strongly $\mathcal{A}$-measurable.*

Proof. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of simple functions converging pointwise to f. Then $\phi \circ f_n$ is a sequence of simple functions converging pointwise to $\phi \circ f$. □

Corollary 3.2.10. *If V is separable, then measurability implies strong measurability.*

Proof. Since $f : \Omega \to V$ is $\mathcal{A}$-measurable, $\langle f, v' \rangle : \Omega \to \mathbb{R}$ is $\mathcal{A}$-measurable for all $v' \in V'$. Hence Pettis measurability theorem implies the claim. □

Remark 3.2.11. Cor. 3.2.10 shows that "$\mathcal{A}$-measurability and separably valued" implies "strong $\mathcal{A}$-measurability". In fact the two are equivalent (exercise).

3.2.2 Strong μ-measurability

In this section $(\Omega, \mathcal{A}, \mu)$ is a σ-finite measure space, that is, μ is a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$.

We say that $f : \Omega \to V$ is **μ-simple** if

$$f = \sum_{j=1}^{n} \mathbb{1}_{A_j} v_j, \tag{3.2.4}$$

where $v_j \in V$ and $A_j \in \mathcal{A}$ such that $\mu(A_j) < \infty$.

We say that a property holds **μ-almost everywhere (a.e.)** (or μ-almost surely) if there exists a **μ-null-set** $N \in \mathcal{A}$, that is, $\mu(N) = 0$, and the property holds on $\Omega \backslash N$.

Definition 3.2.12 (strong μ-measurability)**.** A function $f : \Omega \to V$ is **strongly μ-measurable** iff there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of μ-simple functions converging to f μ-a.e.

We call $\tilde{f}$ a **μ-version** of f if $\tilde{f} = f$ μ-a.e. In case there is a μ-version of f that is $\mathcal{A}$-measurable, we say that f is **μ-measurable**.

Proposition 3.2.13. *For $f : \Omega \to V$ the following are equivalent:*

(i) f is strongly μ-measurable,

(ii) f has a μ-version that is strongly $\mathcal{A}$-measurable.

Proof. (i) ⇒ (ii): With $(f_n)_{n\in\mathbb{N}}$ as in Def. 3.2.12 let $N \subseteq \Omega$ be such that $\mu(N) = 0$ and $\lim_{n\to\infty} f_n = f$ pointwise on $\Omega \backslash N$. Then $\mathbb{1}_{N^c} f_n \to \mathbb{1}_{N^c} f$ pointwise on Ω. Since $\mathbb{1}_{N^c} f_n$ are $\mathcal{A}$-simple functions, this shows that $\tilde{f} := \mathbb{1}_{N^c} f$ is strongly $\mathcal{A}$-measurable, and this function coincides with f μ-a.e.

(ii) ⇒ (i): Let $\tilde{f}$ be a strongly $\mathcal{A}$-measurable μ-version of f and let N be a μ-null set such that $f = \tilde{f}$ on N^c. If $(\tilde{f}_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{A}$-simple functions converging pointwise to $\tilde{f}$, then $\lim_{n\to\infty} \tilde{f}_n = f$ on N^c, i.e. $\lim_{n\to\infty} \tilde{f}_n = f$ μ-a.e. Let $\Omega = \bigcup_{n\in\mathbb{N}} A_j$ with $\mu(A_n) < \infty$ for all n. Then $f_n := \mathbb{1}_{A_n} \tilde{f}_n$ are μ-simple functions and $\lim_{n\to\infty} f_n = f$ μ-a.e. □

We say that $f : \Omega \to V$ is **μ-separably valued** iff there exists a closed separable subspace $V_0 \subseteq V$ such that $f(\omega) \in V_0$ for μ-a.e. $\omega \in \Omega$.

Theorem 3.2.14 (Pettis measurability theorem, second version)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. For $f : \Omega \to V$ the following are equivalent:*

(i) f is strongly μ-measurable,

(ii) f is μ-separably valued and $\langle f, v' \rangle$ is μ-measurable for every $v' \in V'$.

Sketch of proof. (i) $\Rightarrow$ (ii): By Prop. 3.2.13 there exists $\tilde{f}$ such that $f = \tilde{f}$ μ-a.e. and $\tilde{f} : \Omega \to V$ is strongly $\mathcal{A}$-measurable. The statement then follows by Thm. 3.2.7.

(ii) $\Rightarrow$ (i): This direction can be shown analogous to Thm. 3.2.7, with the exception that this time the functions f_n are μ-a.e. equal to functions $\tilde{f}_n$ that are $\mathcal{A}$-simple. □

Prop. 3.2.13 and Corollaries 3.2.9 and 3.2.10 imply:

Corollary 3.2.15. *Let $f_n : \Omega \to V$ be a sequence of strongly μ-measurable functions, and let $\lim_{n\to\infty} f_n = f$ μ-a.e. Then f is strongly μ-measurable.*

Corollary 3.2.16. *Let $f : \Omega \to V$ be strongly μ-measurable and let W be another Banach space. If $\phi : V \to W$ is continuous, then $\phi \circ f : \Omega \to V$ is strongly μ-measurable.*

3.2.3 Bochner integrals

Definition 3.2.17. Let μ be a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$. A function $f : \Omega \to V$ is called **μ-Bochner integrable** iff the following two conditions are met:

(i) there exists a sequence of μ-simple functions $f_n = \sum_{j=1}^{n} \mathbb{1}_{A_{n,j}} v_{n,j}$ such that $\lim_{n\to\infty} f_n = f$ μ-a.e.,

(ii) $\lim_{n\to\infty} \int_\Omega \|f(\omega) - f_n(\omega)\| \, \mathrm{d}\mu(\omega) = 0$.

For a μ-Bochner integrable function we define

$$\int_\Omega f(\omega) \, \mathrm{d}\mu(\omega) := \lim_{n\to\infty} \sum_{j=1}^{n} \mu(A_{n,j}) v_{n,j} \in V. \tag{3.2.5}$$

Exercise 3.2.18. Show that (3.2.5) does not depend on the approximating sequence $(f_n)_{n\in\mathbb{N}}$ and is well-defined (i.e. the limit exists in V).

Lemma 3.2.19. *Let $v' \in V'$ and let $f : \Omega \to V$ be Bochner-integrable. Then*

$$\left\langle \int_\Omega f(\omega) \, \mathrm{d}\mu(\omega), v' \right\rangle = \int_\Omega \left\langle f(\omega), v' \right\rangle \, \mathrm{d}\mu(\omega). \tag{3.2.6}$$

Proof. For a μ-simple function $f_n = \sum_{j=1}^{n} \mathbb{1}_{A_j} v_j$ due to the linearity of the dual product

$$\left\langle \int_\Omega f_n(\omega) \, \mathrm{d}\mu(\omega), v' \right\rangle = \left\langle \sum_{j=1}^{n} v_j \mu(A_j), v' \right\rangle = \sum_{j=1}^{n} \left\langle v_j, v' \right\rangle \mu(A_j) = \int_\Omega \left\langle f_n(\omega), v' \right\rangle \, \mathrm{d}\mu(\omega). \tag{3.2.7}$$

Now let $(f_n)_{n\in\mathbb{N}}$ be as in Def. 3.2.17. Taking the limit $n \to \infty$ on both sides of (3.2.7) yields (3.2.6). Here we use that $v' : V \to \mathbb{R}$ is continuous and that $\int_\Omega f_n \, \mathrm{d}\mu \to \int_\Omega f \, \mathrm{d}\mu$ in V by assumption (which shows that the left-hand side of (3.2.7) converges to the left-hand side of (3.2.6)), and (3.2.5) (which shows that the right-hand side of (3.2.7) converges to the right-hand side of (3.2.6)). □

The next theorem is useful to check for Bochner-integrability of a function.

Theorem 3.2.20. *A strongly μ-measurable function $f : \Omega \to V$ is μ-Bochner integrable iff*

$$\int_\Omega \|f(\omega)\| \, \mathrm{d}\mu(\omega) < \infty$$

(in the sense of the Lebesgue integral) and in this case

$$\left\| \int_\Omega f(\omega) \, \mathrm{d}\mu(\omega) \right\| \leq \int_\Omega \|f(\omega)\| \, \mathrm{d}\mu(\omega). \tag{3.2.8}$$

Proof. If f is μ-Bochner integrable, then for the simple functions $(f_n)_{n\in\mathbb{N}}$ as in Def. 3.2.17 it holds

$$\int_\Omega \|f(\omega)\| \, \mathrm{d}\mu(\omega) \leq \int_\Omega \|f(\omega) - f_n(\omega)\| \, \mathrm{d}\mu(\omega) + \int_\Omega \|f_n(\omega)\| \, \mathrm{d}\mu(\omega).$$

Due to assumption (i) of Def. 3.2.17 the first term is finite for n large enough. The second term is finite since each f_n is a μ-simple function.

To show the other implication let f be strongly μ-measurable such that $\int_\Omega \|f(\omega)\| \, \mathrm{d}\mu(\omega) < \infty$ and let g_n be μ-simple functions satisfying $\lim_{n\to\infty} g_n = f$ μ-a.e. Set

$$f_n := g_n \mathbb{1}_{\|g_n\| \leq 2\|f\|}.$$

Then f_n is μ-simple and $\lim_{n\to\infty} f_n = f$ μ-a.e. Since $\|f_n\| \leq 2\|f\|$ pointwise for every n, by the dominated convergence theorem

$$\lim_{n\to\infty} \int_\Omega \|f - f_n\| \, \mathrm{d}\mu = 0.$$

The inequality claimed in the theorem is trivial for μ-simple functions, and follows by approximation in the general case. □

3.2.4 L^p-spaces

Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. For $1 \leq p < \infty$ we define $L^p(\Omega, \mu; V)$ to be the space of all strongly μ-measurable functions $f : \Omega \to V$ for which

$$\|f\|_{L^p(\Omega,\mu;V)} := \left(\int_\Omega \|f(\omega)\|^p \, \mathrm{d}\mu(\omega) \right)^{1/p} < \infty$$

and identifying μ-a.e. equal functions (i.e. elements of $L^p(\Omega, \mu; V)$ are equivalence classes of μ-a.e. equal functions). In case we wish to emphasize the σ-algebra on Ω we write $L^p(\Omega, \mathcal{A}, \mu; V)$ (note that if $\mathcal{F} \subseteq \mathcal{A}$ is a sub-σ-algebra, in general $L^p(\Omega, \mathcal{A}, \mu; V) \neq L^p(\Omega, \mathcal{F}, \mu; V)$). If there's no confusion about μ or $\mathcal{A}$ we simply write $L^p(\Omega; V)$.

Similarly, $L^\infty(\Omega; V)$ consists of all equivalence classes of strongly measurable μ-a.e. equal functions endowed with the norm

$$\|f\|_{L^\infty(\Omega;V)} := \inf \{ r \geq 0 \, : \, \mu(\{\omega \in \Omega \, : \, \|f(\omega)\| \geq r\}) = 0 \}. \tag{3.2.9}$$

Without proof we mention that $L^p(\Omega; V) = L^p(\Omega, \mathcal{A}, \mu; V)$ is a Banach space for all $1 \leq p \leq \infty$. Note that $L^1(\Omega; V)$ consists of all equivalence classes of Bochner-integrable functions.

3.2.5 Radon-Nikodym derivative

Definition 3.2.21. Given measures μ and ν on $(\Omega, \mathcal{A})$, we say that ν is **absolutely continuous** wrt μ ($\nu \ll \mu$) if for all $A \in \mathcal{A}$ s.t. $\mu(A) = 0$, we have $\nu(A) = 0$. The two measures are called **equivalent** iff $\mu \ll \nu$ and $\nu \ll \mu$.

Suppose that μ, ν are two measures on $(\Omega, \mathcal{A})$. In case there exists an $\mathcal{A}$-measurable $f : \Omega \to \mathbb{R}$ such that for all $A \in \mathcal{A}$

$$\nu(A) = \int_\Omega f(\omega) \mathbb{1}_A(\omega) \, \mathrm{d}\mu(\omega),$$

we call f a **density** of ν w.r.t. μ. If ν is σ-finite, such a density is μ-a.e. unique, and as such this function is called the **Radon-Nikodym derivative** of ν w.r.t. μ. We denote it by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} := f$.

Theorem 3.2.22 (Radon-Nikodym)**.** *Let μ, ν be two σ-finite measures on $(\Omega, \mathcal{A})$. Then*

$$\nu \ll \mu \qquad \Leftrightarrow \qquad \textit{the Radon-Nikodym derivative } \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \textit{ exists.}$$

In this case $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is $\mathcal{A}$-measurable and μ-a.e. finite.

3.2.6 Transformation of measures

Let $(\Omega_1, \mathcal{A}_1, \mu)$ be a measure space and $(\Omega_2, \mathcal{A}_2)$ a measurable space. Let $T : \Omega_1 \to \Omega_2$ be measurable. Then

$$T_\sharp \mu(A_2) := \mu(\underbrace{\{\omega \in \Omega_1 \, : \, T(\omega) \in A_2\}}_{=T^{-1}(A_2)}) \qquad \forall A_2 \in \mathcal{A}_2$$

defines a measure on $(\Omega_2, \mathcal{A}_2)$ (exercise).

Definition 3.2.23. We call $T_\sharp \mu$ the **pushforward measure**.

For real valued measurable functions, we have the usual change of variables formula (for the Lebesgue integrals):

Theorem 3.2.24. *Let $T : \Omega_1 \to \Omega_2$ and $f : \Omega_2 \to \mathbb{R}$ be measurable. Then $\int_{\Omega_2} |f(\omega_2)| \, \mathrm{d}T_\sharp \mu(\omega_2) < \infty$ iff $\int_{\Omega_1} |f \circ T| \, \mathrm{d}\mu(\omega_1) < \infty$ and in this case*

$$\int_{\Omega_1} f \circ T(\omega_1) \, \mathrm{d}\mu(\omega_1) = \int_{\Omega_2} f(\omega_2) \, \mathrm{d}T_\sharp \mu(\omega_2).$$

Remark 3.2.25 (Transformation of densities). Assume that $\mu \ll \lambda$ is a measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with density $f := \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$. In the important case that $T : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1-diffeomorphism, we have for all $A \in \mathcal{B}(\mathbb{R}^d)$

$$T_\sharp \mu(A) = \mu(T^{-1}(A)) = \int_{T^{-1}(A)} f(x) \, \mathrm{d}x = \int_A f(T^{-1}(x)) \det dT^{-1}(x) \, \mathrm{d}x.$$

Hence the density transforms under the pushforward as $\frac{\mathrm{d}T_\sharp \mu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \circ T^{-1} \det dT^{-1}$, where $dT^{-1} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ denotes the Jacobian matrix of $T^{-1} : \mathbb{R}^d \to \mathbb{R}^d$.

[Draft of October 6, 2021. Not for dissemination.]

3.3 Random variables

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Terminology 3.3.1. A set $A \in \mathcal{A}$ is called an **event**. $\mathbb{P}[A]$ is the **probability** of the event A.

Often it is not convenient or possible to work with events. Instead we consider observable quantities of such events. This idea is formalized with the notion of random variables.

Definition 3.3.2 (Random variables)**.** Let $(\Omega, \mathcal{A})$ be a measurable space and V a Banach space. Then a measurable function $X : \Omega \to V$ is called a **V-valued random variable** (RV).

Terminology 3.3.3. (i) It is common practice in probability theory to write X instead of $X(\omega)$, i.e. not to explicitly display the dependence of X on $\omega \in \Omega$.

(ii) For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a RV $X : \Omega \to V$ induces a probability measure $\mathbb{P}_X := X_\sharp \mathbb{P}$ on $(V, \mathcal{B}(V))$, i.e. $\mathbb{P}_X[B] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}]$. For $B \in \mathcal{B}(V)$ we usually write $\mathbb{P}[X \in B]$ to denote $\mathbb{P}_X[B]$, which is the probability of the event $\{\omega \in \Omega : X(\omega) \in B\}$, i.e. the probability that X takes a value in B.

Definition 3.3.4 (distribution)**.** (i) The measure $\mathbb{P}_X$ is the **distribution** of X.

(ii) We write $X \sim \mu$ to express that X has distribution μ, i.e. $\mathbb{P}_X = \mu$.

(iii) A family of V-valued RVs $(X_j)_{j \in I}$ is called **equally distributed** if $\mathbb{P}_{X_i} = \mathbb{P}_{X_j}$ for all $i, j \in I$.

(iv) For a finite family of RVs $X_j : \Omega \to V_j$, $j = 1, \dots, n$, the measure $\mathbb{P}_{X_1,\dots,X_n} := (X_1, \dots, X_n)_\sharp \mathbb{P}$ on $(\times_{j=1}^n V_j, \mathcal{B}(\times_{j=1}^n V_j))$ is the **joint distribution** of the RVs $(X_j)_{j=1}^n$, and $\mathbb{P}_{X_j}$ is the **marginal distribution** of X_j.

Definition 3.3.5 (distribution function and density)**.** Suppose $X : \Omega \to \mathbb{R}$ is a real valued RV.

(i) The function
$$F_X(x) := \mathbb{P}_X[X \leq x]$$
is the **distribution function** of X.

(ii) If there exists a nonnegative integrable function $f : \mathbb{R} \to \mathbb{R}$ such that for all $x \in \mathbb{R}$
$$F(x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t,$$
then f is called the **density** function for X. In this case we also write $f = f_X$.

Note that f is simply the Radon-Nikodym derivative of $\mathbb{P}_X$ w.r.t. the Lebesgue measure (the "Lebesgue density"), i.e. $f = \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\lambda}$. For real valued RVs the last two notions are generalized to n RVs $X_j : \Omega \to \mathbb{R}$ as follows:

(i) we call
$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) := \mathbb{P}_{X_1,\dots,X_n}[X_1 \leq x_1, \dots, X_n \leq x_n]$$
the **joint distribution function**,

(ii) if there exists a nonnegative $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1,\dots,t_n)\,\mathrm{d}t_1 \dots \mathrm{d}t_n$$

then f is the **density** function of $X = (X_1,\dots,X_n)$. In this case we also write $f(x) = f_{X_1,\dots,X_n}(x)$.

Example 3.3.6 (Dice roll.)**.** Let $\Omega = \{1,\dots,6\}$ be equipped with the σ-algebra $\mathcal{A} = 2^\Omega$. We interpret each $\omega \in \Omega$ as the outcome of a dice roll, and set

$$X(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is even} \\ 1 & \text{if } \omega \text{ is odd.} \end{cases}$$

Then $X : \{1,\dots,6\} \to \mathbb{R}$ is an $\mathbb{R}$-valued RV. To model a fair dice, we can define a probability measure $\mathbb{P}$ via $\mathbb{P}[\omega] = \frac{1}{6}$ for each $\omega \in \Omega$.

It is easy to check that for a RV $X : \Omega \to V$,

$$\sigma(X) := \{X^{-1}(B) \,:\, B \in \mathcal{B}(V)\}$$

is a σ-algebra, called the **σ-algebra generated by** X. It is the smallest σ-algebra on Ω w.r.t. which X is measurable, and it can be interpreted as containing all relevant information about the RV X.

Example 3.3.7. Consider $X : \{1,\dots,6\} \to \{0,1\}$ from example 3.3.6. Then

$$\sigma(X) = \{\{1,3,5\},\{2,4,6\},\{1,2,3,4,5,6\},\emptyset\}.$$

This σ-algebra contains all relevant information about X, namely whether the dice shows an odd or an even number.

3.4 Expectation and covariance

Let V be a separable Banach space and $(\Omega,\mathcal{A},\mathbb{P})$ a probability space.

Definition 3.4.1. We say that a RV $X : \Omega \to V$ has finite kth moment, iff $\int_\Omega \|X(\omega)\|^k \,\mathrm{d}\mathbb{P}(\omega) < \infty$.

If $X : \Omega \to V$ has finite first moment, then

$$\mathbb{E}[X] := \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$$

is the expectation of X.

For two separable Hilbert spaces $(H_1,\langle\cdot,\cdot\rangle_{H_1})$, $(H_2,\langle\cdot,\cdot\rangle_{H_2})$ and two random variables $X : \Omega \to H_1$, $Y : \Omega \to H_2$ with finite second moments, we define the **covariance operator** $\mathrm{cov}(X,Y) = C : H_2 \to H_1$ by

$$\langle v, Cw\rangle_{H_1} = \int_\Omega \langle X - \mathbb{E}[X], v\rangle_{H_1} \langle Y - \mathbb{E}[Y], w\rangle_{H_2} \,\mathrm{d}\mathbb{P} \qquad \forall v \in H_1,\ w \in H_2.$$

We also set $\operatorname{cov}(X) := \operatorname{cov}(X, X)$. One can show that $\operatorname{cov}(X)$ is a self-adjoint positive trace-class operator.

In case $H = \mathbb{R}$ the **variance** of $X : \Omega \to \mathbb{R}$ is defined as

$$\begin{aligned}\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] &= \int_{\mathbb{R}} x^2 - 2x\mathbb{E}[X] + \mathbb{E}[X]^2 \,\mathrm{d}\mathbb{P}_X(x) \\ &= \int_{\mathbb{R}} x^2 \,\mathrm{d}\mathbb{P} - 2x\mathbb{E}[X] + \mathbb{E}[X]^2 \,\mathrm{d}\mathbb{P}_X(x) \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

Example 3.4.2. Let $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^m$ be two random variables. Then $\operatorname{cov}(X, Y)$ is represented by the **covariance matrix** $C \in \mathbb{R}^{n \times m}$ with entries

$$C_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])].$$

Under linear transformations the covariance matrix satisfies $\operatorname{cov}(AX, BY) = A\operatorname{cov}(X, Y)B^\top$. In case $X = Y$ we have $C_{ii} = \mathbb{V}(X_i)$.

Expectations can be computed using the following change of variables formula:

Theorem 3.4.3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$ two separable Banach spaces. Let $X : \Omega \to V$ be a RV and $\varphi : V \to W$ a measurable function. Then $\varphi(X) : \Omega \to W$ is a RV. It holds $\varphi \in L^1(V, \mathbb{P}_X; W)$ iff $\varphi(X) \in L^1(\Omega, \mathbb{P}; W)$ and in this case*

$$\mathbb{E}[\varphi(X)] = \int_\Omega \varphi(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega) = \int_V \varphi(v) \,\mathrm{d}\mathbb{P}_X(v).$$

Proof. Both $\varphi(X) : \Omega \to W$ and $\varphi : V \to W$ are measurable, and thus strongly measurable since V and W are separable. $\mathbb{P}$ and $\mathbb{P}_X$ are probability measures, and thus $\varphi(X)$ is strongly $\mathbb{P}$-measurable and φ is strongly $\mathbb{P}_X$-measurable. By Thm. 3.2.24

$$\int_\Omega \|\varphi(X(\omega))\|_W \,\mathrm{d}\mathbb{P}(\omega) = \int_V \|\varphi(v)\|_W \,\mathrm{d}\mathbb{P}_X(v)$$

and hence Thm. 3.2.20 implies $\varphi(X) \in L^1(\Omega, \mathbb{P}; W)$ iff $\varphi \in L^1(V, \mathbb{P}_X; W)$. In this case, Lemma 3.2.19 implies for all $w' \in W'$

$$\left\langle \int_\Omega \varphi(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega), w' \right\rangle = \int_\Omega \left\langle \varphi(X(\omega)), w' \right\rangle \,\mathrm{d}\mathbb{P}(\omega) = \int_V \left\langle \varphi(v), w' \right\rangle \,\mathrm{d}\mathbb{P}_X(v) = \left\langle \int_V \varphi(v) \,\mathrm{d}\mathbb{P}_X(v), w' \right\rangle,$$

where we used again Thm. 3.2.24 for the real-valued measurable function $\omega \mapsto \langle \varphi(X(\omega)), w' \rangle$ (and the fact that the Lebesgue and Bochner integrals coincide for the integral of real-valued measurable functions w.r.t. σ-finite measures). Since this equality holds for all $w' \in W'$, we conclude

$$\int_\Omega \varphi(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega) = \int_V \varphi(v) \,\mathrm{d}\mathbb{P}_X(v). \qquad \square$$

Remark 3.4.4. With $\varphi(v) = v$

$$\mathbb{E}[X] = \int_V v \,\mathrm{d}\mathbb{P}_X(v).$$

3.5 Independence and conditionals

3.5.1 Conditional probability and independence

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let A, $B \in \mathcal{A}$ be two events such that $\mathbb{P}[B] > 0$. For $\omega \in \Omega$, assuming that we already know $\omega \in B$, we want to define the probability that $\omega \in A$—**the probability of A given B**. Since we know $\omega \in B$, we can interpret B together with the σ-algebra $\{C \in \mathcal{A} : C \subseteq B\}$ and the probability measure $\tilde{\mathbb{P}} := \frac{\mathbb{P}}{\mathbb{P}[B]}$ as a new probability space. Then the probability of ω belonging to A equals $\tilde{\mathbb{P}}[A \cap B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$.

Definition 3.5.1 (conditional probability I)**.** The **conditional probability** of A given B is

$$\mathbb{P}[A|B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

If the knowledge of B has no influence on the probability of A, i.e. $\mathbb{P}[A|B] = \mathbb{P}[A]$, we say that the **events are independent**. If $P(B) > 0$, this is equivalent to $\mathbb{P}[A]\mathbb{P}[B] = \mathbb{P}[A \cap B]$. The latter condition is symmetric in A and B, as it should be.

Definition 3.5.2 (independent events)**.** Two events A and B are called **independent** iff

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

Exercise 3.5.3. Show that if A and B are independent, then A^c and B are also independent.

Next we generalize the notion of independence to σ-algebras and RVs.

Definition 3.5.4 (independent σ-algebras)**.** Let $\mathcal{A}_i \subseteq \mathcal{A}$ be σ-algebras on Ω for all $i \in I$. The $(\mathcal{A}_i)_{i \in I}$ are **independent** if for all finite subsets $\{k_1, \dots, k_n\} \subseteq I$ and all events $A_i \in \mathcal{A}_{k_i}$ holds

$$\mathbb{P}[A_1 \cap \cdots \cap A_n] = \mathbb{P}[A_1] \dots \mathbb{P}[A_n].$$

Definition 3.5.5. Let $X_i : \Omega \to V$ for $i \in I$ be a family of RVs for a Banach space V. We say that the X_i are independent if for all finite subsets $\{k_1, \dots, k_n\} \subseteq I$ the σ-algebras $(\sigma(X_{k_i}))_{i=1}^n$ are independent or equivalently for all $B_1, \dots, B_n \in \mathcal{B}(V)$

$$\mathbb{P}[X_{k_1} \in B_1, \dots, X_{k_n} \in B_n] = \mathbb{P}[X_1 \in B_1] \cdots \mathbb{P}[X_{k_n} \in B_n].$$

Exercise 3.5.6. Consider the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$. Define for $\omega \in [0,1]$

$$X_n(\omega) := \begin{cases} 1 & \text{if } \frac{k}{2^n} \le \omega \le \frac{k+1}{2^n},\ k \text{ even} \\ -1 & \text{if } \frac{k}{2^n} \le \omega \le \frac{k+1}{2^n},\ k \text{ odd.} \end{cases}$$

Show that the $(X_n)_{n \in \mathbb{N}}$ are a family of independent random variables.

Exercise 3.5.7 (Bayes' formula)**.** Let $A_1, \dots, A_n$ be disjoint events of positive probability such that $\Omega = \bigcup_{j=1}^n A_j$. Let B be another event with $\mathbb{P}[B] > 0$. Show that for $k \in \{1, \dots, n\}$

$$\mathbb{P}[A_k|B] = \frac{\mathbb{P}[B|A_k]\mathbb{P}[A_k]}{\sum_{j=1}^n \mathbb{P}[B|A_j]\mathbb{P}[A_j]}.$$

Proposition 3.5.8. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space and V a Banach space. Let $X_i : \Omega \to V$ be RVs for $i = 1, \dots, n$. Then the X_i are independent if and only if $\mathbb{P}_{X_1,\dots,X_n} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.*

Proof. Assume that the X_j are independent. Then for all $A_j \in \mathcal{A}$

$$\begin{aligned}\mathbb{P}_{X_1,\dots,X_n}[A_1 \times \cdots \times A_n] &= \mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n] \\ &= \mathbb{P}[X_1 \in A_1] \cdots \mathbb{P}[X_n \in A_n] \\ &= \mathbb{P}_{X_1}(A_1) \cdots \mathbb{P}_{X_n}[A_n].\end{aligned}$$

By Thm. 3.1.13 it holds $\mathbb{P}_{X_1,\dots,X_n} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.

Conversely by definition of the product measure, $\mathbb{P}_{X_1,\dots,X_n} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ implies for all $A_j \in \mathcal{A}$ that $\mathbb{P}_{X_1,\dots,X_n}[A_1 \times \cdots \times A_n] = \mathbb{P}_{X_1}[A_1] \cdots \mathbb{P}_{X_n}[A_n]$. □

For real valued RVs, independence is equivalent to saying that the distribution functions and densities factor.

Theorem 3.5.9. *Let $X_i : \Omega \to \mathbb{R}^m$ be n RVs for $i = 1, \dots, n$.*

(i) The RVs are independent iff for $x = (x_1, \dots, x_n)$

$$F_{X_1,\dots,X_n}(x) = F_{X_1}(x_1) \dots F_{X_n}(x_n).$$

(ii) If the RVs have densities, then they are independent iff

$$f_{X_1,\dots,X_n}(x) = f_{X_1}(x_1) \dots f_{X_n}(x_n).$$

Sketch of Proof. If the X_j are independent, then $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}[X_1 \leq x_1, \dots, X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1] \cdots \mathbb{P}[X_n \leq x_n] = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.

Conversely, let $A_i = X_i^{-1}(B_i)$ for $B_i \in \mathcal{B}(\mathbb{R}^m)$. Then

$$\begin{aligned}\mathbb{P}[A_1 \cap \cdots \cap A_n] &= \mathbb{P}[X_1 \in B_1, \dots, X_n \in B_n] \\ &= \int_{B_1 \times \cdots \times B_n} f_{X_1,\dots,X_n}(x_1, \dots, x_n) \, \mathrm{d}x_1 \dots \, \mathrm{d}x_n \\ &= \prod_{j=1}^{n} \int_{B_j} f_{X_j}(x_j) \, \mathrm{d}x_j \\ &= \prod_{j=1}^{n} \mathbb{P}[X_j \in B_j] = \prod_{j=1}^{n} \mathbb{P}[A_j].\end{aligned}$$

□

Theorem 3.5.10. *Let $X_1, \dots, X_n : \Omega \to \mathbb{R}$ be independent RVs and such that and $\mathbb{E}[|X_i|] < \infty$ for all $i = 1, \dots, n$. Then*

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n] < \infty.$$

Proof. By Thm. 3.4.3 with $\varphi : \mathbb{R}^n \to \mathbb{R} : (x_1, \dots, x_n) \mapsto x_1 \cdots x_n$ (and Fubini's theorem)

$$\begin{aligned}\mathbb{E}[X_1 \cdots X_n] &= \int_\Omega X_1(\omega) \cdots X_n(\omega) \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{\mathbb{R}^n} x_1 \cdots x_n \, \mathrm{d}\mathbb{P}_{X_1,\dots,X_n}(x_1, \dots, x_n) \\ &= \int_{\mathbb{R}} x_1 \, \mathrm{d}\mathbb{P}_{X_1}(x_1) \cdots \int_{\mathbb{R}} x_n \, \mathrm{d}\mathbb{P}_{X_n}(x_n).\end{aligned}$$

□

3.5.2 Conditional expectations

Let X be a random variable on $(\Omega, \mathcal{A})$ and let $B \in \mathcal{A}$. In the previous section we defined $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$ in case $\mathbb{P}[B] > 0$. Given an event B with $\mathbb{P}[B] > 0$, due to $\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$, it is natural to introduce the expectation of X given B as

$$\mathbb{E}[X|B] := \int_B X(\omega)\,\mathrm{d}\mathbb{P}[\omega|B] = \frac{1}{\mathbb{P}[B]} \int_\Omega \mathbb{1}_B(\omega) X(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

Now let X and Y be two random variables. In this section we want to answer the question: How can we define the expectation of X given Y? Since Y is a random variable, this conditional expectation should also be a random variable.

To motivate the following discussion let us start with a simple example. Assume that $X : [0,1] \to \mathbb{R}$ and $Y : [0,1] \to \mathbb{R}$ are two RVs. Additionally let $\bigcup_{j=1}^n A_j = [0,1]$ be a partition of $[0,1]$ and suppose that $Y(\omega) = \sum_{j=1}^n \mathbb{1}_{A_j}(\omega) y_j$ is a simple function with $y_i \neq y_j \in \mathbb{R}$ for all $i \neq j$. Now, if $Y(\omega) = y_j$, then we know $\omega \in A_j$. Hence the expectation for X is the average of X over A_j, i.e.

$$\mathbb{E}[X|A_j] = \frac{1}{\mathbb{P}[A_j]} \int_{A_j} X\,\mathrm{d}\mathbb{P}.$$

We thus set

$$\mathbb{E}[X|Y](\omega) := \frac{1}{\mathbb{P}[A_j]} \int_{A_j} X\,\mathrm{d}\mathbb{P} \qquad \text{if } \omega \in A_j.$$

We make the following observations:

(i) $\mathbb{E}[X|Y] : [0,1] \to \mathbb{R}$ is a random variable that is constant on each A_j.

(ii) The actual values y_j taken by Y are irrelevant for the definition of $\mathbb{E}[X|Y]$, we merely require the sets A_j, or in other words the σ-algebra $\sigma(Y) = \{\emptyset, [0,1]\} \cup \{\bigcup_{i \in I} A_i \,:\, I \subseteq \{1, \dots, n\}\}$ generated by Y.

(iii) $\mathbb{E}[X|Y] : [0,1] \to \mathbb{R}$ is $\sigma(Y)$-measurable.

(iv) $\int \mathbb{1}_A X\,\mathrm{d}\mathbb{P} = \int \mathbb{1}_A \mathbb{E}[X|Y]\,\mathrm{d}\mathbb{P}$ for all $A \in \sigma(Y)$.

The second item motivates us to first introduce expectations of X conditioned on σ-algebras.

Definition 3.5.11 (conditional expectation I)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra, V a separable Banach space and $X : \Omega \to V$ a random variable such that $X \in L^1(\Omega, \mu; V)$. A random variable $Z : \Omega \to V$ is called a **conditional expectation of X given $\mathcal{F}$**, iff

(i) $Z : \Omega \to V$ is $\mathcal{F}$-measurable,

(ii) $\int_\Omega \mathbb{1}_B Z\,\mathrm{d}\mathbb{P} = \int_\Omega \mathbb{1}_B X\,\mathrm{d}\mathbb{P} \in V$ for all $B \in \mathcal{F}$.

In this case we write $\mathbb{E}[X|\mathcal{F}] = Z$.

We next show existence and uniqueness of $\mathbb{E}[X|\mathcal{F}]$ and start with the case $V = \mathbb{R}$.

Theorem 3.5.12. *Let $V = \mathbb{R}$. Then $\mathbb{E}[X|\mathcal{F}]$ exists and is $\mathbb{P}$-a.e. unique.*

Proof. Uniqueness: Assume Z and Z' both satisfy the conditions of Def. 3.5.11. Let $A = \{\omega \in \Omega : Z(\omega) > Z'(\omega)\} \in \mathcal{F}$. Then

$$\int_\Omega \mathbb{1}_A(\omega)(Z(\omega) - Z'(\omega))\,\mathrm{d}\mathbb{P}(\omega) = 0$$

and since $Z - Z' > 0$ on A we have $\mathbb{P}[A] = 0$. Similarly, with $B = \{\omega \in \Omega : Z(\omega) < Z'(\omega)\}$ we get $\mathbb{P}[B] = 0$ and thus $Z = Z'$ $\mathbb{P}$-a.e.

Existence: Set $X^+ := \max\{0, X\}$ and $X^- := -\min\{0, X\}$. For $* \in \{+, -\}$ define

$$\mu^*(A) := \mathbb{E}[X^* \mathbb{1}_A] \qquad \forall A \in \mathcal{F}.$$

Then $\mu^\pm$ are two σ-finite measures on $(\Omega, \mathcal{F})$. By construction $\mu^\pm \ll \mathbb{P}$ and there exist $\mathcal{F}$-measurable Radon-Nikodym derivatives $Z^\pm : \Omega \to \mathbb{R}$ such that

$$\mu^\pm(A) = \int_A Z^\pm \,\mathrm{d}\mathbb{P} \qquad \forall A \in \mathcal{F}.$$

Then $Z := Z^+ - Z^-$ is $\mathcal{F}$-measurable (the difference of $\mathcal{F}$-measurable $\mathbb{R}$-valued functions is again $\mathcal{F}$-measurable), and for all $A \in \mathcal{F}$

$$\int_\Omega \mathbb{1}_A(\omega) Z \,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \mathbb{1}_A(\omega) Z^+ \,\mathrm{d}\mathbb{P}(\omega) - \int_\Omega \mathbb{1}_A(\omega) Z^- \,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \mathbb{1}_A(\omega) X(\omega)\,\mathrm{d}\mathbb{P}(\omega). \qquad \square$$

Remark 3.5.13. The spaces $L^2(\Omega, \mathcal{A}, \mathbb{P}; \mathbb{R})$ and $L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ are Hilbert spaces with the $L^2(\Omega, \mathbb{P})$-inner product. It can be shown that for $X \in L^2(\Omega, \mathcal{A}, \mathbb{P}; \mathbb{R})$, $\mathbb{E}[X|\mathcal{F}]$ is the orthogonal projection onto the closed subspace $L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$, that is for any $\mathcal{F}$-measurable $Z : \Omega \to \mathbb{R}$

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])^2] \le \mathbb{E}[(X - Z)^2]$$

with equality iff $Z = \mathbb{E}[X|\mathcal{F}]$ $\mathbb{P}$-a.e.

Exercise 3.5.14. For $V = \mathbb{R}$ show that

(i) $\mathbb{E}[\mathbb{E}[X|\mathcal{F}]] = \mathbb{E}[X]$,

(ii) $\mathbb{E}[X] = \mathbb{E}[X|\mathcal{F}]$ in case $\mathcal{F} = \{\emptyset, \Omega\}$.

Some further properties of the conditional probability are the following:

Theorem 3.5.15. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, X and Y two real-valued RVs in $L^1(\Omega, \mathcal{A}, \mathbb{P}; \mathbb{R})$, and $\mathcal{G} \subseteq \mathcal{F} \subseteq \mathcal{A}$ sub-σ-algebras. Then*

(i) (linearity) for $\alpha \in \mathbb{R}$, $\mathbb{E}[\alpha X + Y|\mathcal{F}] = \alpha\mathbb{E}[X|\mathcal{F}] + \mathbb{E}[Y|\mathcal{F}]$,

(ii) (montonicity) if $X \ge Y$ $\mathbb{P}$-a.e., then $\mathbb{E}[X|\mathcal{F}] \ge \mathbb{E}[Y|\mathcal{F}]$ $\mathbb{P}$-a.e.,

(iii) (tower property) $\mathbb{E}[\mathbb{E}[X|\mathcal{F}]|\mathcal{G}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{F}] = \mathbb{E}[X|\mathcal{G}]$,

(iv) (triangle inequality) $\mathbb{E}[|X||\mathcal{F}] \ge |\mathbb{E}[X|\mathcal{F}]|$,

(v) (independence) if $\sigma(X)$ and $\mathcal{F}$ are independent, then $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X]$,

(vi) (Lebesgue dominated convergence) if $Y \geq 0$ and $(X_n)_{n\in\mathbb{N}}$ is a sequence of RVs with $|X_n| \leq Y$ for all $n \in \mathbb{N}$ and $X_n \to X$ $\mathbb{P}$-a.e., then

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{F}] = \mathbb{E}[X|\mathcal{F}] \qquad \mathbb{P}-a.e. \text{ and in the sense of } L^1(\Omega,\mathcal{F},\mathbb{P};\mathbb{R}).$$

Sketch of proof. (i) For $\alpha \in \mathbb{R}$, and $X, Y \in L^1(\Omega,\mathbb{P};\mathbb{R})$ the function

$$\mathbb{E}[X|\mathcal{F}] + \alpha\mathbb{E}[Y|\mathcal{F}]$$

is $\mathcal{F}$-measurable and satisfies for every $A \in \mathcal{F}$

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A(\mathbb{E}[X|\mathcal{F}] + \alpha\mathbb{E}[Y|\mathcal{F}])] &= \mathbb{E}[\mathbb{1}_A\mathbb{E}[X|\mathcal{F}]] + \alpha\mathbb{E}[\mathbb{1}_A\mathbb{E}[Y|\mathcal{F}]] \\ &= \mathbb{E}[\mathbb{1}_A X] + \alpha\mathbb{E}[\mathbb{1}_A Y] = \mathbb{E}[\mathbb{1}_A(X+\alpha Y)].\end{aligned}$$

(ii) Let $A = \{\mathbb{E}[X|\mathcal{F}] < \mathbb{E}[Y|\mathcal{F}]\} \in \mathcal{F}$. Due to $X \geq Y$ it holds $\mathbb{E}[\mathbb{1}_A(X-Y)] \geq 0$, and thus $\mathbb{P}[A] = 0$.

(iv) Set $X^+ = \max\{0, X\}$ and $X^- = -\min\{0, X\}$ so that $X = X^+ - X^-$. By (i) and (ii)

$$\mathbb{E}[|X||\mathcal{F}] = \mathbb{E}[X^+|\mathcal{F}] + \mathbb{E}[X^-|\mathcal{F}] \geq \mathbb{E}[-X^+|\mathcal{F}] + \mathbb{E}[X^-|\mathcal{F}] = -\mathbb{E}[X|\mathcal{F}] \qquad \mathbb{P}\text{-a.e.}$$

and similarly $\mathbb{E}[|X||\mathcal{F}] \geq \mathbb{E}[X|\mathcal{F}]$ $\mathbb{P}$-a.e.

□

Lemma 3.5.16. *In the setting of Thm. 3.5.12 denote $T(X) = \mathbb{E}[X|\mathcal{F}]$. Then $T : L^1(\Omega,\mathcal{A},\mathbb{P};\mathbb{R}) \to L^1(\Omega,\mathcal{F},\mathbb{P};\mathbb{R})$ is linear and $\|T\|_{\mathcal{L}(L^1;L^1)} \leq 1$.*

Proof. According to Thm. 3.5.15 (i), T is linear. The bound on the norm of the operator follows by Thm. 3.5.15 (iv) and Exercise 3.5.14:

$$\begin{aligned}\|\mathbb{E}[X|\mathcal{F}]\|_{L^1} &= \mathbb{E}[|\mathbb{E}[X|\mathcal{F}]|] \leq \mathbb{E}[\mathbb{E}[|X||\mathcal{F}]] \\ &= \mathbb{E}[|X|] = \||X|\|_{L^1}.\end{aligned}$$

□

Finally we mention without proof that also a version of Jensen's inequality is satisfied:

Theorem 3.5.17. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and let X be a real-valued random variable on $(\Omega,\mathcal{A},\mathbb{P})$. If $\mathbb{E}[|X|] < \infty$ and $\mathcal{F} \subseteq \mathcal{A}$ is a sub-σ-algebra, then*

$$\varphi(\mathbb{E}[X|\mathcal{F}]) \leq \mathbb{E}[\varphi(X)|\mathcal{F}] \qquad \mathbb{P}-a.e.$$

Theorem 3.5.18. *Let V be a separable Banach space. Then $\mathbb{E}[X|\mathcal{F}]$ exists and is $\mathbb{P}$-a.e. unique.*

Proof. For $\mathcal{A}$-simple functions $Y : \Omega \to V$, $Y = \sum_{j=1}^n \mathbb{1}_{A_j} v_j$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$, define $\tilde{T}(Y)$ via

$$T(Y)(\omega) = \sum_{j=1}^n \mathbb{E}[\mathbb{1}_{A_j}|\mathcal{F}](\omega)v_j = \sum_{j=1}^n T(\mathbb{1}_{A_j})(\omega)v_j,$$

with T from Lemma 3.5.16. Then $\tilde{T}$ is a linear operator on the vector space of V-valued $\mathcal{A}$-simple functions, and we want to show that it can be extended to a bounded operator on all of $L^1(\Omega,\mathbb{P};V)$.

Using linearity of T and the fact that $T(\mathbb{1}_{A_j}) = \mathbb{E}[\mathbb{1}_{A_j}|\mathcal{F}]$ takes nonnegative values $\mathbb{P}$-a.e. (why?),

$$\begin{aligned}
\|\tilde{T}(Y)\|_{L^1(\Omega,\mathcal{F},\mathbb{P};V)} &= \int_\Omega \left\| \sum_{j=1}^n T(\mathbb{1}_{A_j})(\omega) v_j \right\| \mathrm{d}\mathbb{P}(\omega) \\
&\leq \int_\Omega \sum_{j=1}^n |T(\mathbb{1}_{A_j})|(\omega) \|v_j\| \,\mathrm{d}\mathbb{P}(\omega) \\
&= \int_\Omega \left| T\left(\sum_{j=1}^n \mathbb{1}_{A_j} \|v_j\| \right)(\omega) \right| \mathrm{d}\mathbb{P}(\omega) \\
&\leq \|T\|_{\mathcal{L}(L^1;L^1)} \left\| \sum_{j=1}^n \mathbb{1}_{A_j} \|v_j\| \right\|_{L^1(\Omega,\mathbb{P};\mathbb{R})} \\
&= \|T\|_{\mathcal{L}(L^1;L^1)} \|Y\|_{L^1(\Omega,\mathcal{A},\mathbb{P};V)}.
\end{aligned}$$

By density of the $\mathcal{B}(V)$-simple functions in $L^1(\Omega,\mathbb{P};V)$, we conclude that $\tilde{T}$ can be extended to a bounded linear operator $\tilde{T} : L^1(\Omega,\mathcal{A},\mathbb{P};V) \to L^1(\Omega,\mathcal{F},\mathbb{P};V)$ and $\|\tilde{T}\|_{\mathcal{L}(L^1;L^1)} \leq \|T\|_{\mathcal{L}(L^1;L^1)} = 1$.

Now we show that $\tilde{T}(X) = \mathbb{E}[X|\mathcal{F}]$ in the sense of Def. 3.5.11. By definition $\tilde{T}(X)$ is $\mathcal{F}$-measurable. Moreover for $A \in \mathcal{F}$ and $\mathcal{A}$-simple random variables $X : \Omega \to V$ one checks that $\mathbb{E}[\mathbb{1}_A \tilde{T}(X)] = \mathbb{E}[\mathbb{1}_A X]$. By density the equality holds for all $X \in L^1(\Omega,\mathcal{A},\mathbb{P};V)$, and therefore $\tilde{T}(X)$ is a conditional expectation.

Finally we show that $\mathbb{E}[X|\mathcal{F}]$ is $\mathbb{P}$-a.e. unique. Assume that Z and Z' are two conditional expectations. For arbitrary $\varphi \in V'$, $\langle Z,\varphi\rangle$ and $\langle Z',\varphi\rangle$ are (strongly) $\mathcal{F}$-measurable (by Cor. 3.2.9) and satisfy $\mathbb{E}[\mathbb{1}_A \langle Z,\varphi\rangle] = \mathbb{E}[\mathbb{1}_A \langle Z',\varphi\rangle] = \mathbb{E}[\mathbb{1}_A \langle X,\varphi\rangle]$ for all $A \in \mathcal{F}$ (see Lemma 3.2.19). This shows that $\langle Z,\varphi\rangle$ and $\langle Z',\varphi\rangle$ are both conditional expectations of $\langle X,\varphi\rangle$, and by Thm. 3.5.12 there exists a $\mathbb{P}$-null set $N \subseteq \Omega$ such that $\langle Z,\varphi\rangle = \langle Z',\varphi\rangle$ on N^c. Since V is separable, (as shown earlier) there exists a sequence $\varphi_n \in V'$ with $\|\varphi_n\|_{V'} = 1$, and such that $\|v\| = \sup_{n\in\mathbb{N}} \langle v,\varphi_n\rangle$ for all $v \in V$. Let N_n be a $\mathbb{P}$-null set such that $\langle Z, v_n'\rangle = \langle Z', v_n'\rangle$ on N_n^c. Then $N := \bigcup_{n\in\mathbb{N}} N_n$ is a null set and $Z = Z'$ on N^c. □

Now we can introduce the conditional expectation of X given Y.

Definition 3.5.19 (conditional expectation II)**.** Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, V, W two separable Banach spaces and $X : \Omega \to V$, $Y : \Omega \to W$ two random variables such that $X \in L^1(\Omega,\mu;V)$. Then $\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$ is the **conditional expectation of X given Y**.

Example 3.5.20. Let $X : \Omega \to \{0,1\}$ be as in Example 3.3.6, i.e. X is 0 if the dice shows an even number and X is 1 if the dice shows an odd number. Let $Y : \Omega \to \{0,1\}$ with $Y(\omega) = 0$ for $\omega \in \{1,2,3\}$ and $Y(\omega) = 1$ for $\omega \in \{4,5,6\}$. Then

$$\mathbb{E}[X|Y](\omega) = \begin{cases} 1/3 & \omega \in \{1,2,3\} \\ 2/3 & \omega \in \{4,5,6\}. \end{cases}$$

3.5.3 Regular conditional distribution

So far we have defined $\mathbb{P}[A|B] = \frac{\mathbb{P}[A\cap B]}{\mathbb{P}[B]}$ in case $\mathbb{P}[B] > 0$. The goal of this section is to define the conditional probability $\mathbb{P}[A|X = x]$ even if $\mathbb{P}[X = x] = 0$.

Example 3.5.21. Let p be a uniformly distributed RV on $[0,1]$ and let X be a Bernoulli RV, i.e. X takes the value 1 with probability p and the value 0 with probability $1-p$. What is $\mathbb{P}[X=1|p=1/2]$? Our previous definition of conditional probabilities doesn't lead to a meaningful result here since $[p=1/2]$ is an event of probability 0.

Definition 3.5.22 (regular conditional distribution I)**.** Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, and $X:\Omega\to V$ and $Y:\Omega\to W$ two random variables for two separable Banach spaces V and W. A map $\tau_{X|Y}:W\times\mathcal{B}(V)\to[0,1]$ satisfying

(i) $y\mapsto\tau_{X|Y}(y,B)$ is $\mathcal{B}(W)/\mathcal{B}(\mathbb{R})$-measurable for every $B\in\mathcal{B}(V)$,

(ii) $B\mapsto\tau_{X|Y}(y,B)$ is a probability measure on $(V,\mathcal{B}(V))$ for every $y\in\{Y(\omega):\omega\in\Omega\}$,

(iii) $\mathbb{P}[X\in B,Y\in A]=\int_A\tau_{X|Y}(y,B)\,\mathrm{d}\mathbb{P}_Y(y)$ for all $A\in\mathcal{B}(W)$ and all $B\in\mathcal{B}(V)$,

is called a **regular (version of the) conditional distribution of X given Y**. In this case we denote

$$\mathbb{P}[X\in B|Y=y]:=\tau_{X|Y}(y,B).$$

Assuming for the moment that there exists $\tau_{X|Y}$ as in the above definition, we have found a meaningful way to define the probability distribution of X given that $Y=y$, namely the measure $B\mapsto\mathbb{P}[X\in B|Y=y]$. This is well-defined even if $[Y=y]$ is a (nonempty) $\mathbb{P}$-null set. In this sense, $\mathbb{P}[X\in\cdot|Y=y]$ can be interpreted as a well behaved conditional probability.

It remains to show that the conditional distribution exists and is unique (in a suitable sense), to which the rest of this section is dedicated. We emphasize that the existence of regular conditional distributions is not trivial, and indeed not always satisfied. However, in the present setting, where V and W are separable Banach space, existence does hold. In fact, the assumptions that V and W are separable Banach spaces could be significantly weakened, in particular it would suffice for W equipped with some σ-algebra to be a measurable space. Such generalizations are beyond the scope of these lecture notes.

Uniqueness

Uniqueness of the regular conditional distribution holds in the following $\mathbb{P}_Y$-a.e. sense:

Lemma 3.5.23 (uniqueness of the regular conditional distribution)**.** *Assume that τ and $\tilde\tau$ are two functions satisfying the conditions of Def. 3.5.22. Then there exists a $\mathbb{P}_Y$-null set $N\in\mathcal{B}(W)$ such that for all $y\in N^c\cap\{Y(\omega):\omega\in\Omega\}$ holds $\tau(y,\cdot)=\tilde\tau(y,\cdot)$, i.e. these probability measures coincide on $(V,\mathcal{B}(V))$ for all $y\in N^c\cap\{Y(\omega):\omega\in\Omega\}$.*

Proof. Fix $B\in\mathcal{B}(V)$ and let $A_n:=\{y\in W:\tau(y,B)-\tilde\tau(y,B)>\frac{1}{n}\}$. Then $A_n\in\mathcal{B}(W)$. Due to

$$\int_{A_n}\tau(y,B)\,\mathrm{d}\mathbb{P}_Y(y)=\mathbb{P}[X\in B,Y\in A_n]=\int_{A_n}\tilde\tau(y,B)\,\mathrm{d}\mathbb{P}_Y(y),$$

we find $0=\int_{A_n}(\tau(y,B)-\tilde\tau(y,B))\,\mathrm{d}\mathbb{P}_Y(y)\geq\frac{1}{n}\mathbb{P}_Y[A_n]$. Hence $\{y\in Y:\tau(y,B)>\tilde\tau(y,B)\}=\bigcup_{n\in\mathbb{N}}A_n$ is a $\mathbb{P}_Y$-null set, and by symmetry we conclude that $A_B:=\{y\in W:\tau(y,B)\neq\tau(y,B)\}$ is a $\mathbb{P}_Y$-null set.

Now fix a dense sequence $(x_n)_{n\in\mathbb{N}} \subseteq V$, such that with the open balls $B_r(x) := \{v \in V : \|v - x\|_V < r\}$,

$$\tilde{\mathcal{C}} := \{B_{1/n}(x_m) : n, m \in \mathbb{N}\} = \{\tilde{C}_j : j \in \mathbb{N}\}$$

is a countable basis of the topology of $\mathcal{B}(V)$ (i.e. $(\tilde{C}_j)_{j\in\mathbb{N}}$ is some fixed enumeration of the countable set $\tilde{\mathcal{C}}$). Then

$$\mathcal{C} := \{\cap_{i\in I}\tilde{C}_i : I \subseteq \mathbb{N},\ |I| < \infty\} = \{C_j : j \in \mathbb{N}\}$$

is a countable set of open sets (why is $\mathcal{C}$ countable?). Since $\tilde{\mathcal{C}}$ is a basis of the topology on V, it holds $\sigma(\tilde{\mathcal{C}}) = \mathcal{B}(V)$, and in particular $\sigma(\mathcal{C}) = \mathcal{B}(V)$. Furthermore, $\mathcal{C}$ has the property that for any $C_i, C_j \in \mathcal{C}$ also $C_i \cap C_j \in \mathcal{C}$ by definition of $\mathcal{C}$. Now choose for every $i \in \mathbb{N}$ a $\mathbb{P}_Y$-null set $N_i \in \mathcal{B}(W)$ such that $\tau(y, C_i) = \tilde{\tau}(y, C_i)$ for all $y \in W\backslash N_i$. Then $N := \bigcup_{i\in\mathbb{N}} N_i$ is a $\mathbb{P}_Y$-null set and for all $y \in W\backslash N$ and all $i \in \mathbb{N}$ holds $\tau(y, C_i) = \tilde{\tau}(y, C_i)$. Thm. 3.1.12 implies $\tau(y, B) = \tilde{\tau}(y, B)$ for all $y \in \{Y(\omega) : \omega \in \Omega\}\backslash N$ and all $B \in \mathcal{B}(W)$. □

Remark 3.5.24. Due to τ only being unique in the above sense, we speak of regular *versions* of the conditional distribution. Often we will drop this term, and simply say that τ is a regular conditional distribution, with the understanding that such a map is only unique $\mathbb{P}_Y$-a.e.

Existence

Intuitively we expect $\mathbb{P}[X = 1|p = 1/2]$ in Example 3.5.21 to be $1/2$. To make this precise, we now turn to conditional probabilities given a σ-algebra. For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and an event $A \in \mathcal{A}$ it holds $\mathbb{P}[A] = \mathbb{E}[\mathbb{1}_A]$. This motivates:

Definition 3.5.25 (conditional probability II)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra. Then for all $A \in \mathcal{A}$

$$\mathbb{P}[A|\mathcal{F}] := \mathbb{E}[\mathbb{1}_A|\mathcal{F}]$$

is the **conditional probability of A given $\mathcal{F}$**, and if $Y : \Omega \to W$ is a RV

$$\mathbb{P}[A|Y] := \mathbb{E}[\mathbb{1}_A|\sigma(Y)]$$

is the **conditional probability of A given Y**.

Note that $\mathbb{P}[A|Y]$ is again a RV, that is $A \mapsto \mathbb{P}[A|Y]$ is a mapping from events to RVs. Furthermore, one can show that this mapping is σ-additive in the sense $\mathbb{P}[\bigcup_{j\in\mathbb{N}} A_j|Y] = \sum_{j\in\mathbb{N}} \mathbb{P}[A_j|Y]$ $\mathbb{P}$-a.e. for pairwise disjoint $A_j \in \mathcal{A}$. Next we define the second variant of regular conditional distributions, where we condition on $\omega \in \Omega$.

Definition 3.5.26 (regular conditional distribution II)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to V$ a random variable for a Banach space V. Let $\mathcal{F} \subseteq \mathcal{A}$ be a sub-σ-algebra.

A map $\kappa_{X|\mathcal{F}} : \Omega \times \mathcal{B}(V) \to [0, \infty]$ satisfying

(i) $\omega \mapsto \kappa_{X|\mathcal{F}}(\omega, B)$ is $\mathcal{F}$-measurable for each $B \in \mathcal{B}(V)$,

(ii) $B \mapsto \kappa_{X|\mathcal{F}}(\omega, B)$ is a probability measure on $(V, \mathcal{B}(V))$ for each $\omega \in \Omega$,

(iii) for every $B \in \mathcal{B}(V)$ holds $\kappa_{X|\mathcal{F}}(\omega, B) = \mathbb{P}[X \in B|\mathcal{F}](\omega)$ $\mathbb{P}$-a.e., or equivalently

$$\mathbb{P}[A\cap[X \in B]] = \int_\Omega \mathbb{1}_B(X(\omega))\mathbb{1}_A(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \kappa_{X|\mathcal{F}}(\omega, B)\mathbb{1}_A(\omega)\,\mathrm{d}\mathbb{P}(\omega) \qquad \forall A \in \mathcal{F},\ B \in \mathcal{B}(V),$$

is called a **regular (version of the) conditional distribution of X given $\mathcal{F}$.**

The above notions could be introduced for more general spaces V (not necessarily separable Banach spaces). However, as mentioned before, V being a separable Banach space is sufficient to prove existence of regular conditional distributions. The next theorem shows existence in the case $(V, \mathcal{B}(V)) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We will require the following proposition, which we state without proof. A set $\mathcal{D} \subseteq 2^\Omega$ is called a Dynkin-system, iff

- $\Omega \in \mathcal{D}$,
- for A, $B \in \mathcal{D}$ with $A \subset B$ it holds $B \backslash A \in \mathcal{D}$,
- for every countable disjoint pairwise sequence $A_j \in \mathcal{D}$, $j \in \mathbb{N}$, it holds $\bigcup_{j\in\mathbb{N}} A_j \in \mathcal{D}$.

Proposition 3.5.27. *Let $\mathcal{C} \subseteq 2^\Omega$ satisfy that $A \cap B \in \mathcal{C}$ for every A, $B \in \mathcal{C}$. Then the smallest Dynkin-system containing $\mathcal{C}$ exists and is equal to $\sigma(\mathcal{C})$.*

Theorem 3.5.28. *Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a real valued random variable and $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra. Then there exists a regular version $\kappa_{X|\mathcal{F}} : \Omega \times \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ of the conditional distribution of X given $\mathcal{F}$.*

Proof. The proof proceeds as follows: We construct a measurable version of the distribution function of the conditional distribution by first defining it on the countable set of rational numbers, and then extending it to the real numbers. Throughout this proof we write κ instead of $\kappa_{X|\mathcal{F}}$.

Step 1. We construct a function $\tilde{F} : \Omega \times \mathbb{R} \to [0, 1]$ such that $q \mapsto \tilde{F}(\omega, q)$ is the distribution function of the measure $\kappa(\omega, \cdot)$. To this end, for every $q \in \mathbb{Q}$ let $\omega \mapsto F(\cdot, q) : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a fixed version of the conditional probability

$$\mathbb{P}[X \in (-\infty, q]|\mathcal{F}] = \mathbb{E}[\mathbb{1}_{X\in(-\infty,q]}|\mathcal{F}] : \Omega \to \mathbb{R}$$

(remember that the conditional probability is only unique $\mathbb{P}$-a.e.). For any $q \leq r \in \mathbb{Q}$ it holds $\mathbb{1}_{X\in(-\infty,q]} \leq \mathbb{1}_{X\in(-\infty,r]}$ and by the monotonicity of the conditional expectation (Thm. 3.5.15 (ii)) there is a null set $A_{q,r} \in \mathcal{F}$ such that

$$F(\omega, q) \leq F(\omega, r) \qquad \forall \omega \in \Omega \backslash A_{q,r}.$$

By Lebesgue dominated convergence (cp. Thm. 3.5.15 (v)), there are null sets $B_q \in \mathcal{F}$ for every $q \in \mathbb{Q}$ such that

$$\lim_{n\to\infty} F\left(\omega, q + \frac{1}{n}\right) = \lim_{n\to\infty} \mathbb{E}[\mathbb{1}_{X\in(-\infty,q+\frac{1}{n}]}|\mathcal{F}](\omega) = \mathbb{E}[\mathbb{1}_{X\in(-\infty,q]}|\mathcal{F}](\omega) = F(\omega, q) \qquad \forall \omega \in \Omega \backslash B_q,$$

and by the same argument there exists a null set $C \in \mathcal{F}$ such that

$$\begin{aligned} \inf_{n\in\mathbb{N}} F(\omega, -n) &= \lim_{n\to\infty} F(\omega, -n) = \mathbb{E}[0|\mathcal{F}](\omega) = 0 \\ \sup_{n\in\mathbb{N}} F(\omega, n) &= \lim_{n\to\infty} F(\omega, n) = \mathbb{E}[1|\mathcal{F}](\omega) = 1 \end{aligned} \qquad \forall \omega \in \Omega \backslash C.$$

Now set $N := \bigcup_{q,r\in\mathbb{Q}} A_{q,r} \cup \bigcup_{q\in\mathbb{Q}} B_q \cup C$. Then $N \in \mathcal{F}$ and $\mathbb{P}[N] = 0$. Define

$$\tilde{F}(\omega, z) := \inf\{F(\omega, q) \,:\, z < q \in \mathbb{Q}\} \qquad z \in \mathbb{R},\ \omega \in \Omega \backslash N.$$

Then $z \mapsto \tilde{F}(\omega, z)$ is monotonically increasing, right-continuous and satisfies $\lim_{z\to\infty} F(\omega, z) = 1$ and $\lim_{z\to\infty} F(\omega, -z) = 0$. As such it is a distribution function, i.e. $\mu_\omega((a, b]) := F(\omega, b) - F(\omega, a)$ defines a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For $\omega \in N$ set $F(\omega, z) := F_0(z)$ where F_0 is an arbitrary fixed probability distribution function, and again $\mu_\omega((a, b]) := F_0(b) - F_0(a)$ defines a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Step 2. We define κ and show that it possesses the properties (i) and (ii) of Def. 3.5.26. For $B \in \mathcal{B}(\mathbb{R})$ set

$$\kappa(\omega, B) := \mu_\omega(B).$$

By construction, for each $\omega \in \Omega$, $\kappa(\omega, \cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

It remains to show that for each $B \in \mathcal{B}(\mathbb{R})$ the map $\omega \mapsto \kappa(\omega, B)$ is $\mathcal{F}$-measurable. First let $q \in \mathbb{Q}$ and set $B := (-\infty, q]$. Then

$$\kappa(\omega, B) = F(\omega, q)\mathbb{1}_{N^c}(\omega) + F_0(q)\mathbb{1}_N(\omega).$$

Since $N \in \mathcal{F}$ and $\omega \mapsto F(\omega, q)$ is $\mathcal{F}$-measurable by construction, $\omega \mapsto \kappa(\omega, B)$ is $\mathcal{F}$-measurable. Next, note that with

$$\mathcal{C} := \{(-\infty, q] \,:\, q \in \mathbb{Q}\} \tag{3.5.1}$$

it holds $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ (i.e. $\mathcal{C}$ generates the Borel-σ-algebra). We claim that

$$\mathcal{D} := \{B \in \mathcal{B}(\mathbb{R}) \,:\, \omega \mapsto \kappa(\omega, B) \text{ is } \mathcal{F}\text{-measurable}\}$$

is a σ-algebra. In this case $\mathcal{D} \supseteq \sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$, which then shows that $\omega \mapsto \kappa(\omega, B)$ is $\mathcal{F}$-measurable for all $B \in \mathcal{F}$.

To show the claim we first point out that $\mathcal{D}$ is a Dynkin-system:

- $\mathbb{R} \in \mathcal{D}$ since $\omega \mapsto \kappa(\omega, \mathbb{R}) = \mu_\omega(\mathbb{R}) = 1$ is trivially $\mathcal{F}$-measurable,
- for A, $B \in \mathcal{D}$ with $A \subseteq B$ it holds $A \backslash B \in \mathcal{D}$ due to

  $$\kappa(\omega, A\backslash B) = \kappa(\omega, A) - \kappa(\omega, B), \tag{3.5.2}$$

  i.e. $\omega \mapsto \kappa(\omega, A\backslash B)$ is $\mathcal{F}$-measurable since it is the sum of two $\mathcal{F}$-measurable functions (we have used that $A \mapsto \kappa(\omega, A)$ is a probability measure in (3.5.2)),
- for disjoint sets $(A_j)_{j\in\mathbb{N}}$ in $\mathcal{D}$ we have $\bigcup_{j\in\mathbb{N}} A_j \in \mathcal{D}$ since

  $$\kappa\left(\omega, \bigcup_{j\in\mathbb{N}} A_j\right) = \sum_{j\in\mathbb{N}} \kappa(\omega, A_j),$$

  and this sum converges pointwise for every $\omega \in \Omega$ to a number in $[0, 1]$ since $\kappa(\omega, \cdot)$ is a probability measure. Thus $\omega \mapsto \kappa(\omega, \bigcup_{j\in\mathbb{N}} A_j) \in \mathbb{R}$ is $\mathcal{F}$-measurable as the pointwise limit of $\mathcal{F}$-measurable functions (cp. Prop. 3.2.6).

By Prop. 3.5.27 (and because $\mathcal{C}$ satisfies A, $B \in \mathcal{C} \Rightarrow A \cap B \in \mathcal{C}$) we conclude $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C}) \subseteq \mathcal{D}$.

Step 3. Finally we verify that κ satisfies (iii) of Def. 3.5.26 and thus is a regular conditional distribution of $\mathbb{P}[X|\mathcal{F}]$.

By definition of κ, for every $A \in \mathcal{F}$, $q \in \mathbb{Q}$ and $B = (-\infty, q]$

$$\begin{aligned}\int_\Omega \mathbb{1}_A(\omega)\kappa(\omega, B)\,\mathrm{d}\mathbb{P}(\omega) &= \int_\Omega \mathbb{1}_A(\omega)\mathbb{P}[X \in B|\mathcal{F}](\omega)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \int_\Omega \mathbb{1}_A(\omega)\mathbb{E}[\mathbb{1}_{X\in B}|\mathcal{F}](\omega)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \int_\Omega \mathbb{1}_{A\cap\{X\in B\}}(\omega)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \mathbb{P}[A \cap \{X \in B\}]. \end{aligned} \tag{3.5.3}$$

Since $\mathcal{C}$ in (3.5.1) generates $\mathcal{B}(\mathbb{R})$, and because of Thm. 3.1.12 the left and the right-hand side of (3.5.3) coincide in the sense of finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Thus they are equal for all $B \in \mathcal{B}(\mathbb{R})$.

Now fix $B \in \mathcal{B}(\mathbb{R})$ and assume that there exists $A \in \mathcal{F}$ with $\mathbb{P}[A] > 0$ and such that $\kappa(\omega, B) \neq \mathbb{P}[X \in B|\mathcal{F}]$ for all $\omega \in A$. Without loss of generality we can assume that $\kappa(\omega, B) - \mathbb{P}[X \in B|\mathcal{F}] > \varepsilon$ for some $\varepsilon > 0$. But then $\int_\Omega \mathbb{1}_A(\omega)\kappa(\omega, B)\,\mathrm{d}\mathbb{P}(\omega) - \int_\Omega \mathbb{1}_A(\omega)\mathbb{P}[X \in B|\mathcal{F}](\omega)\,\mathrm{d}\mathbb{P}(\omega) \geq \varepsilon\mathbb{P}[A] \neq 0$. Thus such A cannot exist and we conclude that $\kappa(\cdot, B) = \mathbb{P}[X \in B|\mathcal{F}]$ $\mathbb{P}$-a.e. for every $B \in \mathcal{B}(\mathbb{R})$. □

To obtain a version of the above theorem for separabel Banach spaces V, we need the following notion:

Definition 3.5.29. Two measurable spaces $(\Omega, \mathcal{A})$ and $(\tilde{\Omega}, \mathcal{A})$ are **isomorphic** if there exists a bijection $\varphi : \Omega \to \tilde{\Omega}$ such that φ is $\mathcal{A}/\tilde{\mathcal{A}}$-measurable and φ^{-1} is $\tilde{\mathcal{A}}/\mathcal{A}$-measurable. We call $(\Omega, \mathcal{A})$ a **Borel space** if there exists $B \in \mathcal{B}(\mathbb{R})$ such that $(B, \mathcal{B}(B))$ and $(\Omega, \mathcal{A})$ are isomorphic.

We state without proof:

Theorem 3.5.30. *Let V be a separable Banach space. Then $(V, \mathcal{B}(V))$ is a Borel space.*

Exercise 3.5.31. Let $d \in \mathbb{N}$. Show that $([-1, 1]^d, \mathcal{B}([-1, 1]^d)$ is a Borel space. Hint: Use binary representations.

Corollary 3.5.32. *Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to V$ be a RV, V a separable Banach space and $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra. Then there exists a regular version $\kappa_{X|\mathcal{F}} : \Omega \times \mathcal{B}(V) \to \mathbb{R}$ of the conditional distribution of X given $\mathcal{F}$.*

Proof. Let $A \in \mathcal{B}(\mathbb{R})$ and $\varphi : V \to A$ an isomorphism as in Def. 3.5.29, which exists by Thm. 3.5.30. Then $\tilde{X} := \varphi \circ X : \Omega \to \mathbb{R}$ is a real-valued RV, and by Thm. 3.5.28 there exists a regular version $\kappa_{\tilde{X}|\mathcal{F}}$ of the conditional distribution of $\tilde{X}$ given $\mathcal{F}$. Set $\kappa_{X|\mathcal{F}}(\omega, B) := \kappa_{\tilde{X}|\mathcal{F}}(\omega, \varphi(B))$ for all $B \in \mathcal{B}(V)$. Then $\kappa_{X|\mathcal{F}}$ is a regular version of the conditional distribution of X given $\mathcal{F}$. □

Finally, rather than conditioning on $\omega \in \Omega$, we wish to condition on $Y = y$ (i.e. on the event $[Y = y] \subseteq \Omega$). In order to so we need the Doob-Dynkin lemma:

Lemma 3.5.33 (Doob-Dynkin Lemma)**.** *Let Ω be a set and $(\tilde{\Omega}, \tilde{\mathcal{A}})$ a measurable space. Consider the following situation:*

$$\begin{array}{ccc} (\Omega, \sigma(Y)) & \xrightarrow{\quad Y \quad} & (\tilde{\Omega}, \tilde{\mathcal{A}}) \\ & {\scriptstyle\kappa}\searrow \quad \swarrow{\scriptstyle\tau} & \\ & (\mathbb{R}, \mathcal{B}(\mathbb{R})) & \end{array}$$

Then $\kappa : \Omega \to \mathbb{R}$ is $\sigma(Y)/\mathcal{B}(\mathbb{R})$-measurable iff there exists $\tau : \tilde{\Omega} \to \mathbb{R}$ which is $\tilde{\mathcal{A}}/\mathcal{B}(\mathbb{R})$-measurable such that $\kappa = \tau \circ Y$.

Proof. "⇐": If such τ exists then $\tau \circ Y = \kappa$ is measurable as a composition of measurable functions.
The other direction is left as an exercise. □

Exercise 3.5.34. Prove (the other direction of) the Doob-Dynkin Lemma. Proceed as follows: First assume $\kappa : \Omega \to [0, \infty)$ (i.e. κ is nonnegative) and κ is $\sigma(Y)/\mathcal{B}(\mathbb{R})$-measurable.

- Define $\kappa_n := \min\{n, 2^{-n}\lfloor 2^n \kappa \rfloor\}$ and show that κ_n is a simple function and $\kappa_n \nearrow \kappa$.
- Use the κ_n to construct sets $A_j \in \mathcal{A}$ and numbers $\alpha_j \geq 0$ such that $\kappa = \sum_{j \in \mathbb{N}} \alpha_j \mathbb{1}_{A_j}$.
- By definition of $\sigma(Y)$ there exist sets $B_n \in \tilde{\mathcal{A}}$ such that $Y^{-1}(B_n) = A_n$. Use the B_n to define τ such that $\tau \circ Y = \kappa$.
- Conclude that τ also exists under the assumption that $\kappa : \Omega \to \mathbb{R}$ is $\sigma(Y)/\mathcal{B}(\mathbb{R})$-measurable (i.e. is not necessarily nonnegative).

Now we can put everything together:

Theorem 3.5.35. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $X : \Omega \to V$, $Y : \Omega \to W$ two RVs for two separable Banach spaces V and W. Then there exists a regular version of the conditional distribution $\mathbb{P}[X \in \cdot | Y = y]$ (in the sense of Def. 3.5.22). It is unique in the sense of Lemma 3.5.23.*

Proof. By Cor. 3.5.32 there exists a regular version $\kappa_{X|\sigma(Y)}(\omega, B)$ of the conditional distribution. By the Doob-Dynkin Lemma (with $(\tilde{\Omega}, \tilde{\mathcal{A}}) = (W, \mathcal{B}(W))$), for every $B \in \mathcal{B}(V)$, there exists $\tau(\cdot, B) : W \to \mathbb{R}$ such that

$$\kappa(\omega, B) = \tau(Y(\omega), B) \qquad \forall \omega \in \Omega,\ \forall B \in \mathcal{B}(V).$$

Then τ satisfies

(i) $y \mapsto \tau(y, B)$ is $\mathcal{B}(W)/\mathcal{B}(\mathbb{R})$-measurable for every $B \in \mathcal{B}(V)$ by definition of τ (i.e. by Lemma 3.5.33),

(ii) $B \mapsto \tau(y, B)$ is a probability measure on $(V, \mathcal{B}(V))$ for every $y \in \{Y(\omega) : \omega \in \Omega\}$, since this is true for $B \mapsto \kappa(\omega, B) = \tau(Y(\omega), B) = \tau(y, B)$ and $\omega \in Y^{-1}(y)$,

(iii) for any $B \in \mathcal{B}(V)$ and any $A \in \mathcal{B}(W)$ by Thm. 3.2.24 and Def. 3.5.26 (iii) (since here "$\mathcal{F} = \sigma(Y)$" and $[Y \in A] \in \sigma(Y)$)

$$\begin{aligned}
\int_V \mathbb{1}_A(y)\tau(y, B)\, \mathrm{d}\mathbb{P}_Y &= \int_\Omega \mathbb{1}_A(Y(\omega))\tau(Y(\omega), B)\, \mathrm{d}\mathbb{P}(\omega) \\
&= \int_\Omega \mathbb{1}_A(Y(\omega))\kappa(\omega, B)\, \mathrm{d}\mathbb{P}(\omega) \\
&= \int_\Omega \mathbb{1}_A(Y(\omega))\mathbb{1}_B(X(\omega))\, \mathrm{d}\mathbb{P}(\omega) \\
&= \mathbb{P}[X \in B, Y \in A].
\end{aligned}$$

Hence we have shown the existence of a regular version of the conditional distribution as in Def. 3.5.22. □

Conditional densities

Suppose that $X : \Omega \to \mathbb{R}^m$ and $Y : \Omega \to \mathbb{R}^n$ are two RVs on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Assume that $(X, Y) : \Omega \to \mathbb{R}^{m+n}$ has the joint (measurable) density $f_{X,Y} : \mathbb{R}^{m+n} \to [0, \infty)$.

Then for any $A \in \mathcal{B}(\mathbb{R}^n)$, by Fubini's theorem

$$\mathbb{P}[Y \in A] = \mathbb{P}[X \in \mathbb{R}^n, Y \in A] = \int_{\mathbb{R}^n \times A} f_{X,Y}(x, y) \, \mathrm{d}(x, y) = \int_A \int_{\mathbb{R}^n} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Hence the marginal $Y : \Omega \to \mathbb{R}^n$ has a density, which is given by

$$f_Y(y) := \int_{\mathbb{R}} f_{X,Y}(x, y) \, \mathrm{d}x.$$

We also say f_Y is the **marginal density** of Y. We point out that we use here the fact that $y \mapsto \int_{\mathbb{R}^n} f_{X,Y}(x, y) \, \mathrm{d}x$ is measurable, which is also a consequence of Fubini's theorem. Next, let us consider a regular version of the conditional distribution $\mathbb{P}[X \in \cdot | Y = y]$ of X given $Y = y$. It turns out that in the present setting (a version of the) measure $\mathbb{P}[X \in \cdot | Y = y]$ has a density, which we call the **conditional density**, and denote by

$$f_{X|Y}(\cdot|y) := \frac{\mathrm{d}\mathbb{P}[X \in \cdot | Y = y]}{\mathrm{d}\lambda_m}.$$

Proposition 3.5.36. *It holds that*

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) \in (0, \infty) \\ f_0(x) & \text{if } f_Y(y) \in \{0, \infty\} \end{cases} \tag{3.5.4}$$

for some fixed probability density f_0 on $\mathbb{R}^n$, is a density of (a version of) $\mathbb{P}[X \in \cdot | Y = y]$.

Remark 3.5.37. The set $\{y \in \mathbb{R}^n : f_Y(y) = 0\}$ is a $\mathbb{P}_Y$-null set, and also $\{y \in \mathbb{R}^n : f_Y(y) = \infty\}$ is a λ-null set (and thus a $\mathbb{P}_Y$-null set) since otherwise $\int_{\mathbb{R}^2} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = \int_{\mathbb{R}} f_Y(y) \, \mathrm{d}y$ would not be finite. By Lemma 3.5.23, the conditional distribution is only unique $\mathbb{P}_Y$-a.e., hence in (3.5.4) it doesn't matter how we define $f_{X|Y}(x, y)$ for y with $f_Y(y) \in \{0, \infty\}$. Hence (3.5.4) is in agreement with our definition of conditional probabilities. In practice, $f_{X|Y}(\cdot|y)$ is only defined for y with $f_Y(y) > 0$.

Proof of Prop. 3.5.36. By Thm. 3.5.35 the measure $\mathbb{P}[X \in \cdot | Y = y]$ on $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ exists.

For $y \in \mathbb{R}^n$ and $B \in \mathcal{B}(\mathbb{R}^m)$ set

$$\tau(y, B) := \int_B f_{X|Y}(x, y) \, \mathrm{d}x.$$

Then

(i) $y \mapsto \tau(y, B)$ is measurable for every $B \in \mathcal{B}(\mathbb{R})$ (this is a consequence of Fubini's theorem),

(ii) $B \mapsto \tau(y, B)$ is a probability measure for every $y \in \mathbb{R}^n$ since $\int_{\mathbb{R}^m} f_{X|Y}(x, y) \, \mathrm{d}x = 1$ for every $y \in \mathbb{R}^n$,

(iii) for every $B \in \mathcal{B}(\mathbb{R}^m)$ and every $A \in \mathcal{B}(\mathbb{R}^n)$, with the $\mathbb{P}_Y$-null set $N := \{y : f_Y(y) \in \{0, \infty\}\}$,

$$\begin{aligned}\int_A \tau(y, B)\, \mathrm{d}\mathbb{P}_Y(y) &= \int_{A\setminus N} \int_B f_Y(y) f_{X|Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{A\setminus N} \int_B f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{A\times B} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y + \int_N \int_B f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \mathbb{P}[X \in A, Y \in B],\end{aligned}$$

where we used $\int_N \int_B f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \le \int_N f_Y(y)\, \mathrm{d}y = \mathbb{P}_Y[N] = 0$. □

3.6 Some common distributions

3.6.1 Bernoulli

Given a parameter $0 \le p \le 1$, the Bernoulli RV $X \sim \mathrm{Ber}(p)$ is defined such that $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$. This RV can be thought of as representing a coin flip with probability of heads equal to p. It is a special case of the binomial distribution with $n = 1$.

3.6.2 Binomial

For $0 \le p \le 1$ and $n \in \mathbb{N}$, we define the binomial RV $X \sim \mathrm{Bin}(n, p)$ with probability mass function

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0 \ldots n.$$

This mass function can be thought of as the probability of k heads in n independent trials of a Bernoulli RV.

3.6.3 Uniform

For $a < b$, the uniform RV $X \sim \mathrm{uniform}(a, b)$ has probability density function

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

A uniform distribution assigns the same probability mass to all sub-intervals of the same length within its support.

3.6.4 Exponential

For $\lambda > 0$, the exponential RV $X \sim \mathrm{Exp}(\lambda)$ has probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

The exponential RV is memoryless: for $0 \le s < t$,

$$\mathbb{P}[X > s + t \mid X > s] = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}[X > t].$$

3.6.5 Univariate Gaussian

For $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$, the Gaussian RV $X \sim \mathcal{N}(\mu, \sigma^2)$ can be defined by the probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

3.6.6 Multivariate Gaussian

First, let $X = (X_1, \ldots, X_n)$ be a vector of real-valued RVs. We say that $X_1, \ldots, X_n$ are "jointly normal" iff $a^\top X$ is Gaussian for every $a \in \mathbb{R}^n$. Equivalently, we say that X has a multivariate Gaussian distribution, $X \sim N(\mu, \Sigma)$. $\mu \in \mathbb{R}^n$ is the mean of X and $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance of X. If Σ is positive definite, then the probability density of X is

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Jointly normal RVs $X_1, \ldots, X_n$ are independent iff they are uncorrelated. All marginal and conditional distributions of the multivariate Gaussian are (multivariate) Gaussian.

Note that if Σ is not positive definite (in which case it will be positive semi-definite), X can still be multivariate Gaussian. In this case, it is customary to describe X through its *characteristic function*. For any RV X, the characteristic function ϕ_X is:

$$\phi_X(\lambda) = \mathbb{E}[e^{\mathrm{i}\,\lambda^\top X}], \quad \lambda \in \mathbb{R}^n.$$

It is thus a function from the real numbers to the complex numbers; it always exists and completely characterizes the distribution. For a multivariate Gaussian, $X \sim N(\mu, \Sigma)$, we have $\phi_X(\lambda) = e^{\mathrm{i}\,\lambda^\top \mu} e^{-\lambda^\top \Sigma \lambda/2}$. See a more advanced probability text (e.g., Grimmett & Stirzaker) for more information about characteristic functions and how they are useful.

3.6.7 Chi-squared

A chi-squared distributed RV with k degrees of freedom, $X \sim \chi^2(k)$, is the distribution of a sum of the squares of k independent standard normal RVs. Its probability density function (pdf) is

$$f_X(x) = \frac{1}{2^{k/2}\Gamma\left(k/2\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}},$$

where $\Gamma(\cdot)$ is the gamma function.

A chi-squared RV $X \sim \chi^2(k)$ has mean k and variance $2k$. Also note that the sum of chi-squared distributed RVs is also chi-squared distributed. Specifically, if $\{X_i\}_{i=1}^n$ are independent chi-squared variables with $\{k_i\}_{i=1}^n$ degrees of freedom, respectively, then the RV $Y = \sum_{i=1}^n X_i$ is chi-squared distributed with $\sum_{i=1}^n k_i$ degrees of freedom.

3.7 Distances and divergences

Here we consider how to quantify the "difference" between probability measures. Some of these measures of "difference" are distance functions in the proper mathematical sense. Others do not satisfy the triangle inequality and are thus only so-called divergences.

Let $(\Omega, \mathcal{A})$ be a measurable space. The first distance we consider is the total variation distance:

Definition 3.7.1. The **total variation distance** between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{A})$ is defined as:

$$D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{A}} |\mathbb{P}[A] - \mathbb{Q}[A]|.$$

This is the largest possible difference between the probabilities that the two distributions can assign to the same event.

Exercise 3.7.2. If $\mathbb{P} \ll \mu$ and $\mathbb{Q} \ll \mu$, show that $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}(\int_\Omega |\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}| \, \mathrm{d}\mu)$.

Another common distance is the Hellinger distance.

Definition 3.7.3. Consider two probability measures $\mathbb{P}$ and $\mathbb{Q}$ that are absolutely continuous with respect to a third measure μ (such a measure always exists, for example $\frac{1}{2}(\mathbb{P}+\mathbb{Q})$). The **Hellinger distance** between $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = \left(\frac{1}{2} \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}} \right)^2 d\mu \right)^{\frac{1}{2}}. \tag{3.7.1}$$

If $\mu \ll \nu$, then

$$D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = \left(\frac{1}{2} \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}} \right)^2 \frac{\mathrm{d}\mu}{\mathrm{d}\nu} d\nu \right)^{\frac{1}{2}} = \left(\frac{1}{2} \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}} \right)^2 d\nu \right)^{\frac{1}{2}}, \tag{3.7.2}$$

and thus (3.7.1) does not depend on which measure μ was chosen. In particular, if $\Omega = \mathbb{R}^d$ and $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to the Lebesgue measure λ_d, then Hellinger distance in (3.7.2) can be expressed through the probability densities $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda_d}$ and $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\lambda_d}$.

Remark 3.7.4. In case $\mathbb{P} \ll \mathbb{Q}$, $D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = (\frac{1}{2} \int_\Omega (1 - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}})^2 \, \mathrm{d}\mathbb{P})^{1/2}$, and the normalization constant $\frac{1}{2}$ guarantees $D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) \in [0, 1]$. A similar remark can be made about D_{TV}, cp. Exercise 3.7.2.

The Kullback-Leibler (KL) divergence (also called relative entropy) is a measure of how one probability distribution diverges from a second.

Definition 3.7.5. The **Kullback-Leibler divergence** between two probability measures $\mathbb{Q}$ and $\mathbb{P}$ is defined as:

$$D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}) = \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} & \text{if } \mathbb{P} \ll \mathbb{Q} \\ \infty & \text{otherwise.} \end{cases}$$

If $\mathbb{P}$ and $\mathbb{Q}$ are equivalent,

$$D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}) = -\int_\Omega \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right) \mathrm{d}\mathbb{P}.$$

If $\Omega = \mathbb{R}^d$ and $\mathbb{P}$ and $\mathbb{Q}$ have densities $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda_d}$ and $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\lambda_d}$, the KL divergence can be written as

$$D_{\mathrm{KL}}(\mathbb{P} \| \mathbb{Q}) = \int_\Omega \log\left(\frac{p(x)}{q(x)}\right) p(x) dx.$$

We note that the KL divergence is not a distance metric, as it is **not symmetric**. In contrast, the total variation distance and the Hellinger distance are distance metrics. However, the KL divergence is non-negative and takes the value 0 iff $\mathbb{P} = \mathbb{Q}$. This result is known as *Gibb's inequality*.

Compared to the total variation distance and the Hellinger distance, the KL divergence has computational advantages in certain situations. We can write the KL divergence as

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \int \log p(x)p(x)dx - \int \log q(x)p(x)dx,$$

where the second part of the right hand side is called the **cross entropy**,

$$H(\mathbb{P}\|\mathbb{Q}) = -\int \log q(x)p(x)dx.$$

Given a set of samples drawn from $\mathbb{P}$, it is possible to compute the cross entropy for a given $\mathbb{Q}$ with known density function without knowing the density function of $\mathbb{P}$. This is particularly useful for many tasks in computational statistics such as importance sampling and density estimation.

The next proposition summarizes the most important relations between the above divergences. The proof is left as an exercise.

Proposition 3.7.6. *It holds*

(i) $D_{\mathrm{H}}(\mathbb{P},\mathbb{Q})^2 \le D_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) \le \sqrt{2}D_{\mathrm{H}}(\mathbb{P},\mathbb{Q})$.

Moreover, if $\mathbb{P}$ *and* $\mathbb{Q}$ *are equivalent*

(ii) $D_{\mathrm{H}}(\mathbb{P},\mathbb{Q})^2 \le \frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$,

(iii) $D_{\mathrm{TV}}(\mathbb{P},\mathbb{Q})^2 \le \frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$.

Finally, we show how a bound on the Hellinger distance implies a bound on the difference of expectations taken with respect to two different probability measures. To this end in the following we denote $\mathbb{E}_{\mathbb{P}}[f] := \int f \,\mathrm{d}\mathbb{P}$, i.e. the index indicates the probability measure.

Lemma 3.7.7. *Let* $\mathbb{P}$ *and* $\mathbb{Q}$ *be two probability measures on a measurable space* $(\Omega,\mathcal{A})$*. Let* $f : \Omega \to V$ *be a RV for a separable Banach space* V*, such that* f *has finite second moments with respect to both* $\mathbb{P}$ *and* $\mathbb{Q}$*. Then*

$$\|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]\| \le 2(\mathbb{E}_{\mathbb{P}}[\|f\|^2] + \mathbb{E}_{\mathbb{Q}}[\|f\|^2])^{1/2} D_{\mathrm{H}}(\mathbb{P},\mathbb{Q}).$$

Proof. Let $\mathbb{P} \ll \mu$ and $\mathbb{Q} \ll \mu$. Then

$$\begin{aligned}
\|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]\| &\le \int_\Omega \|f\| \left|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}\right| \mathrm{d}\mu \\
&= \int_\Omega \left(\frac{1}{\sqrt{2}}\left|\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}}\right|\right)\left(\sqrt{2}\|f\|\left|\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} + \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}}\right|\right) \mathrm{d}\mu \\
&\le \left(\frac{1}{2}\int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}}\right)^2 \mathrm{d}\mu\right)^{1/2} \left(2\int_\Omega \|f\|^2 \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} + \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}}\right)^2 \mathrm{d}\mu\right)^{1/2} \\
&\le \left(\frac{1}{2}\int_\Omega \left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}} - \sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}}\right)^2 \mathrm{d}\mu\right)^{1/2} \left(4\int_\Omega \|f\|^2 \left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu} + \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}\right) \mathrm{d}\mu\right)^{1/2} \\
&= 2(\mathbb{E}_{\mathbb{P}}[\|f\|^2] + \mathbb{E}_{\mathbb{Q}}[\|f\|^2])^{1/2} D_{\mathrm{H}}(\mathbb{P},\mathbb{Q}).
\end{aligned}$$

□

Chapter 4

Bayesian Inversion

In this chapter we discuss the Bayesian approach towards inverse problems. In contrast to the methods of Chapter 2, in the Bayesian setting all involved quantities are modelled as random variables. As such, the question to be answered is not *what is the value of the unknown variable?*, but rather *what is the distribution of the unknown variable?* It turns out that this is a very powerful viewpoint, leading to a host of numerical methods with broad applications in statistics, applied mathematics and machine learning. Additionally, it has the mathematical advantage of yielding a well-posed inverse problem, as will be discussed in this chapter.

4.1 The Bayesian inverse problem

As in the previous chapter, we will use capital letters to denote RVs. We denote by X the unknown of primary interest which we wish to identify, by Y an observable quantity, and by E a noise term. In the most general form, the model is described by a possibly nonlinear operator Φ such that

$$Y = \Phi(X, E).$$

Thus Φ ties together the three RVs X, Y and E, and their probability distributions are interdependent. The RV X will also be referred to as the *parameter* that we wish to infer. The RV Y is often called the *measurement*, *observation* or *data*, and E can be interpreted as a *measurement error*. The most common model for the measurement error is that of additive noise, i.e.

$$Y = \Phi(X) + E \tag{4.1.1}$$

and we will concentrate on this situation in the following. The interpretation is that we make an observation of $\Phi(X)$ that is polluted by the noise E.

For an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the following assumptions are made throughout this chapter:

(i) $X : \Omega \to V$ is a RV for some separable Banach space V. As such it has a distribution $\mathbb{P}_X$, which is called the **prior distribution** (or simply the prior). The prior is interpreted as the information available on X, *before* observing Y.

(ii) $E : \Omega \to W$ is a RV onto a second separable Banach space W, and E and X are independent.

(iii) $\Phi : V \to W$ is a Borel-measurable function. We will refer to it as the **forward operator**.

From (4.1.1) we observe that $Y : \Omega \to W$ is also a RV. Assuming that we observe Y (i.e. we are given a realization $Y(\omega) = \Phi(X(\omega)) + E(\omega) \in W$ for some $\omega \in \Omega$), the Bayesian inverse problem is then to determine the distribution of X conditioned on the event $[Y = y]$. Under the present assumptions, this distribution can be interpreted as pooling all available information that we (can) have on X.

Terminology 4.1.1. Given a realization $y \in W$ of Y, the solution to the Bayesian inverse problem is the conditional distribution $\mathbb{P}[X \in \cdot | Y = y]$. We call $\mathbb{P}[X \in \cdot | Y = y]$ the **posterior distribution** (or simply the posterior).

The algorithms discussed in Chapter 2 returned a point estimate x for a given value y, for instance assuming the model $y = Ax$ for some matrix A. One advantage of Bayesian methods is, that they do not merely deliver point estimates, but acknowledge the fact, that we cannot know the exact value of x, for instance because there may exist multiple x_j with $Ax_j = y$ due to A being non-regular. This is reflected in the posterior being a distribution, and thus assigning probabilities to events of the type $[X \in B]$, $B \in \mathcal{B}(V)$. The posterior represents our knowledge and uncertainty about X.

Apart from computing point estimates, which we discuss in the next section, the Bayesian approach additionally allows to compute quantities like the variance to investigate uncertainty—a large variance of the parameter w.r.t. the posterior may for instance indicate that the data is not very informative about the parameter. In this chapter we investigate how to determine and explore the posterior.

4.2 Estimators

Even though the posterior contains all available information about X, it is still desirable to have a point estimate, i.e. a concrete value $x \in V$ which can be interpreted as the "most probable" value of X (in a suitable sense) given that we observed some value y for Y. Part of the reason is that the posterior distribution is a measure on V—a possibly high- or even infinite-dimensional Banach space. This precludes visualization of the posterior density and its properties.

Before continuing, we first introduce some shorter notation for the occurring distributions. Furthermore, we will assume in the following that real valued RVs are absolutely continuous w.r.t. the Lebesgue measures, and thus have densities.

- The prior distribution $\mathbb{P}_X$ on $(V, \mathcal{B}(V))$ will be denoted by μ_X. If $V = \mathbb{R}^n$, we write $\pi_X(x)$ for its density.
- The posterior distribution $\mathbb{P}[X \in \cdot | Y = y]$ on $(V, \mathcal{B}(V))$ is denoted by $\mu_{X|y}$. If $V = \mathbb{R}^n$, we write $\pi_{X|Y}(x|y)$ for its density.
- The conditional distribution $\mathbb{P}[Y \in \cdot | X = x]$ is denoted by $\mu_{Y|x}$. If $W = \mathbb{R}^m$ we write $\pi_{Y|X}(y|x)$ for its density.
- The noise distribution $\mathbb{P}_E$ on $(W, \mathcal{B}(W))$ is denoted by μ_E. If $W = \mathbb{R}^m$ we write $\pi_E(e)$ for its density.

Similarly, we will denote the joint density of X and Y by $\pi_{X,Y}(x, y)$, and the joint density of X and E by $\pi_{X,E}(x, e)$. Note that the standing assumption of X and E being independent implies $\pi_{X,E}(x, e) = \pi_X(x)\pi_E(e)$.

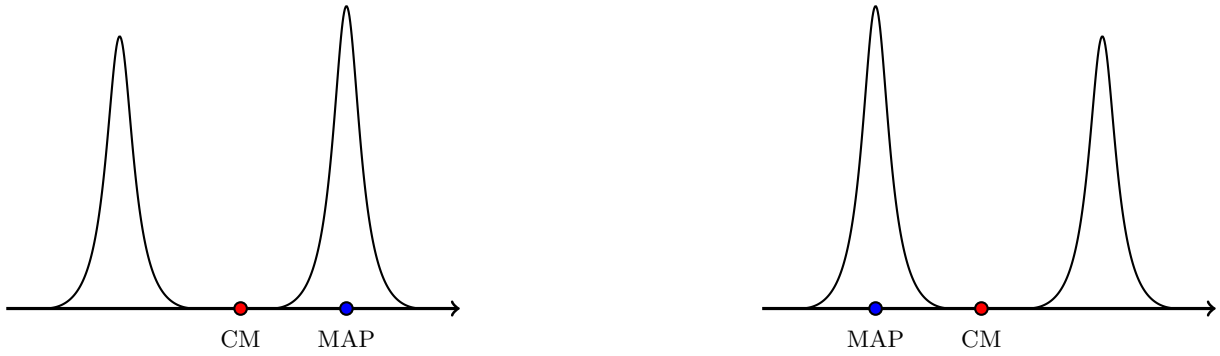


Figure 4.1: MAP and CM for two posterior distributions.

The density $\pi_{Y|X}(y|x)$ is called the **likelihood**. For a fixed $x \in V$, it describes the probability distribution of the observed quantity Y, and thus expresses the likelihood of different measurement outcomes for fixed parameter x. We can also see it as a function of x, in which case the x-value maximizing $\pi(y|x)$ can be interpreted as "best explaining" the observed data y. Let us assume for the moment that all densities exist. Then we consider the following three point estimates:

(i) Maximum likelihood (ML): The maximum likelihood estimate is a popular estimate in statistics. It is defined as a point

$$x_{\mathrm{ML}} \in \operatorname{argmax}_x \pi_{Y|X}(y|x).$$

(ii) Maximum a posteriori (MAP): A MAP point, is a point maximizing the posterior density

$$x_{\mathrm{MAP}} \in \operatorname{argmax}_x \pi_{X|Y}(x|y).$$

(iii) Conditional mean (CM): The conditional mean is the posterior expectation of X, i.e.

$$x_{\mathrm{CM}} := \mathbb{E}[X|Y=y] = \int_V x \,\mathrm{d}\mu_{X|y}(x) = \int_V x \pi_{X|Y}(x|y) \,\mathrm{d}x.$$

We point out that computing x_{ML} and x_{MAP} requires solving an optimization problem, while the computation of x_{CM} requires computing (a high-dimensional) integral. For this reason the computational techniques can differ significantly. However, modern Bayesian techniques are often rooted in a combination of optimization, sampling and integration methods. Moreover, while x_{ML} and x_{MAP} need not be unique, x_{CM} is (in case the expectation exists). The advantage of x_{CM} is that it is not strongly affected by small changes in the posterior measure, and this will be discussed in more detail in the following sections. Such a statement is not true for x_{MAP}. On the other hand, x_{CM} has the disadvantage that it does not necessarily correspond to a point with high posterior density. Such a case is often accompanied by the variance of the posterior being high, indicating that we should not be too confident in our point estimate either way. Figure 4.1 visualizes these statements.

In this lecture we concentrate on the CM estimator.

4.3 Bayes' theorem

We start with the finite dimensional case where $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ and $X : \Omega \to V$ and $Y : \Omega \to W$ are RVs with joint density $\pi_{X,Y}$.

Theorem 4.3.1 (Bayes' theorem)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}^n$, $Y : \Omega \to \mathbb{R}^m$ two RVs with joint density $\pi_{X,Y}$ and marginal densities $\pi_X(x) = \int_{\mathbb{R}^m} \pi_{X,Y}(x,y)\,\mathrm{d}y$ and $\pi_Y(y) = \int_{\mathbb{R}^n} \pi_{X,Y}(x,y)\,\mathrm{d}x$. Let $\pi_{Y|X}(y|x)$ be a conditional density of Y given X. Then μ_Y-a.e.*

$$\pi_{X|Y}(x|y) = \frac{\pi_{Y|X}(y|x)\pi_X(x)}{\pi_Y(y)}. \tag{4.3.1}$$

Proof. By Prop. 3.5.36 and Rmk. 3.5.37 there exists a $\mathbb{P}_X$-null set $N_X \subseteq \mathbb{R}^n$ such that for all $x \in N_X^c$

$$\pi_{Y|X}(y|x) = \frac{\pi_{X,Y}(x,y)}{\pi_X(x)} \qquad \text{for } \lambda_m\text{-a.e. } y \in \mathbb{R}^m, \tag{4.3.2}$$

with the denominator being a positive number. With the $\mathbb{P}_Y$-null set $N_Y := \{y : \pi_Y(y) \in \{0, \infty\}\}$, set for $y \in N_Y^c$ and $B \in \mathcal{B}(\mathbb{R}^n)$ with $\pi_{X|Y}(x|y)$ as in (4.3.1)

$$\tau(y, B) := \int_B \pi_{X|Y}(x|y)\,\mathrm{d}x.$$

Then for any $A \in \mathcal{B}(\mathbb{R}^m)$ and any $B \in \mathcal{B}(\mathbb{R}^n)$, using Fubini's theorem, (4.3.1) as a definition of $\pi_{X|Y}$, and (4.3.2),

$$\begin{aligned}
\int_A \tau(y,B)\,\mathrm{d}\mathbb{P}_Y(y) &= \int_{A\setminus N_Y} \tau(y,B)\,\mathrm{d}\mathbb{P}_Y(y) \\
&= \int_{A\setminus N_Y} \int_B \pi_{X|Y}(x|y)\pi_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{A\setminus N_Y} \int_{B\setminus N_X} \pi_{Y|X}(y|x)\pi_X(x)\,\mathrm{d}x\,\mathrm{d}y + \int_{N_X} \int_{A\setminus N_Y} \pi_{Y|X}(y|x)\,\mathrm{d}y\,\pi_X(x)\,\mathrm{d}x \\
&= \int_{A\setminus N_Y} \int_{B\setminus N_X} \pi_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \mathbb{P}[Y \in A, X \in B].
\end{aligned}$$

Here we have used that $\mathbb{P}_X[N_X] = \mathbb{P}_Y[N_Y] = 0$. □

Bayes' theorem is often referred to in the form

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior} \tag{4.3.3}$$

where $\propto$ signifies equality of two functions up to a constant (independent of the function argument). In our notation the posterior is $\pi_{X|Y}(x|y)$, the likelihood $\pi_{Y|X}(y|x)$ and the prior $\pi_X(x)$. Equality holds up the multiplicative factor $\pi_Y(y)^{-1}$, which does not depend on x—the argument of the conditional density $x \mapsto \pi_{X|Y}(x|y)$. Hence the posterior is proportional to the prior multiplied with the likelihood. The likelihood represents the information obtained through the data and can be interpreted as updating our prior belief (π_X) on the parameter.

Next we deduce explicit expressions of the posterior for the additive noise model (4.1.1). To this end, we first compute the likelihood. With $\Phi : V \to W$, introduce the shift operator $S^{\Phi(x)} : W \to W$ via

$$S^{\Phi(x)}(y) := y + \Phi(x).$$

This function is measurable, and hence for the probability measure μ_E on $(W, \mathcal{B}(W))$, the pushforward $S^{\Phi(x)}_\sharp \mu_E$ also is a probability measure on $(W, \mathcal{B}(W))$.

Lemma 4.3.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, V, W two separable Banach spaces, $\Phi : V \to W$ measurable, and $X : \Omega \to V$ as well as $E : \Omega \to W$ two independent RVs. Assume that $x \mapsto S^{\Phi(x)}_\sharp \mu_E(A)$ is measurable for every $A \in \mathcal{B}(W)$.*

Then with the RV $Y := \Phi(X) + E : \Omega \to W$ it holds μ_X-a.e.

$$\mu_{Y|x} = S^{\Phi(x)}_\sharp \mu_E.$$

Proof. Define for $A \in \mathcal{B}(W)$ and $x \in V$

$$\tau(x, A) := (S^{\Phi(x)}_\sharp \mu_E)(A).$$

Since μ_E and consequently $S^{\Phi(x)}_\sharp \mu_E$ are probability measures, $A \mapsto \tau(x, A) = S^{\Phi(x)}_\sharp \mu_E(A)$ defines a probability measure on W for every x. Measurability of $x \mapsto \tau(x, A)$ holds by assumption. Next, let $B \in \mathcal{B}(V)$. Then

$$\begin{aligned}\mathbb{P}[Y \in A, X \in B] &= \int_\Omega \mathbb{1}_A(Y(\omega))\mathbb{1}_B(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_\Omega \mathbb{1}_A(\Phi(X(\omega)) + E(\omega))\mathbb{1}_B(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega).\end{aligned}$$

Set $\varphi : V \times W \to \mathbb{R}$ via $\varphi(x, e) = \mathbb{1}_A(\Phi(x) + e)\mathbb{1}_B(x)$. Then φ is measurable as a composition of measurable functions, and the integrand equals $\varphi(Z)$ with Z denoting the RV $(X, E) : \Omega \to V \times W$. By Thm. 3.2.24, since $Z_\sharp \mathbb{P} = \mathbb{P}_Z = \mathbb{P}_{X,E} = \mu_X \otimes \mu_E$ due to the independence of X and E,

$$\begin{aligned}\mathbb{P}[Y \in A, X \in B] &= \int_\Omega \varphi(Z(\omega)) \,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_{V \times W} \mathbb{1}_A(\Phi(x) + e)\mathbb{1}_B(x) \,\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_E)(x, e) \\ &= \int_V \mathbb{1}_B(x) \int_W \mathbb{1}_A(\Phi(x) + e) \,\mathrm{d}\mu_E(e) \,\mathrm{d}\mu_X(x).\end{aligned}$$

By a change of variables (again Thm. 3.2.24)

$$\begin{aligned}\mathbb{P}[Y \in A, X \in B] &= \int_V \mathbb{1}_B(x) \int_W \mathbb{1}_A(S^{\Phi(x)}(e)) \,\mathrm{d}\mu_E(e) \,\mathrm{d}\mu_X(x) \\ &= \int_V \mathbb{1}_B(x) \int_W \mathbb{1}_A(e) \,\mathrm{d}S^{\Phi(x)}_\sharp \mu_E(e) \,\mathrm{d}\mu_X(x) \\ &= \int_V \mathbb{1}_B(x)\tau(x, A) \,\mathrm{d}\mathbb{P}_X(x).\end{aligned}$$

Thus $\tau(x, A)$ is a conditional distribution of Y given X. □

Let's again consider the finite dimensional case first.

Corollary 4.3.3. *Let $Y = \Phi(X) + E$ where $X : \Omega \to \mathbb{R}^n$, $E : \Omega \to \mathbb{R}^m$ are independent RVs with densities π_X and π_E and $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is measurable. Then μ_Y-a.e.*

$$\pi_{X|Y}(x|y) = \frac{\pi_E\left(y - \Phi(x)\right)\pi_X(x)}{Z(y)},$$

where

$$Z(y) = \int_{\mathbb{R}^n} \pi_E\left(y - \Phi(x)\right)\pi_X(x)\,\mathrm{d}x.$$

Proof. By definition of $S^{\Phi(x)}(y) = y + \Phi(x)$, we have for any $A \in \mathcal{B}(\mathbb{R}^m)$

$$\begin{aligned} S^{\Phi(x)}_\sharp \mu_E(A) &= \mu_E(\{y \in \mathbb{R}^m \,:\, y + \Phi(x) \in A\}) \\ &= \int_{\mathbb{R}^m} \mathbb{1}_A(y + \Phi(x))\pi_E(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^m} \mathbb{1}_A(y)\pi_E(y - \Phi(x))\,\mathrm{d}y, \end{aligned}$$

and thus $S^{\Phi(x)}_\sharp \mu_E$ has density $y \mapsto \pi_E(y - \Phi(x))$. Hence by Lemma 4.3.2 the conditional density $\pi_{Y|X}(y|x)$ is equal to $\pi_E(y - \Phi(x))$ for μ_X-a.e. $x \in V$. The statement then follows by Thm. 4.3.1 and the observation that by definition of the conditional density for every $A \in \mathcal{B}(\mathbb{R}^m)$

$$\mathbb{P}[Y \in A, X \in \mathbb{R}^n] = \int_{\mathbb{R}^n}\int_A \pi_{Y|X}(y|x)\,\mathrm{d}y\,\pi_X(x)\,\mathrm{d}x = \int_A\int_{\mathbb{R}^n} \pi_E(y - \Phi(x))\pi_X(x)\,\mathrm{d}x\,\mathrm{d}y,$$

so that $Z(y)$ is equal to the marginal density $\pi_Y(y)$ of Y. □

Remark 4.3.4. In particular the μ_Y null set $\{y \in \mathbb{R}^m \,:\, Z(y) = \pi_Y(y) = 0\}$ must be excluded in Cor. 4.3.3.

Let $\Sigma \in \mathbb{R}^{m\times m}$ be a symmetric positive definite (SPD) matrix. In the following $\|x\|^2_\Sigma := x^\top \Sigma^{-1} x$.

Example 4.3.5 (additive Gaussian noise I)**.** Let $E \sim \mathcal{N}(0, \Sigma)$ (which is the most common setting for Bayesian inference problems), then the posterior in Cor. 4.3.3 reads

$$\pi_{X|Y}(x|y) \propto \exp\left(-\frac{1}{2}\|y - \Phi(x)\|^2_\Sigma\right)\pi_X(x). \tag{4.3.4}$$

Example 4.3.6. Let $A \in \mathbb{R}^{m\times n}$ and suppose that for a given $y \in \mathbb{R}^m$ we wish to find $x \in \mathbb{R}^n$ such $Ax = y \in \mathbb{R}^m$. Assume that y is a realization of $Y = AX + E$ with $E \sim \mathcal{N}(0, I_m)$ where $I_m \in \mathbb{R}^{m\times m}$ denotes the identity matrix. As a prior we *choose* $\mu_X \sim \mathcal{N}(0, \frac{1}{\alpha} I_n)$ for some fixed $\alpha > 0$, i.e. $\pi_X(x) = \frac{\alpha^{n/2}}{(2\pi)^{n/2}} \exp(-\frac{\alpha\|x\|^2}{2})$, where $\|x\|^2 = x^\top x$ denotes the squared Euclidean norm. Then

(i) ML: We have

$$\pi_{Y|X}(y|x) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{\|Ax - y\|^2}{2}\right).$$

Maximizing the likelihood is thus equivalent to finding x in $\operatorname{argmin}_x \|Ax - y\|$.

(ii) MAP: By Example 4.3.5, $\pi_{X|Y}(x|y)$ is up to a y-dependent constant equal to

$$\exp\left(-\frac{\|Ax-y\|^2+\alpha\|x\|^2}{2}\right).$$

Therefore a MAP point is a point in $\operatorname{argmin}_x(\|Ax-y\|^2+\alpha\|x\|^2)$.

(iii) CM: The conditional mean is given by

$$\int_{\mathbb{R}^n} x\pi_{X|Y}(x|y)\,\mathrm{d}x = \frac{\int_{\mathbb{R}^n} x\pi_{Y|X}(y|x)\pi_X(x)\,\mathrm{d}x}{\int_{\mathbb{R}^n} \pi_{Y|X}(y|x)\pi_X(x)\,\mathrm{d}x} = \frac{\int_{\mathbb{R}^n} x\exp\left(-\frac{\|Ax-y\|^2+\alpha\|x\|^2}{2}\right)\,\mathrm{d}x}{\int_{\mathbb{R}^n} \exp\left(-\frac{\|Ax-y\|^2+\alpha\|x\|^2}{2}\right)\,\mathrm{d}x}.$$

We make the following observations:

- The ML estimate is *not Bayesian*: The joint distribution can be written as $\pi_{X,Y}(x,y) = \pi_{Y|X}(y|x)\pi_X(x)$ (cp. Prop. 3.5.36). Hence $\pi_{Y|X}(y|x)$ and as a consequence x_{ML} are independent of the (choice of) prior $\pi_X(\cdot)$. In the above example determining x_{ML} amounts to minimizing $\|Ax-y\|$ in x, i.e. to solving the inverse problem without regularization. Therefore the ML estimator is not really interesting in the context of ill-posed inverse problems.
- The MAP estimate in Example 4.3.6 corresponds to the Tikhonov regularized solution. Thus, using prior information can be interpreted as adding a form of regularization.

Example 4.3.7. We consider a logistic differential equation modeling the growth of a population $N(t)$ over time $t \geq 0$, with growth rate $r > 0$ and carrying capacity k:

$$\frac{dN}{dt}(t) = rN(t)(k-N(t)), \qquad N(0) = N_0.$$

The solution is given by

$$N(t) = \frac{k}{1+\exp(-rkt)(\frac{k}{N_0}-1)}.$$

Assume we are given the values $r = 0.25$ and $N_0 = 2$, and wish to infer k. Suppose that we apriorily know $k \in [10,20]$, motivating the prior $K \sim \mathrm{uniform}(10,20)$ for k. At time $t = 0.5$ we observe $N(t)$, however the observation is polluted by a noise term $E \sim \mathcal{N}(0,0.5)$ (independent of K). The forward operator is

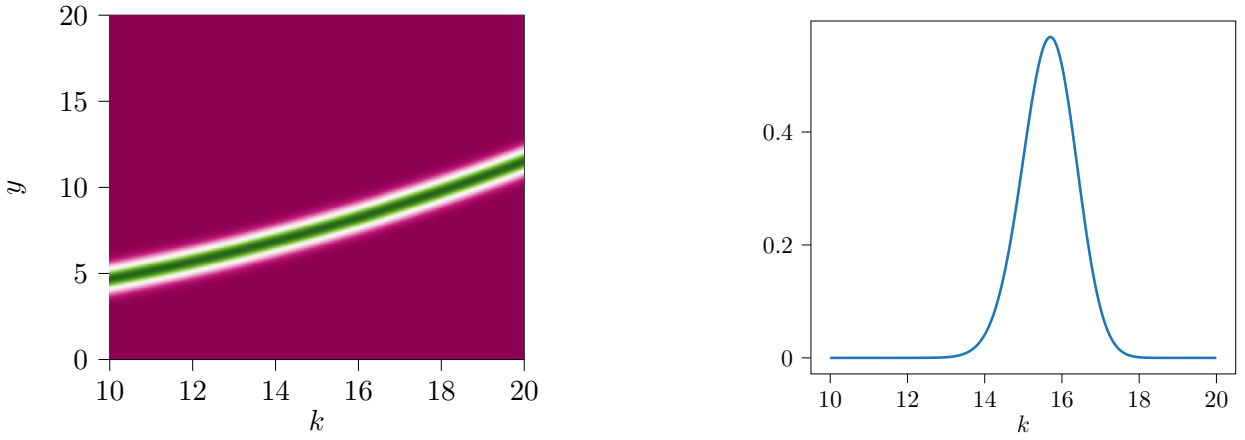
$$\Phi(k) = \frac{k}{1+\exp(-0.125k)(\frac{k}{2}-1)}, \qquad k \in [10,20],$$

the observation is described by

$$Y = \Phi(K) + E, \qquad \mu_{K,E} \sim \mathrm{uniform}(10,20) \otimes \mathcal{N}(0,0.5).$$

The joint and posterior density for $Y = 8$ are depicted in Figure 4.2.

Next we show a version of Bayes' theorem in the infinite dimensional setting. Our goal is to obtain a statement with a computable density function. In infinite dimensional spaces, there is no Lebesgue measure. Hence we have to consider a Radon-Nikodym derivative of the posterior w.r.t. another measure—the prior. Note that it is very natural to assume that the posterior $\mu_{X|y}$ is absolutely continuous w.r.t. the prior μ_X (otherwise the posterior would assign positive measure to an event to which we have apriorily assigned measure 0).



(a) joint density $\pi_{K,Y}(k, y)$

(b) posterior density $\pi_{K|Y}(k|8)$

Figure 4.2: Joint and posterior density in Example 4.3.7.

Theorem 4.3.8 (Bayes' theorem)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, V, W two separable Banach spaces, $X : \Omega \to V$, $E : \Omega \to W$ two RVs and $\Phi : V \to W$ a measurable function. Suppose further that for some σ-finite measure ν on $(W, \mathcal{B}(W))$*

(i) $S_\sharp^{\Phi(x)}\mu_E \ll \nu$ *for all* $x \in V$,

(ii) $(x, y) \mapsto \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y)$ *is measurable for* $(x, y) \in V \times W$.

Then with the RV $Y = \Phi(X) + E : \Omega \to W$ it holds $\mu_{X|y} \ll \mu_X$ for μ_Y-a.e. $y \in W$ and in this case (in particular excluding the μ_Y-null set where $Z(y) = 0$)

$$\frac{\mathrm{d}\mu_{X|y}}{\mathrm{d}\mu_X}(y) = \frac{1}{Z(y)} \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y),$$

where

$$Z(y) = \int_V \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y)\, \mathrm{d}\mu_X(x). \tag{4.3.5}$$

Proof. By definition of a conditional density and due to Lemma 4.3.2, for every $A \in \mathcal{B}(W)$

$$\begin{aligned}
\mathbb{P}[Y \in A, X \in V] &= \int_V \mu_{Y|x}(A)\, \mathrm{d}\mu_X(x) \\
&= \int_V (S_\sharp^{\Phi(x)}\mu_E)(A)\, \mathrm{d}\mu_X(x) \\
&= \int_V \int_W \mathbb{1}_A(e)\, \mathrm{d}(S_\sharp^{\Phi(x)}\mu_E)(e)\ \mathrm{d}\mu_X(x) \\
&= \int_V \int_W \mathbb{1}_A(e) \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(e)\, \mathrm{d}\nu(e)\, \mathrm{d}\mu_X(x) \\
&= \int_W \mathbb{1}_A(e) \int_V \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(e)\, \mathrm{d}\mu_X(x)\ \mathrm{d}\nu(e).
\end{aligned}$$

[Draft of October 6, 2021. Not for dissemination.]

Here we used that $\frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(e)$ is jointly measurable in (x,e) which allowed to use Fubini's theorem. This calculation shows that $Z(y)$ is equal to the Radon-Nikodym derivative $\frac{\mathrm{d}\mu_Y}{\mathrm{d}\nu}(y)$ of μ_Y w.r.t. ν. In particular $N := \{y \in W : Z(y) = 0\}$ is a μ_Y-null set since $\mu_Y[N] = \int_{[Z(y)=0]} Z(y)\,\mathrm{d}\nu(y) = 0$.

Define

$$r(x,y) := \begin{cases} \frac{1}{Z(y)} \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y) & \text{if } y \in N^c \\ 1 & \text{if } y \in N. \end{cases}$$

By (ii) and the fact that $y \mapsto Z_Y(y)$ is measurable as a consequence of the Fubini-Tonelli theorem, we find that r is measurable. Set for $B \in \mathcal{B}(V)$

$$\tau(y,B) := \int_V \mathbb{1}_B(x) r(x,y)\,\mathrm{d}\mu_X(x).$$

By definition of r the map $B \mapsto \tau(y,B)$ is a probability measure for every $y \in W$ (trivially if $y \in N$, and due to the definition of the normalizing factor $Z(y)$ otherwise). Moreover, $y \mapsto \tau(y,B)$ is measurable as a consequence of the Fubini-Tonelli theorem. For each $B \in \mathcal{B}(V)$, $A \in \mathcal{B}(W)$ since $Z(y) = \frac{\mathrm{d}\mu_Y}{\mathrm{d}\nu}$

$$\begin{aligned}
\int_W \mathbb{1}_A(y)\tau(y,B)\,\mathrm{d}\mu_Y(y) &= \int_{W\setminus N}\int_V \mathbb{1}_B(x)\mathbb{1}_A(y) r(x,y)\,\mathrm{d}\mu_X(x)\,\mathrm{d}\mu_Y(y) \\
&= \int_V \mathbb{1}_B(x) \int_{W\setminus N} \mathbb{1}_A(y) \frac{1}{Z(y)} \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y)\,\mathrm{d}\mu_Y(y)\,\mathrm{d}\mu_X(x) \\
&= \int_V \mathbb{1}_B(x) \int_{W\setminus N} \mathbb{1}_A(y) \frac{1}{Z(y)} \frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\nu}(y) Z(y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu_X(x) \\
&= \int_V \mathbb{1}_B(x) \int_{W\setminus N} \mathbb{1}_A(y)\,\mathrm{d}(S_\sharp^{\Phi(x)}\mu_E)(y)\,\mathrm{d}\mu_X(x) \\
&= \int_V \mathbb{1}_B(x) \int_W \mathbb{1}_{A\setminus N}(y)\,\mathrm{d}\mu_{Y|x}(y)\,\mathrm{d}\mu_X(x) \\
&= \mathbb{P}[X \in B, Y \in A\setminus N] = \mathbb{P}[X \in B, Y \in A],
\end{aligned}$$

where we have used Lemma 4.3.2 and that $\mathbb{P}[Y \in N] = 0$. □

Example 4.3.9 (additive Gaussian noise II)**.** Let V be a separable Banach space and $X : \Omega \to V$ a RV. Let $W = \mathbb{R}^m$, $\Phi : V \to \mathbb{R}^m$ measurable, and assume $E : \Omega \to \mathbb{R}^m$ is a RV independent of X and distributed according to $\mathcal{N}(0,\Sigma)$ for an SPD covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Then

$$\pi_E(e) = \frac{\mathrm{d}\mu_E}{\mathrm{d}\lambda_m}(e) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\left(-\frac{1}{2}\|e\|_\Sigma^2\right)$$

and

$$\frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\lambda_m}(e) = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\left(-\frac{1}{2}\|e-\Phi(x)\|_\Sigma^2\right) = \pi_E(e-\Phi(x)).$$

Note that measurability of Φ implies that the last expression is measurable in (e,x). Hence by Thm. 4.3.8

$$\frac{\mathrm{d}\mu_{X|y}}{\mathrm{d}\mu_X}(x) \propto \exp\left(-\frac{1}{2}\|y-\Phi(x)\|_\Sigma^2\right),$$

which (up to a constant) corresponds to the likelihood. This is to be expected, since we took the Radon-Nikodym derivative of the posterior w.r.t. the prior. Thus we have established a form of (4.3.3) in this setting. Up to a constant, the negative log-likelihood

$$\frac{1}{2}\|y - \Phi(x)\|_\Sigma^2$$

is the so-called **data misfit potential**. It tells us how well a parameter x fits the observed value y.

4.4 Stability

In the previous section we have seen that (under certain assumptions), the posterior distribution exists and is unique (μ_Y-a.e.). Thus the Bayesian inverse problem (BIP) possesses a unique solution. To further pursue our investigation of well-posedness, in this section we discuss continuity of the posterior w.r.t. the data.

To this end we focus on the situation where

(i) the noise is finite dimensional: $W = \mathbb{R}^m$, i.e. $E : \Omega \to \mathbb{R}^m$,

(ii) the noise has a density: $\mu_E \ll \lambda_m$.

This covers for example Gaussian noise as discussed in Example 4.3.9. In this case, as we have shown in the proof of Cor. 4.3.3,

$$\frac{\mathrm{d}S_\sharp^{\Phi(x)}\mu_E}{\mathrm{d}\lambda_m}(y) = \pi_E(y - \Phi(x)).$$

We point out that this function is measurable as a composition of Borel measurable functions. Therefore Thm. 4.3.8 implies that for μ_Y-a.e. y and for every $A \in \mathcal{B}(V)$

$$\mu_{X|y}(A) = \frac{1}{Z(y)} \int_V \mathbb{1}_A(x)\pi_E(y - \Phi(x)) \,\mathrm{d}\mu_X(x) \tag{4.4.1a}$$

with normalization constant

$$Z(y) = \int_V \pi_E(y - \Phi(x)) \,\mathrm{d}\mu_X(x). \tag{4.4.1b}$$

We now consider the Hellinger distance between measures of the type (4.4.1a). Under the above assumptions we have:

Theorem 4.4.1. *Let $\sqrt{\pi_E} : \mathbb{R}^m \to [0, L)$ be Lipschitz continuous with Lipschitz constant $L \geq 1$. Let for $i \in \{1, 2\}$*

$$\nu_i(A) := \frac{1}{Z_i(y_i)} \int_A \mathbb{1}_A(x)\pi_E(y_i - \Phi_i(x)) \,\mathrm{d}\mu_X(x)$$

with $y_i \in \mathbb{R}^m$, $\Phi_i \in L^2(V, \mu_X; \mathbb{R}^m)$ and Z_i as in (4.4.1b)*, and where μ_X is a probability measure on $(V, \mathcal{B}(V))$. Then if $\min\{Z_1, Z_2\} > 0$*

$$D_{\mathrm{H}}(\nu_1, \nu_2) \leq \frac{\sqrt{20}L^3}{\min\{Z_1, Z_2\}}(\|y_1 - y_2\| + \|\Phi_1 - \Phi_2\|_{L^2(V,\mu_X;\mathbb{R}^m)}).$$

Proof. Set $\varphi_i(x) := y_i - \Phi_i(x)$. We have

$$\begin{aligned}
D_{\mathrm{H}}(\nu_1,\nu_2)^2 &= \int_V \left(\sqrt{\frac{\mathrm{d}\nu_1}{\mathrm{d}\mu_X}(x)} - \sqrt{\frac{\mathrm{d}\nu_2}{\mathrm{d}\mu_X}(x)}\right)^2 \mathrm{d}\mu_X(x) \\
&= \int_V \left(\sqrt{\frac{1}{Z_1}\pi_E(\varphi_1(x))} - \sqrt{\frac{1}{Z_2}\pi_E(\varphi_2(x))}\right)^2 \mathrm{d}\mu_X(x) \\
&\leq 2\int_V \left(\sqrt{\frac{1}{Z_1}\pi_E(\varphi_1(x))} - \sqrt{\frac{1}{Z_1}\pi_E(\varphi_2(x))}\right)^2 + \left(\sqrt{\frac{1}{Z_1}\pi_E(\varphi_2(x))} - \sqrt{\frac{1}{Z_2}\pi_E(\varphi_2(x))}\right)^2 \mathrm{d}\mu_X(x).
\end{aligned}$$

Denote the integral over the first term by I_1 and the integral over the second term by I_2. Using Lipschitz continuity of $\sqrt{\pi_E}$

$$\begin{aligned}
I_1 &\leq \frac{L}{Z_1}\int_V |\varphi_1(x) - \varphi_2(x)|^2 \,\mathrm{d}\mu_X(x) \\
&\leq \frac{2L}{Z_1}\int_V \|y_1 - y_2\|^2 + (\Phi_1(x) - \Phi_2(x))^2 \,\mathrm{d}\mu_X(x) \\
&\leq \frac{2L}{Z_1}\left(\|y_1 - y_2\|^2 + \|\Phi_1 - \Phi_2\|^2_{L^2(V,\mu_X;\mathbb{R}^m)}\right). \qquad (4.4.2)
\end{aligned}$$

Moreover, due to $t \mapsto \sqrt{t} : t_0 \to \mathbb{R}$ being Lipschitz constant with Lipschitz constant $1/(2\sqrt{t_0})$ for any $t_0 > 0$,

$$I_2 = Z_2\left(\frac{1}{\sqrt{Z_1}} - \frac{1}{\sqrt{Z_2}}\right)^2 = \frac{1}{Z_1}(\sqrt{Z_2} - \sqrt{Z_1})^2 \leq \frac{1}{4Z_1 \min\{Z_1, Z_2\}}(Z_1 - Z_2)^2.$$

We note that π_E is also Lipschitz continuous with Lipschitz constant $2L^2$ due to

$$|\pi_E(a) - \pi_E(b)| = |\sqrt{\pi_E(a)} - \sqrt{\pi_E(b)}||\sqrt{\pi_E(a)} + \sqrt{\pi_E(b)}| \leq 2L^2|a - b|.$$

Next, similar as in (4.4.2),

$$\begin{aligned}
|Z_1 - Z_2|^2 &= \left(\int_V |\pi_E(\varphi_1(x)) - \pi_E(\varphi_2(x))| \,\mathrm{d}\mu_X(x)\right)^2 \\
&\leq 4L^4\left(\int_V |\varphi_1(x) - \varphi_2(x)| \,\mathrm{d}\mu_X(x)\right)^2 \\
&\leq 8L^4\left(\|y_1 - y_2\|^2 + \|\Phi_1 - \Phi_2\|^2_{L^2(V,\mu_X;\mathbb{R}^m)}\right).
\end{aligned}$$

Now, due to $Z_i = \int_V \pi_E(y_i - \Phi_i(x)) \,\mathrm{d}\mu_X(x) \leq L^2$,

$$\frac{1}{Z_1} + \frac{1}{Z_1 \min\{Z_1, Z_2\}} \leq \frac{L^2}{\min\{Z_1, Z_2\}^2}.$$

Putting everything together

$$D_{\mathrm{H}}(\nu_1,\nu_2)^2 \leq 2(I_1 + I_2) \leq \frac{2L^2(2L + 8L^4)}{\min\{Z_1, Z_2\}^2}\left(\|y_1 - y_2\|^2 + \|\Phi_1 - \Phi_2\|^2_{L^2(V,\mu_X;\mathbb{R}^m)}\right). \qquad \square$$

Assume that $Z(y) > 0$ for all $y \in \mathbb{R}^m$ in (4.4.1b). Then the previous proposition tells us that for fixed forward operator $\Phi = \Phi_1 = \Phi_2$, the posterior density depends continuously on the data y; in fact the dependence is locally Lipschitz continuous. As a consequence, also the conditional mean estimate depends continuously on the data, cp. Lemma 3.7.7.

Similarly, for fixed data $y = y_1 = y_2$, the posterior depends continuously on the *forward operator*. Such results are of interest, as in practice, the forward operator needs to be replaced by a numerical approximation $\tilde{\Phi}$ to Φ.

Example 4.4.2 (additive Gaussian noise III)**.** Let $E \sim \mathcal{N}(0, \Sigma)$ for an SPD matrix $\Sigma \in \mathbb{R}^{m\times m}$ then

$$\sqrt{\pi_E(y)} = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{4}\|y\|_\Sigma^2\right)$$

is Lipschitz continuous (because the derivative is uniformly bounded for all $y \in \mathbb{R}^m$) and $\sqrt{\pi_E} : \mathbb{R}^m \to [0, (2\pi \det(\Sigma))^{-1/2}]$. Thus Thm. 4.4.1 shows continuous dependence of the posterior on the forward operator and the data in this case.

To conclude this discussion, we investigate continuity of the posterior w.r.t. the prior.

Theorem 4.4.3. *Let μ_X, $\tilde{\mu}_X$ be two probability measures on the separable Banach space V, $y \in \mathbb{R}^m$ and $\Phi : V \to \mathbb{R}^m$ measurable. Let the noise density $\pi_E : \mathbb{R}^m \to [0, L]$ for some $L < \infty$. For $A \in \mathcal{B}(V)$ set*

$$\nu(A) := \frac{1}{Z}\int_V \mathbb{1}_A(x)\pi_E(y - \Phi(x))\,\mathrm{d}\mu_X(x)$$

as well as

$$Z := \int_V \pi_E(y - \Phi(x))\,\mathrm{d}\mu_X(x)$$

and define $\tilde{\nu}$, $\tilde{Z}$ analogous, but with μ_X replaced by $\tilde{\mu}_X$.

Then if $\min\{Z, \tilde{Z}\} > 0$

$$D_{\mathrm{H}}(\nu, \tilde{\nu}) \le \frac{2L}{\min\{Z, \tilde{Z}\}} D_{\mathrm{H}}(\mu_X, \tilde{\mu}_X).$$

Proof. Let η be a probability measure on V such that $\mu_X \ll \eta$ and $\tilde{\mu}_X \ll \eta$ (e.g. $\eta = \frac{\mu_X + \tilde{\mu}_X}{2}$). Then, using $\pi_E \le L$,

$$\begin{aligned}
D_{\mathrm{H}}(\nu, \tilde{\nu})^2 &= \frac{1}{2}\int_V \left(\sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\eta}(x)} - \sqrt{\frac{\mathrm{d}\tilde{\nu}}{\mathrm{d}\eta}(x)}\right)^2 \mathrm{d}\eta(x) \\
&= \frac{1}{2}\int_V \pi_E(y - \Phi(x)) \left(\sqrt{\frac{1}{Z}\frac{\mathrm{d}\mu_X}{\mathrm{d}\eta}(x)} - \sqrt{\frac{1}{\tilde{Z}}\frac{\mathrm{d}\tilde{\mu}_X}{\mathrm{d}\eta}(x)}\right)^2 \mathrm{d}\eta(x) \\
&\le \int_V L \left(\sqrt{\frac{1}{Z}\frac{\mathrm{d}\mu_X}{\mathrm{d}\eta}(x)} - \sqrt{\frac{1}{Z}\frac{\mathrm{d}\tilde{\mu}_X}{\mathrm{d}\eta}(x)}\right)^2 \mathrm{d}\eta(x) \\
&\quad + \int_V \pi_E(y - \Phi(x)) \left(\sqrt{\frac{1}{Z}\frac{\mathrm{d}\tilde{\mu}_X}{\mathrm{d}\eta}(x)} - \sqrt{\frac{1}{\tilde{Z}}\frac{\mathrm{d}\tilde{\mu}_X}{\mathrm{d}\eta}(x)}\right)^2 \mathrm{d}\eta(x). \qquad (4.4.3)
\end{aligned}$$

As in the proof of Thm. 4.4.1, the second part of the integral can be bounded by

$$\tilde{Z}\left(\sqrt{\frac{1}{Z}}-\sqrt{\frac{1}{\tilde{Z}}}\right)^2=\frac{1}{Z}(\sqrt{\tilde{Z}}-\sqrt{Z})^2\leq\frac{1}{4Z\min\{Z,\tilde{Z}\}}(Z-\tilde{Z})^2$$

and with Prop. 3.7.6

$$|Z-\tilde{Z}|\leq L\int_V\left|\frac{\mathrm{d}\mu_X}{\mathrm{d}\eta}-\frac{\mathrm{d}\tilde{\mu}_X}{\mathrm{d}\eta}\right|\mathrm{d}\eta\leq 2LD_{\mathrm{TV}}(\mu_X,\tilde{\mu}_X)\leq 2LD_{\mathrm{H}}(\mu_X,\tilde{\mu}_X).$$

Hence by (4.4.3)

$$D_{\mathrm{H}}(\nu,\tilde{\nu})^2\leq\frac{2L}{Z}D_{\mathrm{H}}(\mu_X,\tilde{\mu}_X)^2+\frac{L^2}{2Z\min\{Z,\tilde{Z}\}}D_{\mathrm{H}}(\mu_X,\tilde{\mu}_X)^2.$$ □

In all, we have seen that under the stated assumptions (in particular $\sqrt{\pi_E}$ is Lipschitz and bounded, and the normalization constant Z is positive for all y) the posterior density defined in (4.4.1) depends w.r.t. the Hellinger distance continuously on

- the forward operator $\Phi\in L^2_{\mu_X}$,
- the data $y\in\mathbb{R}^m$,
- the prior μ_X.

In this sense the inverse problem is well-posed.

4.5 Prior measures

The choice of prior plays an important role in Bayesian inference, and is what distinguishes it from the frequentist approach. In the (finite dim.) Gaussian setting, see Ex. 4.3.4, the prior and the posterior are equivalent measures. In particular, if the prior assigns the value 0 to an event, then the same holds for the posterior. Hence, as a general rule of thumb, apart from excluding physically impossible events, the prior should not be too restrictive.

In this section we discuss a few techniques to construct suitable measures in separable Banach spaces. As a motivation, we first we look at a PDE driven inverse problem.

Example 4.5.1. Let $D\subseteq\mathbb{R}^d$ be a bounded Lipschitz domain and consider the elliptic PDE

$$-\mathrm{div}(a\nabla u)=f,\qquad u|_{\partial D}=0. \tag{4.5.1}$$

Here we assume $f\in H^{-1}(D)$ and $a\in L^\infty(D)$ with $\mathrm{essinf}_{x\in D}a(x)>0$. Then there exists a unique weak solution $u\in H^1_0(D)$ of (4.5.1). For a bounded linear operator $B:H^1_0(D)\to\mathbb{R}^m$ define the forward operator $\Phi(a):=Bu\in\mathbb{R}^m$; note that the solution $u\in H^1_0(D)$ of (4.5.1) depends on a, so that $\Phi(a)$ is well-defined. One can show that Φ is a continuous function from

$$\{a\in L^\infty(D)\,:\,\mathrm{essinf}_{x\in D}a(x)>0\}\to\mathbb{R}^m$$

and thus Φ is also measurable.

The inverse problem is to find the diffusion coefficient $a \in L^\infty$ from a noisy measurement

$$Y = \Phi(a) + E, \tag{4.5.2}$$

with $E \sim \mathcal{N}(0, \Gamma)$ for an SPD matrix $\Gamma \in \mathbb{R}^{m\times m}$. In order to so, we proceed as outlined in Sec. 4.1: $a : \Omega \to L^\infty(D)$ is modelled as a RV for a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, i.e.

$$a(\omega, x) \in \mathbb{R}, \qquad \omega \in \Omega,\ x \in D.$$

Once we have constructed a prior measure μ_a on L^∞ (μ_a is the distribution of the RV a), the posterior can be determined with Thm. 4.3.8.

4.5.1 Karhunen-Loève expansion

Let in this section $D \subseteq \mathbb{R}^d$ be a bounded (closed) Lipschitz domain and $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space.

The map $a : \Omega \times D \to \mathbb{R}$ in Example 4.5.1, which models a random diffusion coefficient in (4.5.1), allows for two interpretations:

- $\omega \mapsto a(\omega, \cdot) \in L^\infty$ is an L^∞-valued RV,
- $(a(\cdot, x))_{x\in D}$ is a stochastic process, that is, a family of real-valued RVs indexed over $x \in D$.

For now we adopt the second viewpoint, and use the notation $a_x : \Omega \to \mathbb{R}$ instead of $a(\cdot, x)$. Hence $(a_x)_{x\in D}$ is a collection of RVs. Assuming that each $a_x : \Omega \to \mathbb{R}$ has finite first and second moments, set for $x \in D$

$$m_x := \mathbb{E}[a_x] \in \mathbb{R}$$

and for x, $y \in D$

$$c(x, y) := \operatorname{cov}(a_x, a_y) = \mathbb{E}[(a_x - m_x)(a_y - m_y)] \in \mathbb{R}.$$

We call c the **covariance function**. It is symmetric, i.e. $c(x, y) = c(y, x)$, and satisfies for all $n \in \mathbb{N}$

$$\sum_{i,j=1}^{n} s_i s_j c(x_i, x_j) \geq 0 \qquad \forall x_j \in D,\ \forall x_j \in \mathbb{R} \tag{4.5.3}$$

since $\sum_{i,j=1}^n s_i s_j c(x_i, x_j) = \mathbb{E}[(\sum_{j=1}^n s_j(a_{x_j} - m_{x_j}))^2]$. In other words, the matrix $c(x_i, x_j)$ is positive semi-definite. Therefore, we say that a function $c : D \times D \to \mathbb{R}$ is **positive semi-definite** if it satisfies (4.5.3). Moreover, a stochastic process is called **centered** if $m_x = 0$ for all $x \in D$. In the following it will be convenient to work with centered processes, but we emphasize that the following discussion also applies to non-centered processes $(a_x)_{x\in D}$ by considering $\tilde{a}_x := a_x - m_x$.

Let us begin our discussion by relating continuity of the covariance function to a form of continuity of the stochastic process.

Definition 4.5.2. We say that $(a_x)_{x\in D}$ is **mean-square continuous**, if for all $x \in D$ holds $a_x \in L^2(\Omega, \mathbb{P}; \mathbb{R})$ and

$$\lim_{y\to x} \mathbb{E}[(a_x - a_y)^2] = 0.$$

This condition can equivalently be stated as $\lim_{y\to x} \|a_x - a_y\|_{L^2(\Omega,\mathbb{P})} = 0$.

Lemma 4.5.3. *Assume that a_x has finite second moment and $\mathbb{E}[a_x] = 0$ for all $x \in D$.*

Then the covariance function $c : D \times D \to \mathbb{R}$ is continuous iff the stochastic process $(a_x)_{x\in D}$ is mean-square continuous.

Proof. We have

$$\mathbb{E}[(a_x - a_y)^2] = \mathbb{E}[a_x^2] - 2\mathbb{E}[a_x a_y] + \mathbb{E}[a_y^2] = c(x,x) - 2c(x,y) + c(y,y).$$

Hence continuity of c implies mean-square continuity of the stochastic process.

Conversely, let $(a_x)_{x\in D}$ be mean-square continuous. Then

$$\begin{aligned}
|c(x+s, y+t) - c(x,y)| &= |\mathbb{E}[a_{x+s}a_{y+t}] - \mathbb{E}[a_x a_y]| \\
&= |\mathbb{E}[(a_{x+s} - a_x)(a_{y+t} - a_y)] + \mathbb{E}[(a_{x+s} - a_x)a_y] + \mathbb{E}[a_x(a_{y+t} - a_y)]| \\
&\le \|a_{x+s} - a_x\|_{L^2(\Omega,\mathbb{P})}\|a_{y+t} - a_y\|_{L^2(\Omega,\mathbb{P})} + \|a_{x+s} - a_x\|_{L^2(\Omega,\mathbb{P})}\|a_y\|_{L^2(\Omega,\mathbb{P})} \\
&\quad + \|a_x\|_{L^2(\Omega,\mathbb{P})}\|a_{y+t} - a_y\|_{L^2(\Omega,\mathbb{P})}.
\end{aligned}$$

This term tends to 0 as s, $t \to 0$ due to the mean-square continuity of $(a_x)_{x\in D}$. □

Remark 4.5.4. More generally, the smoothness of the covariance function can be related to the smoothness of *paths* of the random process; a path is a function $a(\omega, \cdot) : D \to \mathbb{R}$ for fixed $\omega \in \Omega$. This will be further discussed in the exercises.

For a function $c \in L^2(D \times D)$ in the following we consider the *Hilbert-Schmidt integral operator* defined as

$$T_c f(x) = \int_D c(x,y) f(y)\,\mathrm{d}y \qquad x \in D.$$

Lemma 4.5.5 (Hilbert-Schmidt integral operator)**.** *Let $c \in L^2(D \times D, \mathbb{R})$ be symmetric. Then $T_c : L^2(D) \to L^2(D)$ is a compact, self-adjoint operator.*

Proof. For every $f \in L^2(D)$

$$\begin{aligned}
\int_D T_c f(x)^2\,\mathrm{d}x &= \int_D \left(\int_D c(x,y) f(y)\,\mathrm{d}y\right)^2 \mathrm{d}x \\
&\le \int_D \int_D c(x,y)^2\,\mathrm{d}y \int_D f(y)^2\,\mathrm{d}y\,\mathrm{d}x \\
&= \|c\|^2_{L^2(D\times D)}\|f\|^2_{L^2(D)},
\end{aligned}$$

which shows $\|T_c\|_{\mathcal{L}(L^2,L^2)} \le \|c\|_{L^2(D\times D)}$.

If additionally $g \in L^2(D)$, then due to the symmetry of c

$$\langle T_c f, g\rangle_{L^2(D)} = \int_D \int_D c(x,y) f(y) f(x)\,\mathrm{d}y\,\mathrm{d}x = \langle f, T_c g\rangle_{L^2(D)},$$

so that T_c is self-adjoint.

Finally we show compactness. Let $(\varphi_j)_{j\in\mathbb{N}}$ be an orthonormal basis of the separable Hilbert space $L^2(D)$. The functions $D\times D \ni (x,y) \mapsto \varphi_i(x)\varphi_j(y)$ yield an orthonormal basis $(\varphi_i(x)\varphi_i(y))_{i,j\in\mathbb{N}}$ of $L^2(D\times D)$ (exercise; hint: use Fubini's theorem). Hence with $c_{i,j} = \int_D \int_D c(x,y)\varphi_i(x)\varphi_j(y)\,\mathrm{d}x\,\mathrm{d}y$ it holds

$$c(x,y) = \sum_{i,j\in\mathbb{N}} c_{i,j}\varphi_i(x)\varphi_j(y)$$

[Draft of October 6, 2021. Not for dissemination.]

with convergence in $L^2(D \times D)$. Set

$$c_n(x,y) = \sum_{i,j=1}^{n} \varphi_i(x) \sum_{j \in \mathbb{N}} c_{i,j} \varphi_j(y) \in L^2(D \times D).$$

Then $T_c - T_{c_n} = T_{c-c_n}$ and thus by Parseval's identity

$$\|T_c - T_{c_n}\|^2_{\mathcal{L}(L^2,L^2)} \leq \|c - c_n\|^2_{L^2(D\times D)} = \sum_{i>n} \sum_{j \in \mathbb{N}} c_{i,j}^2 \to 0$$

as $n \to \infty$ since $\sum_{i,j} c_{i,j}^2 < \infty$. Moreover, for all $f \in L^2(D)$

$$T_{c_n} f(x) = \sum_{i=1}^{n} \varphi_i(x) \int_D f(y) \sum_{j \in \mathbb{N}} c_{i,j} \varphi_j(y) \, \mathrm{d}y \in \operatorname{span}\{\varphi_1, \ldots, \varphi_n\},$$

so that T_{c_n} has finite range and is therefore compact. Hence $T_c = \lim_{n\to\infty} T_{c_n}$ is compact. □

Exercise 4.5.6. For a symmetric positive semi-definite $c \in L^2(D \times D; \mathbb{R})$ show that $T_c : L^2(D) \to L^2(D)$ is a positive operator, i.e. $\langle T_c f, f \rangle \geq 0$ for all $f \in L^2(D)$.

Since $T_c : L^2(D) \to L^2(D)$ is a compact, self-adjoint, positive operator, there's an orthonormal system $(\varphi_j)_{j\in\mathbb{N}}$ of eigenvectors of T_c with corresponding positive eigenvalues $(\ell_j)_{j\in\mathbb{N}}$ and such that $T_c f = \sum_{j\in\mathbb{N}} \ell_j \langle f, \varphi_j \rangle_{L^2(D)} \varphi_j$ (or $T_c f = \sum_{j=1}^{n} \ell_j \langle f, \varphi_j \rangle_{L^2(D)} \varphi_j$ in case T_c has $n \in \mathbb{N}$ dimensional range) is the spectral decomposition of T_c; to avoid distinguishing two cases, for simplicity we assume in the following that c is such that the range of T_c is infinite dimensional.

Definition 4.5.7. Let H be a separable Hilbert space and $A \in \mathcal{L}(H, H)$ a bounded positive linear operator. We say that A is a **trace-class operator** (or A is of trace class) if for an ONB $(\varphi_j)_{j\in\mathbb{N}}$ of H holds

$$\operatorname{tr}(A) := \sum_{j \in \mathbb{N}} \langle A\varphi_j, \varphi_j \rangle_H < \infty.$$

One can show that the definition trace $\operatorname{tr}(A)$ does not depend on the choice of ONB $(\varphi_j)_{j\in\mathbb{N}}$ (exercise).

Theorem 4.5.8 (Mercer's theorem)**.** *Let $c : D \times D \to \mathbb{R}$ be continuous, positive semi-definite and symmetric. Then there exists an orthonormal system $(\varphi_j)_{j\in\mathbb{N}}$ in $L^2(D)$ such that $\varphi_j \in C^0(D)$, $T_c\varphi_j = \ell_j\varphi_j$ for a sequence of nonnegative numbers satisfying $\ell_j \to 0$ as $j \to \infty$, and*

$$c(x,y) = \sum_{j \in \mathbb{N}} \ell_j \varphi_j(x) \varphi_j(y)$$

in the sense of absolute and uniform convergence for all x, $y \in D$. Moreover T_c is a trace-class operator and

$$\operatorname{tr}(T_c) = \int_D c(x,x) \, \mathrm{d}x < \infty.$$

Sketch of proof. Let $(\varphi_j)_{j\in\mathbb{N}}$ be an orthonormal basis of $L^2(D)$ such that $T_c\varphi_j = \ell_j\varphi_j$ for some $\ell_j \geq 0$; this is possible by Lemma 4.5.5, and by extending an orthonormal system of eigenvectors forming a basis of the range of T_c to an orthonormal basis of $L^2(D)$.

As pointed out earlier, the functions $(\varphi_i(x)\varphi_j(y))_{i,j\in\mathbb{N}}$ yield an orthonormal basis of $L^2(D\times D)$. Hence

$$c(x,y) = \sum_{i,j\in\mathbb{N}} \ell_{i,j}\varphi_i(x)\varphi_j(y) \in L^2(D\times D),$$

with coefficients $\ell_{i,j} = \int_D\int_D c(x,y)\varphi_i(x)\varphi_j(y)\,\mathrm{d}x\,\mathrm{d}y$. For all i, j, using that $T_c\varphi_j = \ell_j\varphi_j$,

$$\ell_{i,j} = \int_D\int_D c(x,y)\varphi_i(x)\varphi_j(y)\,\mathrm{d}x\,\mathrm{d}y = \ell_i\int_D \varphi_i(y)\varphi_j(y)\,\mathrm{d}y = \ell_i\delta_{ij}.$$

Thus

$$c(x,y) = \sum_{j\in\mathbb{N}} \ell_j\varphi_j(x)\varphi_j(y) \in L^2(D\times D).$$

Wlog we assume in the following $\ell_j > 0$ for all $j\in\mathbb{N}$ since the other terms do not contribute to the series. Then due to $\ell_j\varphi_j(x) = \int_D c(x,y)\varphi_j(y)\,\mathrm{d}y$ and because c is continuous (and thus uniformly continuous) on the compact set $D\times D$, it follows that $\varphi_j : D\to\mathbb{R}$ is continuous.

Set

$$c_n(x,y) := \sum_{j=1}^{n} \ell_j\varphi_j(x)\varphi_j(y).$$

The orthonormality of the φ_j implies that $T_{c_n} : L^2(D)\to L^2(D)$ has the spectral decomposition $T_{c_n}f = \sum_{j=1}^n \ell_j\langle f,\varphi_j\rangle_{L^2(D)}\varphi_j$. We conclude that for all $f\in L^2(D)$

$$(T_c - T_{c_n})f = \sum_{j>n} \ell_j\langle f,\varphi_j\rangle_{L^2(D)}\varphi_j$$

and thus $\langle (T_c - T_{c_n})f, f\rangle_{L^2(D)} = \sum_{j>n}\ell_j\langle f,\varphi_j\rangle^2_{L^2(D)} \geq 0$. As a consequence the continuous function $c(x,x) - c_n(x,x) : D\to\mathbb{R}$ is nonnegative (why?), and thus $c_n(x,x)$ is bounded from above by $c(x,x)$ for all $n\in\mathbb{N}$. By definition $x\mapsto c_n(x,x)$ is monotonically increasing in n, and therefore $c_n(x,x)\to g(x)\leq c(x)$ for some $g : D\to\mathbb{R}$ as $n\to\infty$.

It can be shown that the convergence $c_n(x,x)\to g(x)$ is uniform on D, i.e.

$$\lim_{n\to\infty}\sup_{x\in D}\left|\sum_{j>n}\ell_j\varphi_j(x)^2\right| \to 0. \tag{4.5.4}$$

We don't provide the argument here; see for example p. 245 in [F. Riesz and B. Sz.-Nagy, Functional Analysis, 1955] or Theorem 3.a.1 in [H. König, Eigenvalue Distribution of Compact Operators, 1986] for a complete proof. For all $m\geq n$, by the Cauchy-Schwarz inequality

$$\sup_{x,y\in D}|c_m(x,y) - c_n(x,y)| = \sup_{x,y\in D}\left|\sum_{j=n+1}^{m}\ell_j\varphi_j(x)\varphi_j(y)\right| \leq \sup_{x,y\in D}\left(\sum_{j=n+1}^{m}\ell_j\varphi_j(x)^2\sum_{j=n+1}^{m}\ell_j\varphi_j(y)^2\right)^{1/2},$$

which implies due to (4.5.4) that $(c_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $C^0(D\times D)$. Therefore its limit belongs to $C^0(D\times D)$ as well, and since $\lim_{n\to\infty} c_n = c$ in the sense of $L^2(D\times D)$, we obtain uniform convergence $\lim_{n\to\infty} c_n(x,y) = c(x,y)$ for all x, $y\in D$. Finally,

$$\int_D c(x,x)\,\mathrm{d}x = \int_D \sum_{j\in\mathbb{N}} \ell_j \varphi_j(x)^2\,\mathrm{d}x = \sum_{j\in\mathbb{N}} \ell_j = \operatorname{tr}(T_c) < \infty$$

because $c : D\times D\to\mathbb{R}$ is continuous and thus bounded. □

For every ω, we can formally expand $a_x(\omega) = a(\omega,x)$ as a function of x in the $L^2(D)$ orthonormal basis $(\varphi_j)_{j\in\mathbb{N}}$:

$$a_x(\omega) = a(\omega,x) = \sum_{j\in\mathbb{N}} a_j(\omega)\varphi_j(x) \tag{4.5.5}$$

with the coefficients defined as

$$a_j(\omega) = \int_D a(\omega,x)\varphi_j(x)\,\mathrm{d}x$$

being real-valued RVs. Such an expansion is called a Karhunen-Loève expansion. We next show its convergence.

Theorem 4.5.9 (Karhunen-Loève expansion)**.** *Let $a : \Omega\times D\to\mathbb{R}$ be a measurable centered mean-square continuous stochastic process with $a\in L^2(\Omega\times D, \mathbb{P}\otimes\lambda_d;\mathbb{R})$. There exists an orthonormal basis $(\varphi_j)_{j\in\mathbb{N}}\subseteq L^2(D)$ and nonnegative numbers $(\ell_j)_{j\in\mathbb{N}}$ (we allow for $\ell_j = 0$) such that with $a_j(\omega) = \int_D a(\omega,x)\varphi_j(x)\,\mathrm{d}x$ and $a_x(\omega) = a(\omega,x)$*

$$\lim_{n\to\infty} \sup_{x\in D} \mathbb{E}\left[\left(a_x - \sum_{j=1}^{n} a_j\varphi_j(x)\right)^2\right] = 0. \tag{4.5.6}$$

The coefficients a_j satisfy for all j with $\ell_j > 0$

(i) $\mathbb{E}[a_j] = 0$,

(ii) $\mathbb{E}[a_i a_j] = \delta_{ij}\ell_j$ *and hence* $\mathbb{V}[a_j] = \ell_j$.

Proof. In the following $(\varphi_j)_{j\in\mathbb{N}}$ is an orthonormal basis of $L^2(D)$ that is obtained by extending an orthonormal sequence of eigenvectors φ_j of T_c with eigenvalues $\ell_j > 0$ to an ONB. We extend the sequence of eigenvalues by zeros, i.e. $\ell_j = 0$ for all j for which φ_j belongs to the kernel of T_c.

Joint measurability and Fubini's theorem imply that $a_j : \Omega\to\mathbb{R}$ is measurable and $a_j\in L^2(\Omega,\mathbb{P};\mathbb{R})$ since

$$\begin{aligned}\int_\Omega |a_j(\omega)|^2\,\mathrm{d}\mathbb{P}(\omega) &= \int_\Omega \left|\int_D a(\omega,x)\varphi_j(x)\,\mathrm{d}x\right|^2 \mathrm{d}\mathbb{P}(\omega)\\ &\le \int_\Omega\int_D a(\omega,x)^2\,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega) = \|a\|^2_{L^2(\Omega\times D)},\end{aligned}$$

where we used $\int_D \varphi_j(x)^2\,\mathrm{d}x = 1$. Hence $a_j : \Omega\to\mathbb{R}$ is a RV with finite first and second moment for each j.

Since $(a_x)_{x\in D}$ is centered

$$\mathbb{E}[a_j] = \int_\Omega a_j(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_D \varphi_j(x)\int_\Omega a(\omega,x)\,\mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}x = 0.$$

Moreover since $c(x,y) = \int_\Omega a(\omega,x)a(\omega,y)\,\mathrm{d}\mathbb{P}(\omega)$

$$\begin{aligned}\mathbb{E}[a_i a_j] = \int_\Omega a_i(\omega)a_j(\omega)\,\mathrm{d}\mathbb{P}(\omega) &= \int_\Omega\int_D a(\omega,x)\varphi_i(x)\,\mathrm{d}x\int_D a(\omega,y)\varphi_j(y)\,\mathrm{d}y\,\mathrm{d}\mathbb{P}(\omega)\\ &= \int_D\int_D \varphi_i(x)\varphi_j(y)c(x,y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_D \ell_i\varphi_i(y)\varphi_j(y)\,\mathrm{d}y\\ &= \delta_{ij}\ell_i,\end{aligned}$$

where we used that φ_i is an eigenvector of T_c with eigenvalue ℓ_i, and the $(\varphi_j)_{j\in\mathbb{N}}$ are orthonormal in $L^2(D)$.

Next we show (4.5.6). For $x \in D$ define

$$\epsilon_n(x) := \mathbb{E}\left[\left(a_x - \sum_{j=1}^n a_j\varphi_j(x)\right)^2\right].$$

We need to prove $\sup_{x\in D}\epsilon_n(x) \to 0$. It holds

$$\epsilon_n(x) = \mathbb{E}\left[a_x^2\right] - 2\mathbb{E}\left[a_x\sum_{j=1}^n a_j\varphi_j(x)\right] + \mathbb{E}\left[\sum_{i,j=1}^n a_i a_j\varphi_i(x)\varphi_j(x)\right] \tag{4.5.7}$$

Now

$$\begin{aligned}\mathbb{E}\left[a_x\sum_{j=1}^n a_j\varphi_j(x)\right] &= \int_\Omega a_x(\omega)\sum_{j=1}^n\int_D a_y(\omega)\varphi_j(y)\,\mathrm{d}y\varphi_j(x)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \sum_{j=1}^n\int_D \varphi_j(x)\varphi_j(y)\underbrace{\int_\Omega a_x(\omega)a_y(\omega)\,\mathrm{d}\mathbb{P}(\omega)}_{=c(x,y)}\,\mathrm{d}y\\ &= \sum_{j=1}^n \ell_j\varphi_j(x)^2.\end{aligned}$$

Furthermore by (ii)

$$\mathbb{E}\left[\sum_{i,j=1}^n a_i a_j\varphi_i(x)\varphi_j(x)\right] = \sum_{i,j=1}^n \varphi_i(x)\varphi_j(x)\mathbb{E}[a_i a_j] = \sum_{j=1}^n \ell_j\varphi_j(x)^2.$$

Since $\mathbb{E}[a_x^2] = c(x,x)$ by (4.5.7) we find

$$\epsilon_n(x) = c(x,x) - \sum_{j=1}^n \ell_j\varphi_j(x)^2$$

and an application of Mercer's theorem concludes the proof. □

4.5.2 Uniform measures

We let again $D \subseteq \mathbb{R}^d$ be a bounded compact Lipschitz domain. The Karhunen-Loève expansion provides a method to construct measures on function spaces such as $L^2(D)$, respectively to sample from a $L^2(D)$-valued RV: We adopt now again the viewpoint that $\omega \mapsto a(\omega, \cdot)$ is an $L^2(D)$ valued RV. A sample from this RV can be drawn by first sampling from the RVs $(a_j)_{j\in\mathbb{N}}$, and then computing

$$\sum_{j\in\mathbb{N}} a_j \varphi_j(x).$$

In practice, the sum is truncated after s terms and $\sum_{j=1}^{s} a_j\varphi_j(x)$ is obtained as an approximation.

To continue the discussion of Example 4.5.1, we construct an $L^\infty(D)$ random field through an expansion with real-valued RVs.

Remark 4.5.10. Let V be a separable Banach space and let $X_j : (\Omega, \mathcal{A}) \to (V, \mathcal{B}(V))$, $j \in \mathbb{N}$, be a sequence of RVs converging pointwise to $X : \Omega \to V$. Then for any closed $C \subseteq V$, with $C_\epsilon := \bigcup_{v\in C} B_\epsilon(v)$ where $B_\epsilon(v)$ denotes the open ball of radius $\epsilon > 0$ and center v, we have

$$X^{-1}(C) = \bigcap_{k\in\mathbb{N}} \bigcup_{s\in\mathbb{N}} \bigcap_{j\geq s} X_j^{-1}(C_{1/k}).$$

This set belongs to $\mathcal{A}$ since each X_j is measurable. Thus $X^{-1}(C^c) = (X^{-1}(C))^c \in \mathcal{A}$, which shows that the pre-image of open sets are measurable, and thus X is measurable.

Proposition 4.5.11. *Let $(\varphi_j)_{j\in\mathbb{N}} \subseteq L^\infty(D)$ such that $\|\varphi_j\|_{L^\infty(D)} = 1$ for all j, $m \in L^\infty(D)$ and $(\ell_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$. Furthermore, let $(\xi_j)_{j\in\mathbb{N}}$ be a sequence of iid RVs $\xi_j : \Omega \to [-1, 1]$ such that $\xi_j \sim \text{uniform}(-1, 1)$.*

Then

$$a(\omega, x) := m(x) + \sum_{j\in\mathbb{N}} \ell_j \xi_j(\omega)\varphi_j(x)$$

defines a RV $\omega \mapsto a(\omega, \cdot) \in L^\infty(D)$ satisfying $\mathbb{E}[a] = m$ and for all $\omega \in \Omega$

$$\|a(\omega, \cdot)\|_{L^\infty(D)} \leq \|m\|_{L^\infty(D)} + \sum_{j\in\mathbb{N}} \ell_j, \qquad \operatorname*{ess\,inf}_{x\in D} a(\omega, x) \geq \operatorname*{ess\,inf}_{x\in D} m(x) - \sum_{j\in\mathbb{N}} \ell_j. \tag{4.5.8}$$

Proof. Since each $\xi_j : \Omega \to \mathbb{R}$ is measurable, also $a_n := \sum_{j=1}^{n} \ell_j\xi_j\varphi_j : \Omega \to L^\infty(D)$ is measurable as a sum of measurable functions.

Moreover by the triangle inequality $\|a_n(\omega, \cdot)\|_{L^\infty(D)} \leq \|m\|_{L^\infty(D)} + \sum_{j\in\mathbb{N}} \ell_j < \infty$ for every n, and thus a_n converges pointwise to a measurable function $a : \Omega \to L^\infty(D)$ by Rmk. 4.5.10. This also implies the first bound in (4.5.8), and the second bound follows similarly. □

Example 4.5.12 (Continuation of Example 4.5.1)**.** Assume that the sequence $(\ell_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$ and $m \in L^\infty(D)$ are chosen such that

$$\operatorname*{ess\,inf}_{x\in D} m(x) > \sum_{j\in\mathbb{N}} \ell_j.$$

Let $\varphi_j \in L^\infty(D)$ with $\|\varphi_j\|_{L^\infty(D)} = 1$ for all $j \in \mathbb{N}$ and set $a(\omega, x) = m(x) + \sum_{j\in\mathbb{N}} \ell_j\xi_j(\omega)\varphi_j(x)$.

Then for every $\omega \in \Omega$

$$\operatorname*{ess\,inf}_{x\in D} a(\omega, x) \geq \operatorname*{ess\,inf}_{x\in D} m(x) - \sum_{j\in\mathbb{N}} \ell_j =: a_- > 0 \tag{4.5.9}$$

and

$$\|a(\omega,\cdot)\|_{L^\infty(D)} \le \|m\|_{L^\infty(D)} + \sum_{j\in\mathbb{N}} \ell_j < \infty. \tag{4.5.10}$$

Fix $\omega \in \Omega$. The weak formulation of (4.5.1) is: Find $u \in H^1_0(D)$ such that

$$\int_D a\nabla u(x)^\top \nabla v(x)\,\mathrm{d}x = {}_{H^{-1}}\langle f, v\rangle_{H^1_0} \qquad \forall v \in H^1_0(D), \tag{4.5.11}$$

where the last bracket denotes the application of $f \in H^{-1} = H^1_0(D)'$ to $v \in H^1_0(D)$. Equation (4.5.9) implies coercivity of the bilinear form on the left-hand side, and (4.5.10) implies its boundedness. Therefore, by the Lax-Milgram Lemma, for each $\omega \in \Omega$ there is a unique weak solution to (4.5.11) for the diffusion coefficient $a(\omega,\cdot) \in L^\infty(D)$. Since the solution depends on $a(\omega,\cdot)$ we also write $u(a(\omega),\cdot)$ (or simply $u(\omega,\cdot)$) to emphasize this dependence.

Recall that the inverse problem (4.5.2) is to determine $a(\omega,\cdot) \in L^\infty(D)$ from the measurement $\Phi(a(\omega,\cdot)) + E \in \mathbb{R}^m$. The prior measure is in this case the distribution of the RV $a : \Omega \to L^\infty(D)$. Let

$$T := \begin{cases} [-1,1]^{\mathbb{N}} \to L^\infty(D) \\ (\zeta_j)_{j\in\mathbb{N}} \mapsto m + \sum_{j\in\mathbb{N}} \zeta_j \ell_j \varphi_j(x), \end{cases} \tag{4.5.12}$$

with the $\varphi_j \in L^\infty(D)$ as in Prop. 4.5.11. With the $\xi_j \sim \mathrm{uniform}(-1,1)$ iid as in Prop. 4.5.11, the RV $(\xi_j)_{j\in\mathbb{N}} : \Omega \to [-1,1]^{\mathbb{N}}$ is distributed according to the probability measure $\mu := \otimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ on $[-1,1]^{\mathbb{N}}$, where $\frac{\lambda}{2}$ is the uniform probability measure on $[-1,1]$ (i.e. 1/2 times the Lebesgue measure). Therefore the RV

$$\omega \mapsto a(\omega,\cdot) = T((\xi_j)_{j\in\mathbb{N}})(\omega,\cdot)$$

is distributed according to $T_\sharp\mu$, which is a measure on $L^\infty(D)$. This is the prior.

For computational purposes, it is much more convenient, to formulate the inverse problem in terms of the RV $(\xi_j)_{j\in\mathbb{N}} : \Omega \to [-1,1]^{\mathbb{N}}$: With the uniform prior μ on $[-1,1]^{\mathbb{N}}$, and the measurement $\Phi(u(\omega,\cdot))+E \in \mathbb{R}^m$, the goal is to determine $(\xi_j)_{j\in\mathbb{N}}$. The diffusion coefficient can then be obtained as $a(\omega,x) = T((\xi_j(\omega))_{j\in\mathbb{N}})$.

So far we have constructed a uniform RV on $L^\infty(D)$. In practice, one sometimes assumes knowledge of $\mathbb{E}[a]$ and $\mathrm{cov}(a)$, and wishes to have a random field with the given expectation and covariance. Note that such a random field is not unique, since the expectation and covariance do not uniquely determine a RV. We next give one possible construction. To begin with we show that T_c from the previous section corresponds to the covariance operator, so that the functions φ_j are the eigenvectors of the covariance operator.

Lemma 4.5.13. *Let $a : \Omega \to L^2(D)$ be a RV with finite second moment and such that $\mathbb{E}[a] = m \in L^2(D)$. Then $\mathrm{cov}(a) : L^2(D) \to L^2(D)$ is given by T_c where $c(x,y) = \mathbb{E}[(a(\cdot,x) - m(x))(a(\cdot,y) - m(y))]$ and $c \in L^2(D\times D)$.*

Proof. By definition of the covariance we have for all f, $g \in L^2(D)$ with $C := \mathrm{cov}(a)$

$$\begin{aligned} \langle g, Cf\rangle_{L^2(D)} &= \int_\Omega \langle a(\omega,\cdot) - m(\cdot), f(\cdot)\rangle_{L^2(D)} \langle a(\omega,\cdot) - m(\cdot), g(\cdot)\rangle_{L^2(D)}\,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_D \int_D \mathbb{E}[(a(\cdot,x) - m(x))(a(\cdot,y) - m(y))] f(x)\,\mathrm{d}x g(y)\,\mathrm{d}y \\ &= \langle g, T_c f\rangle_{L^2(D)}. \end{aligned}$$

Furthermore since $\int_\Omega \int_D (a(\omega,x)-m(x))^2 \,\mathrm{d}x\,\mathrm{d}\mathbb{P}(\omega) = \mathbb{E}[\|a-m\|_{L^2}^2] < \infty$

$$\begin{aligned}\int_D \int_D c(x,y)^2 \,\mathrm{d}x\,\mathrm{d}y &= \int_D \int_D \mathbb{E}[(a(\cdot,x)-m(x))(a(\cdot,y)-m(y))]^2 \,\mathrm{d}x\,\mathrm{d}y \\ &\le \int_D \int_\Omega (a(\omega,x)-m(x))^2 \,\mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}x \int_D \int_\Omega (a(\omega,y)-m(y))^2 \,\mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}y \\ &< \infty. \end{aligned}$$

□

Theorem 4.5.9 and Lemma 4.5.13 show that a RV can be expanded in terms the eigenfunctions of the covariance operator (under the assumptions of Thm. 4.5.9). This also works the other way around in the following sense:

Proposition 4.5.14. *Let $(\varphi_j)_{j\in\mathbb{N}}$ be an orthonormal system in $L^2(D)$. Let $\xi_j : \Omega \to [-1,1]$ be a sequence of iid RVs with $\xi_j \sim \mathrm{uniform}(-1,1)$. Then for a sequence $(\ell_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N})$ and $m \in L^2(D)$*

$$a(\omega,x) := m(x) + \sum_{j\in\mathbb{N}} \ell_j \xi_j(\omega)\varphi_j(x)$$

defines a RV $a : \Omega \to L^2(D)$ where $\mathbb{E}[a] = m$ and $\mathrm{Cov}(a) = T_c$ with $c(x,y) = \frac{1}{3}\sum_{j\in\mathbb{N}} \ell_j^2 \varphi_j(x)\varphi_j(y) \in L^2(D\times D)$.

Proof. With $a_s(\omega,x) := m(x) + \sum_{j=1}^s \ell_j\xi_j(\omega)\varphi_j(x)$ it holds for every $\omega\in\Omega$ that $a_s(\omega,\cdot) \to a(\omega,\cdot) \in L^2(D)$ as $s\to\infty$. Measurability of each $\xi_j : \Omega\to\mathbb{R}$ implies measurability of $a_s : \Omega \to L^2(D)$. Thus $a : \Omega \to L^2(D)$ is measurable by Rmk. 4.5.10.

Note that $a_s \to a$ in the topology of $L^2(\Omega,\mathbb{P};L^2(D))$:

$$\lim_{s\to\infty} \mathbb{E}\left[\|a-a_s\|_{L^2(D)}^2\right] = \mathbb{E}\left[\sum_{j>s} \ell_j^2 \xi_j(\omega)^2\right] \le \mathbb{E}\left[\sum_{j>s}\ell_j^2\right] = 0. \tag{4.5.13}$$

The Cauchy-Schwarz inequality implies in particular $a_s \to a$ in the topology of $L^1(\Omega,\mathbb{P};L^2(D))$. Therefore $\mathbb{E}[a] = \lim_{s\to\infty}\mathbb{E}[a_s] = m$.

Next

$$\begin{aligned} c_s(x,y) &:= \mathbb{E}[(a_s(\cdot,x)-m(x))(a_s(\cdot,y)-m(y))] \\ &= \sum_{i,j=1}^s \ell_i\ell_j\varphi_i(x)\varphi_j(x)\mathbb{E}[\xi_i\xi_j] \\ &= \frac{1}{3}\sum_{i=1}^s \ell_i^2\varphi_i(x)\varphi_i(y), \end{aligned}$$

where we used that ξ_i, ξ_j are independent if $i\neq j$ so that in this case $\mathbb{E}[\xi_i\xi_j]=0$, and additionally $\mathbb{E}[\xi_i^2] = \frac{1}{2}\int_{-1}^1 x^2\,\mathrm{d}x = \frac{1}{3}$. From (4.5.13) it follows that $c_s \to c = \mathbb{E}[(a(\cdot,x)-m(x))(a(\cdot,y)-m(y))]$ in $L^2(D\times D)$ as $s\to\infty$. An application of Lemma 4.5.13 concludes the proof. □

4.5.3 Gaussian measures

Definition 4.5.15. Let V be a separable Banach space. A Borel measure μ on $(V, \mathcal{B}(V))$ (i.e. a locally finite measure) is called a **Gaussian probability measure** iff for every $f \in V'$ the measure $f_\sharp \mu$ is Gaussian; here Dirac measures are considered to be Gaussian with zero variance. The measure is said to be **centered** if $\int_{\mathbb{R}} x \, \mathrm{d} f_\sharp \mu(x) = 0$ for all $f \in V'$.

Remark 4.5.16. It can be shown ("Fernique's theorem") that for every Gaussian measure μ there exists $\alpha > 0$ such that $\int_V \exp(\alpha \|x\|_V^2) \, \mathrm{d}\mu(x) < \infty$. Thus a RV with Gaussian distribution has finite moments of all orders.

We concentrate on the case of separable Hilbert spaces H. The expectation and covariance of a probability measure μ are understood as the expectation and covariance of a RV with distribution μ: Assuming they exists, the expectation of μ is $m := \int_H x \, \mathrm{d}\mu(x) \in H$ and the covariance operator $C : H \to H$ of μ is the operator satisfying

$$\langle x, Cy \rangle_H = \int_H \langle h - m, x \rangle_H \langle h - m, y \rangle_H \, \mathrm{d}\mu(h) \qquad \forall x, y \in H.$$

We state the next result without proof.

Theorem 4.5.17. *Let H be a separable Hilbert space. Every Gaussian measure μ on $(H, \mathcal{B}(H))$ has a positive covariance operator $C_\mu : H \to H$ which is of trace-class and satisfies*

$$\operatorname{tr}(C_\mu) = \int_H \|x\|_H^2 \, \mathrm{d}\mu(x) < \infty.$$

Conversely, for every positive trace-class symmetric operator $K : H \to H$ there exists a Gaussian measure μ on H with covariance operator K.

Definition 4.5.18. A RV $X : \Omega \to H$ is called Gaussian, if its distribution is a Gaussian measure on H with expectation $m \in H$ and covariance operator $C \in CL(H, H)$. In this case we write $X \sim \mathcal{N}(m, C)$.

Exercise 4.5.19. Check that if $X \sim \mathcal{N}(m, C)$ and $h \in H$, then $\langle X, h \rangle \sim \mathcal{N}(\langle m, h \rangle, \langle h, Ch, \rangle)_H$.

We mention that, as in the finite dimensional case, Gaussian measures are uniquely determined through their expectation and covariance operator.

There holds the following Karhunen-Loève expansion for Gaussian measures:

Theorem 4.5.20 (Karhunen-Loève expansion)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let $D \subseteq \mathbb{R}^d$ be a compact set and let $a : \Omega \to L^2(D)$ be a RV with distribution $\mathcal{N}(m, C)$. Let $(\varphi_j)_{j \in \mathbb{N}} \subseteq L^2(D)$ be an orthonormal system of eigenvectors of C with positive eigenvalues $(\ell_j)_{j \in \mathbb{N}}$ such that $Cf = \sum_{j \in \mathbb{N}} \ell_j \langle f, \varphi_j \rangle_{L^2(D)} \varphi_j$.*

Then

$$a(\omega, x) = m(x) + \sum_{j \in \mathbb{N}} a_j(\omega) \, \varphi_j(x), \qquad a_j \sim \mathcal{N}(0, \ell_j),$$

in the sense of $L^2(\Omega, \mathbb{P}; L^2(D))$ convergence, where the $(a_j)_{j \in \mathbb{N}}$ are independent real-valued RVs.

Proof. We first show that $a(\omega) - m$ is $\mathbb{P}$-a.e. in $H := \overline{\operatorname{span}\{\varphi_j : j \in \mathbb{N}\}} \subseteq L^2(D)$. Fix $h \in H^\perp$. Then

$$Ch = \sum_{j\in\mathbb{N}} \ell_j \underbrace{\langle h, \varphi_j\rangle_{L^2(D)}}_{=0} \varphi_j = 0.$$

By Exercise 4.5.19

$$\langle a - m, h\rangle_{L^2(D)} \sim \mathcal{N}(0, \langle h, Ch\rangle_{L^2(D)})) = \mathcal{N}(0,0) = \delta_0,$$

where δ_0 denotes the Dirac measure at $0 \in \mathbb{R}$. Hence $\langle a - m, h\rangle_{L^2(D)} = 0$ $\mathbb{P}$-a.e. Therefore $\mathbb{P}$-a.e.

$$(a(\omega) - m) \perp H^\perp \qquad \Leftrightarrow \qquad a(\omega) - m \in H.$$

We conclude that $\mathbb{P}$-a.e. in the sense of $L^2(D)$

$$a(\omega) = m + \sum_{j\in\mathbb{N}} a_j(\omega)\varphi_j \qquad a_j(\omega) = \langle a(\omega) - m, \varphi_j\rangle_{L^2(D)}. \tag{4.5.14}$$

Again by Exercise 4.5.19 holds $a_j \sim \mathcal{N}(0, \ell_j)$, where we used $\langle \varphi_j, C\varphi_j\rangle_{L^2(D)} = \ell_j$. Independence of the $(a_j)_{j\in\mathbb{N}}$ follows by the fact that they are uncorrelated:

$$\begin{aligned}\operatorname{cov}(a_i, a_j) &= \mathbb{E}[\langle a - m, \varphi_i\rangle_{L^2(D)} \langle a - m, \varphi_j\rangle_{L^2(D)}] \\ &= \langle \varphi_i, C\varphi_j\rangle_{L^2(D)} \\ &= \ell_i \delta_{ij}.\end{aligned}$$

Convergence in $L^2(\Omega, \mathbb{P}; L^2(D))$ follows by (4.5.14) and Parseval's identity, which gives

$$\begin{aligned}\left\| a - \left(m + \sum_{j=1}^{s} a_j\varphi_j \right) \right\|^2_{L^2(\Omega;L^2(D))} &= \int_\Omega \left\| a(\omega) - \left(m + \sum_{j=1}^{s} a_j(\omega)\varphi_j \right) \right\|^2_{L^2(D)} \mathrm{d}\mathbb{P}(\omega) \\ &= \sum_{j>s} \mathbb{E}[a_j^2] \\ &= \sum_{j>s} \ell_j \to 0\end{aligned}$$

since C is of trace class by Thm. 4.5.17, which implies $\sum_{j\in\mathbb{N}} \ell_j < \infty$. □

There holds the following converse of the previous theorem, which provides a method to construct Gaussian RVs. The proof is left as an exercise.

Theorem 4.5.21. *Let $D \subseteq \mathbb{R}^d$ be compact. Let $a : \Omega \times D \to \mathbb{R}$ via*

$$a(\omega, x) = m(x) + \sum_{j\in\mathbb{N}} a_j(\omega)\varphi_j(x)$$

where $m \in L^2(D)$, $(\varphi_j)_{j\in\mathbb{N}}$ is an ONS of $L^2(D)$, $(\ell_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$ and the $a_j \sim \mathcal{N}(0, \ell_j)$ are independent.

Then $a \sim \mathcal{N}(m, T_c)$ with covariance function $c \in L^2(D \times D)$ given by

$$c(x, y) := \sum_{j\in\mathbb{N}} \ell_j \varphi_j(x)\varphi_j(y) \qquad \forall x, y \in D.$$

4.5.4 Uninformative priors

Suppose again that we wish to determine X from the measurement Y. If we have no prior information about X, it is tempting to choose a uniform distribution as a prior. However, for example in case $X : \Omega \to \mathbb{R}$, this leads to an **improper** prior with density $\pi_X \equiv 1$, i.e. π_X does not satisfy $\int_{\mathbb{R}} \pi_X(x)\,\mathrm{d}x = 1$. Improper priors may still be used, but are not in line with the theory discussed in this lecture. Furthermore, a uniform distribution should not be interpreted as being "uninformative":

Example 4.5.22. Suppose that we wish to find a parameter X. Assume that we know (apriori) that X belongs to $[0,1]$, but we know nothing else about X. We may choose the prior $X \sim \mathrm{uniform}(0,1)$. Finding $X \in [0,1]$ is equivalent to finding $X^2 \in [0,1]$. Note that the RV X^2 is *not* uniformly distributed on $[0,1]$: $\mathbb{P}[X^2 \le a] = \mathbb{P}[X \le \sqrt{a}] = \sqrt{a}$ and thus X^2 has density $\pi_{X^2}(x) = \frac{1}{2}\frac{1}{\sqrt{x}}$. Hence this prior "favours" smaller values of X^2 over larger values of X^2. This is counterintuitive: If we have no information about $X \in [0,1]$, then we also shouldn't have any information about $X^2 \in [0,1]$.

Definition 4.5.23 (Jeffreys prior)**.** Given a likelihood function $L(y,x) := \pi_{Y|X}(y|x)$ with $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, *Jeffreys prior* is defined as

$$\pi_X(x) \propto \sqrt{\det(I_X(x))},$$

where $I_X(x) \in \mathbb{R}^{n\times n}$ is the *expected Fisher information* of X

$$I_X(x) := \int_{\mathbb{R}^m} \nabla_x \ell(y,x) \cdot \nabla_x \ell(y,x)^\top \; \pi_{Y|X}(y|x)\,\mathrm{d}y, \qquad \ell(y,x) := \log L(y,x).$$

Jeffreys prior satisfies the following form of "invariance": Suppose that X is a $\mathbb{R}^n$ valued RV and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism with nonnegative Jacobian determinant $\det Dg : \mathbb{R}^n \to (0,\infty)$. Then $\tilde{X} := g(X)$ is a RV representing another parametrization of X. To obtain a prior for the reparametrization $\tilde{X}$, we could now proceed in two ways: (i) given the prior $\pi_X(x) \propto \sqrt{\det(I_X(x))}$, the density of $\tilde{X}$ is obtained after a change of variables as $\pi_X(g^{-1}(\tilde{x}))\det Dg^{-1}(\tilde{x})$ (ii) we may set $\pi_{\tilde{X}}(\tilde{x}) = \sqrt{\det(I_{\tilde{X}}(\tilde{x}))}$ obtained with the reparametrized likelihood $\pi_{Y|\tilde{X}}(y|\tilde{x}) = \pi_{Y|X}(y|g^{-1}(\tilde{x}))$. It can be shown that both constructions lead to the same prior.

[Draft of October 6, 2021. Not for dissemination.]

Chapter 5

Numerical Methods

After having carefully defined and analysed the well-posedness and stability of Bayesian inverse problems in a very general setting, we can now address the core task of the course, namely to consider some specific problems and solve them numerically.

We will start by looking at some typical examples of inverse problems in applications that we want to consider, which all have an infinite-dimensional state space and often also an infinite-dimensional (or at least high-dimensional) parameter space. Recall from Chapter 2 that finite dimensional inverse problems, while still possibly leading to non-existence and non-uniqueness problems, do typically not violate Hadamard's third condition of stability in Definition 1.0.1 and – while still interesting – are fundamentally not as challenging.

We then analyse the effect of numerical approximation on the posterior distribution, such as the finite element discretisation of the elliptic PDE problem (4.5.1), before considering the Gaussian case (for prior and additive noise) with a linear forward operator, which can be solved in closed form. In general however, the conditional mean and other statistically interesting quantites need to be estimated via quadrature. In Section 5.4, we recall some classical and more advanced sampling-based quadrature methods for high dimensions and in a first attempt apply them directly to compute the conditional mean with respect to the posterior distribution in Bayesian inverse problems.

However, finally in Sections 5.6.2-5.8 we present the main numerical methods applied in general to solve Bayesian inverse problems in practice: the prevalent Markov chain Monte Carlo (MCMC) method (Sect. 5.6.2-5.6.4), variational methods (Sect. 5.7) and sequential Monte Carlo (Sect. 5.8).

5.1 Examples

Our main focus will be on PDE-constrained Bayesian inverse problems, but more classical examples are typically in the context of integral equations, such as the problem of **X-ray tomography**, or from **spatial statistics**. We will first briefly discuss those before returning to the two PDE-constrained model problems that we have already seen in earlier chapters.

X-ray tomography. Given a bounded domain D, for simplicity $D \subset \mathbb{R}^2$, representing a cross-sectional slice of the object to be studied. Assume that a pointlike X-ray source is placed on one side of the object. The radiation passes through the object and is detected on the other side by an X-ray film or a digital sensor (see Fig. 5.1). It is common to assume that the scattering of the

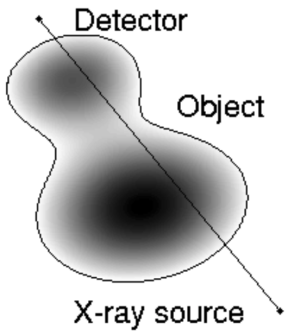


Figure 5.1: X-ray measurement setting.

X-rays by the traversed material is insignificant, i.e., only absorption occurs, and that rays are not deflected through interaction with the material. If we further assume that the **mass absorption coefficient** is proportional to the density of the material, the attenuation $\mathrm{d}I$ of the intensity $I(x)$ along a line segment $\mathrm{d}s$ at a point $x \in D$ is given by

$$\mathrm{d}I = -I(x)\theta(x)\mathrm{d}s$$

where $\theta(x) \geq 0$ is the mass absorption coefficient of the material. We assume that θ is compactly supported in $\overline{D}$ and bounded. If an X-ray is transmitted with intensity I_0^ℓ along a straight line ℓ towards a receiver, the received intensity I_r^ℓ can be obtained from the equation

$$\log I_r^\ell - \log I_0^\ell = \int_{I_0^\ell}^{I_r^\ell} \frac{\mathrm{d}I}{I} = -\int_\ell \theta(x)\,\mathrm{d}s\,. \tag{5.1.1}$$

The inverse problem of X-ray tomography can thus be stated as a problem of integral geometry: Estimate the function $\theta : D \to \mathbb{R}_+$ from the values of its integrals along a set of straight lines $\{\ell(n,s) : n \in \mathbb{R}^2,\ \|n\|_2 = 1,\ s \in \mathbb{R}\}$ passing through D, parametrised by their normal vector n and their distance $s > 0$ from the origin. Denoting the data by $y(n,s) := \log\left(I_r^{\ell(n,s)}/I_0^{\ell(n,s)}\right)$, equation (5.1.1) leads to the linear operator equation

$$y = \mathcal{R}\theta$$

with compact integral operator $\mathcal{R}$, the so-called **Radon transform**.

The nature of the X-ray tomography problem depends on how many lines of integration are available. In the ideal case, we have data along all possible lines passing through the object. The classical results are based on the availability of this complete data. The problem can then be solved explicitly using the **inverse Radon transform**. However, it involves differentiating the data, which is an ill-posed problem in the sense of Hadamard, such that small errors in the data lead to large errors in the solution.

In practice, often only limited-angle data is available and the data is polluted by electronic noise that can be assumed to be additive Gaussian noise $E \sim \mathcal{N}(0,\Sigma)$ to a good approximation. Upon putting a prior measure μ_Θ on the parameter $\Theta \in L^\infty(D)$, this can be formulated as an infinite-dimensional (linear) Bayesian inverse problem of the form (4.1.1),

$$Y = \mathcal{R}\Theta + E,$$

and a typical computational task would be to estimate the conditional mean of the mass absorption coefficient, i.e.

$$\theta_{\mathrm{CM}} = \mathbb{E}[\Theta|Y = y].$$

Pointwise data for a random field - Gaussian process regression or kriging. Many problems in spatial statistics are of the form that a functional quantity is to be estimated from a few point evaluations, a form of statistical interpolation also called **kriging**.

Let $D \subset \mathbb{R}^d$ be a bounded open set. Consider a field $u \in \mathcal{H} = L^2(D; \mathbb{R}^n)$. Assume that we are given noisy observations $\{y_k\}_{k=1}^q$ of a function $g : \mathbb{R}^n \to \mathbb{R}^\ell$ of the field u at a set of points $\{x_k\}_{k=1}^q$. Thus

$$y_k = g(u(x_k)) + \eta_k,$$

where the $\{\eta_k\}_{k=1}^q$ describe the observational noise. Concatenating data, we have

$$y = \mathcal{G}(u) + \eta,$$

where $y = \left(y_1^\top, \ldots, y_q^\top\right)^\top \in \mathbb{R}^{\ell q}$ and $\eta = \left(\eta_1^\top, \ldots, \eta_q^\top\right)^\top \in \mathbb{R}^{\ell q}$. The observation operator $\mathcal{G}$ maps $V = \left(C(\overline{D})\right)^n \subset \mathcal{H}$ to $W = \mathbb{R}^{\ell q}$. The inverse problem is to reconstruct the field u from the data y.

We further assume that the observational noise η is Gaussian $\mathcal{N}(0, \Sigma)$ and specify a prior measure μ_U on the random field U, which is Gaussian $\mathcal{N}(m_0, \mathcal{C}_0)$ and determine the posterior measure $\mu_{U|y}$ for U given y. This is exactly the problem considered in Example 4.3.9. We recall that, provided g and thus $\mathcal{G}$ are measurable, we get

$$\frac{\mathrm{d}\mu_{U|y}}{\mathrm{d}\mu_U}(u) \propto \exp\left(-\frac{1}{2}\|y - \mathcal{G}(u)\|_\Sigma^2\right)$$

which (up to a constant) corresponds to the likelihood. The negative log-likelihood

$$\frac{1}{2}\|y - \mathcal{G}(u)\|_\Sigma^2$$

is the so-called **data misfit potential** (up to the factor $\frac{1}{2}$, which tells us how well function u fits the observed data y.

If $g : \mathbb{R}^n \to \mathbb{R}^\ell$ is linear, so that $\mathcal{G}(u) = Au$ for some linear operator $A : V \to W$, then the posterior measure $\mu_{U|y}$ is also Gaussian with a mean and covariance operator that can be computed explicitly, as we will do in Section 5.3. This is called a **Gaussian process** and the data fitting approach is called **Gaussian process regression**. It has many interesting and favourable properties and is very popular in computational statistics.

Inverse heat equation. Let us return to the motivating example in Section 1.1, the inverse heat equation. Recall the one-dimensional heat equation

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}, \qquad \text{for } \ 0 < x < 1, \ t > 0,$$

with thermal diffusivity $\alpha > 0$ and boundary/inital conditions

$$u(0,t) = u(1,t) = 0, \quad \text{for } \ t > 0, \quad \text{and} \quad u(x,0) = u_0(x), \quad \text{for } \ 0 < x < 1.$$

The inverse problem we considered in Section 1.1 was to find the initial temperature profile u_0 given the profile $u(\cdot, T)$ at some time $T > 0$. As shown, the forward problem admitted the explicit solution

$$u(x,t) = \sum_{n=1}^{\infty} \theta_n e^{-(n\pi)^2 \alpha t} \sin(n\pi x) \,,$$

where $\theta_n = \langle u_0, \sin(n\pi\cdot)\rangle_{L^2(0,1)}$ are the Fourier-sine-coefficients of the initial condition u_0.

Denoting by y_n the Fourier-sine coefficients of the measured data at time $T > 0$, assuming an additive measurement noise, we obtain the following infinite-dimensional (linear) Bayesian inverse problem: to find posterior distribution $\mu_{\Theta|y}$ for the Fourier coefficients Θ of u_0 (understood as a random sequence in ℓ^1) such that

$$Y = \Lambda\Theta + E$$

with the linear operator $\Lambda : \ell^1 \to \ell^1$ that can be represented as an infinite diagonal matrix with diagonal entries $\Lambda_{nn} := e^{-(n\pi)^2 \alpha T}$.

This viewpoint shows very clearly how the inversion leads to an exponential amplification of any errors in the higher frequencies. In the case of a Gaussian prior and a Gaussian measurement error the posterior distribution is again Gaussian and the problem can be solved explicitly. We will come back to this example in Section 5.3.

Subsurface flow, heat conduction, impedance tomography. However, the main model problem we will consider in the rest of this chapter, is the problem to identify the diffusion coefficient a in a stationary diffusion problem, the elliptic PDE defined in Example 4.5.1.

This model problem is ubiquitous in many fields of mathematics due to the many important applications it appears in centrally. Whether we are interested in heat conduction, electrostatics, magnetostatics, porous media flow or even radiation shielding, the central mechanism in all those physical processes (in practically relevant regimes) is diffusion and a central question in practice is often how to estimate the diffusion coefficient non-destructively (or with minimal "destruction") from indirect measurements.

The Bayesian inverse problem associated with this problem has already been extensively described and analysed in Section 4.5. The only important point we want to add here is that even when the forward operator Φ in (4.5.2) is linear as a map of u, i.e., $\Phi(a) = Bu$ as defined in Example 4.5.1, since u depends nonlinearly on a the forward operator Φ will always be nonlinear making it the pre-eminent infinite-dimensional non-linear (Bayesian) inverse problems.

There are many more examples of important inverse problems in applications – some of them also studied in our groups in Heidelberg – such as **inverse scattering** (geophysics, MRT), **inverse source problems** (Tsunami prediction, subsurface pollution), **data assimilation** (weather prediction), **parameter estimation** (pattern formation in developmental biology) or **epidemiology** (COVID-19 modelling and prediction).

For more examples see [Stuart, 2010, Chap. 3] and [Kaipio & Somersalo, 2004, Chap. 6].

5.2 Discretisation

To numerically solve such an infinite-dimensional inverse problem it is of course necessary to discretise the problem. The approximation error then leads to a bias in the posterior distribution and in

any derived quantities, such as the conditional mean, that needs to be estimated. Let us consider the model problems from Section 5.1

For the X-ray tomography problem, the domain D is commonly subdivided using a uniform Cartesian grid into pixels (or voxels in three dimensions) and the absorption coefficient θ is then approximated by a piecewise constant approximation θ_h, where h denotes the mesh size, see [Kaipio & Somersalo, 2004, Sect. 6.1]. The data is typically already given in discrete form. The same holds for Gaussian process regression.

In the inverse heat equation problem, also described Section 5.1, the forward problem was already represented in an orthonormal system. In that case the discretisation is very naturally (and in some sense optimally) achieved by truncating the infinite series expansions after a suitable number of terms N. In general, however, finding the initial condition of an evolution equation, e.g. in the context of the Navier-Stokes, Euler or shallow water equations in data assimilation for weather forecasting or for tsunami prediction, the forward problem can not be solved explicitly and the forward operator needs to be discretised, e.g. via finite elements.

Thus, let us now discuss the final and main model problem from Section 5.1, the elliptic PDE with an unknown diffusion coefficient.

5.2.1 Finite element analysis of the elliptic model problem

The references for the theoretical results quoted in this section are:

- J. Charrier, R. Scheichl, A.L. Teckentrup, Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods, *SIAM J. Num. Anal.* **51**:322–352, 2013.
- I.G. Graham, F.Y. Kuo, J. Nichols, R. Scheichl, C. Schwab, I.H. Sloan, Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients, *Numer. Math.* **131**:329–368, 2015.
- V.H. Hoang, C. Schwab, A.M. Stuart, Complexity analysis of accelerated MCMC methods for Bayesian inversion, *Inverse Probl.* **29**:085010, 2013.
- R. Scheichl, A.M. Stuart, A.L. Teckentrup, Quasi-Monte Carlo and multilevel Monte Carlo methods for computing posterior expectations in elliptic inverse problems, *SIAM J. Uncertain. Quantif.* **5**:493–518, 2017.
- A.L. Teckentrup, R. Scheichl, M.B. Giles, E. Ullmann, Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients, *Numer. Math.* **125**:569–600, 2013.

For completeness we recall the definition of the model problem from Examples 4.5.1 and 4.5.12:

Let $D \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain and consider the weak formulation of the elliptic PDE (4.5.1): Find $u \in H_0^1(D)$ such that

$$\int_D a\nabla u(x)^\top \nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x \qquad \forall v \in H_0^1(D), \tag{5.2.1}$$

where for simplicity we assume that $f \in L^2(D)$ and that it is known and not random. Furthermore, as above $a \in L^\infty(D)$ with $\operatorname{essinf}_{x\in D} a(x) > 0\}$ and the forward (observation) operator $\Phi(a) := Bu \in \mathbb{R}^m$ for some bounded linear operator $B : H_0^1(D) \to \mathbb{R}^m$ – although for some of the results below

the uniform ellipticity is not needed and we could also consider nonlinear functionals. It is also possible to extend the analysis to a random or less regular source term f. The inverse problem is to find the diffusion coefficient $a \in L^\infty(D)$ from noisy measurements

$$Y = \Phi(a) + E, \tag{5.2.2}$$

with $E \sim \mathcal{N}(0, \Sigma)$ for an SPD matrix $\Sigma \in \mathbb{R}^{m\times m}$. To solve this inverse problem we model a as a RV from Ω to $L^\infty(D)$, for a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and associated prior measure μ_a. The posterior can then be defined via Thm. 4.3.8.

However, to solve this problem numerically, we need to discretise (5.2.1). Let us recall without proofs some of the results from Chapter II.7 of the lecture *"High-Dimensional Approximation and Applications in Uncertainty Quantification"* (HDAUQ), taught in SS20 at Heidelberg. (For more details and proofs see the notes on Moodle.): We assume that for almost all $\omega \in \Omega$ ($\mathbb{P}$-a.s.), realizations $a(\cdot, \omega)$ of the coefficient function a are strictly positive, lie in $L^\infty(D)$ and satisfy

$$0 < a_{\min}(\omega) \le a(x, \omega) \le a_{\max}(\omega) < \infty \quad \text{for a.e. } x \in D, \tag{5.2.3}$$

where

$$a_{\min}(\omega) := \operatorname*{ess\,inf}_{x\in D} a(x, \omega), \qquad a_{\max}(\omega) := \operatorname*{ess\,sup}_{x\in D} a(x, \omega). \tag{5.2.4}$$

The following theorem is a simple consequence of the Lax-Milgram Lemma [HDAUQ, Lem. B.5].

Theorem 5.2.1. *$\mathbb{P}$-a.s. problem (5.2.1) has a unique solution $u(\cdot, \omega) \in H^1_0(D)$ and*

$$|u(\cdot, \omega)|_{H^1(D)} \le C a^{-1}_{\min}(\omega) \|f\|_{L^2(D)}\,.$$

If $a^{-1}_{\min} \in L^p(\Omega)$, for some $p \in [1, \infty]$, then

$$\|u\|_{L^p(\Omega; H^1_0(D))} \le C \|a^{-1}_{\min}\|_{L^p(\Omega;\mathbb{R})} \|f\|_{L^2(D)}\,.$$

Let $U_h \subset H^1_0(D)$ denote a closed subspace, e.g., the finite element (FE) space of piecewise polynomial functions with respect to a triangulation $\mathcal{T}_h$ of D with mesh width $h > 0$ [HDAUQ, App. B]. Suppose $u_h : \Omega \to U_h$ satisfies $\mathbb{P}$-a.s.

$$\int_D a(x, \omega) \nabla u_h(x, \omega)^\top \nabla v_h(x)\, \mathrm{d}x = \int_D f(x) v_h(x)\, \mathrm{d}x \quad \forall v_h \in U_h\,. \tag{5.2.5}$$

Since U_h is a closed subspace of $H^1_0(D)$ with norm $|\cdot|_{H^1(D)}$ all the above results hold in an identical form also for u_h; in particular:

Theorem 5.2.2. *The results and the bounds in Theorem 5.2.1 hold under the same assumptions on a and f also for the FE system (5.2.5) and its solution u_h.*

To bound the FE error we also need a regularity assumption.

Assumption 5.2.3. $\mathbb{P}$-a.s. for all $\omega \in \Omega$, $u(\cdot, \omega) \in H^2(D)$ and there exists a $q \in [1, \infty]$ such that

$$\|u\|_{L^q(\Omega; H^2(D))} \le C \|f\|_{L^2(D)}\,.$$

Remark 5.2.4. Sufficient conditions for $u(\cdot,\omega) \in H^2(D)$ are that D is convex, that $a(\cdot,\omega)$ is Lipschitz continuous and that (5.2.3) holds (see, e.g., P. Bastian, "Scientific Computing with PDEs", Lecture Notes, U. Heidelberg, 2019, Sect. 8.3). For a uniform distribution, as in Section 4.5.2, Assumption 5.2.3 holds with $q = \infty$. However, a more commonly used type of prior distribution, especially in subsurface flow, is a **lognormal distribution** for a with **Matérn covariance**. In that case, $\log a$ is a Gaussian field that admits a Karhunen-Loève expansion as in Theorem 4.5.20, and it can be shown [Charrier et al, 2013] that Assumption 5.2.3 holds for all $q < \infty$.

Theorem 5.2.5 (FE error bounds [Charrier et al, 2013], [Teckentrup et al, 2013])**.** *Let Assumption 5.2.3 hold and let* $\sqrt{a_{\max}/a_{\min}} \in L^r(\Omega)$ *with* $q, r \in [1,\infty]$*. Suppose* $U_h \subset H_0^1(D)$ *is the piecewise linear FE space associated with a triangulation* $\mathcal{T}_h$ *of* D*. Then, for any* $p \in [1,\infty]$ *with* $\frac{1}{p} \geq \frac{1}{q} + \frac{1}{r}$*,*

$$\|u - u_h\|_{L^p(\Omega;H_0^1(D))} \leq Ch\|f\|_{L^2(D)}\,.$$

Moreover, for any bounded linear operator $B : H_0^1(D) \to \mathbb{R}^m$*,*

$$\|Bu - Bu_h\|_{L^p(\Omega;\mathbb{R}^m)} \leq Ch^2\,. \tag{5.2.6}$$

We are now in a position to extend these results to bound the bias in the posterior measure and in the conditional mean of any derived quantities of interest due to the FE approximation.

Theorem 5.2.6. *Let us assume that* $a_{\min}^{-1} \in L^2(\Omega)$ *and the assumptions of Theorem 5.2.5 hold for* $p = 2$*. Suppose* $B : H_0^1(D) \to \mathbb{R}^m$ *is a bounded linear operator and let the (exact) forward operator* $\Phi : L^\infty(D) \to \mathbb{R}^m$ *be defined by* $\Phi(a) := Bu$*. Under the noise model* (5.2.2) *this induces the posterior measure* $\nu := \mu_{a|y}$ *on the diffusion coefficient* $a \in L^\infty(D)$*, as described in Chapter 4. In the same way, the discretised observation operator* $\Phi_h(a) := Bu_h$ *induces an approximate posterior measure* ν_h *on* a *and*

$$D_H(\nu,\nu_h) \leq Ch^2\,. \tag{5.2.7}$$

Moreover, for any bounded linear operator $G : H_0^1(D) \to \mathbb{R}$*, the approximation error in the posterior expectation of the functional* $\Psi(a) := Gu$*, which is approximated by* $\Psi_h(a) := Gu_h$ *can be bounded as*

$$\left|\mathbb{E}_\nu[\Psi(a)] - \mathbb{E}_{\nu_h}[\Psi_h(a)]\right| \leq Ch^2\,. \tag{5.2.8}$$

Proof. The two measures ν and ν_h satisfy the assumptions of Theorem 4.4.1. Thus, the bound on the Hellinger distance follows directly by applying the bound (5.2.6) in Theorem 4.4.1.

For the bound on the posterior expectations, we note first that due to the additive Gaussian noise assumption, for any measurable function $f : L^\infty(D) \to \mathbb{R}^m$ and for $q \in \mathbb{N}$,

$$\left|\mathbb{E}_\nu[f(a)^q]\right| \leq \mathbb{E}_\nu[|f(a)|^q] \leq \frac{1}{Z_\nu}\mathbb{E}_{\mu_a}[|f(a)|^q], \tag{5.2.9}$$

which follows immediately by taking the supremum of the Radon-Nikodym derivative out of the integral and bounding it by $\frac{1}{Z_\nu}$, where Z_ν is the normalization constant for ν (as in (4.4.1b)). The analogous bound holds for ν_h.

Now we use triangle inequality to separate the error into the FE error in the posterior measure and the FE error in approximating the target functional, i.e.

$$\left|\mathbb{E}_\nu[\Psi(a)] - \mathbb{E}_{\nu_h}[\Psi_h(a)]\right| \leq \left|\mathbb{E}_{\nu_h}[\Psi(a) - \Psi_h(a)]\right| + \left|\mathbb{E}_\nu[\Psi(a)] - \mathbb{E}_{\nu_h}[\Psi(a)]\right| \tag{5.2.10}$$

The bound on the first term follows from (5.2.9) and (5.2.6) (with G instead of B). For the second term, we proceed as in the proof of Lemma 3.7.7, i.e.

$$\begin{aligned}\left|\mathbb{E}_\nu[\Psi(a)] - \mathbb{E}_{\nu_h}[\Psi(a)]\right| &\le 2D_{\mathrm{H}}(\nu,\nu_h)\left(\int_\Omega |\Psi(a)|^2\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_a}+\frac{\mathrm{d}\nu_h}{\mathrm{d}\mu_a}\right)\mathrm{d}\mu_a\right)^{1/2}\\ &\le Ch^2\|G\|_{H^{-1}(D)}\|u\|_{L^2(\Omega;H_0^1(D))}\end{aligned}$$

where in the last step we have used (5.2.7), (5.2.9) and the boundedness of G. The result then follows from Theorem 5.2.1. □

Remark 5.2.7. (a) Similar results can be proved for Fréchet-differentiable nonlinear functionals $G : H_0^1(D) \to \mathbb{R}$ of the PDE solution with $\Psi(a) := G(u)$ or for other Fréchet-differentiable nonlinear functionals $\Psi : L^\infty \to \mathbb{R}$ (with measurable Fréchet derivative). For the former see [Scheichl et al, 2017]; the latter can be proved in a similar way.

(b) Another approximation error that we have not discussed so far concerns the numerical approximation of the prior distribution.

In case of the Karhunen-Loève expansion (cf. Sect. 4.5.1), a natural way to discretise the prior is truncation of the series expansion (4.5.5) at a suitably large index $s \in \mathbb{N}$. For the case of Matérn covariances, both in the uniform and in the lognormal case it can be shown, e.g., in [Graham et al, 2015] that there exists a $\chi > 0$ such that

$$\|Bu_h - Bu_{s,h}\|_{L^p(\Omega)} \le Cs^{-\chi}\,.$$

where $u_{s,h}$ is the solution to a FE system like (5.2.5) but with truncated coefficient function a_s instead of a and the value of χ depends on the smoothness parameter in the Matérn covariance. From this it can again be deduced [Hoang et al, 2013] that

$$D_{\mathrm{H}}(\nu_h,\nu_{s,h}) \le Cs^{-\chi}\,.$$

5.3 Linear problems and the Laplace approximation

The key references for this section are:

- S. Ghosal, A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*, Cambridge University Press, 2017.
- C. Schillings, B. Sprungk, P. Wacker, On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems, *Numer. Math.* **145**:915–971, 2020.
- R. Wong, *Asymptotic Approximations of Integrals*, SIAM, Philadelphia, 2001.

If the observational noise is additive and Gaussian, the prior μ_X is Gaussian and the forward operator Φ is linear, then the posterior measure $\mu_{X|y}$ is also Gaussian and can be given explicitly.

Theorem 5.3.1. *Let H be a separable Hilbert space and $X : \Omega \to H$ a RV with prior distribution $\mathcal{N}(\overline{x}, C)$ with positive covariance operator C. Let $W = \mathbb{R}^m$ and assume $E : \Omega \to \mathbb{R}^m$ is a Gaussian RV independent of X that is distributed according to $\mathcal{N}(0, \Sigma)$ with SPD covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Suppose furthermore that the forward operator $\Phi : H \to \mathbb{R}^m$ is linear, i.e. $\Phi(x) = Ax$ and $Y = AX + E$. Then the posterior measure $\mu_{X|y}$ is Gaussian $\mathcal{N}(x_{\mathrm{CM}}, C_{X|y})$ with*

$$x_{\mathrm{CM}} := \overline{x} + CA^*(\Sigma + ACA^*)^{-1}(y - A\overline{x}) \tag{5.3.1}$$

$$C_{X|y} := C - CA^*(\Sigma + ACA^*)^{-1}AC \tag{5.3.2}$$

Proof. For the proof we restirct ourselves to $H = \mathbb{R}^s$, but most steps extend straightforwardly also to infinite dimensions. The proof of this extension is left as an exercise.

Given the additive model (4.1.1) for the observable RV Y, the RV $Z := \binom{X}{Y}$ is jointly Gaussian

$$Z = \begin{pmatrix} I & 0 \\ A & I \end{pmatrix} \begin{pmatrix} X \\ E \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \overline{x} \\ A\overline{x} \end{pmatrix}, \begin{pmatrix} C & CA^* \\ AC & \Sigma + ACA^* \end{pmatrix} \right) =: \mathcal{N}(m_Z, C_Z) \tag{5.3.3}$$

This follows directly from the fact that for any random variables X_1 and X_2 under linear transformations A_1 and A_2 the covariance operator satisfies

$$\mathrm{cov}(A_1X_1, A_2X_2) = A_1\mathrm{cov}(X_1, X_2)A_2^*,$$

as pointed out in finite dimensions already in Example 3.4.2.

Since Σ and C are SPD, the bottom-right-block $C_Y := \Sigma + ACA^*$ is also SPD and we can block-LDL^T factorise the covariance matrix of (X, Y) to give

$$\begin{pmatrix} C & CA^* \\ AC & C_Y \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -C_Y^{-1}AC & I \end{pmatrix} \begin{pmatrix} \left(C - CA^*C_Y^{-1}AC\right)^{-1} & 0 \\ 0 & C_Y^{-1} \end{pmatrix} \begin{pmatrix} I & -CA^*C_Y^{-1} \\ 0 & I \end{pmatrix}$$

and thus, since

$$\begin{pmatrix} I & -CA^*C_Y^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x - \overline{x} \\ y - A\overline{x} \end{pmatrix} = \begin{pmatrix} x - x_{\mathrm{CM}} \\ Y - A\overline{x} \end{pmatrix}$$

with x_{CM} as defined in (5.3.1) and with $C_{X|y} := C - CA^*C_Y^{-1}AC$, it follows that

$$\begin{aligned} (z - m_Z)^*C_Z^{-1}(z - m_Z) &= \begin{pmatrix} x - \overline{x} \\ y - A\overline{x} \end{pmatrix}^* \begin{pmatrix} C & CA^* \\ AC & C_Y \end{pmatrix}^{-1} \begin{pmatrix} x - \overline{x} \\ y - A\overline{x} \end{pmatrix} \\ &= \begin{pmatrix} x - x_{\mathrm{CM}} \\ y - A\overline{x} \end{pmatrix}^* \begin{pmatrix} C_{X|y}^{-1} & 0 \\ 0 & C_Y^{-1} \end{pmatrix} \begin{pmatrix} x - x_{\mathrm{CM}} \\ y - A\overline{x} \end{pmatrix} \end{aligned}$$

Due to the diagonal structure of the covariance and the formula for conditional densities, $\pi_{X,Y}(x, y) = \pi_{X|Y}(x|y)\pi(y)$, this completes the proof. The employed technique – although often presented in a very complicated way – is sometimes called *"completing the square"*. □

In principle, this solves the problem in the linear Gaussian case and for low- to intermediate-dimensional problems. In that case, it is possible to assemble and factorise $C_{X|y}$ and to perform inference from this posterior distribution. In high dimensions, factorisation might be prohibitive and it is necessary to consider alternatives – this will be the focus of the next three sections.

However, due to its importance and explicit tractability, there is a large body of literature on efficient numerical methods specifically for the Gaussian case.

Note that in the linear Gaussian case, the conditional mean x_{CM} is in fact identical to the MAP point, which for $H = \mathbb{R}^s$ can be computed as

$$x_{\mathrm{MAP}} = \mathrm{argmax}_{x\in\mathbb{R}^s}\, \pi_{X|Y}(x|y) = \mathrm{argmin}_{x\in H}\, \frac{1}{2}\|y - Ax\|_\Sigma^2 + \frac{1}{2}\|x - \overline{x}\|_C^2$$

and is in fact also identical to the solution of a generalised Tikhonov-regularised system as discussed in Section 2.6 (with suitable norms on the parameter space X and on the observation space Y, in the notation there).

As in numerical optimisation for general, nonlinear deterministic problems, and in particular for the solution of Tikhonov-regularised, nonlinear inverse problems, such as (2.6.3), a powerful way to accelerate numerical methods is a change of metric, also referred to as **preconditioning**. For simplicity, let us restrict ourselves to a finite dimensional parameter space $H = \mathbb{R}^s$ and to additive Gaussian noise E independent of X and distributed according to $\mathcal{N}(0, \Sigma)$ again. A simple and popular preconditioning technique that is easy to understand and to apply to general Bayesian inverse problems is the so-called Laplace approximation of the posterior distribution $\mu_{X|y}$.

Definition 5.3.2 (Laplace, 1774)**.** Suppose the forward operator $\Phi : \mathbb{R}^s \to \mathbb{R}^m$ and the prior distribution $\mu_X(\,\mathrm{d}x) = \pi_X(\,\mathrm{d}x)$ are such that $\Phi, \pi_0 \in C^2(\mathcal{S}_X)$, for $\mathcal{S}_X := \{x \in \mathbb{R}^s \colon \pi_X(x) > 0\}$. Let $\Psi : \mathcal{S}_X \to \mathbb{R}$ be given by $\Psi(x) := \frac{1}{2}\|y - \Phi(x)\|_\Sigma^2 - \log \pi_X(x)$ and assume that Ψ has a unique minimiser $x_{\mathrm{MAP}} \in \mathcal{S}_X$ satisfying

$$\nabla\Psi(x_{\mathrm{MAP}}) = 0 \quad \text{and} \quad \nabla^2\Psi(x_{\mathrm{MAP}}) \ \text{ is SPD.}$$

Then, the **Laplace approximation** of $\mu_{X|y}$ is given by the Gaussian measure

$$\mathcal{L}_{\mu_{X|y}} := \mathcal{N}(x_{\mathrm{MAP}}, C_{\mathrm{MAP}}) \quad \text{with} \quad C_{\mathrm{MAP}}^{-1} := \nabla^2\Psi(x_{\mathrm{MAP}})\,. \tag{5.3.4}$$

Finding the minimiser x_{MAP} can be achieved with classical unconstrained, nonlinear minimisation methods, e.g. quasi-Newton methods with low rank updates (such as SR1 or BFGS) and a globalisation strategy (such as a line search or trust region method).

A common and desirable situation in Bayesian inverse problems is concentration of the posterior around the true parameter, especially in the small noise or large data limit. However, often the posterior concentrates more or less strongly in different directions. Optimisation works well in that context and can thus be used to efficiently construct the Laplace approximation as a suitable reference measure to remove the degeneracy and to precondition sampling or quadrature algorithms, such as importance sampling or Markov chain Monte Carlo (see Sect. 5.4 and 5.6.4 below).

To see why the Laplace approximation is useful in this limit, let us first consider a scaled version $n\Psi_n$ of the posterior log-likelihood Ψ with

$$\Psi_n(x) := \frac{1}{2}\|y - \Phi(x)\|_\Sigma^2 - \frac{1}{n}\log \pi_X(x)\,, \tag{5.3.5}$$

e.g. if the measurement error E_n is assumed to decrease as n increases, such that $E_n \sim \mathcal{N}(0, \frac{1}{n}\Sigma)$. The scaled posterior measure is then

$$\nu_n(\,\mathrm{d}x) = \frac{1}{Z_n}\exp\left(-n\left(\tfrac{1}{2}\|y - \Phi(x)\|_\Sigma^2\right)\right)\mu_x(\,\mathrm{d}x), \quad Z_n := \int_{\mathbb{R}^s} \exp\left(-n\left(\tfrac{1}{2}\|y - \Phi(x)\|_\Sigma^2\right)\right)\mu_x(\,\mathrm{d}x).$$

We can see that the weight function concentrates more and more around the MAP point $x_{\mathrm{MAP},n}$ as $n \to \infty$. It is a classical result [Wong, 2001] that integrals with respect to such a measure ν_n can be well approximated via integrals with respect to the Laplace approximation, but even a stronger convergence result can be proved. We will only state the result informally and refer to the original paper for a complete statement and for the proof.

Theorem 5.3.3 (Schillings et al, 2020)**.** *Suppose $\Psi_n \in C^3(\mathcal{S}_X)$ and satisfies further technical conditions. Suppose further that $\lim_{n\to\infty} x_{MAP,n} \in \mathcal{S}_X$ and $\lim_{n\to\infty} \nabla^2 \Psi_n(x_{MAP,n})$ exist. Then*

$$D_{\mathrm{H}}(\nu_n, \mathcal{L}_{\nu_n}) \leq C n^{-1/2}.$$

Remark 5.3.4. To finish let us draw another link to generalised Tikhonov regularisation. The scaled posterior log-likelihood in (5.3.5) is the same as the generalised Tikhonov functional (cf. Sect. 2.6)

$$\Psi_{\alpha,\delta}(x) := \frac{1}{2}\|y - \Phi(x)\|_\Sigma - \alpha \log \pi_X(x) \tag{5.3.6}$$

with penalisation functional $\log \pi_X : \mathbb{R}^s \to \mathbb{R}$, with noise level $\delta = \frac{1}{n}$ and with regularisation parameter $\alpha = \frac{1}{n}$. In the Bayesian setting, adding the regularisation parameter α corresponds to "flattening" the prior distribution π_X to $\pi_X^{1/n}$ as $n \to \infty$, thus reducing its influence.

Note also that the choice $\alpha = 1/n$ for a Gaussian prior in (5.3.6) and for $\delta = 1/n$ leads to a convergent regularisation method in the sense of Sect. 2.4, since $\delta/\sqrt{\alpha} \to 0$ as $\delta \to 0$.

There is more rigorous mathematical theory on the topic of posterior consistency [Ghosal, van der Vaart, 2017], but we will not discuss this any further. The explicit form of the posterior measure in the linear Gaussian case and the Laplace approximation play a central role in filtering, in the (extended) Kalman filter and we will come back to this point in Section 5.8.

5.4 High-dimensional quadrature

Even though mathematically speaking the solution to a Bayesian inverse problem is the posterior distribution $\mu_{X|y}$, it is of little practical value (especially in high dimensions). As highlighted already, the central task in Bayesian inference is the computation of expectations of certain functionals of the parameter with respect to the posterior, so called **statistics** or **quantities of interest**.

Care is required, when designing quadrature algorithms in high dimensions; the computational cost of simple tensor product rules of standard 1D quadrature rules (as presented for example in Numerik 0) explodes as the dimension $s \to \infty$. The workhorses in high dimensions are sampling based methods that do not suffer from this so-called **"curse of dimensionality"**, and in particular Monte Carlo type methods.

Let us recall some of the main methods for high dimensional quadrature. For details we refer again to the notes of the HDAUQ course that was taught in the SS 2020 at Heidelberg.

5.4.1 Monte Carlo quadrature

Consider again the general setting of a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a RV $X : \Omega \to V$ mapping to a seperable Banach space V with measure μ_X, for the moment assumed to be available explicitly.

For any measurable $F\colon V \to \mathbb{R}$ (for simplicity one-dimensional), given N realisations $x^{(1)}, \dots, x^{(N)}$ of **independent** RVs $X^{(i)} \sim \mu_X$, a practical method to compute the high-dimensional integral

$$\mathbb{E}[F(X)] = \int_V F(x)\ \mathrm{d}\mu(x), \quad \text{with} \quad \mathbb{E} := \mathbb{E}_{\mu_X}$$

is the **Monte Carlo (MC) method**

$$\frac{1}{N}\sum_{i=1}^{N} F(x^{(i)}) \ \approx\ \mathbb{E}[F(X)]$$

For the associated estimator $\widehat{F(X)}_N := \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} F(X^{(i)})$ we have the following result.

Proposition 5.4.1. *Let $X \in L^2(\Omega; V)$ and iid $X^{(i)} \sim \mu_X$, $i \in \mathbb{N}$. Then*

$$\widehat{F(X)}_N = \frac{1}{N}\sum_{i=1}^{N} F(X^{(i)}) \ \xrightarrow[N\to\infty]{\mathbb{P}\text{-a.s.}}\ \mathbb{E}[F(X)], \tag{5.4.1}$$

as well as

$$\sqrt{N}\left(\widehat{F(X)}_N - \mathbb{E}[F(X)]\right) \ \xrightarrow[N\to\infty]{d}\ \mathcal{N}\Big(0, \mathbb{V}(F(X))\Big) \tag{5.4.2}$$

and

$$\mathbb{E}\left[\left|\widehat{F(X)}_N - \mathbb{E}[F(X)]\right|^2\right] = \frac{\mathbb{V}(F(X))}{N}. \tag{5.4.3}$$

Proof. The convergence results (5.4.1) and (5.4.2) follow directly from the Law of Large Numbers and the Central Limit Theorem. To see (5.4.3), note that due to the independence of the $X^{(i)}$,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} F(X^{(i)}) - \mathbb{E}\left[F(X)\right]\right|^2\right] &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^{N}\Big(F(X^{(i)}) - \mathbb{E}\left[F(X)\right]\Big)\right)^2\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\Big(F(X^{(i)}) - \mathbb{E}\left[F(X)\right]\Big)\Big[F(X^{(j)}) - \mathbb{E}\left[F(X)\right]\Big)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\Big(F(X^{(i)}) - \mathbb{E}\left[F(X)\right]\Big)^2\right] \\
&\quad + \frac{1}{N^2}\sum_{i\neq j}^{N}\underbrace{\mathbb{E}\left[\Big(F(X^{(i)}) - \mathbb{E}\left[F(X)\right]\Big)\right]}_{=0}\underbrace{\mathbb{E}\left[\Big(F(X^{(j)}) - \mathbb{E}\left[F(X)\right]\Big)\right]}_{=0} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(F(X^{(i)})) = \frac{1}{N}\mathrm{Var}(F(X)).
\end{aligned}$$

□

As discussed in Section 5.2, if V is infinite dimensional or F is given as $F(X) = \Psi(\mathcal{G}(X))$, for some operator $\mathcal{G} : V \to W$ and an infinite dimensional latent space W with $\Psi : W \to \mathbb{R}$, it is necessary in practice to discretise the problem. The operator $\mathcal{G}$ could be the Radon transform and Ψ the restriction operator to the measurement along a single line, or $\mathcal{G}$ could be the solution operator for the ellitic PDE that takes the diffusion coefficient a to the solution u and Ψ could be a point evaluation of u at some point in the domain.

The general setting is then that $X_s : \Omega \to V_s := \mathbb{R}^s$ and $F_h : V_s \to \mathbb{R}$ are approximations of X and F with $\mu_{X_s} \ll \mu_X$, parametrised by some parameters $h > 0$ and $s \in \mathbb{N}$, e.g. the FE mesh width and the truncation dimension for the Karhunen-Loève expansion. Thus, we need to analyse not only the Monte Carlo sampling error but also the bias in the estimator due to the discretisation.

For simplicity, let us consider only the case $X_s = X$ and denote by $Q := F(X)$ and $Q_h := F_h(X)$ the quantity of interest and its approximation.

Lemma 5.4.2 (Bias-Variance Decomposition)**.** *Let $X \in L^2(\Omega; V)$ and iid $X^{(i)} \sim \mu_X$, $i \in \mathbb{N}$. Then*

$$\mathbb{E}\left[\left|\widehat{Q}^{\mathrm{MC}}_{h,N} - \mathbb{E}[Q]\right|^2\right] = \left(\mathbb{E}\left[Q_h - Q\right]\right)^2 + \frac{\mathbb{V}(Q_h)}{N} \quad \textit{with} \quad \widehat{Q}^{\mathrm{MC}}_{h,N} := \frac{1}{N}\sum_{i=1}^{N} F_h(X^{(i)})\,. \tag{5.4.4}$$

In fact, we also have

$$\sqrt{N}\left(\widehat{Q}^{\mathrm{MC}}_{h,N} - \mathbb{E}[Q]\right) \xrightarrow[N\to\infty]{d} \mathcal{N}\Big(\mathbb{E}\left[Q_h - Q\right], \mathbb{V}(Q_h)\Big). \tag{5.4.5}$$

Proof. This is a classical result in statistics and left as an exercise (cf. [HDAUQ, Lemma II.4.1]). □

An important concept in the analysis of the computational complexity of various numerical methods is the so-called ε-cost, the cost (measured in CPU time or arithmetic operations) to achieve an error less than some tolerance $\varepsilon > 0$.

Definition 5.4.3 (ε-cost)**.** The ε-cost $\mathcal{C}_\varepsilon(\widehat{Q})$ for any estimator $\widehat{Q}$ of $\mathbb{E}[Q]$ is defined to be the total number of arithmetic operations to achieve

$$\|\widehat{Q} - \mathbb{E}[Q]\|^2_{L^2(\Omega)} = E\left[\left|\widehat{Q} - \mathbb{E}[Q]\right|^2\right] \leq \varepsilon^2.$$

Theorem 5.4.4 (Complexity Theorem for MC)**.** *Suppose there are constants $\alpha, \gamma > 0$, such that*

$$\left|\mathbb{E}\left[Q_h - Q\right]\right| \;\leq\; Ch^\alpha, \tag{5.4.6}$$

$$\mathcal{C}(Q_h) \;\leq\; Ch^{-\gamma}, \tag{5.4.7}$$

as $h \to 0$, where $\mathcal{C}(Y)$ denotes the cost to compute one sample of a RV Y. Then, for any $\varepsilon > 0$, there exists $h = h(\varepsilon) > 0$ and $N_{\mathrm{MC}} := N_{\mathrm{MC}}(\varepsilon) \in \mathbb{N}$ such that

$$\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{MC}}_{h,N_{\mathrm{MC}}}) \leq C\varepsilon^{-2-\gamma/\alpha}.$$

Proof. See [HDAUQ, Thm. II.5.1]. □

Using (5.4.5), a similar result follows also for ε-cost defined with respect to convergence in probability (cf. [HDAUQ, Thm. II.5.1]). We can deduce the following corollary [HDAUQ, Cor. II.7.6].

Corollary 5.4.5. *Let us consider the elliptic PDE in Section 5.2.1 under the assumptions of Theorem 5.2.5 with $p = 2$ and with $a_{\min}^{-1} \in L^2(\Omega)$. Let $Q := Bu$ and $Q_h := Bu_h$. Suppose the FE solution is computed with an optimal multigrid method, such that $\mathcal{C}(Q_h) \leq Ch^{-d}$. Then, for any $\varepsilon > 0$, there exists $h = h(\varepsilon) > 0$ and $N_{\mathrm{MC}} := N_{\mathrm{MC}}(\varepsilon) \in \mathbb{N}$ such that*

$$\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{MC}}_{h,N_{\mathrm{MC}}}) \leq C\varepsilon^{-2-d/2}.$$

5.4.2 Multilevel Monte Carlo

The key idea in multilevel Monte Carlo is to use samples of Q on a hierarchy of different **discretisation levels**, i.e., for different values $h_0 > h_1 > \ldots h_L =: h > 0$ of the discretization parameter with $L \in \mathbb{N}$, and to decompose

$$\mathbb{E}\left[Q_h\right] = \mathbb{E}\left[Q_{h_0}\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right] =: \sum_{\ell=0}^{L} \mathbb{E}\left[Y_\ell\right]. \tag{5.4.8}$$

For simplicity, we will choose

$$h_{\ell-1} = m\, h_\ell, \quad \ell = 1, \ldots, L, \quad \text{for some } \ m \in \mathbb{N} \setminus \{1\} \text{ and } \ h_0 > 0, \tag{5.4.9}$$

i.e. uniform grid refinement for the elliptic PDE. With iid $X_\ell^{(i)} \sim \mu_X$, $\ell, i \in \mathbb{N}$, we define the **multilevel Monte Carlo (MLMC)** estimator for $\mathbb{E}[Q]$ as

$$\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} := \sum_{\ell=0}^{L} \widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} \ = \ \frac{1}{N_0} \sum_{i=1}^{N_0} F_{h_0}(X_0^{(i)}) \ + \ \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \Big(F_{h_\ell}(X_\ell^{(i)}) - F_{h_{\ell-1}}(X_\ell^{(i)}) \Big) \tag{5.4.10}$$

As in Lemma 5.4.2 for standard MC, the following bias-variance decomposition is a simple consequence of (5.4.8) and the independence of the RVs $X_\ell^{(i)}$:

$$\mathbb{E}\left[\left(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} - \mathbb{E}\left[Q\right] \right)^2 \right] = \left(\mathbb{E}\left[Q_h - Q\right] \right)^2 + \sum_{\ell=0}^{L} \frac{\mathbb{V}(Y_\ell)}{N_\ell}. \tag{5.4.11}$$

If $\|Q_h - Q\|_{L^2(\Omega)} \to 0$ as $h \to 0$, i.e., if the RV Q_h converges strongly (samplewise) to Q, then $\mathbb{V}(Y_\ell) \to 0$ as $\ell \to \infty$, leading to a huge variance reduction in the MLMC estimator compared to standard MC. In particular, a significantly smaller number $N_L \ll N_{\mathrm{MC}}$ of expensive samples on the finest level L with $h = h_L$ are sufficient to achieve a prescribed tolerance ε and the slightly larger number of samples $N_0 > N_{\mathrm{MC}}$ necessary on the coarsest level 0 are significantly cheaper than samples on level L under assumption (5.4.7).

It is possible to prove the following general complexity theorem [HDAUQ, Thm. II.5.2].

Theorem 5.4.6 (MLMC Complexity)**.** *Let $\varepsilon < e^{-1}$ and let $\alpha, \beta, \gamma > 0$ be such that $\alpha \geq \frac{1}{2}\min\{\beta, \gamma\}$ and such that for all $\ell \in \mathbb{N}_0$*

$$\textbf{(M1)} \ \ |\mathbb{E}\left[Q_{h_\ell}\right] - \mathbb{E}\left[Q\right]| \leq Ch_\ell^\alpha, \quad \textbf{(M2)} \ \ \mathrm{Var}[Y_\ell] \leq Ch_\ell^\beta, \quad \textbf{(M3)} \ \ \mathcal{C}(Y_\ell) \leq Ch_\ell^{-\gamma}.$$

Then there are $L \in \mathbb{N}$ and $\{N_\ell\}_{\ell=0}^L \subset \mathbb{N}$ such that

$$\mathcal{C}_\varepsilon\Big(\widehat{Q}^{ML}_{L,\{N_\ell\}}\Big) \leq C \begin{cases} \varepsilon^{-2}, & \text{if } \beta > \gamma, \\ \varepsilon^{-2}\,|\log\varepsilon|^2, & \text{if } \beta = \gamma, \\ \varepsilon^{-2-(\gamma-\beta)/\alpha}, & \text{if } \beta < \gamma. \end{cases}$$

A sufficient condition to achieve this asymptotic ε-cost is that

$$N_\ell \;\propto\; \left(m^{\frac{\beta+\gamma}{2}}\right)^{-\ell} \tag{5.4.12}$$

with L and N_0 chosen such that both of the two terms on the right hand side of (5.4.11) are equal to $\varepsilon^2/2$. For the elliptic PDE, we can again deduce the following corollary [HDAUQ, Cor. II.7.7]

Corollary 5.4.7. *Consider again the elliptic PDE in Section 5.2.1 under the assumptions of Theorem 5.2.5 with $d = 1, 2, 3$, $p = 2$ and $a_{\min}^{-1} \in L^2(\Omega)$. Let $Q := Bu$ and, for $\ell \in \mathbb{N}_0$, let $Q_{h_\ell} := Bu_{h_\ell}$. Suppose the FE solution on each level is computed with an optimal multigrid method, such that $\mathcal{C}(Q_{h_\ell}) \leq Ch_\ell^{-d}$. Then, for any $0 < \varepsilon < e^{-1}$, there exists $L \in \mathbb{N}$ and $\{N_\ell\}_{\ell=0}^L \subset \mathbb{N}$ such that*

$$\mathcal{C}_\varepsilon\left(\widehat{Q}^{ML}_{L,\{N_\ell\}}\right) \leq C\varepsilon^{-2}.$$

In the case of uniform mesh refinement in two spatial dimensions with $d = 2$ and $m = 2$, since the variance of Y_ℓ decreases with a rate $\beta = 4$, it follows from (5.4.12) that the number of samples can be reduced by a factor of $2^{(4+2)/2} = 8$ from level to level.

5.4.3 Quasi-Monte Carlo

We only consider quasi-Monte Carlo (QMC) methods in the context of the elliptic PDE in Section 5.2.1 with uniform diffusion coefficient a, as described in Example 4.5.12. For more details and rigorous proofs we refer to [HDAUQ, Chapter III] and

- J. Dick, F.Y. Kuo, I.H. Sloan, High-dimensional integration: The quasi-Monte Carlo way, *Acta Numer.* **22**:133–288, 2013.
- J. Dick, F.Y. Kuo, Q.T. Le Gia, D. Nuyens, C. Schwab, Higher order QMCPetrov–Galerkin discretization for affine parametric operator equations with random field inputs, *SIAM J. Numer. Anal.* **52**:2676–2702, 2014.
- I.G. Graham, F.Y. Kuo, J. Nichols, R. Scheichl, C. Schwab, I.H. Sloan, Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients, *Numer. Math.* **131**:329–368, 2015.
- F.Y. Kuo, C. Schwab, I.H. Sloan, Quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients, *SIAM J. Numer. Anal.* **50**:3351-3374, 2012.
- F.Y. Kuo, C. Schwab, I.H. Sloan, Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients, *Found. Comput. Math.* **15**:411-449, 2015.
- F.Y. Kuo, R. Scheichl, C. Schwab, I.H. Sloan, E. Ullmann, Multilevel quasi-Monte Carlo methods for lognormal diffusion problems, *Math. Comp.* **86**:2827-2860, 2017.

The Karhunen-Loève expansion is truncated after s terms, and we set $\Xi := (\xi_j)_{j=1}^s$ with iid $\xi_j \sim$ uniform$(-1, 1)$, such that μ_Ξ is the product uniform measure on $V = [-1, 1]^s$. As quantity of interest, we consider a linear functional $B : H_0^1(D) \to \mathbb{R}$ of the PDE solution u, which, as a

functional of the parameter vector Ξ we denote by $Q := F(\Xi)$. The functional $F : V \to \mathbb{R}$ can then be written as the composition

$$F := B \circ \mathcal{G} \circ T, \quad \text{such that} \quad \Xi \xrightarrow{T} a \xrightarrow{\mathcal{G}} u \xrightarrow{B} Q,$$

where $T : V \to L^\infty(D)$ is the operator defined in (4.5.12) and $\mathcal{G} : L^\infty(D) \to H_0^1(D)$ is the solution operator, mapping the coefficient a to the PDE solution u. Similarly, we denote by $Q_h := F_h(\Xi)$ with $F_h = B \circ \mathcal{G}_h \circ T$ the FE approximation of Q, where $\mathcal{G}_h : L^\infty(D) \to U_h$ is the FE solution operator, mapping the coefficient a to the FE solution u_h.

Quasi-Monte Carlo methods are formulated as quadrature rules over the unit cube $[0,1]^s$. Treating Ξ as a deterministic parameter vector ξ distributed according to product uniform measure,

$$\mathbb{E}[Q] \approx \int_{[-1,1]^s} F_h(x)\, \mathrm{d}\mu_\Xi(x) = \int_{[0,1]^s} F_h(2v - \mathbf{1})\, \mathrm{d}v, \tag{5.4.13}$$

where we used the simple change of variables $x = 2v - \mathbf{1}$ from $[0,1]$ to $[-1,1]$. We will use a **randomly shifted rank-1 lattice rule** to approximate (5.4.13). This takes the form

$$\widehat{Q}_{h,N}^{\mathrm{QMC}} = \frac{1}{N} \sum_{i=1}^{N} F_h\big(\widetilde{\Xi}^{(i)}\big), \quad \text{where} \quad \widetilde{\Xi}^{(i)} := 2\,\mathrm{frac}\left(\frac{iz}{N} + \Delta\right) - \mathbf{1}, \tag{5.4.14}$$

$z \in \{1, \ldots, N-1\}^J$ is a so-called *generating vector*, Δ is a uniformly distributed *random shift* on $[0,1]^s$, and "frac" denotes the fractional part function, applied componentwise [Dick et al, 2013]. To ensure that every one-dimensional projection of the lattice rule has N distinct values we furthermore assume that each component z_j of z satisfies $\gcd(z_j, N) = 1$. See Figure 5.2 for an example of a lattice rule in two dimensions.

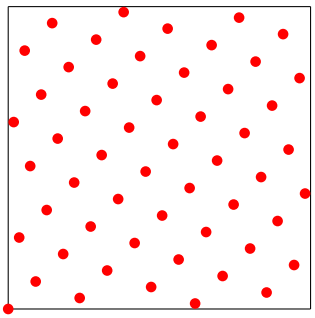


Figure 5.2: Two dimensional lattice rule with $N = 55$, $z = (1, 34)^\top$, $\Delta = (0, 0)^\top$.

Due to the random shift, (5.4.14) is an unbiased estimator of $\mathbb{E}_{\mu_\Xi}[F_h(\Xi)]$ an thus we have – as for MC and MLMC - again

$$\mathbb{E}\left[\left(\widehat{Q}_{h,N}^{\mathrm{QMC}} - \mathbb{E}[Q]\right)^2\right] = \left(\mathbb{E}\left[Q_h - Q\right]\right)^2 + \mathbb{V}\left(\widehat{Q}_{h,N}^{\mathrm{QMC}}\right), \tag{5.4.15}$$

where the variance of the QMC estimator is given by

$$\mathbb{V}\left(\widehat{Q}_{h,N}^{\mathrm{QMC}}\right) = \mathbb{E}_\Delta\left[\left(\widehat{Q}_{h,N}^{\mathrm{QMC}} - \mathbb{E}_{\mu_\Xi}[F_h(\Xi)]\right)^2\right]. \tag{5.4.16}$$

To bound it, we make the following assumption on the integrand $F_h(\Xi)$.

Assumption 5.4.8. Let $C > 0$ be a constant independent of s and let $(\ell_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$ be as defined in Example 4.5.12. We assume that, for any multi-index $\boldsymbol{\nu} \in \{0,1\}^s$ with $|\boldsymbol{\nu}| = \sum_{j\le s} \nu_j$,

$$\left|\frac{\partial^{|\boldsymbol{\nu}|} F(\xi)}{\partial \xi^{\boldsymbol{\nu}}}\right| \le C \frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} \prod_{j=1}^{J} \ell_j^{\nu_j} .$$

For linear functionals B on $H_0^1(D)$, this assumption has been proved in the uniform case. However, it can also be shown for nonlinear functionals.

Lemma 5.4.9 (Kuo et al, 2012)**.** *Suppose Assumption 5.4.8 holds and $(\ell_j)_{j\in\mathbb{N}} \in \ell^r(\mathbb{N})$, for some $r \in (0,1)$. Then, a randomly shifted lattice rule can be constructed via a component-by-component algorithm in $\mathcal{O}(sN \log N)$ cost, such that*

$$\mathbb{V}\Big(\widehat{Q}_{h,N}^{\mathrm{QMC}}\Big) \le C \begin{cases} N^{-1/\delta}, & \text{if } r \in (0, 2/3], \\ N^{-(1/r-1/2)}, & \text{if } r \in (2/3, 1), \end{cases}$$

for any $\delta \in (1/2, 1]$, independently of s.

This estimate can again be combined with (5.4.15) to bound the computational complexity of QMC estimators for the elliptic PDE.

Corollary 5.4.10. *Suppose Assumption 5.4.8 holds and $(\ell_j)_{j\in\mathbb{N}} \in \ell^r(\mathbb{N})$, for some $r \in (0, 2/3]$, and let z be the generating vector for the randomly shifted lattice rule that achieves the optimal rate in Lemma 5.4.9. Suppose further that the piecewise linear FE solution is computed with an optimal multigrid method, such that $\mathcal{C}(Q_h) \le Ch^{-d}$. Then, for any $\varepsilon > 0$, there exists $h > 0$ and $N \in \mathbb{N}$ such that*

$$\mathcal{C}_\varepsilon(\widehat{Q}_{h,N}^{\mathrm{QMC}}) \le C\varepsilon^{-2\delta - d/2}, \quad \text{for any } \ \delta \in (1/2, 1].$$

Remark 5.4.11. (a) It is even possible to combine quasi-Monte Carlo sampling and multilevel estimation and the gains are complementary [Kuo et al, 2015; Kuo et al, 2017], but we will not include these estimators or their analysis here.

(b) For smooth random fields, e.g. fast decay of the ℓ_j in Example 4.5.12, even faster convergence rates are possible with higher-order QMC rules [Dick et al, 2014] or with stochastic collocation and sparse grid quadrature rules [HDAUQ, Chapter IV].

(c) Note that due to Remark 5.2.7 and the comments before Lemma 5.4.9, the statements of Corollaries 5.4.5, 5.4.7 and 5.4.10 also hold for Fréchet-differentiable nonlinear functionals $Q := B(u)$ and for nonuniform measures μ_a.

Example 5.4.12 (Comparsion of sampling methods in the lognormal case)**.** To compare the approaches, let us consider the elliptic PDE (5.2.1) for $D = (0,1)^2$ (i.e., $d = 2$) and $f \equiv 1$, with lognormal diffusion coefficient $a \in L^\infty(D)$, i.e., $\log a \sim \mathcal{N}(m, C_{\nu,\sigma^2,\lambda})$, with Matérn covariance $C_{\nu,\sigma^2,\lambda}$. The Matérn covariance function is defined, for any $x, y \in D$, as

$$c(x,y) := \sigma^2 \, \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,|x-y|}{\lambda}\right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\,|x-y|}{\lambda}\right),$$

[Draft of October 6, 2021. Not for dissemination.]

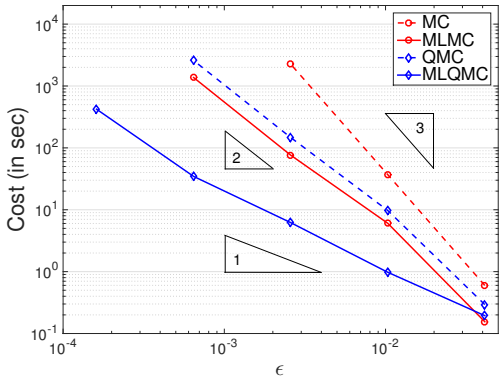


Figure 5.3: Comparison of measured ε-costs for $\nu = 2.5$, $\sigma^2 = 0.25$ and $\lambda = 1$. The points on each of the graphs correspond to the choices $L = 1, \dots, 5$ with $h_0 := \sqrt{2}/8$.

where Γ and K_ν are the Gamma-function and the modified Bessel function (of second-kind) of order ν, and where ν, σ^2 and λ are the so-called *smoothness parameter*, *total variance* and *correlation length*, respectively. The quantity of interest is

$$Q(\omega) := \frac{1}{|D^*|} \int_{D^*} u(x, \omega) \, \mathrm{d}x, \quad \text{with} \quad D^* := \left(\tfrac{3}{4}, \tfrac{7}{8}\right) \times \left(\tfrac{7}{8}, 1\right).$$

For a comparison of the sampling approaches discussed above, we use piecewise linear FEs on a uniform simplicial mesh to discretise the PDE and a truncated Karhunen-Loève expansion to sample from $\log a$ for $\nu = 2.5$. In that case, all the relevant assumptions in Corollaries 5.4.5 and 5.4.7 are satisfied and the assumptions of (the lognormal equivalent of) Corollary 5.4.10 hold with $(\ell_j)_{j \in \mathbb{N}} \in \ell^r(\mathbb{N})$ and $r < 2/3$. We collect the theoretical complexity bounds, as well as the theoretical bound for multilevel QMC (MLQMC) in Table 5.1. We see that in dimension $d \geq 2$, the cost of MLQMC is asymptotically optimal, in the sense that even to compute a single sample to accuracy ε has the same asymptotic complexity.

Table 5.1: Theoretical bounds on the order of growth of the ε-cost with respect to ε^{-1}, for the lognormal case for $\nu = 2.5$ (ignoring log-factors and choosing $\delta = 1/2$ in Corollary 5.4.10).

d	MC	MLMC	QMC	MLQMC	One sample
1	2.5	2	1.5	1	0.5
2	3	2	2	1	1
3	3.5	2	2.5	1.5	1.5

In Figure 5.3, we can see that these theoretical bounds are attained in practice. For the QMC methods we used an embedded lattice rule with weights $\gamma_j = j^{-2}$ with generating vector taken from the file `lattice-39102-1024-1048576.3600.txt` on Frances Kuo's webpage (UNSW Sydney).

5.5 Importance sampling estimators for posterior expectations

However, crucially, in Bayesian inverse problems we typically only have access to the posterior distribution in unnormalised form, i.e.

$$\frac{\mathrm{d}\mu_{X|y}}{\mathrm{d}\mu_X}(x) \propto \exp\left(-\frac{1}{2}\|y-\Phi(x)\|_\Sigma^2\right) \quad \text{or} \quad \pi_{X|Y}(x|y) \propto \exp\left(-\frac{1}{2}\|y-\Phi(x)\|_\Sigma^2\right)\pi_X(x),$$

in the infinite/finite dimensional case for an additive Gaussian noise model, respectively. For simplicity, we will only focus on the case of a finite dimensional parameter $X : \Omega \to \mathbb{R}^s$.

5.5.1 Importance sampling and ratio estimators

A classical technique to sample from a distribution that is only given in unnormalised form and a method that can also be used to reduce the variance in the estimator if we have an approximation for the (normalised or unnormalised) density is **importance sampling**. The following is based on the lecture notes of A. Owen (Stanford).

Suppose again that we are interested in computing

$$\mathbb{E}_p[F(X)] = \int_{\mathcal{S}} F(x)\ \mathrm{d}\nu(x) = \int_{\mathcal{S}} F(x)p(x)\ \mathrm{d}x,$$

for some RV $X : \Omega \to \mathcal{S} \subset \mathbb{R}^s$ where ν is a probability measure on $\mathcal{S} \subset \mathbb{R}^s$ with density p and the functional F is ν-measurable, for example if ν is the posterior measure on X. We take $p(x) = 0$ for all $x \notin \mathcal{S}$. If q is a positive probability density function on $\mathbb{R}^s$, then

$$\mathbb{E}_p[F(X)] = \int_{\mathcal{S}} F(x)p(x)\,\mathrm{d}x = \int_{\mathbb{R}^s} \frac{F(x)p(x)}{q(x)}q(x)\,\mathrm{d}x = \mathbb{E}_q\left[\frac{F(X)p(X)}{q(X)}\right] \tag{5.5.1}$$

By making a multiplicative adjustment to the integrand we compensate for sampling from q instead of p. The adjustment factor $w(x) = p(x)/q(x)$ is called the **likelihood ratio** or **importance weight**. The distribution q is the **importance distribution** and p is the **nominal distribution**. The importance distribution q does not have to be positive everywhere. It is enough to have $q(x) > 0$ whenever $F(x)p(x) \neq 0$.

The **importance sampling estimator** for $m := \mathbb{E}_p[F(X)]$ is

$$\widehat{Q}^{\mathrm{IS}}_{q,N} := \frac{1}{N}\sum_{i=1}^N \frac{F\big(X^{(i)}\big)p\big(X^{(i)}\big)}{q\big(X^{(i)}\big)} \quad \text{with} \quad \text{iid} \ \ X^{(i)} \sim q. \tag{5.5.2}$$

To use (5.5.2) we must be able to compute Fp/q, and in particular $p(x)/q(x)$ at any x we might sample. When p or q has an unknown normalization constant, then we will resort to a ratio estimate (see below). For now, we assume that p/q is computable, and study the variance of $\widehat{Q}^{\mathrm{IS}}_{q,N}$.

Theorem 5.5.1. *Let $q(x) > 0$ whenever $F(x)p(x) \neq 0$. For any $N \in \mathbb{N}$, $\mathbb{E}_q\big[\widehat{Q}^{\mathrm{IS}}_{q,N}\big] = m = \mathbb{E}_p[F(X)]$ and $\mathbb{V}_q\big(\widehat{Q}^{\mathrm{IS}}_{q,N}\big) = \sigma_q^2/N$ where*

$$\sigma_q^2 = \int_{\mathcal{S}} \frac{\Big(F(x)p(x)\Big)^2}{q(x)}\,\mathrm{d}x - m^2 = \int_{\mathcal{S}} \frac{\Big(F(x)p(x) - mq(x)\Big)^2}{q(x)}\,\mathrm{d}x\,. \tag{5.5.3}$$

Proof. Left as an exercise. □

Theorem 5.5.1 guides us in selecting a good importance sampling rule. From the first expression in (5.5.3) we see that a better q is one that gives a smaller value of $\int_{\mathcal{S}}(Fp)^2/q\,\mathrm{d}x$.

Lemma 5.5.2. *Let $\mathbb{E}_p\big[|F(X)|\big] \neq 0$. The probability density q^* with $q^*(x) \propto |F(x)|p(x)$ minimises σ_q^2 over all densities q that are positive when $Fp \neq 0$, i.e. $\sigma_{q^*}^2 \leq \sigma_q^2$.*

Proof. Let $q^*(x) = |F(x)|p(x)/\mathbb{E}_p\big[|F(X)|\big]$ and let q be an arbitrary density such that $q(x) > 0$ when $F(x)p(x) \neq 0$. Then

$$m^2 + \sigma_{q^*}^2 = \int_{\mathcal{S}} \frac{F(x)^2p(x)^2}{q^*(x)}\,\mathrm{d}x = \int_{\mathcal{S}} \frac{F(x)^2p(x)^2}{|F(x)|p(x)/\mathbb{E}_p\big[|F(X)|\big]}\,\mathrm{d}x =$$

$$= \big(\mathbb{E}_p\big[|F(X)|\big]\big)^2 = \left(\mathbb{E}_q\left[\frac{|F(X)|p(X)}{q(X)}\right]\right)^2 \leq \mathbb{E}_q\left[\frac{F(X)^2p(X)^2}{q(X)^2}\right] = m^2 + \sigma_q^2.$$

□

Remark 5.5.3. If $F(x) > 0$ is positive where $p(x) > 0$ and $m > 0$, then the optimal density $q^* = \frac{1}{m}Fp$ has $\sigma_{q^*}^2 = 0$, cf. the second form of $\sigma_{q^*}^2$ in (5.5.3), but it is of no pratical interest, because each of the samples in $\widehat{Q}_{q^*,N}^{\mathrm{IS}}$ becomes $F\big(X^{(i)}\big)p\big(X^{(i)}\big)/q^*\big(X^{(i)}\big) = m$, which is available only if we know the final result anyway in the first place. Although zero-variance importance sampling densities are not usable, they provide insight into the design of a good importance sampling scheme. It may be good for q to have spikes in the same places that $|F|$ does, or where p does, but it is even better to have them where $|F|p$ does. The appearance of q in the denominator of $w = p/q$, means that light-tailed importance densities q are dangerous. If we are clever or lucky, then F might be small just where it needs to be to offset the small denominator. But we often need to use the same sample with multiple integrands F, and so as a rule q should have tails at least as heavy as p does.

As mentioned at the beginning, in the Bayesian setting we can only sample from an unnormalized version of p, $p_u(x) = cp(x)$, where $c > 0$ is unknown. The same may be true of q, e.g., if we can compute $q_u(x) = bq(x)$ and $b > 0$ might be unknown. In general, $b \neq c$ and thus $p(x)/q(x) \neq p_u(x)/q_u(x)$. However, we may compute the ratio $w_u(x) = p_u(x)/q_u(x) = (c/b)p(x)/q(x)$ and consider the **self-normalized importance sampling estimator** or **ratio estimator**

$$\widehat{Q}_{q,N}^{\mathrm{RE}} := \frac{\sum_{i=1}^N F\big(X^{(i)}\big)w_u\big(X^{(i)}\big)}{\sum_{i=1}^N w_u\big(X^{(i)}\big)} = \frac{\frac{1}{N}\sum_{i=1}^N F\big(X^{(i)}\big)w\big(X^{(i)}\big)}{\frac{1}{N}\sum_{i=1}^N w\big(X^{(i)}\big)} \quad \text{with} \quad \text{iid} \;\; X^{(i)} \sim q. \tag{5.5.4}$$

To obtain iid samples of q it suffices to know q_u and the factor b/c cancels from the numerator and the denominator in (5.5.4), leading to the same estimate as if we had used the desired ratio $w(x) = p(x)/q(x)$ instead of the computable alternative $w_u(x)$.

Lemma 5.5.4. *Let p, q be two probability densities on $\mathbb{R}^s$ with $q(x) > 0$ whenever $p(x) > 0$. Then,*

$$\widehat{Q}_{q,N}^{\mathrm{RE}} \xrightarrow[N\to\infty]{\mathbb{P}\text{-a.s.}} \mathbb{E}_p[F(X)] =: m, \tag{5.5.5}$$

but in general $\mathbb{E}_q\big[\widehat{Q}^{\mathrm{RE}}_{q,N}\big] \neq m$*, i.e. the estimator is biased. We also have*

$$\sqrt{N}\left(\widehat{Q}^{\mathrm{RE}}_{q,N} - m\right) \xrightarrow[N\to\infty]{d} \mathcal{N}\Big(0, \sigma_q^2\Big), \tag{5.5.6}$$

with **asymptotic variance** σ_q^2*, as defined in* (5.5.3)*.*

Proof. Consider the second form of the definition of $\widehat{Q}^{\mathrm{RE}}_{q,N}$ in (5.5.4). The numerator is equal to $\widehat{Q}^{\mathrm{IS}}_{q,N}$, which we have already seen is an unbiased estimator of m. The strong law of large numbers gives $\mathbb{P}\left(\lim_{N\to\infty} \widehat{Q}^{\mathrm{IS}}_{q,N} = m\right) = 1$. Using the same arguments as for the numerator also for the denominator, but with the constant functional $F \equiv 1$, we see that the denominator converges in probability to 1, which implies (5.5.5).

To see that in general $\widehat{Q}^{\mathrm{RE}}_{q,N}$ is biased, consider $N = 1$, $p \neq q$ and $F(x) = x$. Then,

$$\mathbb{E}_q\big[\widehat{Q}^{\mathrm{RE}}_{q,N}\big] = \mathbb{E}_q\big[X^{(i)}\big] \neq \mathbb{E}_p\big[X^{(i)}\big] = m.$$

The result in (5.5.6) can be shown using again the Central Limit Theorem. □

Note that the condition on q for the ratio estimator are slightly stronger than for the importance sampling estimator, i.e., we need $q(x) > 0$ whenever $p(x) > 0$, rather than whenever $F(x)p(x) \neq 0$.

5.5.2 Estimating posterior expectations

Now let us return to the Bayesian inverse problem with additive Gaussian noise and assume that p is the density of the posterior $\nu = \mu_{X|y}$ such that

$$p_u(x) = \exp\left(-\frac{1}{2}\|y - \Phi(x)\|^2_\Sigma\right)\pi_X(x)$$

and we have access to a family of approximations $F_h(X)$ of $F(X)$ and $\Phi_h(X)$ of $\Phi(X)$, parametrised by $h > 0$, such that $F_h \to F$ and $\Phi_h \to \Phi$ as $h \to 0$. For the remainder of this section, we will analyse the accuracy and complexity of various ratio estimators for $\mathbb{E}_p\big[F(X)\big]$ based on samples $F_h(X^{(i)})$ with $X^{(i)}$ drawn from some distribution q, again possibly given only in unnormalised form.

To simplify the notation let $w_u(x) = Zw(x)$ with $w = p/q$ and $Z := \mathbb{E}_q[w_u(X)]$, such that

$$\mathbb{E}_p[F(X)] = \frac{\mathbb{E}_q[Q_w]}{Z} \quad \text{with} \quad Q_w := F(X)w_u(X),$$

the quantity of interest times the (unnormalised) weight. Similarly, we write $w_{u,h}(x) = Z_h w_h(x)$ with $w_h = p_h/q$ and $Z_h := \mathbb{E}_q[w_{u,h}(X)]$, and consider the ratio estimator

$$\widehat{Q}^{\mathrm{RE}}_{q,h,N} := \frac{\widehat{Q}_{w,h}}{\widehat{Z}_h}, \tag{5.5.7}$$

for $\mathbb{E}_p[Q]$ where again $Q = F(X)$ and $\widehat{Q}_{w,h}$ and $\widehat{Z}_h$ are estimators of MC-type (as discussed above) for $\mathbb{E}_q[Q_w]$ and for Z, respectively

Lemma 5.5.5 (MSE of the ratio estimator)**.** *If $q(x) > 0$ when $p(x) > 0$ and $\|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\|_{L^\infty(\Omega)} < \infty$, then*

$$\mathbb{E}\left[\left(\widehat{Q}^{\mathrm{RE}}_{q,h,N} - \mathbb{E}_p[Q]\right)^2\right] \leq CZ^{-2}\left(\mathbb{E}\left[\left(\widehat{Q}_{w,h} - \mathbb{E}_q[Q_w]\right)^2\right] + \mathbb{E}\left[\left(\widehat{Z}_h - Z\right)^2\right]\right), \tag{5.5.8}$$

where $C := 2\max\left(1, \|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\|^2_{L^\infty(\Omega)}\right)$. The expected value is with respect to q in the case of MC and MLMC and with respect to the random shift $\Delta \sim \mathrm{uniform}(0,1)^s$ in QMC.

Proof. Rearranging the MSE and using triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{Q}^{\mathrm{RE}}_{q,h,N} - \mathbb{E}_p[Q]\right)^2\right] &= \frac{1}{Z^2}\mathbb{E}\left[\left(\widehat{Q}_{w,h} - \mathbb{E}_q[Q_w] + \left(\widehat{Q}_{w,h}/\widehat{Z}_h\right)\left(\widehat{Z}_h - Z\right)\right)^2\right] \\
&\leq \frac{2}{Z^2}\mathbb{E}\left[\left(\widehat{Q}_{w,h} - \mathbb{E}_q[Q_w]\right)^2 + \left(\widehat{Q}^{\mathrm{RE}}_{q,h,N}\right)^2\left(\widehat{Z}_h - Z\right)^2\right] \\
&\leq \frac{2}{Z^2}\max\left(1, \|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\|^2_{L^\infty(\Omega)}\right)\left(\mathbb{E}\left[\left(\widehat{Q}_{w,h} - \mathbb{E}_q[Q_w]\right)^2\right] + \mathbb{E}\left[\left(\widehat{Z}_h - Z\right)^2\right]\right).
\end{aligned}$$

□

In the following, let $p = \pi_{X|Y}$ and denote by

$$\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}} = \widehat{Q}^{\mathrm{typ}}_{w,h} \Big/ \widehat{Z}^{\mathrm{typ}}_h \quad \text{with} \quad \mathrm{typ} = \mathrm{MC},\ \mathrm{ML},\ \mathrm{QMC},$$

the ratio estimator defined in (5.5.7) for the posterior expectation $\mathbb{E}_p[Q]$ with $\widehat{Q}^{\mathrm{typ}}_{w,h}$ chosen to be the MC estimator, the MLMC estimator or the QMC estimator for $\mathbb{E}_q[Q_w]$, respectively, and let $\widehat{Z}^{\mathrm{typ}}_h$ be the corresponding estimator for the normalization constant. Then, Lemma 5.5.5 implies that the convergence and the computational complexity of the ratio estimator (5.5.7) follow directly from the results on the basic estimators in Section 5.4.

Theorem 5.5.6 (Complexity of the ratio estimator)**.** *Let* typ = MC *or* ML *and let q be a probability distribution on $\mathbb{R}^s$ with $q(x) > 0$ whenever $p(x) > 0$. Suppose $\|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\|_{L^\infty(\Omega)} < \infty$ and the assumptions of Theorems 5.4.4 and 5.4.6 hold with $\alpha, \beta \neq \gamma > 0$. Then, for any $0 < \varepsilon < e^{-1}$ there exists an $h > 0$ and an $N \in \mathbb{N}$, resp. $\{N_\ell\} \subset \mathbb{N}$, such that*

$$\mathcal{C}_\varepsilon\left(\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}}\right) \leq C\begin{cases} (Z\varepsilon)^{-2-\gamma/\alpha}, & \text{if typ} = \mathrm{MC}, \\ (Z\varepsilon)^{-2-\max(0,(\gamma-\beta)/\alpha)}, & \text{if typ} = \mathrm{ML}. \end{cases}$$

A similar result can be proved also for the QMC-based ratio estimator.

Remark 5.5.7. The asymptotic order of the ε-cost is independent of the choice of the importance distribution q. However, due to the appearance of the extra factor $Z^{-\eta}$, for some $\eta \geq 2$, the asymptotic constant will **strongly** depend on the choice of q, as we will discuss in Section 5.5.3.

Let us give some more details for all three estimators in the case of the elliptic PDE with uniform diffusion coefficient a. As in Section 5.4.3, we assume that a is discretised via a truncated Karhunen-Loève expansion parametrised via $\Xi \sim \mathrm{uniform}(-1,1)^s$. The first and most obvious choice for the importance distribution is the prior distribution, i.e. $q = \pi_\Xi$.

Corollary 5.5.8. *Consider the elliptic PDE in Section 5.4.3 with $p = \pi_{\Xi|Y}$ and $q = \pi_\Xi$. Let* typ = MC, QMC *or* ML, *and consider the ratio estimator $\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}}$ for the posterior expectation $\mathbb{E}_p[Q]$ of $Q = F(\Xi)$ under the assumptions of Corollaries 5.4.5, 5.4.7, or 5.4.10, respectively. In the QMC case, let $(\ell_j)_{j\in\mathbb{N}} \in \ell^r(\mathbb{N})$ with $r < 2/3$; in the MLMC case, let h_0 be sufficiently small.*

Let $Q = F(\Xi) := B(u)$ and $\Phi(\Xi) := H(u)$ with B and H two bounded (and sufficiently smooth) functionals of the PDE solution from $H^1_0(D)$ to $\mathbb{R}$ and $\mathbb{R}^m$, respectively. Then, for any $0 < \varepsilon < e^{-1}$ there exists an $h > 0$ and an $N \in \mathbb{N}$, resp. $\{N_\ell\} \subset \mathbb{N}$, such that

$$\mathcal{C}_\varepsilon\Big(\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}}\Big) \le C \begin{cases} (Z\varepsilon)^{-2-d/2}, & \text{if } \ \mathrm{typ} = \mathrm{MC}, \\ (Z\varepsilon)^{-2}, & \text{if } \ \mathrm{typ} = \mathrm{ML}, \\ (Z\varepsilon)^{-1+\delta-d/2}, & \text{if } \ \mathrm{typ} = \mathrm{QMC}, \quad \text{for any } \ \delta > 0. \end{cases}$$

Proof. Since the unnormalised weight function $w_{u,h}(\xi) = \exp\big(-\frac{1}{2}\|y - H(u_h(\xi))\|^2_\Sigma\big)$ and the product $F_h(\xi)w_{u,h}(\xi) = B(u_h(\xi))\exp\big(-\frac{1}{2}\|y - H(u_h(\xi))\|^2_\Sigma\big)$ are both sufficiently smooth, nonlinear functionals of the PDE solution, the extension of Theorem 5.2.5 referred to in Remark 5.2.7(a) applies and it is possible to prove analogues of Corollaries 5.4.5, 5.4.7 and 5.4.10 for nonlinear functionals to bound the right hand side of (5.5.8). For a full proof of this extension see [Scheichl, Stuart, Teckentrup, 2017].

Since by definition $q(\xi) > 0$ when $p(\xi) > 0$, we have $Z > 0$. Thus, provided $\big\|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\big\|_{L^\infty(\Omega)} < \infty$, we can apply Lemma 5.5.5 and deduce that

$$\mathcal{C}_\varepsilon\Big(\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}}\Big) \le C\left(\mathcal{C}_{Z\varepsilon}\big(\widehat{Q}^{\mathrm{typ}}_{w,h}\big) + \mathcal{C}_{Z\varepsilon}\big(\widehat{Z}^{\mathrm{typ}}_h\big)\right).$$

Note that to compensate the factor Z^{-2} in (5.5.8) we need to scale the required tolerances ε for the individual estimators for the numerator and the denominator by Z.

It remains to verify $\big\|\widehat{Q}^{\mathrm{RE}}_{q,h,N}\big\|_{L^\infty(\Omega)} < \infty$. From the assumptions on B and H we deduce that there exist two constants $M_F, M_\Phi < \infty$, such that $|F_h(\xi)| \le M_F$ and $\|\Phi_h(\xi)\|_\Sigma \le M_\Phi$, for any $\xi \in [-1,1]^s$ and for any $h > 0$. Thus, recalling that $w_{u,h}(\xi) \le 1$, we have

$$\big|\widehat{Q}^{\mathrm{MC}}_{w,h}\big| = \left|\frac{1}{N}\sum_{i=1}^N F_h(\xi^{(i)})w_{u,h}(\xi^{(i)})\right| \le M_F \quad \text{and}$$
$$\widehat{Z}^{\mathrm{MC}}_h = \frac{1}{N}\sum_{i=1}^N w_{u,h}(\xi^{(i)}) \ge \exp\big(-\tfrac{1}{2}\big(\|y\|^2_\Sigma + M^2_\Phi\big)\big) =: M_Z > 0,$$

and thus $\big|\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{MC}}\big| \le M_\Phi/M_Z < \infty$. The proof for typ = QMC is identical.

For typ = ML, the upper bound follows in the same way. On the other hand, to bound $\widehat{Z}^{\mathrm{ML}}_h$ we can use the nonlinear extension of Theorem 5.2.5 again to obtain, with $Y_\ell = w_{u,h_\ell} - w_{u,h_{\ell-1}}$, that

$$\widehat{Z}^{\mathrm{ML}}_h \ \ge \ \widehat{Z}^{\mathrm{MC}}_{h_0} - \sum_{\ell=1}^L \widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} \ \ge \ M_Z - C\sum_{\ell=1}^L h^2_\ell\,,$$

with a constant C independent of $\{h_\ell\}$. Thus, if h_0 is sufficiently small, such that $\sum_{\ell=1}^L h^2_\ell < M_Z/C$, we also have $\big|\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{ML}}\big| < \infty$. □

For $\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{ML}}$, a similar result can also be proved for the case of a lognormal PDE coefficient a (see again [Scheichl, Stuart, Teckentrup, 2017]).

Example 5.5.9 (Continuation of Example 5.4.12)**.** For a numerical comparison we return again to the lognormal case of the elliptic PDE with Matérn covariance on $D = (0,1)^2$ discretised by piecewise linear FEs. However, in the following experiments we choose $\nu = 1/2$, $\sigma^2 = 1$ and $\lambda = 0.3$, which is a significantly harder case than the one considered in Example 5.4.12. Due to the low regularity, in that case it is only possible to prove

$$\|B(u) - B(u_h)\|_{L^p(\Omega;\mathbb{R}^m)} \leq Ch$$

for sufficiently smooth (nonlinear) functionals $B : H^1_0(D) \to \mathbb{R}^m$. Thus, the assumptions in Section 5.4 only hold with $\alpha = 1$, $\beta = 2$ and $\gamma = 2$. Thus, the theoretically expected ε-costs in the following experiments are $\mathcal{O}(\varepsilon^{-4})$, $\mathcal{O}(\varepsilon^{-3})$ and $\mathcal{O}(\varepsilon^{-2})$ for typ = MC, QMC and ML, respectively.

We consider the case of $f \equiv 0$ and mixed boundary conditions, such that

$$u(x) = 1, \text{ for } x_1 = 0, \quad u(x) = 0, \text{ for } x_1 = 1, \quad \text{and} \quad \frac{\partial u}{\partial x_2}(x) = 0, \text{ on the rest of the boundary,}$$

leading to a flow of heat (or fluid) from $x_1 = 0$ to $x_1 = 1$. The quantity of interest is the outflow over the boundary at $x_1 = 1$, which can be computed as

$$Q_h = Gu_h = -\int_D a(x,\omega)\nabla u_h(x,\omega)^\top \nabla w_h(x)\,\mathrm{d}x,$$

for a suitably chosen weight function w_h with $w_h|_{x_1=0} = 0$ and $w_h|_{x_1=1} = 1$. The observation functional $\Phi : [-1,1]^s \to \mathbb{R}^m$ consists of m (local) averages of the PDE solution u at m uniformly distributed points in D. The data $y \in \mathbb{R}^m$ is generated (synthetically) from the solution of (5.2.5) with $h^* = 1/256$, adding noise in the form of a realisation of $E \sim \mathcal{N}(0,\Sigma)$ with $\Sigma = \sigma_E^2 I$. For more details see [Scheichl, Stuart, Teckentrup, 2017].

In Figure 5.4, we compare the computational ε-cost for the three ratio estimators – here measured in terms of arithmetic operations – for $h = 1/16, \ldots, 1/256$, $m = 9$ and $\sigma_E^2 = 0.09$. The left figure shows results for ratio estimators with the same random samples used in $\widehat{Q}^{\mathrm{typ}}_{w,h}$ and $\widehat{Z}^{\mathrm{typ}}_h$, referred to as *dependent* estimators, while the right figure shows ratio estimators with different random samples used in $\widehat{Q}^{\mathrm{typ}}_{w,h}$ and $\widehat{Z}^{\mathrm{typ}}_h$, referred to as *independent* estimators.

In Figure 5.5, we show estimates of the discretisation error and of the MC sampling error, for the numerator and denominator in (5.5.7) individually, as well as for the ratio estimate itself. While we see clearly that they all converge with the same (predicted) rates, the ratio estimate is several orders of magnitude bigger. This is due to the factor Z^{-2} on the right hand side of (5.5.8). Note that $Z \to 0$ as $\sigma_E^2 \to 0$ or as $m \to \infty$. In Figure 5.6, we can see that this is also reflected in the asymptotic variance $\tilde{\sigma}^2_{pq}$ of the dependent ratio estimators as $\sigma_E^2 \to 0$ or as $m \to \infty$. The growth in the independent ratio estimators is significantly faster.

5.5.3 Data-informed importance distributions – preconditioning

The lack of robustness of the ratio estimator with respect to the prior density $q = \pi_X$ observed in Example 5.5.9 can be clearly seen when considering again the scaled posterior log-likelihood

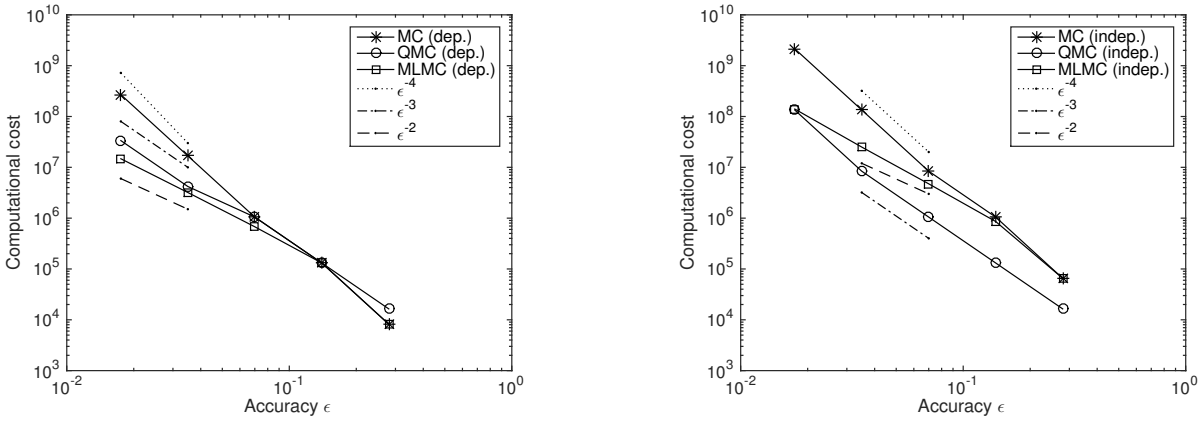


Figure 5.4: Lognormal problem (with $\nu = 0.5$): Comparison of ε-costs (in arithmetic Ops.) for the ratio estimators $\widehat{Q}^{\mathrm{RE}}_{q,h,\mathrm{typ}}$ with dependent (left) and independent (right) estimators $\widehat{Q}^{\mathrm{typ}}_{w,h}$ and $\widehat{Z}^{\mathrm{typ}}_{h}$.

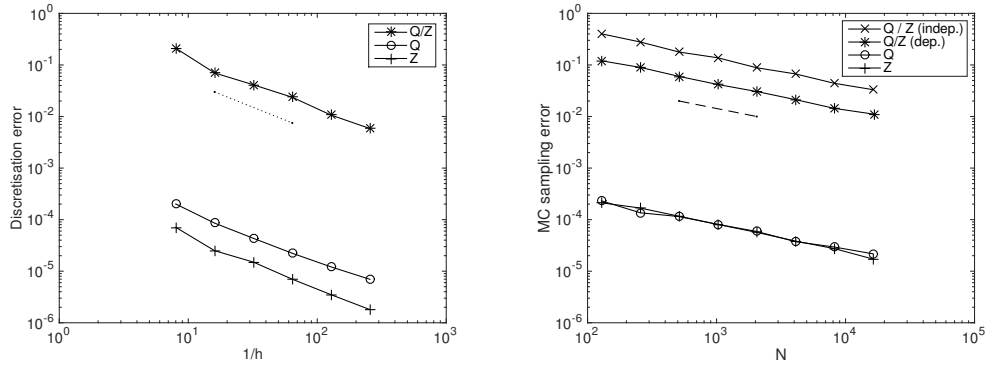


Figure 5.5: FE discretisation errors vs. $1/h$ (left) and MC sampling errors vs. N (right) for $\widehat{Q}_{w,h}$, $\widehat{Z}_h$ and $\widehat{Q}^{\mathrm{RE}}_{q,h,N}$. The dotted and dashed lines reference slopes are -1 and $-1/2$, respectively.

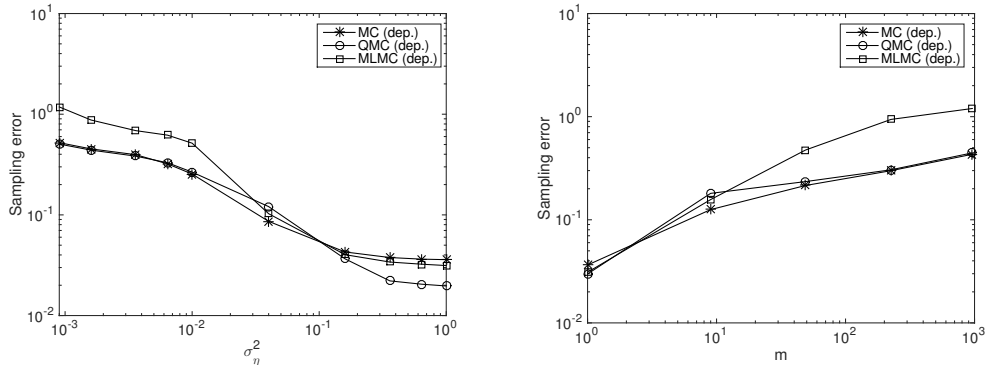


Figure 5.6: Sampling errors $\mathbb{E}_q\big[(\widehat{Q}^{\mathrm{RE}}_{q,h,N} - \mathbb{E}_p[Q_h])^2\big]^{1/2}$ for $N = 256$, as functions of noise level σ^2_E (left) and of number of observations m (right).

$n\Psi_n(x)$ with Ψ_n defined in (5.3.5). It was used there to study posterior concentration in the small noise limit. Note that in that case

$$Z = \mathbb{E}_q[w_u(X)] = \int_{\mathbb{R}^s} \exp\left(-\frac{n}{2}\|y - \Phi(x)\|_\Sigma^2\right) \pi_X(x)\,\mathrm{d}x \;\to\; 0 \quad \text{as} \quad n \to \infty,$$

and so clearly the bound on the MSE of the ratio estimator in Lemma 5.5.5 explodes with $n \to \infty$.

The following lemma shows that not only the bound but in fact the asymptotic variance of the prior-based ratio estimator explodes as $n \to \infty$. We do not include the proof.

Lemma 5.5.10 (Schillings, Sprungk, Wacker, 2020)**.** *For a RV $X : \Omega \to \mathbb{R}^s$ and a sufficiently smooth and measurable $F : \mathbb{R}^s \to \mathbb{R}$, consider the scaled posterior log-likelihood $n\Psi_n(x)$ with Ψ_n defined in* (5.3.5)*. Under the assumptions of Theorem 5.3.3 with $p = \pi_{X|Y}$ and $q = \pi_X$, there exist $0 < c < C$ such that the asymptotic variance σ_q^2 of $\widehat{Q}_{q,N}^{\mathrm{RE}}$ satisfies*

$$cn^{s/2}\mathbb{V}_p(F(X)) \;\le\; \sigma_q^2 \;\le\; Cn^{s/2}\mathbb{V}_p(F(X))\,.$$

Let us now instead consider as the importance distribution the Laplace approximation of the posterior, i.e. q_u is the unnormalised density of $\mathcal{L}_{\mu_{X|y}}$. It can be shown using Theorem 5.3.3 that

$$Z = \mathbb{E}_q[w_u(X)] = \mathbb{E}_q\left[\frac{p_u(X)}{q_u(X)}\right] \to 1 \;\text{ as }\; n \to \infty \quad \text{and} \tag{5.5.9}$$

Again this result can be sharpened leading to the following theorem.

Theorem 5.5.11 (Schillings, Sprungk, Wacker, 2020)**.** *Under the assumptions of Lemma 5.5.10 with $p = \pi_{X|Y}$ and q the density of the Laplace approximation $\mathcal{L}_{\mu_{X|y}}$, for any $N \in \mathbb{N}$ and $\delta \in [0, 1/2)$,*

$$n^\delta \left|\widehat{Q}_{q,N}^{\mathrm{RE}} - \mathbb{E}_p[F(X)]\right| \;\xrightarrow[N\to\infty]{\mathbb{P}}\; 0\,,$$

i.e. the error of the Laplace-based ratio estimator converges in probability to zero as $n \to \infty$, independently of the sample size N and with a rate arbitrarily close to $n^{-1/2}$.

Example 5.5.12. The following numerical experiment for the elliptic (P)DE on $(0, 1)$ (i.e. for $d = 1$), with $u|_{x=0} = u|_{x=1} = 0$ and $f(x) = 100x$, is taken from [Schillings, Sprungk, Wacker, 2020]. It is a toy example with uniform a, with $\Xi \sim \mathrm{uniform}(-1, 1)^s$, for $s = 1, 2, 3$, and

$$\ell_j\varphi_j(x) = (10j)^{-1} sin(j\pi x).$$

The data are $m = 2$ (resp. 7) measurements $y_k = u(x_k^*)$ of the solution at equally spaced points $x_k^* \in (0, 1)$ for $s = 1, 2$ (resp. 3) with measurement noise $E_n \sim \mathcal{N}(0, (100n)^{-1}I)$ for $n \in \mathbb{N}$. The quantity of interest is $Q = u(0.5)$.

In Figure 5.7, we see the lack of robustness of the prior-based ratio estimator, as well as the convergence and robust behaviour of the Laplace-based estimator as $n \to \infty$.

It is also possible to use other preconditioners. For example, in

- S. Dolgov, K. Anaya-Izquierdo, C. Fox, R. Scheichl, Approximation and sampling of multivariate probability distributions in tensor train decomposition, *Stat. Comput.* **30**:603, 2020,

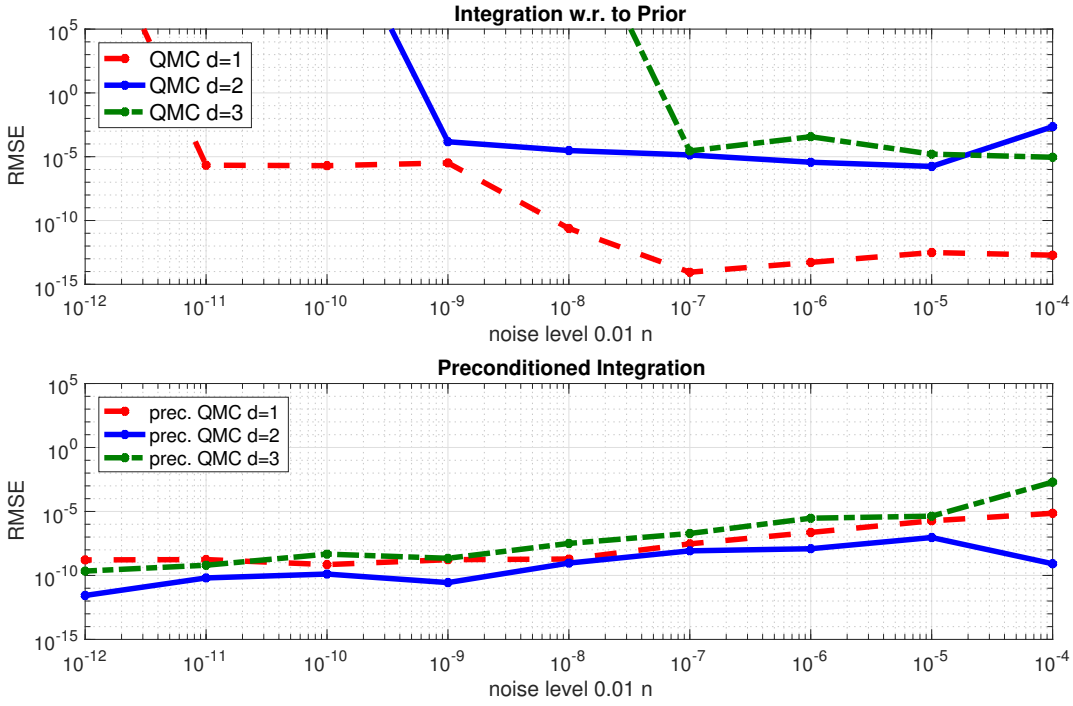


Figure 5.7: The estimated root mean square error (RMSE) of $\widehat{Q}^{\mathrm{RE}}_{q,\mathrm{QMC}}$ for $q = \pi_X$ (top) and Laplace-based importance sampling (bottom) using a randomised lattice rule with 8192 points averaged over 64 random shifts, for $s = 1, 2, 3$ and for $n = 10^2, 10^3, \ldots, 10^{10}$.

we have used **TT-cross approximations**, i.e. low-rank tensor approximations described in [HDAUQ, Chap. 7], of the unnormalised posterior density $p_u \propto \pi_{X|Y}$ as the unnormalised importance distribution q_u. By increasing the ranks in the low-rank approximation, it is possible to make $w_u(x)$ arbitrarily close to 1. This will be discussed in more detail in Section 5.7.

Example 5.5.13. Here, we just show how the TT-cross approximation improves the efficiency of the ratio estimator for the elliptic PDE over a prior-based ratio estimator and how it compares to MCMC-based estimators (more details on those in Section 5.6.2). The setup is almost identical to that in Example 5.5.9, except that $\log a$ is modelled as a (Karhunen-Loève like) expansion with independent uniform instead of independent Gaussian coefficients. The observation operator $\Phi : H_0^1(D) \to \mathbb{R}^m$ and the quantity of interest are the same. We use the same randomised lattice rule, $m = 9$ measurements and a noise $E \sim \mathcal{N}(0, \frac{1}{100} I)$. For details see [Dolgov et al., 2020].

In Figure 5.8, we compare the relative sampling errors of various estimators, plotted against the number of samples and also against CPU time. In particular, we compare the TT-based and the prior-based ratio estimators with QMC rules, qRE(prior) and qRE(TT) resp., with three Markov chain Monte Carlo (MCMC) estimators: DRAM [Haario et al, 2001], MALA [Roberts, Tweedie, 1996] and a Metropolis-Hastings algorithm with independent proposals drawn from the TT approximation of the posterior distribution, MetH(TT). We observe the better rate of convergence of almost $\mathcal{O}(N^{-1})$ for the QMC-based ratio estimators and also how much the TT-based precondi-

[Draft of October 6, 2021. Not for dissemination.]

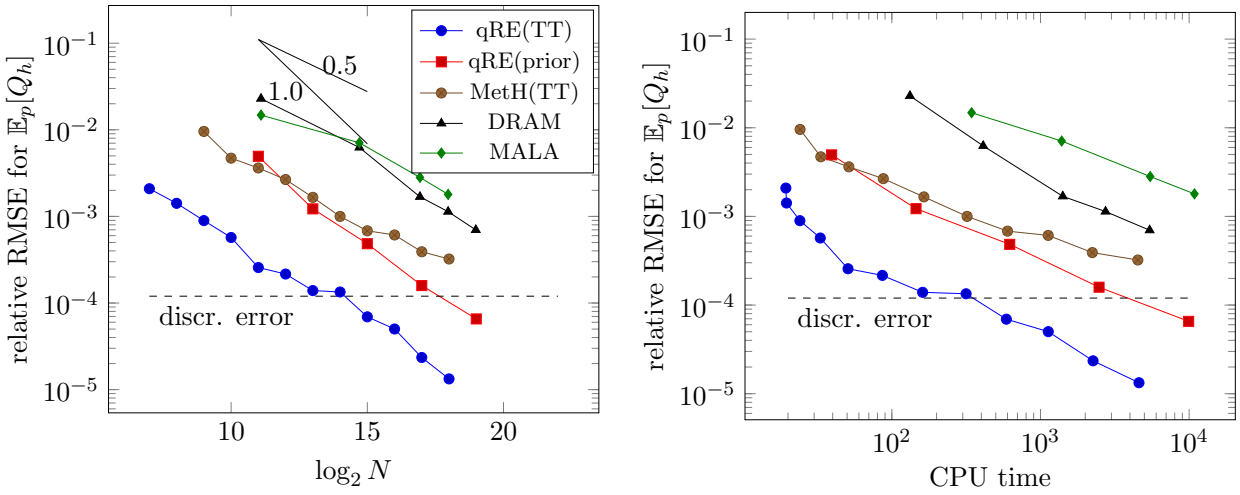


Figure 5.8: Estimated RMSEs of $\widehat{Q}^{\mathrm{RE}}_{q,\mathrm{QMC}}$ for $q = \pi_X$ (qRE(prior)) and TT-based importance sampling (qRE(TT)), plotted against number of samples N (left) and CPU time in seconds (right). The estimators are also compared to MCMC estimators based on TT-proposals (MetH(TT)), DRAM and MALA. (The dashed line indicates the discretisation error for the chosen FE mesh size h.)

tioning helps both in the case of the ratio estimator and in the MCMC case.

5.6 The Markov chain Monte Carlo method

The basic idea of this method is to compute a sequence of RVs $(X_j)_{j\in\mathbb{N}}$, such that

$$X_j \xrightarrow[i\to\infty]{d} X \sim \mu^{X|y} \tag{5.6.1}$$

Under appropriate conditions it is possible in this setting to prove a strong law of large numbers and a central limit theorem (as for iid samples in Prop. 5.4.1), in particular if the sequence of RVs comes from a **Markov chain** (a well-studied class of **time-discrete stochastic processes**).

5.6.1 Basic concepts of Markov chain theory

Let H be a separable Hilbert space.

Definition 5.6.1. A **Markov chain** in H is a sequence of H-valued RVs $X_j : \Omega \to H$, satisfying the **Markov property**, i.e.,

$$\mathbb{P}(X_{j+1} \in A | X_1, \dots, X_j) = \mathbb{P}(X_{j+1} \in A | X_j) \quad \mathbb{P}\text{-a.s.},$$

for each $j \in \mathbb{N}$ and $A \in \mathcal{B}(H)$ (i.e., the state X_{j+1} of a Markov chain only depends on the previous state X_j and not on the entire history of the chain).

An important special case are homogeneous Markov chains.

Definition 5.6.2. (a) A map $K : H \times \mathcal{B}(H) \to [0,1]$ is called **transition** or **Markov kernel** if

(i) for all $x \in H$ ist $K(x, \cdot)$ is a probability measure on $(H, \mathcal{B}(H))$, and

(ii) for all $A \in \mathcal{B}(H)$ ist $K(\cdot, A)$ is a measurable functional from H to $[0, 1]$.

(b) A Markov chain $(X_j)_{j\in\mathbb{N}}$ is **homogeneous**, if there exists a transition kernel K, such that for all $j \in \mathbb{N}$, $x \in H$ and $A \in \mathcal{B}(H)$

$$K(x, A) = \mathbb{P}(X_{j+1} \in A \mid X_j = x) \qquad \mathbb{P}\text{-a.s.}$$

Thus, the transition kernel K of a homogeneous Markov chain provides the transition probability from the jth state X_j to the next state X_{j+1} of the Markov chain.

We will only consider homogeneous Markov chains. Their properties can be expressed as properties of their transition kernel. We therefore introduce the following notions.

Definition 5.6.3. Let K be a Markov kernel on H and let $\nu \in \mathcal{P}(H)$. Then we denote by νK the probability measure on $(H, \mathcal{B}(H))$ given by

$$(\nu K)(A) := \int_H K(x, A)\nu(\,\mathrm{d}x) \quad \text{for all} \quad A \in \mathcal{B}(H). \tag{5.6.2}$$

Moreover, we define recursively, for $j \in \mathbb{N}$, the Markov kernel K^j on H by

$$K^j(x, A) := \int_H K^{j-1}(x', A)K(x,\, \mathrm{d}x') \quad \text{for all} \quad x \in H,\ A \in \mathcal{B}(H). \tag{5.6.3}$$

In this notation, we can express the distribution of the jth state X_j of a Markov chain with transition kernel K and initial distribution $X_1 \sim \nu$ simply by $X_j \sim \nu K^{j-1}$.

Remark 5.6.4. The definition of νK is an abuse of notation, since K denotes a Markov kernel but in νK it plays the role of a mapping from $\mathcal{P}(H)$ to $\mathcal{P}(H)$. Also, it seems odd to place ν on the left hand side of K instead of writing $K\nu$. The reason for this is that, in the special case of a discrete state space H, with $|H| = M$ – where Markov chains have been studied first – the transition kernel K is simply a **(row) stochastic matrix** $K \in [0, 1]^{M\times M}$ and thus for a (column) vector $\nu \in [0, 1]^M$ of initial probabilities, the vector given by νK describes the distribution of the next state of the Markov chain.

Definition 5.6.5. (a) Let $\mu \in \mathcal{P}(H)$ and let K be the transition kernel of a Markov chain $(X_j)_{j\in\mathbb{N}}$ in H. The measure μ is called an **invariant measure** of the Markov chain or **invariant** with respect to K if

$$\mu = \mu K \tag{5.6.4}$$

(b) The transition kernel K and the corresponding Markov chain $(X_j)_{j\in\mathbb{N}}$ are called μ-**reversible** if they satisfy the so-called **detailed balance condition** for all $x, x' \in H$, i.e.,

$$K(x,\, \mathrm{d}x')\,\mu(\,\mathrm{d}x) \;=\; K(x',\, \mathrm{d}x)\,\mu(\,\mathrm{d}x')\,, \tag{5.6.5}$$

where equality holds in the sense of measures on $H \times H$.

Reversibility means that provided $X_j \sim \mu$ the jump from $X_j = x$ to $X_{j+1} = x'$ has the same probability as the reverse jump from $X_j = x'$ to $X_{j+1} = x$, and (5.6.5) is equivalent to

$$\mathbb{P}(X_j \in A, X_{j+1} \in B) = \mathbb{P}(X_j \in B, X_{j+1} \in A), \quad \text{for all} \quad A, B \in \mathcal{P}(H).$$

This property is easier to verify than invariance (5.6.4) itself and it we have the following result.

Proposition 5.6.6. *Let $\mu \in \mathcal{P}(H)$ and let $K\colon H \times \mathcal{B}(H) \to [0,1]$ be a μ-reversible transition kernel. Then μ is invariant with respect to K.*

Proof. Let $A \in \mathcal{B}(H)$. Then, it follows from (5.6.5) that

$$(\mu K)(A) = \int_H K(x,A)\,\mu(\mathrm{d}x) = \int_H \int_A \underbrace{K(x,\mathrm{d}x')\,\mu(\mathrm{d}x)}_{=K(x',\mathrm{d}x)\,\mu(\mathrm{d}x')} = \int_H \int_A K(x',\mathrm{d}x)\,\mu(\mathrm{d}x')$$
$$= \int_A \int_H K(x',\mathrm{d}x)\,\mu(\mathrm{d}x') = \int_A \underbrace{K(x',H)}_{=1}\,\mu(\mathrm{d}x') = \int_A 1\;\mu(\mathrm{d}x') = \mu(A).$$

□

If the transition kernel $K\colon H \times \mathcal{B}(H) \to [0,1]$ of a Markov chain is understood as a linear operator from $\mathcal{P}(H)$ to $\mathcal{P}(H)$, then (5.6.4) simply means that μ is a fixed point of K and the Markov chain is a fixed point iteration. Classical convergence results for Markov chains rely on this point of view. They show that K is a contraction and apply the Banach Fixed Point Theorem.

However, instead we now introduce a notion of geometric convergence of Markov chains to their invariant distribution.

Definition 5.6.7. A Markov chain $(X_j)_{j\in\mathbb{N}}$ in H with transition kernel K is **$L^2_\mu(H)$-geometrically ergodic** if there exists a number $r \in [0,1)$ such that for any probability measure ν which has a density $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^2_\mu(H)$ w.r.t. μ

$$D_{\mathrm{TV}}(\nu K^j, \mu) \le C_\nu r^j \quad \text{for all} \quad j \in \mathbb{N}\,.$$

Example 5.6.8. Consider again a Markov chain in a discrete state space with M states and a row-stochastic matrix $K \in [0,1]^{M\times M}$ representing its transition kernel. Then, $\sum_{j=1}^M K_{ij} = 1$ for all $i = 1,\dots,M$, and the vector of all ones e is a right eigenvector of K to the eigenvalue 1. The corresponding left eigenvector μ is the invariant measure. In fact, it is easy to see that the spectral radius of K is 1. If all entries of K are strictly between 0 and 1 then K is called **irreducible** and it follows from the **Perron-Frobenius Theorem** that 1 is in fact **dominant**, i.e. a simple eigenvalue strictly larger in modulus than all other eigenvalues of K.

Let the distribution of the initial state of the Markov chain be $\nu \in \mathbb{R}^M$. Then, the distribution $\nu_j := \nu K^{j-1}$ of the jth state represents the jth iterate of the **power method** to find the eigenvector corresponding to the dominant eigenvalue of K, normalised such that $e^\top \nu_j = 1$. It is easy to see that the power method converges geometrically to μ:

Let $\lambda_1,\dots,\lambda_M$ be the eigenvalues of K with $1 = \lambda_1 > |\lambda_2| \ge |\lambda_3| \ge \dots \ge |\lambda_M|$ and corresponding left eigenvectors $\mu = v_1, v_2, \dots, v_M$. They form an orthonormal basis of $\mathbb{R}^M$ and thus $\nu = \sum_{m=1}^M \alpha_m v_m$ for some $\alpha_1,\dots,\alpha_M \in \mathbb{R}$. If we assume that $\alpha_1 > 0$, then we see easily that

$$\nu_{j+1} = \frac{\nu K^j}{\|\nu K^j\|_1} = \frac{\sum_{m=1}^M \lambda_m^j \alpha_m v_m}{\sum_{m=1}^M |\lambda_m|^j |\alpha_m|} = \frac{\alpha_1\mu + \sum_{m=2}^M \lambda_m^j \alpha_m v_m}{|\alpha_1|\left(1 + \sum_{m=2}^M |\lambda_m|^j |\alpha_m/\alpha_1|\right)} \;\to\; \mu \quad \text{as} \quad j\to\infty,$$

since $|\lambda_m| < 1$ for $m \ge 2$. The denominator can be bounded below by α_1. Thus,

$$\|\nu_{j+1} - \mu\|_1 \le \sum_{m=2}^M |\lambda_m|^j \left|\frac{\alpha_m}{\alpha_1}\right| \le \left(\sum_{m=2}^M \left|\frac{\alpha_m}{\alpha_1}\right|\right) |\lambda_2|^j$$

and the Markov chain converges geometrically with rate $r = |\lambda_2|$.

[Draft of October 6, 2021. Not for dissemination.]

Remark 5.6.9. In fact, this link between the spectrum of the transition kernel and of the associated Markov operator and the geometric ergodicity of the Markov chain is also a key concept in the convergenve analysis of Markov chains on general state spaces, leading to the concept of the so-called **spectral gap**, which we will come back to below. However, already for Markov chains in continuous state space, such as $\mathbb{R}^n$, we can distinguish between geometric ergodicity as in Def. 5.6.7 with a constant C_ν that depends on the initial distribution ν and **uniform ergodicity** as in the example above, i.e., that there exists a (uniform) constant $C < \infty$ for all initial distributions ν.

If the distribution of X_j converges to μ, then the Markov chain $(X_j)_{j\in\mathbb{N}}$ can be used for approximate sampling from μ, leading to the very powerful concept of **Markov chain Monte Carlo** methods for the computation of expectations. In particular, the expectation $\mathbb{E}_\mu[F(X)]$ of a function $F : H \to \mathbb{R}$ of X w.r.t. μ can be approximated by

$$\widehat{Q}^{\mathrm{MCMC}}_{N,N_0} := \frac{1}{N}\sum_{j=1}^{N} F(X_{j+N_0}), \tag{5.6.6}$$

where N is the sample size and N_0 is a so-called **burn-in parameter** to decrease the influence of the initial distribution. In fact, a strong law of large numbers and also a central limit theorem hold for $\widehat{Q}^{\mathrm{MCMC}}_{N,N_0}$ under appropriate assumptions.

Theorem 5.6.10 (Central Limit Theorem for reversible Markov chains)**.** *Let $(X_j)_{j\in\mathbb{N}}$ be a μ-reversible and $L^2_\mu(H)$-geometrically ergodic Markov chain and let F be μ-measurable. Then*

$$\sqrt{N}\left(\widehat{Q}^{\mathrm{MCMC}}_{N,N_0} - \mathbb{E}_\mu[F(X)]\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma_F^2)$$

*where $\sigma_F^2 := \lim_{N\to\infty} N\mathbb{V}\Big(\widehat{Q}^{\mathrm{MCMC}}_{N,N_0}\Big)$ denotes the **asymptotic variance**, which in this case satisfies*

$$\sigma_F^2 := \mathbb{V}\big(F(X_1)\big) + 2\sum_{j=1}^{\infty} \mathrm{cov}\big(F(X_1), F(X_{1+j})\big) \;<\; \infty. \tag{5.6.7}$$

A proof of Theorem 5.6.10 is beyond the scope of this course, but we can motivate the specific form (5.6.7) of the asymptotic variance. We have

$$\begin{aligned}\mathbb{V}\Big(\widehat{Q}^{\mathrm{MCMC}}_{N,N_0}\Big) &= \mathbb{V}\left(\frac{1}{N}\sum_{j=1}^{N} F(X_{j+N_0})\right) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathrm{cov}\big(F(X_{j+N_0}), F(X_{k+N_0})\big)\\ &= \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{V}\big(F(X_{j+N_0})\big) + \frac{1}{N^2}\sum_{j\neq k}\mathrm{cov}\big(F(X_{j+N_0}), F(X_{k+N_0})\big).\end{aligned}$$

If we now assume that $(X_j)_{j\in\mathbb{N}}$ is μ-reversible and $X_1 \sim \mu$, then $X_j \sim \mu$ for any $j \in \mathbb{N}$, which further implies that (X_j, X_{j+k}) follows the same distribution as (X_1, X_{1+k}), for all $j, k \in \mathbb{N}$. Hence,

$$\mathbb{V}(F(X_j)) = \mathbb{V}(F(X_1)) \quad \text{and} \quad \mathrm{cov}\big(F(X_{j+N_0}), F(X_{k+N_0})\big) = \mathrm{cov}\big(F(X_1), F(X_{1+|j-k|})\big)$$

and we get

$$\mathbb{V}\Big(\widehat{Q}^{\mathrm{MCMC}}_{N,N_0}\Big) = \frac{1}{N}\mathbb{V}\big(F(X_1)\big) + \frac{2}{N}\sum_{j=1}^{N}\mathrm{cov}\big(F(X_1), F(X_{1+j})\big).$$

Of course, the assumption $X_1 \sim \mu$ is rather academic and, in general, not given in practice. However, since the Markov chain in Theorem 5.6.10 is assumed to be $L^2_\mu(H)$-geometrically ergodic, the distribution of its jth state X_j converges exponentially fast to μ as $j \to \infty$.

5.6.2 The Metropolis-Hastings Markov chain Monte Carlo method

We will now describe a generic algorithm to (approximately) sample from distributions μ that are difficult to sample from directly, e.g. because they are only known in unnormalised form, such as the posterior distribution $\mu_{X|y}$ in a Bayesian inverse problem. It is based on a Markov chain of proposals and rejection sampling, and was first proposed in 1953 by Metropolis and co-authors before being generalised in 1970 by Hastings. It is considered to be one of the ten most important algoritms of the 20th century.

We focus again first on finite dimensional $H = \mathbb{R}^n$. Let $\mu \in \mathcal{P}(\mathbb{R}^n)$ with $\mu(\,\mathrm{d}x) \propto p(x)\,\mathrm{d}x$.

Definition 5.6.11 (Proposal distribution)**.** Let $Q\colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ be a Markov kernel on $\mathbb{R}^n$ with a transition density $q\colon \mathbb{R}^n \times \mathbb{R}^n \to [0,\infty)$ such that

$$Q(x,A) = \int_A q(x,x')\,\mathrm{d}x' \quad \text{for all} \quad A \in \mathcal{B}(\mathbb{R}^n).$$

The Markov kernel Q is called the **proposal kernel**.

Given this proposal kernel we are now able to define the **Metropolis-Hastings algorithm** that allows to sample from the target distribution μ. It is defined in Algorithm 1. The function $\alpha\colon \mathbb{R}^n \times \mathbb{R}^n \to [0,1]$ in (5.6.8) is called the **acceptance probability**. The special choice of α has the desired effect that the resulting Markov chain has μ as its invariant measure, as the following proposition shows.

Algorithm 1 (Metropolis–Hastings Algorithm).

Input: Proposal kernel Q with transition density q, initial distribution $\nu \in \mathcal{P}(\mathbb{R}^n)$.
Output: Realisations $(x_j)_{j\in\mathbb{N}}$ of a Markov chain $(X_j)_{j\in\mathbb{N}}$.

1 Draw a realisation $x_1 \sim \nu$.
2 **for** $j = 1, 2, \ldots$ **do**
3 Given the current state $X_j = x_j$, draw a realisation x' from $Q(x_j, \cdot)$.
4 Compute the acceptance probability

$$\alpha(x_j, x') := \min\left(1, \frac{p(x')\,q(x', x_j)}{p(x_j)\,q(x_j, x')}\right). \tag{5.6.8}$$

5 Draw an independent sample $u_{j+1} \sim \text{uniform}[0,1]$ and set

$$x_{j+1} = \begin{cases} x', & \text{if } u_{j+1} \le \alpha(x_j, x'), \\ x_j, & \text{otherwise.} \end{cases}$$

6 **end for**

Proposition 5.6.12. *The transition kernel $K\colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ of the Markov chain $(X_j)_{j\in\mathbb{N}}$ produced by Algorithm 1 with proposal kernel Q and acceptance probability α in* (5.6.8) *is given by*

$$K(x,\,\mathrm{d}x') = \alpha(x,x')Q(x,\,\mathrm{d}x') + \left(1 - \int_{\mathbb{R}^n} \alpha(x,x'')\,Q(x,\,\mathrm{d}x'')\right)\,\delta_x(\,\mathrm{d}x'), \tag{5.6.9}$$

where $\delta_x \in \mathcal{P}(\mathbb{R}^n)$ denotes the ***Dirac-measure*** *at $x \in \mathbb{R}^n$, that is, $\delta_x(A) = \mathbf{1}_A(x)$ for all $A \in \mathcal{B}(\mathbb{R}^n)$. The* ***Metropolis kernel*** *K is μ-reversible, and thus μ is invariant with respect to K.*

Proof. We first show that the Metropolis kernel K is of the form (5.6.9). By definition

$$K(x, A) = \mathbb{P}(X_{j+1} \in A \mid X_j = x).$$

We only consider proposal kernels Q with smooth density q and note that in that case the probability that $x' = x$, i.e. that we propose x given $X_j = x$ is zero. In that case it suffices to study the cases $\mathbb{P}(X_{j+1} = x \mid X_j = x)$, i.e., the probability that the proposal is rejected, and $\mathbb{P}(X_{j+1} = A \mid X_j = x)$ for $x \notin A$, i.e., the proposal $x' \in A$ and x' is accepted.

The rejection probability for a proposal is exactly $1 - \alpha(x, x')$ with $x' \sim Q(x, \mathrm{d}x')$. Thus,

$$\mathbb{P}(X_{j+1} = x \mid X_j = x) = \int_{\mathbb{R}^n} (1 - \alpha(x, x'))\, Q(x, \mathrm{d}x') = 1 - \int_{\mathbb{R}^n} \alpha(x, x')\, Q(x, \mathrm{d}x').$$

On the other hand, the probability that $\mathbb{P}(X_{j+1} = A \mid X_j = x)$ for $x \notin A$ is

$$\mathbb{P}(X_{j+1} = A \mid X_j = x) = \int_A \alpha(x, x')\, Q(x, \mathrm{d}x').$$

Combining these two cases we obtain (5.6.9).

To show detailed balance, we consider first $A, B \in \mathcal{B}(\mathbb{R}^n)$ with $A \cap B = \emptyset$. W.l.o.g. we can assume that $p(x)q(x, x') > 0$ for all $x, x' \in \mathbb{R}^n$ (otherwise we simply have to restrict the integrations below accordingly). Since $A \cap B = \emptyset$ we have

$$\begin{aligned}
\int_{A\times B} K(x, \mathrm{d}x')\, \mu(\mathrm{d}x) &= \int_A \int_B \alpha(x, x') Q(x, \mathrm{d}x')\, \mu(\mathrm{d}x) \\
&= \frac{1}{c} \int_A \int_B \min\left(1, \frac{p(x')\, q(x', x)}{p(x)\, q(x, x')}\right) p(x)\, q(x, x')\, \mathrm{d}x'\, \mathrm{d}x \\
&= \frac{1}{c} \int_A \int_B \min\big(p(x)\, q(x, x'),\ p(x')\, q(x', x)\big)\, \mathrm{d}x'\, \mathrm{d}x \\
&= \frac{1}{c} \int_A \int_B \min\left(\frac{p(x)\, q(x, x')}{p(x')\, q(x', x)}, 1\right) p(x')\, q(x', x)\, \mathrm{d}x'\, \mathrm{d}x \\
&= \int_A \int_B \alpha(x', x)\, \mu(\mathrm{d}x')\, Q(x', \mathrm{d}x) \\
&= \int_B \int_A \alpha(x', x)\, Q(x', \mathrm{d}x)\, \mu(\mathrm{d}x') \\
&= \int_{A\times B} K(x', \mathrm{d}x) \mu(\mathrm{d}x')
\end{aligned}$$

If $A \cap B \neq \emptyset$ we also need to consider rejections, but detailed balance can again be shown similarly, since $\mathbb{P}(X_{j+1} = x \mid X_j = x)$ is clearly symmetric. □

The big advantage of the Metropolis-Hastings (MH) algorithm is that we only need to be able to evaluate the unnormalised density p of the target measure μ and the density q of the proposal kernel Q. The proposal density is often chosen to be very simple, e.g., **symmetric**, such that

$$q(x, x') = q(x', x) \qquad \forall x, x' \in \mathbb{R}^n. \tag{5.6.10}$$

In that special case, the acceptance probability simplifies to

$$\alpha(x, x') := \min\left(1, \frac{p(x')}{p(x)}\right). \tag{5.6.11}$$

The 'rule' (5.6.11) can be interpreted such that x' is definitely accepted (i.e. with probability 1) if $p(x') \geq p(x)$, and if $p(x') < p(x)$ it is accepted with probability $\frac{p(x')}{p(x)}$. However, importantly, in comparison to pure rejection sampling, when a proposal is rejected we include the previous state x_j again as state x_{j+1}, i.e. we increase the 'weight' of that state due to its relatively high probability density (at least when compared to the density of the proposed state x').

A poopular proposal kernel is the following.

Example 5.6.13 (Gaussian random walk)**.** The proposal kernel $Q\colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ is chosen to be

$$Q(s; x, \cdot) = \mathcal{N}(x, s^2 I), \tag{5.6.12}$$

where $s > 0$ is the **step size** parameter that can be optimised or calibrated. A well established rule of thumb (which also has some theoretical foundations) is that

$$s \text{ should be chosen such that } \quad \bar{\alpha} := \int_{\mathbb{R}^n} \alpha(x, x')\, Q(s; x, \mathrm{d}x')\, \mu(\mathrm{d}x) \approx 0,21, \tag{5.6.13}$$

where $\bar{\alpha}$ is called the *mean acceptance rate* which can be estimated on the basis of a short trial run of the Markov chain in practice.

The proposal kernel $Q(s; \cdot, \cdot)$ has a transition density

$$q(s; x, x') \propto \exp\left(-\frac{1}{2s^2}\|x' - x\|^2\right)$$

and is thus clearly symmetric, i.e., it satisfies (5.6.10).

Theorem 5.6.14. *Let p be the unnormalised density of the target measure $\mu \in \mathcal{P}(\mathbb{R}^n)$. If the proposal kernel Q in Algorithm 1 is such that*

$$p(x') > 0 \quad \textit{implies} \quad q(x, x') > 0, \quad \textit{for all} \ \ x \in \mathbb{R}^n, \tag{5.6.14}$$

and if

$$\mathbb{P}\big(\alpha(x_j, X') = 1\big) < 1, \quad \textit{for all} \ \ j \in \mathbb{N}, \tag{5.6.15}$$

then the Markov chain $(X_j)_{j\in\mathbb{N}}$ produced by the MH algorithm is $L^2_\mu(\mathbb{R}^n)$-geometrically ergodic and the Central Limit Theorem 5.6.10 applies.

Remark 5.6.15. Note that condition (5.6.14) on the proposal distribution is similar to the condition required on the importance distribution q for the ratio estimator in Lemma 5.5.4. Condition (5.6.15), on the other hand, guarantees that the Markov chain is **aperiodic**. However, it is somewhat academic, since there is not much point in applying the MH algorithm, when the proposal distribution is so good that proposed states are accepted a.s.

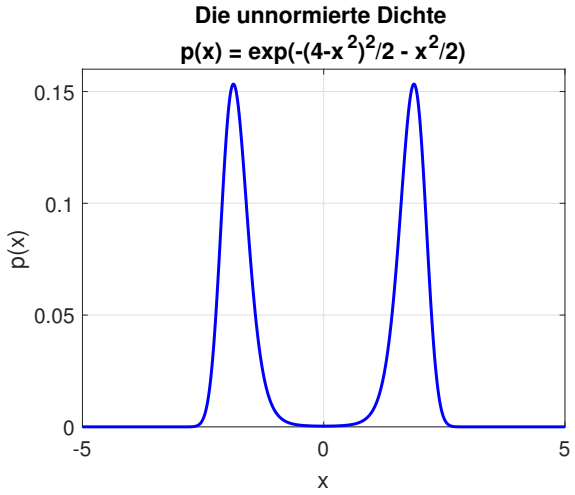


Figure 5.9: The unnormalised posterior density from Example 5.6.16.

Example 5.6.16. Let us consider the MH algorithm for a simple one dimensional posterior distribution. In particular, we consider a RV $X : \Omega \to \mathbb{R}$ with prior distribution $X \sim \mu_X = \mathcal{N}(0,1)$, conditioned on the observation $y = 4$ of $Y = X^2 + E$ with $E \sim \mathcal{N}(0,1)$. Thus,

$$\pi_{X|Y}(x|y) \propto \exp\left(-\tfrac{1}{2}(y - x^2)^2\right) \exp\left(-\tfrac{1}{2}x^2\right) = \exp\left(-\tfrac{1}{2}[(y - x^2)^2 + x^2]\right),$$

see also Figure 5.9.

Let us use the MH algorithm with random walk proposal kernel $Q(s; \cdot, \cdot)$ defined in (5.6.12) to sample from $\pi_{X|Y}(x|y)$. The acceptance probability is

$$\alpha(x, x') = \min\left(1, \frac{\pi_{X|Y}(x'|y)}{\pi_{X|Y}(x|y)}\right) = \min\left(1, \frac{\exp\left(-\tfrac{1}{2}[(4 - (x')^2)^2 + (x')^2]\right)}{\exp\left(-\tfrac{1}{2}[(4 - x^2)^2 + x^2]\right)}\right).$$

The criterion (5.6.13) is satisfied for a step size of roughly $s = 1.5$. As the initial state, we choose $x_1 = 0$, i.e. $\nu = \delta_0$, and show in Figure 5.10 a realisation $(x_j)_{j\in\mathbb{N}}$ of the Markov chain produced by the resulting MH algorithm. In Figure 5.11, we also show a histogram of relative frequencies averaged over the path $(x_j)_{j\in\mathbb{N}}$ of the Markov chain compared to the true, normalised posterior density and observe a very good agreement.

5.6.3 Extension to infinite dimensions

Let us briefly discuss the extension to a general, possibly infinite-dimensional, separable Hilbert space H. Except for the acceptance probability α in (5.6.8), all the elements of the MH algorithm in Algorithm 1 were not specific to $\mathbb{R}^n$.

In particular, if $\mu \in \mathcal{P}(H)$, $\nu \in \mathcal{P}(H)$ and $Q\colon H \times \mathcal{B}(H) \to [0,1]$ are the target measure, a measure for the initial state and a proposal kernel on H, respectively, the only remaining question is how to choose α, such that the Markov chain produced by Algorithm 1 is μ-reversible. With the probability measures $\rho, \rho^\top \in \mathcal{P}(H \times H)$ defined as

$$\rho(\mathrm{d}x, \mathrm{d}x') := Q(x, \mathrm{d}x')\mu(\mathrm{d}x) \quad \text{and} \quad \rho^\top(\mathrm{d}x, \mathrm{d}x') := \rho(\mathrm{d}x', \mathrm{d}x), \tag{5.6.16}$$

the following proposition can be proved similarly to Proposition 5.6.12.

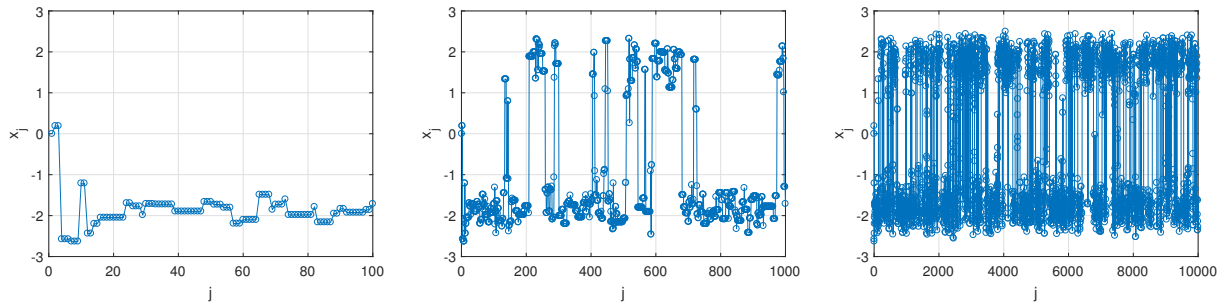


Figure 5.10: A path of the Markov chain produced in Example 5.6.16.

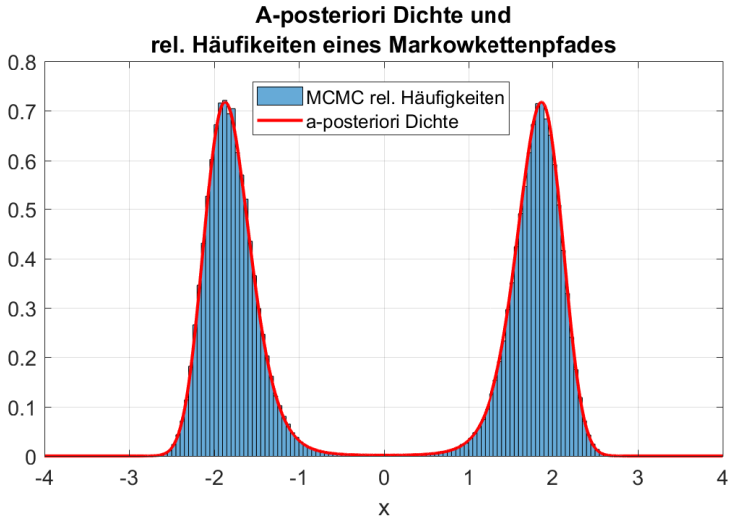


Figure 5.11: The normalised posterior density for Example 5.6.16 (red curve) and the histogram of the realisation of the Markov chain from Figure 5.10.

Proposition 5.6.17. *If the Radon-Nikodym derivative* $\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}\colon H\times H\to[0,\infty)$ *exists and we replace the acceptace probability* (5.6.8) *in Algorithm 1 by*

$$\alpha(x_j,x')=\min\left(1,\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}(x_j,x')\right), \tag{5.6.17}$$

then the transition kernel $K\colon H\times\mathcal{B}(H)\to[0,1]$ *of the Markov chain* $(X_j)_{j\in\mathbb{N}}$ *that is produced by Algorithm 1 with proposal kernel* Q *is given by*

$$K(x,\,\mathrm{d}x')=\alpha(x,x')Q(x,\,\mathrm{d}x')+\int_H(1-\alpha(x,x''))\,Q(x,\,\mathrm{d}x'')\,\delta_x(\,\mathrm{d}x')$$

and it is μ*-reversible.*

In infinite-dimensional spaces the existence of $\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}$ is not guaranteed. For $H=\mathbb{R}^n$,

$$\rho(\,\mathrm{d}x,\,\mathrm{d}x')=Q(x,\,\mathrm{d}x')\mu(\,\mathrm{d}x)=q(x,x')\,p(x)\,\,\mathrm{d}x'\,\mathrm{d}x\,,$$

so that, provided $q(x,x')\,p(x)>0$, we have

$$\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}(x,x')=\frac{q(x',x)\,p(x')}{q(x,x')\,p(x)},$$

i.e., the two definitions of α in (5.6.17) and (5.6.8) agree.

A possible way to ensure the existence of $\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}$ in infinite dimensions in the case of a posterior measure $\mu=\mu_{X|y}$ with

$$\frac{\mathrm{d}\mu_{X|y}}{\mathrm{d}\mu_X}(x)\propto\exp\left(-\frac{1}{2}\|y-\Phi(x)\|_\Sigma^2\right)=:\exp\big(-\mathcal{M}(x)\big),$$

is to choose a proposal kernel Q that is **prior-reversible**, i.e.,

$$\eta(\,\mathrm{d}x,\,\mathrm{d}x'):=Q(x,\,\mathrm{d}x')\mu_X(\,\mathrm{d}x)=Q(x',\,\mathrm{d}x)\mu_X(\,\mathrm{d}x')=:\eta^\top(\,\mathrm{d}x,\,\mathrm{d}x')\,. \tag{5.6.18}$$

Multiplying both sides with $\exp\big[-\mathcal{M}(x)-\mathcal{M}(x')\big]$, this implies

$$\exp(-\mathcal{M}(x'))\underbrace{Q(x,\,\mathrm{d}x')\mu_{X|y}(\,\mathrm{d}x)}_{=\rho(\,\mathrm{d}x,\,\mathrm{d}x')}=\exp(-\mathcal{M}(x))\underbrace{Q(x',\,\mathrm{d}x)\mu_{X|y}(\,\mathrm{d}x')}_{=\rho^\top(\,\mathrm{d}x,\,\mathrm{d}x')}\,.$$

Thus, $\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}$ is well-defined (under resonable conditions on the observation operator Φ) and

$$\alpha(x,x')=\min\left(1,\frac{\mathrm{d}\rho^\top}{\mathrm{d}\rho}(x,x')\right)=\min\Big(1,\exp\big(\mathcal{M}(x)-\mathcal{M}(x')\big)\Big) \tag{5.6.19}$$

One way to trivially obtain prior-reversibility is an **independence sampler** with $Q(x,\,\mathrm{d}x')=\mu_X(\,\mathrm{d}x')$, independently of x, which uses independent draws from the prior as proposals in Algorithm 1. However, this will not work well in practice if the data is informative and the posterior concentrates only on part of the support of μ_X.

A more efficient alternative can be obtained through a slight modification of the Gaussian random walk proposal kernel from Example 5.6.13.

Proposition 5.6.18 (pCN proposals)**.** *Let X be a RV on H with Gaussian prior $\mu_X = \mathcal{N}(0, C)$ and let $\mu_{X|y} \in \mathcal{P}(H)$ be a posterior distribution of the usual form with additive Gaussian likelihood, such that*

$$\mu_{X|y}(\,\mathrm{d}x) \propto \exp(-\mathcal{M}(x))\,\mu_X(\,\mathrm{d}x), \quad \textit{with} \quad \mathcal{M}(x) = \frac{1}{2}\|y - \Phi(x)\|_\Sigma^2 .$$

Then, Algorithm 1 with the so-called **preconditioned Crank-Nicolson (pCN) proposal kernel**

$$Q(s; x, \cdot) := \mathcal{N}\left(\sqrt{1-s^2}\,x, s^2 C\right), \quad \textit{for} \;\; s \in (0,1), \tag{5.6.20}$$

and acceptance probability

$$\alpha(x, x') = \min\left(1, \exp\left(\mathcal{M}(x) - \mathcal{M}(x')\right)\right) \tag{5.6.21}$$

produces a $\mu_{X|y}$-reversible Markov chain.

Proof. To see that $Q(s; x, \cdot)$ in (5.6.20) is prior-reversible, consider $\eta(\,\mathrm{d}x,\,\mathrm{d}x') = Q(s; x,\,\mathrm{d}x')\mu_X(\,\mathrm{d}x)$ and let X, W be two independent samples from $\mu_X = \mathcal{N}(0, C)$. Then,

$$\begin{pmatrix} X \\ X' \end{pmatrix} := \begin{bmatrix} I & 0 \\ \sqrt{1-s^2}I & sI \end{bmatrix} \begin{pmatrix} X \\ W \end{pmatrix} = \begin{pmatrix} X \\ \sqrt{1-s^2}X + sW \end{pmatrix} \;\sim\; \eta\,.$$

As a linear combination of Gaussians, the RV (X, X') is jointly Gaussian, and as in (5.3.3), it follows that

$$\eta \;=\; \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} C & \sqrt{1-s^2}C \\ \sqrt{1-s^2}C & C \end{bmatrix}\right),$$

which is symmetric and independent of the order of the two RVs X and X'. Thus, $\eta = \eta^\top$. □

The assumption that $\mu_X = \mathcal{N}(0, C)$ in Prop. 5.6.18 is crucial. The result is not true in general.

Remark 5.6.19. In practice, we never work with infinite-dimensional distributions. However, the classical Gaussian random walk in Example 5.6.13 is not prior-reversible and its acceptance probability is not well-defined in $H = \mathbb{R}^n$ in the limit as $n \to \infty$. In fact, the step size s to achieve $\bar{\alpha} \approx 0.21$ tends to 0 as $n \to \infty$, so that the proposal distribution degenerates and the convergence of the algorithm becomes very poor.

In contrast, MH algorithms that are well-defined also in the infinite-dimensional limit, such as the pCN algorithm in Proposition 5.6.18, typically lead to a dimension-independent convergence and are therefore suitable also for very high dimensions n of the state space.

5.6.4 Efficient proposal kernels and multilevel MCMC

There are a number of more efficient and more cutting-edge proposal distributions, but to describe those would go beyond the scope of this course. Research in these directions is also at the centre of interest in both our research groups in Heidelberg.

Another promising direction which is at the heart of our research is the extension of the multilevel idea to MCMC, but again we will not have the tim to cover this; for details see

- T.J. Dodwell, C. Ketelsen, R. Scheichl, A.L. Teckentrup, Multilevel Markov chain Monte Carlo, *SIAM Review* **61**:509–545, 2019.

5.7 Variational methods

Contrary to MCMC methods, variational inference is based on optimization instead of sampling. The general idea can be described as follows: Let $\mathcal{H}$ be a **variational family** of probability measures on $\mathbb{R}^n$. To approximate the posterior $\mu_{X|y}$, we determine as a surrogate the best approximation within the class $\mathcal{H}$ w.r.t. the KL-divergence

$$\rho^* \in \operatorname{argmin}_{\rho\in\mathcal{H}} D_{\mathrm{KL}}(\rho\|\mu_{X|y}). \tag{5.7.1}$$

Depending on the choice of $\mathcal{H}$, in general such ρ^* need not exist or be unique. However, if it does exist, it can be used in place of $\mu|_{X|y}$ approximate the quantities we are interested in, such as the conditional mean $\mathbb{E}_{\mu_{X|y}}[X] \approx \mathbb{E}_{\rho^*}[X]$. For this reason the family $\mathcal{H}$ has to be chosen such that expectations $\mathbb{E}_\rho[f]$ for $\rho \in \mathcal{H}$ are easy and cheap to compute (this is losely referred to as being "tractable"). This is for example the case, if we have a method of computing iid samples from ρ, as we may then approximate $\mathbb{E}_\rho[f]$ with a Monte Carlo estimate.

Example 5.7.1. Set $\mathcal{H} := \{\mathcal{N}(\mu,\Sigma) : \mu \in \mathbb{R}^m,\ \Sigma \in \mathbb{R}^{m\times m} \text{ SPD}\}$. Then (5.7.1) corresponds to fitting a Gaussian to the posterior w.r.t. the KL-divergence. Since a Gaussian is uniquely determined through its expectation and covariance, we merely need to determine $\mu \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$. In practice this is done by minimizing $D_{\mathrm{KL}}(\rho\|\mu_{X|y})$ with optimization methods such as gradient descent or—in particular for large datasets and high-dimensional parameters—stochastic gradient descent. Note that (5.7.1) will in general not yield the same result as the Laplace approximation (5.3.4).

In this section we concentrate on the finite dimensional case and let the parameter $X \in \mathbb{R}^n$, the data $y \in \mathbb{R}^m$, and the posterior $\mu_{X|y} \ll \lambda_n$ with density $\pi_{X|y}(x) = \frac{\pi_{X,Y}(x,y)}{Z(y)} = \frac{\pi_{Y|x}(y)\pi_X(x)}{Z(y)}$, and normalization constant

$$Z(y) = \int_{\mathbb{R}^n} \pi_{X,Y}(x,y)\,\mathrm{d}x = \int_{\mathbb{R}^n} \pi_{Y|x}(x)\pi_X(x)\,\mathrm{d}x \tag{5.7.2}$$

as in Chapter 4.

5.7.1 ELBO

The normalization constant $Z(y)$ is also referred to as the **model evidence**. Recall that $y \mapsto Z(y) = \pi_Y(y)$ is the marginal density of the data. Assume that $\rho \ll \lambda_n$ for all $\rho \in \mathcal{H}$ and denote $f_\rho(x) = \frac{\mathrm{d}\rho}{\mathrm{d}\lambda_n}(x)$. The **objective function** to be minimized in (5.7.1) then equals

$$D_{\mathrm{KL}}(\rho\|\mu_{X|y}) = \mathbb{E}_\rho\left[\log\left(\frac{f_\rho}{\pi_{X|y}}\right)\right] = \mathbb{E}_\rho[\log(f_\rho)] - \mathbb{E}_\rho[\log(\pi_{X,Y}(\cdot,y))] + \log(Z(y)), \tag{5.7.3}$$

where, as earlier, we use the notation $\mathbb{E}_\rho[F] = \int F(x)\,\mathrm{d}\rho(x)$. With

$$\mathrm{ELBO}(\rho) := \mathbb{E}_\rho[\log(\pi_{X,Y}(\cdot,y))] - \mathbb{E}_\rho[\log(f_\rho)],$$

the optimization problem (5.7.1) can be reformulated as

$$\operatorname{argmin}_{\rho\in\mathcal{H}} D_{\mathrm{KL}}(\rho\|\mu_{X|y}) = \operatorname{argmax}_{\rho\in\mathcal{H}} \mathrm{ELBO}(\rho),$$

with the equality being an equality of sets in case there are multiple minimizers and maximizers. By Jensen's inequality for concave functions,

$$\mathrm{ELBO}(\rho) = \mathbb{E}_\rho\left[\log\left(\frac{\pi_{X,Y}(\cdot,y)}{f_\rho}\right)\right] \leq \log\left(\mathbb{E}_\rho\left[\frac{\pi_{X,Y}(\cdot,y)}{f_\rho}\right]\right) = \log(Z(y)).$$

Therefore $\mathrm{ELBO}(\rho)$ is a lower bound of the logarithm of the model evidence; hence the acronym ELBO (evidence lower bound). This could also be deduced from $0 \leq D_{\mathrm{KL}}(\rho\|\mu_{X|y}) = \log(Z(y)) - \mathrm{ELBO}(\rho)$.

Remark 5.7.2. In principle another distance or divergence apart from the KL-divergence could be used in (5.7.1), but the KL-divergence has the advantage that the resulting optimization problem can be formulated as maximizing $\mathrm{ELBO}(\rho)$, which is independent of (the in practice unknown constant) $Z(y)$.

5.7.2 CAVI

In this section we consider coordinate ascent mean-field variational inference (CAVI).

To simplify the optimization problem (5.7.1), one can choose a variational family $\mathcal{H}$ which factorizes over individual variables: $\mathcal{H} = \{\otimes_{j=1}^n \rho_j : \rho_j \in \mathcal{H}_j\}$ for certain classes $\mathcal{H}_j$ of probability measures on $\mathbb{R}$. Note that this corresponds to the assumption of the unknown parameters $(X_j)_{j=1}^n$ being independent. Assuming $\rho_j \ll \lambda$ and setting $f_{\rho_j} := \frac{\mathrm{d}\rho_j}{\mathrm{d}\lambda}$, the density function f_ρ of $\rho = \otimes_{j=1}^n \rho_j$ becomes

$$f_\rho(x_1,\dots,x_n) = \prod_{j=1}^n f_{\rho_j}(x_j).$$

The surrogate ρ^* of $\mu_{X|y}$ in (5.7.1) can in this case capture marginal densities of the posterior, but it cannot capture correlation between the different parameters. Due to the type of ansatz, the method is referred to as **mean field variational inference**.

Remark 5.7.3. More generally, one can partition $\{1,\dots,n\}$ via $1 = i_1 < \cdots < i_m = n+1$ and consider $f_\rho(x_1,\dots,x_n) = \prod_{j=1}^{m-1} f_j(x_{i_j},\dots,x_{i_{j+1}-1})$.

The coordinate ascent algorithm tries to optimize $\mathrm{ELBO}(\rho) = \mathrm{ELBO}(\otimes_{j=1}^n \rho_j)$ by repeatedly iterating through all $j = 1,\dots,n$, each time only updating (i.e. maximizing in) ρ_j. To describe the procedure in more detail let us first introduce the notation

$$\mathbb{E}_{-j}[f](x_j) := \int_{\mathbb{R}^{n-1}} f(x)\,\mathrm{d}\rho_1(x_1)\dots\,\mathrm{d}\rho_{j-1}(x_{j-1})\,\mathrm{d}\rho_{j+1}(x_{j+1})\dots\,\mathrm{d}\rho_n(x_n)$$

for $f : \mathbb{R}^n \to \mathbb{R}$ and $x = (x_1,\dots,x_n)^\top \in \mathbb{R}^n$. Then $\mathbb{E}_{-j}[f]$ is a function of x_j. Fixing ρ_i for all $i \neq j$, we write

$$\rho_j^* := \mathrm{argmax}_{\rho_j \in \mathcal{H}_j}\,\mathrm{ELBO}(\rho) = \mathrm{argmax}_{\rho_j \in \mathcal{H}_j}\,\mathbb{E}_\rho[\log(\pi_{X,Y})] - \mathbb{E}_\rho[\log(f_\rho)].$$

If $\mathcal{H}_j$ is chosen as the set of all probability measures on $\mathbb{R}$ which have a density (i.e. are absolutely continuous w.r.t. λ), then the argmax can be expressed explicitly:

Lemma 5.7.4. *With the above choice of $\mathcal{H}_j$ it holds for $f_{\rho_j^*} := \frac{\mathrm{d}\rho_j^*}{\mathrm{d}\lambda}$ that*

$$f_{\rho_j^*}(x_j) \propto \exp\left(\mathbb{E}_{-j}[\log(\pi_{X,Y})]\right)$$

in case $\exp(\mathbb{E}_{-j}[\log(\pi_{X,Y})]) \in L^1(\mathbb{R})$.

Proof. Due to $f_\rho(x) = \prod_{j=1}^n f_{\rho_j}(x_j)$ holds $\mathbb{E}_\rho[\log(f_\rho)] = \sum_{j=1}^n \mathbb{E}_{\rho_j}[\log(f_{\rho_j})]$ and therefore

$$\begin{aligned}
\rho_j^* &= \operatorname{argmax}_{\rho_j \in \mathcal{H}_j} \mathbb{E}_\rho[\log(\pi_{X,Y})] - \mathbb{E}_{\rho_j}[\log(f_{\rho_j})] \\
&= \operatorname{argmax}_{\rho_j \in \mathcal{H}_j} \mathbb{E}_{\rho_j}\left[\mathbb{E}_{-j}[\log(\pi_{X,Y})]\right] - \mathbb{E}_{\rho_j}[\log(f_{\rho_j})] \\
&= \operatorname{argmax}_{\rho_j \in \mathcal{H}_j} \mathbb{E}_{\rho_j}\left[\log(\exp(\mathbb{E}_{-j}[\log(\pi_{X,Y})]))\right] - \mathbb{E}_{\rho_j}[\log(f_{\rho_j})] \\
&= \operatorname{argmax}_{\rho_j \in \mathcal{H}_j} -\mathbb{E}_{\rho_j}\left[\log\left(\frac{f_{\rho_j}}{\exp(\mathbb{E}_{-j}[\log(\pi_{X,Y})])}\right)\right]. \qquad (5.7.4)
\end{aligned}$$

The last expression in (5.7.4) is up to a constant equal to the negative KL-divergence between ρ_j and the probability measure with density proportional to $\exp(\mathbb{E}_{-j}[\log(\pi_{X,Y})])$ (the constant is $\log(\int_{\mathbb{R}} \exp(\mathbb{E}_{-j}[\log(\pi_{X,Y})])\,\mathrm{d}x_j)$ and does not depend on ρ_j). Since the KL-divergence between two measures is nonnegative, and it is equal to 0 if and only if they are the same, this concludes the proof. □

This leads to Alg. 2.

Algorithm 2 CAVI; input: tolerance, $\pi_{X,Y}$

```
while ELBO(ρ) > tolerance do
  for j = 1, ..., n do
    set f_{ρ_j} ∝ exp(E_{-j}[log(π_{X,Y})])
  end for
  compute ELBO(ρ)
end while
```

5.7.3 Transport Maps

As mentioned before, the optimization problem (5.7.1) only yields a useful result ρ^*, in case the probability measures $\rho \in \mathcal{H}$ are such that they allow for simple computation of quantities like $\mathbb{E}_{\mu_{X|y}}[X] \approx \mathbb{E}_{\rho^*}[X]$. For this reason, variational methods are often applied with somewhat simple variational families $\mathcal{H}$ such as in Example 5.7.1, which are not able to capture more complex features of the posterior.

Transport maps provide a general approach which is in principle suitable for arbitrarily complex posteriors. Set

$$\mathcal{H} := \{T_\sharp \eta \,:\, T \in \mathcal{T}\}, \qquad (5.7.5)$$

where $\eta \ll \lambda^n$ is a fixed **reference probability measure** on $\mathbb{R}^n$ (typically $\eta \sim \mathcal{N}(0, I)$), and $\mathcal{T}$ is a family of **transport maps**, which we here assume to be bijective maps $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T \in C^1$ and also $T^{-1} \in C^1$ (i.e. T is a diffeomorphism). Recall that the pushforward measure is defined as $T_\sharp\eta(A) := \eta(T^{-1}(A))$, and in this section we'll also use the **pullback** measure defined via $T^\sharp\eta(A) := \eta(T(A))$. The optimization problem (5.7.1) can then be equivalently stated as finding

$$T^* := \operatorname{argmin}_{T \in \mathcal{T}} D_{\mathrm{KL}}(T_\sharp\eta \| \mu_{X|y}) = \operatorname{argmin}_{T \in \mathcal{T}} D_{\mathrm{KL}}(\eta \| T^\sharp \mu_{X|y}),$$

where the second inequality will be shown in (5.7.7). The desired quantity in (5.7.1) is then $\rho^* = T^*_\sharp\eta$. We will next discuss how T^* can be used to compute or approximate $\mathbb{E}_{\rho^*}[X] \approx \mathbb{E}_{\mu_{X|y}}[X]$—our main goal in Bayesian inference. Afterwards we will show that T satisfying $T_\sharp\eta = \mu_{X|y}$ exists (under certain assumptions on η and $\mu_{X|y}$), thus justifying this approach.

Sampling using measure transport

Note that for any $A \in \mathcal{B}(\mathbb{R}^n)$, a RV $S \sim \eta$ and a bijection $T : \mathbb{R}^n \to \mathbb{R}^n$

$$\mathbb{P}[T(S) \in A] = \mathbb{P}[S \in T^{-1}(A)] = \eta(T^{-1}(A)) = T_\sharp \eta(A).$$

Thus

$$S \sim \eta \qquad \Rightarrow \qquad T(S) \sim T_\sharp \eta. \tag{5.7.6}$$

Hence, with the minimizer T^* in (5.7.5), an approximation to the conditional mean $\mathbb{E}_{\mu_{X|y}}[X]$ is obtained by the Monte Carlo estimate

$$\mathbb{E}_{\mu_{X|y}}[X] \approx \mathbb{E}_{T^*_\sharp \eta}[X] \approx \frac{1}{N} \sum_{j=1}^{N} T^*(S_j), \qquad S_j \sim \eta.$$

Having computed T^*, it is therefore easy to approximate the conditional mean.

Next, let us write down the optimization problem (5.7.5) in terms of densities. To this end assume $\eta \ll \lambda_n$ and denote $f_\eta = \frac{\mathrm{d}\eta}{\mathrm{d}\lambda_n}$. As pointed out in Rmk. 3.2.25 we have

$$\frac{\mathrm{d}T_\sharp \rho}{\mathrm{d}\lambda_n}(x) = f_\eta(T^{-1}(x)) \det dT^{-1}(x),$$

where $dT : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ denotes the Jacobian matrix of T. Similarly with the density $\pi_{X|y}$ of $\mu_{X|y}$

$$\frac{\mathrm{d}T^\sharp \mu_{X|y}}{\mathrm{d}\lambda_n}(x) = \frac{\mathrm{d}(T^{-1}_\sharp \mu_{X|y})}{\mathrm{d}\lambda_n}(x) = \pi_{X|y}(T(x)) \det dT(x).$$

Hence, using Thm. 3.2.24,

$$\begin{aligned} D_{\mathrm{KL}}(T_\sharp \eta \| \mu_{X|y}) &= \int_{\mathbb{R}^n} \log\left(\frac{f_\eta(T^{-1}(x)) \det dT^{-1}(x)}{\pi_{X|y}(x)} \right) \mathrm{d}T_\sharp \eta(x) \\ &= \int_{\mathbb{R}^n} \log\left(\frac{f_\eta(x) \det dT^{-1}(T(x))}{\pi_{X|y}(T(x))} \right) \mathrm{d}\eta(x) \\ &= \int_{\mathbb{R}^n} \log\left(\frac{f_\eta(x)}{\pi_{X|y}(T(x)) \det dT(x)} \right) \mathrm{d}\eta(x) \\ &= D_{\mathrm{KL}}(\eta \| T^\sharp \mu_{X|y}). \end{aligned} \tag{5.7.7}$$

With $\pi_{X,Y}(x, y) = \pi_{X|y}(x) Z(y)$ the optimization problem reads: Find

$$\begin{aligned} &\mathrm{argmin}_{T \in \mathcal{T}} \int_{\mathbb{R}^n} \log(f_\eta(x)) - \log(\pi_{X,Y}(T(x), y)) - \log(\det dT(x)) + \log(Z(y)) \, \mathrm{d}\eta(x) \\ &= \mathrm{argmin}_{T \in \mathcal{T}} \int_{\mathbb{R}^n} \log(f_\eta(x)) - \log(\pi_{X,Y}(T(x), y)) - \log(\det dT(x)) \, \mathrm{d}\eta(x), \end{aligned}$$

where this optimization problem is again independent of the constant $Z(y)$. We emphasize once more, that as earlier depending on the choice of $\mathcal{T}$ this argmin need neither exist nor be unique in general.

In practice $\mathcal{T}$ is chosen as some parametrization of possible transport maps, for instance using polynomial expansions or neural networks with a suitable network architecture. The problem is then solved by performing gradient descent (or other optimization techniques) on the approximate objective

$$\frac{1}{N}\sum_{j=1}^{N} \log(f_\eta(S_j)) - \log(\pi_{X,Y}(T(S_j), y)) - \log(\det dT(S_j))$$

with iid samples $S_j \sim \eta$. This optimization problem is in general nonconvex and highly nontrivial to solve. We summarize this strategy in Alg. 3.

Algorithm 3 Approximate CM computation using transport; input: f_η, $\pi_{X,Y}$, n

$\tilde{T} \leftarrow \operatorname{argmin}_{T\in\mathcal{T}} \mathbb{E}_{x\sim\eta}[\log(f_\eta) - \log(\pi_{X,Y}(T(x), y)) - \log(\det dT(x))]$
$S_j \sim \eta$ iid for $j = 1, \dots, n$
return $\frac{1}{n}\sum_{j=1}^{n} \tilde{T}(S_j)$

Remark 5.7.5. Note that in the general form (5.7.1), every $\rho \in \mathcal{H}$ needs to be such that we can easily sample from it in order to compute a Monte Carlo approximation. Using the transport maps approach (5.7.5), due to (5.7.6) this automatically holds as long as we can sample from η.

Triangular transports

In this section we show the existence of transport maps pushing forward a reference measure to a target. We denote in the following by μ a target measure on $\mathbb{R}^n$. For the moment we can think of μ as the posterior $\mu_{X|y}$, however we'll later also consider the target $\mu_{Y,X}$.

In the following lemma, we write $F^{[-1]} : [0,1] \to \mathbb{R}$ for the inverse CDF, i.e.

$$F^{[-1]}(a) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq a\}.$$

As earlier, for a set $A \in \mathcal{B}(\mathbb{R})$ we write $F^{-1}(A) = \{x \,:\, F(x) \in A\}$.

Lemma 5.7.6 (Monotone transport in 1d)**.** *Let η, μ be two probability measures on $\mathbb{R}$ with CDFs $F_\eta : \mathbb{R} \to [0,1]$ and $F_\mu : \mathbb{R} \to [0,1]$, and let η be atomless. Then $T := F_\mu^{[-1]} \circ F_\eta$ is nondecreasing and satisfies $T_\sharp \eta = \mu$.*

Proof. We have $F_\eta(x) = \eta((-\infty, x])$ and $F_\mu(x) = \mu((-\infty, x])$. Since η is atomless, $F_\eta : \mathbb{R} \to [0,1]$ is continuous with $\lim_{x\to-\infty} F_\eta(x) = 0$ and $\lim_{x\to\infty} F_\eta(x) = 1$. Hence for $a \in (0,1)$, $F_\eta^{-1}([0,a]) = \{x \in \mathbb{R} \,:\, F_\eta(x) \leq a\}$ is the closed interval $(-\infty, x_a]$ where $x_a = \max\{x \in \mathbb{R} \,:\, F_\eta(x) = a\}$. Thus $\eta(F_\eta^{-1}([0,a])) = \eta((-\infty, x_a]) = F_\eta(x_a) = a$ for all $a \in (0,1)$, implying that $(F_\eta)_\sharp \eta = \lambda|_{[0,1]}$ (the Lebesgue measure on $[0,1]$).

Next we show $(F_\mu^{[-1]})_\sharp \lambda|_{[0,1]} = \mu$. We have $F_\mu^{[-1]} : [0,1] \to \mathbb{R}$, and for every $x \in \mathbb{R}$

$$\begin{aligned}(F_\mu^{[-1]})^{-1}((-\infty, x]) &= \{a \in [0,1] \,:\, F_\mu^{[-1]}(a) \in (-\infty, x]\} \\ &= \{a \in [0,1] \,:\, F_\mu^{[-1]}(a) \leq x\} \\ &= \{a \in [0,1] \,:\, a \leq F_\mu(x)\}.\end{aligned}$$

The last equality follows by equivalence of $F_\mu^{[-1]}(a) \le x$ and $a \le F_\mu(x)$, and this implies $(F_\mu^{[-1]})_\sharp \lambda|_{[0,1]} = \mu$. In all $(F_\mu^{[-1]})_\sharp (F_\eta)_\sharp \eta = \mu$, i.e. $T_\sharp \eta = \mu$. The map T is nondecreasing as a composition of two nondecreasing functions. □

Remark 5.7.7. One can show that the nondecreasing transport in Lemma 5.7.6 satisfying $T_\sharp \eta = \mu$ is η-a.e. unique. If we drop the assumption of T being nondecreasing, then T satisfying $T_\sharp \eta = \mu$ is in general not η-a.e. unique (exercise: come up with a counterexample).

The above lemma shows the existence of a transport map pushing forward η to μ for two measures on $\mathbb{R}$. Next, we generalize this construction to the case of two measures on $\mathbb{R}^n$. This then proves the existence of a transport map T such that $T_\sharp \eta = \mu$, and thus justifies the approach (5.7.5). Again, we emphasize that in general there exist many different T satisfying $T_\sharp \eta = \mu$, and we only construct one of them.

To do so, we first need to introduce some notation. For simplicity, we assume in the following that $\eta \ll \lambda_n$ and $\mu \ll \lambda_n$ with *continuous and positive* probability densities

$$f := \frac{\mathrm{d}\eta}{\mathrm{d}\lambda_n} \in C^0(\mathbb{R}^n; (0,\infty)), \qquad g := \frac{\mathrm{d}\mu}{\mathrm{d}\lambda_n} \in C^0(\mathbb{R}^n; (0,\infty)).$$

The assumption that both densities are positive and continuous is not necessary, but allows to avoid some technicalities. For a vector $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^n$, throughout we will use the notation

$$\bar{x}_k := (x_1, \dots, x_k)^\top \in \mathbb{R}^k,$$

in particular $x = \bar{x}_n$.

Let $f^n = f$, $g^n = g$ and for every $k \in \{1, \dots, n-1\}$

$$f^k(\bar{x}_k) := \int_{\mathbb{R}^{n-k}} f(\bar{x}_n)\, \mathrm{d}x_{k+1} \dots \mathrm{d}x_n \tag{5.7.8a}$$

$$g^k(\bar{x}_k) := \int_{\mathbb{R}^{n-k}} g(\bar{x}_n)\, \mathrm{d}x_{k+1} \dots \mathrm{d}x_n, \tag{5.7.8b}$$

and with the convention $f^0 \equiv 1$

$$f^k_{\bar{x}_{k-1}}(x_k) := \frac{f^k(\bar{x}_k)}{f^{k-1}(\bar{x}_{k-1})} \tag{5.7.8c}$$

$$g^k_{\bar{x}_{k-1}}(x_k) := \frac{g^k(\bar{x}_k)}{g^{k-1}(\bar{x}_{k-1})}. \tag{5.7.8d}$$

Note that $f^k : \mathbb{R}^k \to (0,\infty)$ is simply the marginal density of η in the first k variables $\bar{x}_k$, and $f^k_{\bar{x}_{k-1}} : \mathbb{R} \to (0,\infty)$ is the density of η in x_k conditioned on $\bar{x}_{k-1}$. Due to $\int_{\mathbb{R}^k} f^k = \int_{\mathbb{R}^n} f = 1$, it holds $0 < f^{k-1}(\bar{x}_{k-1}) < \infty$ for every $\bar{x}_{k-1}$ so that indeed $f^k_{\bar{x}_{k-1}} : \mathbb{R} \to \mathbb{R}$ is a probability density, and the same holds true for $g^k_{\bar{x}_{k-1}} : \mathbb{R} \to \mathbb{R}$. Here we used that f and g are continuous and positive. We denote in the following by η^k, μ^k the corresponding measures on $\mathbb{R}^k$, and by $\eta^k_{\bar{x}_{k-1}}$, $\mu^k_{\bar{x}_{k-1}}$ the corresponding measures on $\mathbb{R}$, that is

$$f^k(\bar{x}_k) = \frac{\mathrm{d}\eta^k}{\mathrm{d}\lambda_k}(\bar{x}_k), \qquad g^k(\bar{x}_k) = \frac{\mathrm{d}\mu^k}{\mathrm{d}\lambda_k}(\bar{x}_k). \tag{5.7.9a}$$

and

$$f^k_{\bar{x}_{k-1}}(x_k) = \frac{\mathrm{d}\eta^k_{\bar{x}_{k-1}}}{\mathrm{d}\lambda}(x_k), \qquad g^k_{\bar{x}_{k-1}}(x_k) = \frac{\mathrm{d}\mu^k_{\bar{x}_{k-1}}}{\mathrm{d}\lambda}(x_k). \tag{5.7.9b}$$

Since $f_n = f$ and $g_n = g$

$$\eta^n = \eta \qquad \text{and} \qquad \mu^n = \mu.$$

We next construct $T = (T_1, \dots, T_n)$. Let $T_1 : \mathbb{R} \to \mathbb{R}$ be such that

$$(T_1)_\sharp \eta^1 = \mu^1. \tag{5.7.10}$$

Thus T_1 pushes forward the marginal of η in the first variable to the marginal of μ in the first variable. Both of these marginals are probability measures on $\mathbb{R}$, and T_1 exists by Lemma 5.7.6. We set $T^1 : \mathbb{R} \to \mathbb{R}$ via $T^1(x_1) := T_1(x_1)$.

Inductively, for each $k = 2, \dots, d$ and for each $(\bar{x}_{k-1}) \in \mathbb{R}^{k-1}$ we let $T_k(\bar{x}_{k-1}, \cdot) : \mathbb{R} \to \mathbb{R}$ be the transport satisfying

$$(T_k(\bar{x}_{k-1}, \cdot))_\sharp \eta^k_{\bar{x}_{k-1}} = \mu^k_{T^{k-1}(\bar{x}_{k-1})} \tag{5.7.11}$$

and set

$$T^k := \begin{cases} \mathbb{R}^k \to \mathbb{R}^k \\ \bar{x}_k \mapsto (T_1(x_1), \dots, T_k(\bar{x}_{k-1}))^\top. \end{cases}$$

Note that $T_k(\bar{x}_{k-1}, \cdot) : \mathbb{R} \to \mathbb{R}$ pushes forward the marginal of η in the kth variable conditioned on $\bar{x}_{k-1}$, to the marginal of μ in the kth variable conditioned on $T^{k-1}(\bar{x}_{k-1})$. Existence of $T_k(\bar{x}_{k-1}, \cdot) : \mathbb{R} \to \mathbb{R}$ follows again by Lemma 5.7.6.

In all this yields a map $T := T^n = (T_1, \dots, T_n)^\top : \mathbb{R}^n \to \mathbb{R}^n$ which is triangular in the sense that the kth component T_k depends only on $\bar{x}_k = (x_1, \dots, x_k)^\top$ i.e.

$$T(x_1, \dots, x_n) = \begin{pmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_n(x_1, \dots, x_n) \end{pmatrix}.$$

Theorem 5.7.8 (Knothe-Rosenblatt transport)**.** *Under the above conditions it holds $T_\sharp \eta = \mu$.*

Proof. We will show inductively that

$$T^k_\sharp \eta^k = \mu^k. \tag{5.7.12}$$

For $k = n$ this proves $T_\sharp \eta = \mu$.

For $k = 1$, (5.7.12) holds by definition of $T_1 = T^1$ in (5.7.10). To show the induction step, by Thm. 3.1.12 it suffices to show that for all $A = \times_{j=1}^k A_j$ with $A_j \in \mathcal{B}(\mathbb{R})$ holds

$$T^k_\sharp \eta^k(A) = \mu^k(A) \qquad \Leftrightarrow \qquad \eta^k(\{\bar{x}_k \in \mathbb{R}^k \,:\, T^k(\bar{x}_k) \in A\}) = \mu^k(A)$$

which is equivalent to

$$\int_{\mathbb{R}^k} \mathbb{1}_A(T^k(\bar{x}_k)) \,\mathrm{d}\eta^k(\bar{x}_k) = \int_{\mathbb{R}^k} \mathbb{1}_A(\bar{y}_k) \,\mathrm{d}\mu^k(\bar{y}_k). \tag{5.7.13}$$

For the rest of the proof we show (5.7.13) under the induction hypothesis that (5.7.13) holds for $k-1 \geq 1$, which by density of indicator functions in $L^1(\mathbb{R}^{k-1}, \mu^{k-1}; \mathbb{R})$ is equivalent to

$$\int_{\mathbb{R}^{k-1}} \psi(T^{k-1}(\bar{x}_{k-1}))\, \mathrm{d}\eta^{k-1}(\bar{x}_{k-1}) = \int_{\mathbb{R}^{k-1}} \psi(\bar{y}_{k-1})\, \mathrm{d}\mu^{k-1}(\bar{y}_{k-1}) \tag{5.7.14}$$

for all $\psi \in L^1(\mathbb{R}^{k-1}, \mu^{k-1}; \mathbb{R})$.

With $\psi(\bar{y}_{k-1}) := \mathbb{1}_{\times_{j=1}^{k-1} A_j}(\bar{y}_{k-1})$ and $\phi(y_k) := \mathbb{1}_{A_k}(y_k)$ (5.7.13) reads

$$\int_{\mathbb{R}^k} \psi(T^{k-1}(\bar{x}_{k-1}))\phi(T_k(\bar{x}_k))\, \mathrm{d}\eta^k(\bar{x}_k) = \int_{\mathbb{R}^k} \psi(\bar{y}_{k-1})\phi(y_k)\, \mathrm{d}\mu^k(\bar{y}_k).$$

Using the definition of the densities in (5.7.8)

$$\begin{aligned}
\int_{\mathbb{R}^k} \psi(T^{k-1}(\bar{x}_{k-1}))\phi(T_k(\bar{x}_k))\, \mathrm{d}\eta^k(\bar{x}_k) &= \int_{\mathbb{R}^k} \psi(T^{k-1}(\bar{x}_{k-1}))\phi(T_k(\bar{x}_k)) f^k(\bar{x}_k)\, \mathrm{d}\bar{x}_k \\
&= \int_{\mathbb{R}^{k-1}} \psi(T^{k-1}(\bar{x}_{k-1})) f^{k-1}(\bar{x}_{k-1}) \int_{\mathbb{R}} \phi(T_k(\bar{x}_k)) \frac{f^k(\bar{x}_k)}{f^{k-1}(\bar{x}_{k-1})}\, \mathrm{d}x_k\, \mathrm{d}\bar{x}_{k-1} \\
&= \int_{\mathbb{R}^{k-1}} \psi(T^{k-1}(\bar{x}_{k-1})) f^{k-1}(\bar{x}_{k-1}) \left(\int_{\mathbb{R}} \phi(T_k(\bar{x}_k))\, \mathrm{d}\eta^k_{\bar{x}_{k-1}}(x_k) \right) \mathrm{d}\bar{x}_{k-1}.
\end{aligned}$$

Using the definition of T_k in (5.7.11) we have $(T_k(\bar{x}_{k-1}, \cdot)_\sharp \eta^k_{\bar{x}_{k-1}} = \mu^k_{\bar{T}^{k-1}(\bar{x}_{k-1})}$ and thus

$$\begin{aligned}
\int_{\mathbb{R}} \phi(T_k(\bar{x}_k))\, \mathrm{d}\eta^k_{\bar{x}_{k-1}}(x_k) &= \int_{\mathbb{R}} \mathbb{1}_{A_k}(T_k(\bar{x}_k))\, \mathrm{d}\eta^k_{\bar{x}_{k-1}}(x_k) \\
&= \mu^k_{\bar{T}^{k-1}(\bar{x}_{k-1})}(A_k) \\
&= \int_{\mathbb{R}} \phi(y_k) g^k_{\bar{T}^{k-1}(\bar{x}_{k-1})}(y_k)\, \mathrm{d}y_k \\
&=: G(\bar{T}^{k-1}(\bar{x}_{k-1})).
\end{aligned}$$

By the induction hypthesis (5.7.14) and the definition of G

$$\begin{aligned}
\int_{\mathbb{R}^k} \psi(T^{k-1}(\bar{x}_{k-1}))\phi(T_k(\bar{x}_k))\, \mathrm{d}\eta^k(\bar{x}_k) &= \int_{\mathbb{R}^{k-1}} \psi(T^{k-1}(\bar{x}_{k-1})) G(\bar{T}^{k-1}(\bar{x}_{k-1}))\, \mathrm{d}\eta^{k-1}(\bar{x}_{k-1}) \\
&= \int_{\mathbb{R}^{k-1}} \psi(\bar{y}_{k-1}) G(\bar{y}_{k-1})\, \mathrm{d}\mu^{k-1}(\bar{y}_{k-1}) \\
&= \int_{\mathbb{R}^{k-1}} \psi(\bar{y}_{k-1}) \int_{\mathbb{R}} \phi(y_k) g^k_{\bar{y}_{k-1}}(y_k)\, \mathrm{d}y_k\, g^{k-1}(\bar{y}_{k-1})\, \mathrm{d}\bar{y}_{k-1} \\
&= \int_{\mathbb{R}^k} \psi(\bar{y}_{k-1})\phi(y_k)\, \mathrm{d}\mu^k,
\end{aligned}$$

where for the last equality we used $\frac{\mathrm{d}\mu^k}{\mathrm{d}\lambda_k} = g^k$ by (5.7.9), and $g^k_{\bar{y}_{k-1}}(y_k) g^{k-1}(\bar{y}_{k-1}) = g^k(\bar{y}_k)$ by (5.7.8). This shows (5.7.13) and concludes the proof. □

Conditional sampling using triangular transports

In the previous section we interpreted the target μ as the posterior density $\mu_{X|y}$ so that $T_\sharp \eta = \mu_{X|y}$, with the reference η being a measure on $\mathbb{R}^n$ if $X \in \mathbb{R}^n$. For Bayesian inference it is also interesting to consider the case $\mu = \mu_{Y,X}$, i.e. the target is the joint measure of the data and the parameter, and $T_\sharp \eta = \mu_{Y,X}$ for some reference measure η on $\mathbb{R}^{m+n}$ with $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. The reason is, that, as we saw in the construction of the Knothe-Rosenblatt map, its components push forward conditional densities to conditional densities. Since the posterior is a conditional density, this yields a method to sample from the posterior. We emphasize that this is a specific feature of the triangular Knothe-Rosenblatt map (and does not hold for other types of transport maps).

To illustrate the idea, we consider the simplest case where $m = n = 1$, i.e. the parameter $X \in \mathbb{R}$ and the data $Y \in \mathbb{R}$ are one-dimensional. Suppose that $T : \mathbb{R}^2 \to \mathbb{R}^2$ pushes forward a reference $\eta = \eta_1 \otimes \eta_2$ (e.g. $\eta_j \sim \mathcal{N}(0,1)$) to the joint $\mu_{Y,X}$. Then $T = (T_1, T_2)$ as in the previous subsection satisfies (cp. (5.7.10), (5.7.11))

$$(T_1)_\sharp \eta_1 = \mu_Y, \qquad (T_2(y,\cdot))_\sharp \eta_2 = \mu_{X|T_1(y)}.$$

Thus for a RV $S \in \mathbb{R}$

$$S \sim \eta_2 \qquad \Rightarrow \qquad T_2(T_1^{-1}(y), S) \sim \mu_{X|y}.$$

In other words, if we have T as in (5.7.10)-(5.7.11) such that $T_\sharp(\eta_1 \otimes \eta_2) = \mu_{Y,X}$, we can use it to construct iid samples from the posterior $\mu_{X|y}$ by sending iid samples $S_j \sim \eta_2$ through the map $x \mapsto T_2(T_1^{-1}(y), x)$ (keep in mind that η_2 can be chosen at will, and we choose it such that we can easily sample from it, e.g. $\eta_2 \sim \mathcal{N}(0,1)$). A similar construction also works in the more general case where $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$.

5.8 Sequential Monte Carlo methods & Bayesian filtering

In this section, we are going to present an outlook to data assimilation problems. This part is mainly based on the references listed below.

- P. Del Moral, A. Doucet, A. Jasra, *Sequential Monte Carlo samplers*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68: 411-436, 2006.
- G. Evensen, *The Ensemble Kalman filter: theoretical formulation and practical implementation*, Ocean Dynamics, 53(4):343-367, 2003.
- K. Law, A. M. Stuart and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 2016.
- S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, 2015.
- T. Sullivan, *Introduction to Uncertainty Quantification*, Springer, 2015.

In the data assimilation problem we deal with the combination of two information sources:

- **Dynamical system:** We consider a time-dependent physical system described through our mathematical model. In particular, let $Z = (Z_j)_{j\in\mathbb{N}}$ be a Markov chain describing the dynamical system through
$$Z_{j+1} = H_j(Z_j) + \xi_j, \quad j \in \mathbb{N}, \tag{5.8.1}$$
with $Z_0 \sim \pi_0$ for some probability distribution π_0 on $\mathbb{R}^n$. The dynamics are driven by the possibly nonlinear mappings $H_j : \mathbb{R}^n \to \mathbb{R}^n$ and perturbed by additive Gaussian noise. The noise is modelled as i.i.d. sequence $\xi = (\xi_i)_{i\in\mathbb{N}}$ of random variables with $\xi_1 \sim \mathcal{N}(0, \Sigma)$ for some symmetric and positive definite $\Sigma \in \mathbb{R}^{n\times n}$. Further, we assume that Z_0 and ξ_0 are stochastically independent. We refer to the equation (5.8.1) as the **(stochastic) dynamical system** and denote its current state Z_j as **signal**.

- **Observations:** We assume to have access to a time series of **observations** of the underlying stochastic dynamical system. The time series of observations $Y = (Y_i)_{i\in\mathbb{N}}$ are described through the observation model
$$Y_{j+1} = h_{j+1}(Z_{j+1}) + \eta_{j+1}, \quad j \in \mathbb{N}, \tag{5.8.2}$$
where $h_j : \mathbb{R}^n \to \mathbb{R}^K$ are mapping the signal to the observation space $\mathbb{R}^K$. The measurement is assumed to be perturbed by additive Gaussian noise given as i.i.d. sequence $\eta = (\eta_i)_{i\in\mathbb{N}}$ of random variables with $\eta_1 \sim \mathcal{N}(0, \Gamma)$ for some symmetric and positive definite $\Gamma \in \mathbb{R}^{K\times K}$.

We aim to use both the dynamical system as well as the incoming observations to construct sequential estimates of the current signal or even to predict the future signal. We call the task of determining information about the signal Z, given the observation Y, **data assimilation problem**. The common tools in data assimilation are based on Bayesian models.

Example 5.8.1. We consider a $\mathbb{R}^n$-valued stochastic differential equation described by
$$\mathrm{d}Z_t = b(Z_t)\,\mathrm{d}t + \sigma(Z_t)\,\mathrm{d}W_t, \quad Z_0 \sim \pi_0, \quad t \in [0, T],$$
where $b : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift coefficient and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ denotes the diffusion coefficient. The diffusion is driven by the $\mathbb{R}^n$-valued Brownian motion $W = (W_t)_{t\geq 0}$. For simplicity, we assume that b and σ are both global Lipschitz continuous, such that unique existence of strong solutions is verified. A common numerical approximation method is the Euler–Maruyama method, which approximates the solution of the SDE recusively by
$$Z_{j+1} = Z_k + (t_{j+1} - t_j) \cdot b(Z_j) + \sigma(Z_j) \cdot (W_{t_{j+1}} - W_{t_j}).$$
Here, we have used a partition of the time interval $[0, T]$ given by
$$0 = t_0 < t_1 < \cdots < t_N = T.$$
We note that the increments of the Brownian motion are multivariate Gaussian distributed
$$W_{t_{j+1}} - W_{t_j} \sim \mathcal{N}(0, (t_{j+1} - t_j)Id).$$
Suppose that $\Delta = t_{j+1} - t_j$ for all j and $\sigma(z) = R \in \mathbb{R}^{n\times n}$ for all $z \in \mathbb{R}^n$, then our stochastic dynamical system is described by
$$Z_{j+1} = Z_j + \Delta \cdot b(Z_j) + \xi_j, \quad \xi_j \sim \mathcal{N}(0, \Delta(RR^\top)).$$

[Draft of October 6, 2021. Not for dissemination.]

The observation model might be described by an observation matrix

$$h(\cdot) = \mathcal{O}\,\cdot, \quad \mathcal{O} \in \mathbb{R}^{K\times n}.$$

For example, we can choose an index subset $\mathcal{I} = \{i_1, \dots, i_K\} \subset \{1, \dots, n\}$ and define

$$\mathcal{O}_{l,j} = \begin{cases} 1, & j = i_l \\ 0, & j \neq i_l \end{cases}, \quad l = 1, \dots, K,\ j = 1, \dots, n.$$

This special choice of $\mathcal{O}$ observes the components $\mathcal{I}$ of the vector Z.

5.8.1 Prediction, filtering and smoothing

We assume to have access to the prior information about the unknown signal given by the probability density function π_0. With the application of the Chapman–Kolmogorov equation for the Markov chain constructed in (5.8.1), we can compute the marginal distribution π_{Z_j} of Z_j sequentially by

$$\pi_{Z_{j+1}}(\mathrm{d}z') = \mathbb{P}(Z_{j+1} \in \mathrm{d}z') = \int_{R^n} \pi_j(\mathrm{d}z' \mid z)\pi_{Z_j}(\mathrm{d}z).$$

Since we assume that the distribution of Z_0 has Lebesgue density π_0 and the underlying noise is Gaussian, the transition density function can be derived explicitly by

$$\pi_j(\mathrm{d}z' \mid z) = \frac{1}{\det(2\pi\Sigma)} \exp(-\frac{1}{2}\|z' - H_j(z)\|_\Sigma^2)\,\mathrm{d}z'.$$

Similarly we can derive the marginal pdf of the observation Y_j conditioned on the state $Z_j = z$

$$\pi_{Y_j}(\mathrm{d}y \mid z) = \frac{1}{\det(2\pi\Gamma)} \exp(-\frac{1}{2}\|y - h_j(z)\|_\Gamma^2)\,\mathrm{d}y.$$

Given a realization $y^{[1:N_{\mathrm{obs}}]} = (y_1, \dots, y_{N_{\mathrm{obs}}})$ of the time series of observations $Y^{[1:N_{\mathrm{obs}}]} = (Y_1, \dots, Y_{N_{\mathrm{obs}}})$, $N_{\mathrm{obs}} \geq 1$, the data assimilation problem is the computation of the conditional distribution of Z_j given $y^{[1:N_{\mathrm{obs}}]}$:

$$\pi_{Z_j|y^{[1:N_{\mathrm{obs}}]}}(\mathrm{d}z) = \mathbb{P}(Z_j \in \mathrm{d}z \mid Y^{[1:N_{\mathrm{obs}}]} = y^{[1:N_{\mathrm{obs}}]}). \tag{5.8.3}$$

Definition 5.8.2. We call the task of computing (5.8.3)

(i) **prediction problem** if $j > N_{\mathrm{obs}}$,

(ii) **filtering problem** if $j = N_{\mathrm{obs}}$,

(iii) and **smoothing problem** if $j < N_{\mathrm{obs}}$.

Depending on the corresponding case, we denote the distribution in (5.8.3) as **prediction, filtering and smoothing distribution**.

Through the connection to Bayesian inverse problems, we will focus on filtering problems. The filtering problem splits into two steps, where for the first step we are updating the filtering distribution using only the stochastic dynamical system (5.8.1), while in the second step we apply Bayes' theorem to incorporate information from the incoming data.

Definition 5.8.3. Given the filtering distribution $\pi_{Z_j|y^{[1:j]}}$, we refer the **prediction step** to the computation of the marginal distribution of the next state through

$$\pi_{Z_{j+1}|y^{[1:j]}}(\mathrm{d}z) = \mathbb{P}(Z_{j+1} \in \mathrm{d}z \mid Y^{[1:j]} = y^{[1:j]}) = \int_{\mathbb{R}^n} \pi_{j+1}(\mathrm{d}z \mid z')\pi_{Z_j|y^{[1:j]}}(\mathrm{d}z').$$

We call the second step **Bayesian assimilation step**, which is the computation of the filtering distribution $\pi_{Z_{j+1}|y^{[1:j+1]}}$ via Bayes' theorem

$$\pi_{Z_{j+1}|y^{[1:j+1]}}(\mathrm{d}z) = \mathbb{P}(Z_{j+1} \in \mathrm{d}z \mid Y^{[1:j+1]} = y^{[1:j+1]}) = \frac{\pi_{Y_{j+1}}(y_{j+1} \mid z)\pi_{Z_{j+1}|y^{[1:j]}}(\mathrm{d}z)}{\int_{\mathbb{R}^n} \pi_{Y_{j+1}}(y_{j+1} \mid z)\pi_{Z_{j+1}|y^{[1:j]}}(\mathrm{d}z)}.$$

Summarizing, given the current filtering distribution $\pi_{Z_j|y^{[1:j]}}$, we construct a prior distribution $\pi_{Z_{j+1}|y^{[1:j]}}$ using our knowledge about the stochastic dynamical system and update w.r.t. the incoming data $Y_{j+1} = y_{j+1}$ via Bayes' theorem.

5.8.2 Linear Kalman filter

Under linear and Gaussian assumptions on the underlying stochastic dynamical system and the corresponding observations, the **Kalman filter** solves the filtering problem exactly. We consider the signal described through

$$Z_{j+1} = FZ_j + \xi_j, \quad j \in \mathbb{N} \tag{5.8.4}$$

and the observations

$$Y_{j+1} = AZ_{j+1} + \eta_j, \quad j \in \mathbb{N}, \tag{5.8.5}$$

where $F \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^K)$. Furthermore, we assume that the initial distribution is Gaussian, i.e. $\pi_0 = \mathcal{N}(m_0, C_0)$. Since the forward maps are assumed to be linear and the noise to be Gaussian, the filtering distribution remains Gaussian

$$\pi_{Z_j|y^{[1:j]}} = \mathcal{N}(m_j, C_j).$$

Given the initial mean $m_0 \in \mathbb{R}^n$ and symmetric, positive definite covariance $C_0 \in \mathbb{R}^{n\times n}$, the Kalman filter computes the mean m_j and covariance C_j of the filtering distribution recursively.

(i) **prediction step:** Given the mean m_j and covariance C_j of iteration j, we first update based on the stochastic dynamical system (5.8.4). Since we have assumed that ξ_j is independent of $Z_j \sim \mathcal{N}(m_j, C_j)$, the prediction step computes

$$\widehat{m}_{j+1} = Fm_j, \quad \widehat{C}_{j+1} = FC_jF^\top + \Sigma.$$

(ii) **Bayesian assimilation step:** We set the prior distribution $\pi_Z = \mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1})$ and update the mean and the covariance according to Bayes' Theorem (compare Theorem 5.3.1.)

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + \widehat{C}_{j+1}A^\top(A\widehat{C}_{j+1}A^\top + \Gamma)^{-1}(y_{j+1} - A\widehat{m}_{j+1}) \\ C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1}A^\top(A\widehat{C}_{j+1}A^\top + \Gamma)^{-1}A\widehat{C}_{j+1} \end{aligned} \tag{5.8.6}$$

Defining the Kalman gain

$$K_j = \widehat{C}_j A^\top (A\widehat{C}_j A^\top + \Gamma)^{-1}$$

we can write the Bayesian update step as

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + K_{j+1}(y_{j+1} - A\widehat{m}_{j+1}) \\ C_{j+1} &= \widehat{C}_{j+1} - K_{j+1}A\widehat{C}_{j+1} \end{aligned}$$

As we describe the filtering distribution through $\mathcal{N}(m_j, C_j)$ we need to ensure that C_j stays positive definite.

Lemma 5.8.4. *Assume that $Z_0 \sim \mathcal{N}(m_0, C_0)$ for some symmetric and positive definite covariance matrix $C_0 \in \mathbb{R}^{n\times n}$. Then the matrix C_j resulting from* (5.8.6) *is symmetric and positive definite.*

Proof. The proof is left as an exercise (Hint: apply Woodbury matrix identity). □

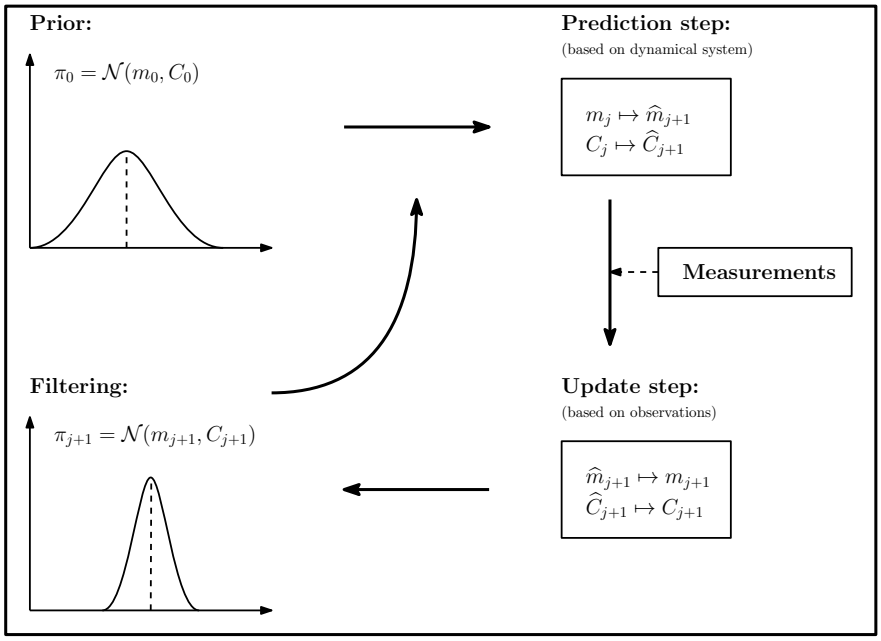


Figure 5.12: Summary of the linear Kalman filter method.

5.8.3 Extended Kalman filter

The extended Kalman filter is a generalization of the linear Kalman filter to nonlinear dynamical systems. In order to apply the introduced Kalman filter, we firstly linearize the nonlinear dynamical system and then apply the Kalman filter to the resulting linear system. This method results in a Gaussian approximation to the filtering distribution. For strongly nonlinear dynamical systems the resulting filtering distribution might be poorly approximated through Gaussian measures. However, the extended Kalman filter can be viewed as best linear unbiased estimate for the linearized dynamical system, which can lead to a good approximation of the original nonlinear system.

We assume that the signal and the observations are described by

$$Z_{j+1} = H(Z_j) + \xi_j, \quad Y_{j+1} = AZ_{j+1} + \eta_j, \quad j \in \mathbb{N},$$

where $H : \mathbb{R}^n \to \mathbb{R}^n$ is a possibly nonlinear mapping and $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^K)$. We note the the following ideas can straightforwardly be extended to a nonlinear observation model. Given the initial distribution $\pi_0 = \mathcal{N}(m_0, C_0)$, the extended Kalman filter approximates the filtering distribution by $\mathcal{N}(m_j, C_j)$ as follows

(i) **linearization step:** Given the mean m_j and covariance C_j, we build the linearized approximation

$$Z_{j+1} = F_j Z_j + b_j + \xi_j, \quad j \in \mathbb{N},$$

with

$$F_j := \mathrm{D}H(m_j) \quad \text{and} \quad b_j = H(m_j) - F_j m_j.$$

(ii) **prediction step:** We predict the mean and the covariance by

$$\widehat{m}_{j+1} = F_j m_j + b_j, \quad \widehat{C}_{j+1} = F_j C_j F_j^\top + \Sigma.$$

(iii) **Bayesian assimilation step:** We again set the prior distribution $\pi_Z = \mathcal{N}(\widehat{m}_{j+1}, \widehat{C}_{j+1})$ and update the mean and the covariance according to Bayes' Theorem

$$\begin{aligned} m_{j+1} &= \widehat{m}_{j+1} + \widehat{C}_{j+1} A^\top (A\widehat{C}_{j+1}A^\top + \Gamma)^{-1}(y_{j+1} - A\widehat{m}_{j+1}) \\ C_{j+1} &= \widehat{C}_{j+1} - \widehat{C}_{j+1} A^\top (A\widehat{C}_{j+1}A^\top + \Gamma)^{-1} A\widehat{C}_{j+1} \end{aligned}$$

5.8.4 Ensemble Kalman filter

An alternative method to overcome the nonlinearity in the dynamical system is the application of the ensemble Kalman filter (EnKF). The EnKF has been originally introduced by G. Evensen (2003) and can be viewed as a Monte Carlo approximation of the Kalman filter. The basic idea is to use a particle system, initialized by a sample of the prior distribution $Z_0 \sim \pi_0$, which will then be updated according to the Kalman filter. Since we do not use any Gaussian assumptions and approximate the filtering distribution with the particle system empirically, we are now able to apply the EnKF in nonlinear dynamical systems (5.8.1). For simplicity, we again consider a linear observation model (5.8.5).

Given the current particle system $(v_j^{(m)})_{m=1,\dots,M}$ of size M, we proceed as follows.

(i) **prediction step:** We apply the dynamical system to predict the system's state by

$$\widehat{v}_{j+1}^{(m)} = H(v_j^{(m)}) + \xi_j^{(m)}, \quad m = 1, \dots, M,$$

where $(\xi_j^{(m)})_{m=1,\dots,M}$ is an i.i.d. sample of $\mathcal{N}(0, \Sigma)$. The empirical mean and the empirical covariance are given by

$$\widehat{m}_{j+1} = \frac{1}{M}\sum_{m=1}^{M} \widehat{v}_{j+1}^{(m)}, \quad \widehat{C}_{j+1} = \frac{1}{M}\sum_{m=1}^{M} (\widehat{v}_{j+1}^{(m)} - \widehat{m}_{j+1})(\widehat{v}_{j+1}^{(m)} - \widehat{m}_{j+1})^\top. \tag{5.8.7}$$

(ii) **analysis step:** We apply to each particle the linear Kalman filter update corresponding to a Gaussian approximation. The particles are updated by

$$\begin{aligned} v_{j+1}^{(m)} &= \widehat{v}_{j+1}^{(m)} + K_{j+1}(\widetilde{y}_{j+1}^{(m)} - A\widehat{v}_{j+1}^{(m)}), \\ \widetilde{y}_{j+1}^{(m)} &= y_j + \eta_{j+1}^{(m)}, \quad \eta_{j+1}^{(m)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma). \\ K_j &= \widehat{C}_j A^\top (A\widehat{C}_j A^\top + \Gamma)^{-1}, \end{aligned}$$

where we denote $\widetilde{y}_{j+1}^{(m)}$ as perturbed observation and K_j is again the introduced Kalman gain.

The filtering distribution is approximated empirically by

$$\pi_{Z_j|y^{1:j}}(\mathrm{y}) \approx \widehat{\pi}_j(\mathrm{y}) = \frac{1}{M}\sum_{m=1}^{M} \delta_{v_j^{(m)}}(\mathrm{y}),$$

As stated above one advantage of the EnKF is the application in nonlinear dynamical systems. Furthermore, through the computation of the empirical covariance we save computational costs compared to updating the covariance in each iteration according to (5.8.6).

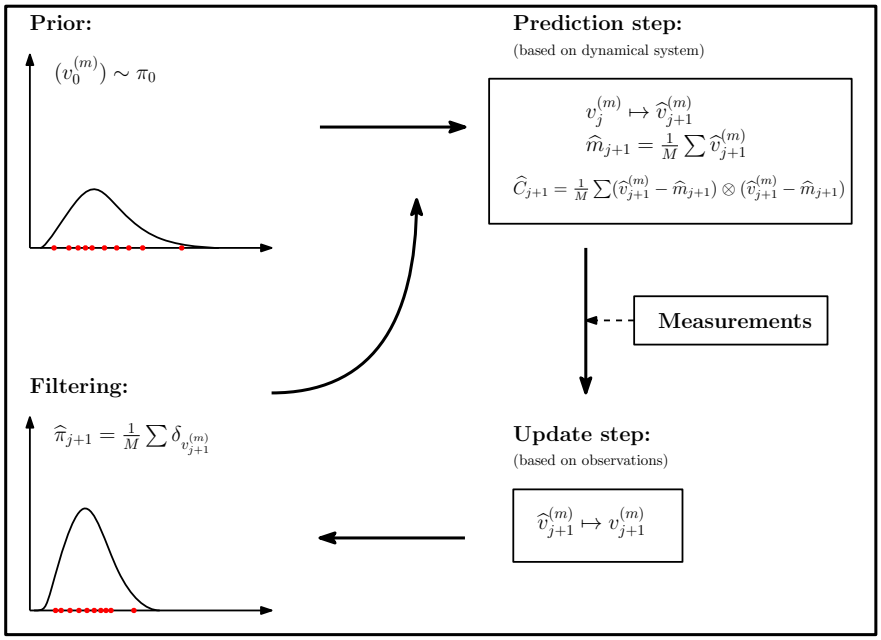


Figure 5.13: Summary of the ensemble Kalman filter method.

5.8.5 Particle filters - Sequential Monte Carlo methods

As alternative to the different presented variants of Kalman filters, we briefly introduce the class of particle filters which can be seen as sequential Monte Carlo method of the filtering distribution without including Gaussian approximations. To derive the scheme we consider again a $\mathbb{R}^n$-valued state driven by a stochastic differential equation of the form

$$\mathrm{d}Z_t = b(Z_t)\,\mathrm{d}t + R\,\mathrm{d}W_t, \quad Z_0 \sim \pi_0, \quad t \in [0, T],$$

[Draft of October 6, 2021. Not for dissemination.]

and its discrete time approximation

$$Z_{j+1} = Z_k + \Delta \cdot b(Z_j) + \xi_j, \tag{5.8.8}$$

where $b : \mathbb{R}^n \to \mathbb{R}^n$ is the drift coefficient,, $\Delta > 0$ is the time increment, $R \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite and $\xi_j \sim \mathcal{N}(0, \Delta \cdot RR^\top)$. Furthermore, we consider a possibly nonlinear observation model

$$Y_{j+1} = h(Z_j) + \eta_j, \quad \eta_j \sim \mathcal{N}(0, \Gamma).$$

We can then describe the marginal distribution of the state Z_j through its Markov transition kernel

$$K_j(z, \mathrm{d}z') = \mathbb{P}(Z_{j+1} \in \mathrm{d}z' \mid Z_j = z).$$

The aim of particle filters is to approximate the filtering distribution empirically with a weighted particle system by combining the prediction step with ideas from importance sampling. The method proceeds as follows. We generate an initial particle system $Z_0^{(m)} \sim \pi_0$, $m = 1, \dots, M$ and define initial weights $w_0^{(m)} = \frac{1}{M}$, $m = 1, \dots, M$ weighting each particle equally. Using the discrete time dynamical system we generate prediction for the next state by

$$\widehat{Z}_1^{(m)} = Z_0^{(m)} + \Delta \cdot b(Z_0^{(m)}) + \xi_1^{(m)}, \quad \xi_1^{(m)} \sim N(0, \Delta \cdot RR^\top)$$

which follows the marginal distribution $K_1(Z_0^{(m)}, \mathrm{d}z)$. With the help of these predictions we approximate the marginals of Z_1 empirically by

$$\pi_{Z_1}(\mathrm{d}z) = \int_{\mathbb{R}^n} \pi_0(\mathrm{d}z') K_1(z', \mathrm{d}z) \approx \sum_{m=1}^{M} w_0^{(m)} \delta_{\widehat{Z}_1^{(m)}}(\mathrm{d}z).$$

Given this empirical approximation we apply an importance sampling step following Bayes' theorem

$$\pi_{Z_1|y^{[1]}}(\mathrm{d}z) = \frac{\pi_{Y_1}(y_1 \mid z)}{\int_{\mathbb{R}^n} \pi_{Y_1}(y_1 \mid z) \pi_{Z_1|y^{[1]}}(\mathrm{d}z)} \pi_{Z_1}(\mathrm{d}z) \approx \sum_{m=1}^{M} w_1^{(m)} \delta_{\widehat{Z}_1^{(m)}}(\mathrm{d}z),$$

where we have updated and normalized the weights

$$w_1^{(m)} = \frac{w_0^{(m)} \pi_{Y_1}(y_1 \mid \widehat{Z}_1^{(m)})}{\sum_{m=1}^{M} w_0^{(m)} \pi_{Y_1}(y_1 \mid \widehat{Z}_1^{(m)})}.$$

Hence, given the current weighted particle system $(w_j^{(m)}, \widehat{Z}_j^{(m)})_{m=1,\dots,M}$ we can divide the update scheme again in a prediction step followed by a Bayesian assimilation step described as follows.

(i) **prediction step:** Given the current state approximations $\widehat{Z}_j^{(m)}$, we first update the particles based on the stochastic dynamical system (5.8.8) by

$$\widehat{Z}_{j+1}^{(m)} = \widehat{Z}_j^{(m)} + \Delta \cdot b(\widehat{Z}_j^{(m)}) + \xi_{j+1}^{(m)}, \quad \xi_{j+1}^{(m)} \sim N(0, \Delta \cdot RR^\top)$$

according to the marginal distribution $K_{j+1}(\widehat{Z}_j^{(m)}, \mathrm{d}z)$ such that the marginal distribution of the state Z_{j+1} can be approximated by

$$\pi_{Z_{j+1}}(\mathrm{d}z) = \int_{\mathbb{R}^n} \pi_{Z_j|y^{[1:j]}}(\mathrm{d}z') K_{j+1}(z', \mathrm{d}z) \approx \sum_{m=1}^{M} w_j^{(m)} \delta_{\widehat{Z}_{j+1}^{(m)}}(\mathrm{d}z).$$

(ii) **Bayesian assimilation step:** Following Bayes' Theorem we approximate the filtering distribution by

$$\pi_{Z_{j+1}|y^{[1:j+1]}}(\mathrm{d}z) = \frac{\pi_{Y_{j+1}}(y_{j+1} \mid z)}{\int_{\mathbb{R}^n} \pi_{Y_{j+1}}(y_{j+1} \mid z)\pi_{Z_{j+1}|y^{[1:j]}}(\mathrm{d}z)}\pi_{Z_{j+1}}(\mathrm{d}z) \approx \sum_{m=1}^{M} w_{j+1}^{(m)}\delta_{\widehat{Z}_{j+1}^{(m)}}(\mathrm{d}z),$$

where we have updated and normalized the weights

$$w_{j+1}^{(m)} = \frac{w_j^{(m)}\pi_{Y_{j+1}}(y_{j+1} \mid \widehat{Z}_{j+1}^{(m)})}{\sum_{m=1}^{M} w_j^{(m)}\pi_{Y_{j+1}}(y_{j+1} \mid \widehat{Z}_{j+1}^{(m)})}.$$

Given the weighted particle system $(w_{j+1}^{(m)}, \widehat{Z}_{j+1}^{(m)})_{m=1,\dots,M}$ in iteration $j+1$, we are able to approximate expectation values for functionals $F : \mathbb{R}^n \to \mathbb{R}$ of the following kind

$$\mathbb{E}[F(Z_{j+1}) \mid Y^{[1:j]} = y^{[1:j]}] \approx \sum_{m=1}^{M} w_j^{(m)} F(\widehat{Z}_{j+1}^{(m)}),$$
$$\mathbb{E}[F(Z_{j+1}) \mid Y^{[1:j]} = y^{[1:j+1]}] \approx \sum_{m=1}^{M} w_{j+1}^{(m)} F(\widehat{Z}_{j+1}^{(m)}).$$

In general, particle filters of this form can be viewed as sequential importance sampling method, where existing consistency results are based on perfect models and M approaching infinity. In practical implementations, the generated weights tend to degenerate for small choices of the number of particles M. To overcome this issue, resampling methods based on the effective sample size have been considered in the literature.

Appendices

Appendix A

Basic Concepts in Functional Analysis

In this appendix we put together the main concepts from functional analysis that will be needed in this lecture. We recommend the following supplementary references:

- H. Alt, *Funktionalanalysis*, 6. Auflage, Springer, Berlin, 2012.
- W. Rudin, *Functional Analysis*, 2nd ed., Mc-Graw-Hill, New York, 1991.
- D. Werner, *Funktionalanalysis*, 6. Auflage, Springer, Berlin, 2007.

A.1 Normed spaces and bounded linear operators

A.1.1 Normed spaces

In the following X will denote a vector space. We restrict ourselves to vector spaces over the field $\mathbb{K} = \mathbb{R}$. A map $\|\cdot\| : X \to [0.\infty)$ is a **norm** on X, if

(i) $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in K$, $x \in X$,

(ii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

(iii) $\|x\| = 0$ iff $x = 0$.

The norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are **equivalent**, if there are constants $c_1, c_2 > 0$, such that

$$c_1\|x\|_\beta \leq \|x\|_\alpha \leq c_2\|x\|_\beta \qquad \text{for all } x \in X\,.$$

If $\dim(X) < \infty$ all norms on X are equivalent. The constants c_1, c_2 depend on the dimension of X.

Example A.1.1. The following maps are norms on

(i) $X = \mathbb{R}^n$, $n \in \mathbb{N}$:

$$\|x\|_p = \left(\sum_{j=1}^{n} |x|^p\right)^{1/p}, \quad 1 \leq p < \infty, \quad \text{and} \quad \|x\|_\infty = \max_{j=1,\ldots,n} |x_j|\,.$$

(ii) $X = l^p$ $(:= \{(t_n) : t_n \in \mathbb{R}, \sum_{n=1}^{\infty} |t_n|^p < \infty\})$:

$$\|x\|_p = \left(\sum_{j=1}^{\infty} |x|^p\right)^{1/p}, \quad 1 \leq p < \infty, \quad \text{and} \quad \|x\|_\infty = \max_{j=1,\ldots,\infty} |x_j|\,.$$

[Draft of October 6, 2021. Not for dissemination.]

(iii) $X = L^p(\Omega)$ $(:= \{f : \Omega \to \mathbb{K} : f \text{ messbar}, \int_\Omega |f|^p \mathrm{d}\lambda < \infty\})$ where $\Omega \subset \mathbb{R}^n$:

$$\|f\|_p = \left(\int_\Omega |f|^p \mathrm{d}\lambda\right)^{1/p}, \quad 1 \le p < \infty, \quad \text{and} \quad \|f\|_\infty = \operatorname*{ess\,sup}_{x\in\Omega} |f(x)| .$$

If $\|\cdot\|$ is a norm on X, we call the pair $(X, \|\cdot\|)$ a **normed space**.

A normed space $(X, |\cdot\|_X)$ with $X \subset Y$ is said to be **continuously embedded** in $(Y, \|\cdot\|_Y)$, denoted by $X \hookrightarrow Y$, if there is a constant $C > 0$ such that

$$\|x\|_Y \le C\|x\|_X \qquad \text{for all } x \in X .$$

A sequence $(x_n) \subset X$ **converges strongly** in X to $x \in X$, denoted $x_n \to x$ as $n \to \infty$, if

$$\lim_{n\to\infty} \|x_n - x\|_X = 0 .$$

A subset $U \subset X$ is called

- **closed**, if the limit of any convergent sequence $(x_n) \subset U$ lies in U, i.e., $x \in U$;
- **compact**, if any sequence $(x_n) \subset U$ has a convergent subsequence $(x_{n_k})_{k\ge1}$ with limit $x \in U$;
- **dense in** X, if for any $x \in X$ there exists a sequence $(x_n) \subset U$ with $x_n \to x$.

The union of U with the set of all limits of convergent sequences in U is called the **closure** $\overline{U}$ of U. It follows that U is dense in $\overline{U}$.

A normed space X is said to be **complete**, if every Cauchy sequence in X converges. Such a space X is also called a **Banach space**. If X is not complete, we denote by $\overline{X}$ its **completion** (w.r.t. the norm $\|\cdot\|_X$).

For $x \in X$ and $r > 0$ we define

- the **open ball** $U_r = z \in X : \|x - z\|_X < r$ and
- the **closed ball** $B_r = z \in X : \|x - z\|_X \le r$.

The closed ball at 0 with radius 1 is called the **unit ball** B_X in X. Furthermore, the set $U \subset X$ is called

- **open**, if for all $x \in U$ there exists a $r > 0$ such that $U_r(x) \subset U$;
- **bounded**, if there exists an $r > 0$ such that U is contained in the closed ball $B_r(0)$;
- **convex**, if for all $x, y \in U$ and $\lambda \in (0, 1)$ we have $\lambda x + (1 - \lambda)y \in U$.

The complement of an open set in a normed space is closed and vice versa.

As a consequence of the norm axioms, all open and closed balls are convex.

A.1.2 Bounded operators

Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be normed spaces, $U \subset X$ und $F : U \to Y$ a map. We denote by

- $\mathcal{D}(F) := U$ the **domain** of F,

- $\mathcal{N}(F) := \{x \in U : F(x) = 0\}$ the **kernel** of F,
- $\mathcal{R}(F) := \{F(x) \in Y : x \in U\}$ the **range** of F.

Furthermore, we say that F is

- **continuous** in $x \in U$, if for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\|F(x) - F(y)\|_Y < \epsilon \qquad \text{for all } z \in U \text{ with } \|x - z\|_X < \delta;$$

- **Lipschitz continuous**, if there exists a $L > 0$ such that

$$\|F(x_1) - F(x_2)\|_Y \leq L\|x_1 - x_2\|_X \qquad \text{for all } x_1, x_2 \in U\,.$$

A map $F : X \to Y$ is continuous iff $x_n \to x$ implies $F(x_n) \to F(x)$, and **closed**, if for any sequence $x_n \to x$ with $F(x_n) \to y$ it follows that $F(x) = y$.

If $F : X \to Y$ is linear, i.e., $F(\lambda_1 x + \lambda_2 x_2) = \lambda_1 F(x_1) + \lambda_2 F(x_2)$ for all $x_1, x_2 \in X,\ \lambda_1, \lambda_2 \in \mathbb{R}$, then the continuity of F is equivalent with the condition that there exists a constant $C > 0$ such that

$$\|Fx\|_Y \leq C\|x\|_X \qquad \text{for all } x \in X\,. \tag{A.1}$$

For that reason continuous, linear maps are also called **bounded** and we speak about a bounded, linear operator. (To stress this in the following, we will denote such maps with A.)

If Y is complete, then the space of all bounded, linear operators from X to Y, denoted by $\mathcal{L}(X, Y)$, is a Banach space with the **operator norm**

$$\|A\|_{\mathcal{L}(X,Y)} = \sup_{x \in X \setminus \{0\}} \frac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X \leq 1} \|Ax\|_Y$$

It is equal to the smallest constant C in the definition of continuity in (A.1). As for linear operators in $\mathbb{R}^n$ we say that A is

- **injective**, if $\mathcal{N}(A) = \{0\}$,
- **surjective**, if $\mathcal{R}(A) = Y$,
- **bijective**, if A is injective and surjective.

If $A \in \mathcal{L}(X, Y)$ is bijective, the inverse $A^{-1} : Y \to X$ is bounded iff there exists a $c > 0$ such that

$$c\|x\|_X \leq \|Ax\|_Y \qquad \text{für alle } x \in X\,,$$

and $\|A^{-1}\|_{\mathcal{L}(Y,X)} = c^{-1}$ for the largest possible c. When this is the case follows from the following fundamental theorem of functional analysis.

Theorem A.1.2 (Closed Graph Theorem)**.** *Let X, Y be Banach spaces. A map $F : X \to Y$ is continuous iff F is closed.*

Corollary A.1.3. *Let X, Y be Banach spaces and $A \in \mathcal{L}(X, Y)$ bijective. Then $A^{-1} : Y \to X$ is continuous.*

We consider now sequences of linear operators and distinguish two different notions of convergence:

A sequence $(A_n) \subset \mathcal{L}(X,Y)$ converges to $A \in \mathcal{L}(X,Y)$

- **pointwise**, if $A_n x \to Ax$ for all $x \in X$ (strong convergence in Y);
- **uniformly**, if $A_n \to A$ (strong convergence in $\mathcal{L}(X,Y)$).

Uniform convergence implies pointwise convergence.

Theorem A.1.4 (Banach-Steinhaus)**.** *Let X be a Banach space and Y a normed vector space, and suppose that $(A_i)_{i\in I} \subset \mathcal{L}(X,Y)$ is a family of pointwise bounded, linear operators, i.e., for all $x \in X$ there exists $M_x > 0$ such that $\sup_{i\in I} \|A_i x\| \le M_x$. Then*

$$\sup_{i\in I} \|A_i\|_{\mathcal{L}(X,Y)} < \infty\,.$$

Corollary A.1.5. *Let X, Y be Banach spaces and $(A_n) \subset \mathcal{L}(X,Y)$. Then the following three statements are equivalent:*

(i) *(A_n) converges uniformly on compact subsets of X.*

(ii) *(A_n) converges pointwise on X,*

(iii) *(A_n) converges pointwise on a dense subset $U \subset X$ and $\sup_{n\in\mathbb{N}} \|A_n\|_{\mathcal{L}(X,Y)} < \infty$.*

Also, if A_n converges pointwise to $A : X \to Y$ then A is bounded.

A.2 Hilbert spaces, compact operators and the Spectral Theorem

Inverse problems can be analysed in Banach spaces, but the theory can be presented more comprehensively in Hilbert spaces. It also provides a clearer link to underdetermined or ill-conditioned linear equation systems in $\mathbb{R}^n$, which have been covered, e.g., in introductory numerical analysis courses.

A.2.1 Scalar product and weak convergence

Hilbert spaces distinguish themselves from Banach spaces by having one additional structure: a map $\langle\cdot,\cdot\rangle : X \times X \to \mathbb{R}$, called a **scalar product**, with the properties

(i) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x,z\rangle + \beta\langle y,z\rangle$ for all $x,y,z \in X$, $\alpha,\beta \in \mathbb{R}$,

(i) $\langle x,y\rangle = \langle y,x\rangle$ for all $x,y \in X$,

(i) $\langle x,x\rangle \ge 0$ for all $x \in X$, with $\langle x,x\rangle = 0$ iff $x = 0$.

The skalar product induces a norm

$$\|x\|_X := \sqrt{\langle x,x\rangle_X}$$

that satisfies the Cauchy-Schwarz inequality

$$|\langle x,y\rangle_X| \le \|x\|_X \|y\|_X\,.$$

A Banach space with a skalar product $(X, \langle\cdot,\cdot\rangle_X)$ is called a **Hilbert space**.

Example A.2.1. We can define the following scalar products on

(i) $X = \mathbb{R}^n$, $n \in \mathbb{N}$:
$$\langle x, y\rangle = \sum_{j=1}^{n} x_i y_i\,, \qquad \text{for all } x, y \in X;$$

(ii) $X = l^2$:
$$\langle x, y\rangle = \sum_{j=1}^{\infty} x_i y_i \qquad \text{for all } x, y \in X;$$

(iii) $X = L^2(\Omega)$, $\Omega \subset \mathbb{R}^n$:
$$\langle f, g\rangle = \int_\Omega fg \,\mathrm{d}\lambda \qquad \text{for all } f, g \in X\,.$$

In all cases the scalar product also induces a canonical norm on X.

The scalar product also allows to define a further notion of convergence: a sequence $(x_n) \subset X$ **converges weakly** (in X) to $x \in X$ – we write $x_n \rightharpoonup x$ – if

$$\langle x_n, z\rangle_X \to \langle x, z\rangle_X \qquad \text{for all } z \in X\,.$$

It generalises coordinatewise convergence in $\mathbb{R}^n$. In finite dimensional spaces strong and weak convergence are equivalent. In infinite dimensional spaces, strong convergence implies weak convergence, but the converse is not true. However, if a sequence (x_n) converges weakly to $x \in X$ and in addition $\|x_n\|_X \to \|x\|_X$, then (x_n) converges also strongly to x.

Theorem A.2.2 (Bolzano-Weierstrass)**.** *Every bounded sequence in a Hilbert space has a weakly convergent subsequence.*

Conversely, every weakly convergent sequence is bounded.

Let us now consider bounded, linear operators $A \in \mathcal{L}(X, Y)$ on Hilbert spaces X, Y. Of particular interest is the special case $Y = \mathbb{R}$, i.e., the space $\mathcal{L}(X, \mathbb{R})$ of **bounded, linear functionals** on X.

Theorem A.2.3 (Riesz-Fischer)**.** *Let X be a Hilbert space. For every functional $\lambda \in \mathcal{L}(X, \mathbb{R})$ there exists a unique $z_\lambda \in X$ with $\|z_\lambda\|_X = \|\lambda\|_{\mathcal{L}(X,\mathbb{R})}$ such that*

$$\lambda(x) = \langle z_\lambda, x\rangle_X \qquad \textit{für alle } x \in X\,.$$

This theorem allows to define an **adjoint operator** $A^* \in \mathcal{L}(Y, X)$ for every linear operator $A \in \mathcal{L}(X, Y)$ such that

$$\langle A^* y, x\rangle_X = \langle y, Ax\rangle_Y \qquad \text{for all } x \in X, y \in Y\,,$$

and

(i) $(A^*)^* = A$,

(ii) $\|A^*\|_{\mathcal{L}(Y,X)} = \|A\|_{\mathcal{L}(X,Y)}$,

(iii) $\|A^*A\|_{\mathcal{L}(X,X)} = \|A\|^2_{\mathcal{L}(X,Y)}$.

If $A^* = A$, then A is called **self-adjoint**.

[Draft of October 6, 2021. Not for dissemination.]

A.2.2 Orthogonality and orthogonal systems

A scalar product allows to coin the notion of orthogonality: if X is a Hilbert space, then two elements $x, y \in X$ are **orthogonal**, if $\langle x, y\rangle_X = 0$.

For any subset $U \subset X$, the set

$$U^\perp := \{x \in X : \langle x, u\rangle_X = 0 \text{ for all } u \in U\}$$

is called **orthogonal complement** of U in X. It follows from the definition that $U^\perp$ is a closed subspace of X. In particular, we have $X^\perp = \{0\}$ and $U \subset (U^\perp)^\perp$.

If U is a closed subspace of X, then $U = (U^\perp)^\perp$ (and thus also $\{0\}^\perp = X$). In this case, there exists an **orthogonal decomposition**

$$X = U \oplus U^\perp\,,$$

i.e., each element $x \in X$ can be uniquely decomposed as

$$x = u + u_\perp, \qquad u \in U,\ u_\perp \in U^\perp\,.$$

The assignment $x \mapsto u$ defines a linear operator $P_U \in \mathcal{L}(X, X)$, the **orthogonal projection** onto U. It has the following properties:

(i) P_U is self-adjoint;

(ii) $\|P_U\|_{\mathcal{L}(X,U)} = 1$ for $U \neq \{0\}$;

(iii) $Id - P_U = P_{U^\perp}$;

(iv) $\|x - P_U x\|_X = \min_{u\in U} \|x - u\|_X$;

(v) $z = P_U x$ iff $z \in U$ and $z - u \in U^\perp$.

If the subspace U is not closed, we only have$(U^\perp)^\perp = \overline{U} \supset U$. Thus, for any $A \in \mathcal{L}(X, Y)$,

(i) $\mathcal{R}(A)^\perp = \mathcal{N}(A^*)$ and thus $\mathcal{N}(A^*)^\perp = \overline{\mathcal{R}(A)}$,

(ii) $\mathcal{R}(A^*)^\perp = \mathcal{N}(A)$ and thus $\mathcal{N}(A)^\perp = \overline{\mathcal{R}(A^*)}$.

The kernel of a bounded linear operator is always closed and A is injective iff $\mathcal{R}(A^*)$ is dense in X.

A set $U \subset X$, consisting of pairwise orthogonal elements, is called an **orthogonal system** in X. If

$$\langle x, y\rangle_X = \begin{cases} 1 & \text{for } x = y, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for all } \ x, y \in U,$$

we speak of an **orthonormal system**. An orthonormal system is **complete**, if there exists no orthonormal system $V \subset X$ with $U \subsetneq V$.

Every orthonormal system $U \subset X$ satisfies the **Bessel inequality**

$$\sum_{y\in U} |\langle x, y\rangle_X|^2 \leq \|x\|_X^2 \qquad \text{for all } x \in X\,, \tag{A.1}$$

with only countably many nonzero terms in the sum. In the case of equality, U is complete and called an **orthonormal basis (ONB)** and

$$x = \sum_{y \in U} \langle x, y \rangle_X \, y \qquad \text{for all } x \in X \,.$$

Every Hilbert space has an ONB. If the ONB is countable, the Hilbert space is called **separable**. In that case, there exists a sequence $(u_n) \subset U$, such that $U = \operatorname{span}(u_n)$. It follows from the Bessel inequality that the sequence (u_n) converges weakly to zero (but not strongly, since $\|u_n\|_X = 1$).

Example A.2.4. Let $X = L^2([0,1])$. An ONB (u_n) for X is given by

$$u_n = \begin{cases} \sqrt{2}\sin(\pi(n+1)x) & n > 0 \text{ odd} \\ \sqrt{2}\cos(\pi n x) & n > 0 \text{ even} \\ 1 & n = 0 \,. \end{cases}$$

Every closed subspace $U \subset X$ has an ONB (u_n), which can be used to define the orthogonal projection onto U by

$$P_U x = \sum_{j=1}^{\infty} \langle x, u_j \rangle_X \, u_j \,.$$

A.2.3 Compact operators and the Spectral Theorem

In the same way as Hilbert spaces are a generalisation of finite dimensional vector spaces, compact operators are the infinite dimensional analogon of matrices.

An operator $A : X \to Y$ is said to be **compact**, if the image of any bounded sequence $(x_n) \subset X$ has a convergent subsequence $(Ax_{n_k})_{k \geq 1} \subset Y$. An equivalent chatacterisation is as follows: A is compact iff A maps weakly convergent sequences in X to strongly convergent sequences in Y. (Such an operator is also called **completely continuous**.) In general, compact operators will be denoted by K. Clearly every linear operator is compact if Y is finite dimensional. In particular, the identity operator $Id : X \to X$ is compact iff $\dim(X) < \infty$.

Furthermore, the space $\mathcal{K}(X,Y)$ of all compact operators from X to Y is a closed subspace of $\mathcal{L}(X,Y)$ (and hence a Banach space with the operator norm), and the limit of a sequence of linear operators with finite dimensional range is also compact. If $A, S \in \mathcal{L}(X,Y)$ and at least one of the two operators is compact, then $S \circ A$ is also compact. Finally, A^* is compact iff A is compact (Schauder Fixed-Point Theorem).

Example A.2.5. A canonical example for compact operators are **integral operators**.

Let $X = Y = L^2([0,1])$ and, for a given kernel function $k \in L^2([0,1] \times [0,1])$, consider the linear operator $K : L^2([0,1]) \to L^2([0,1])$, pointwise defined by

$$[Kx](t) = \int_0^1 k(s,t)x(s)\mathrm{d}s \qquad \text{for almost all } t \in [0,1] \,.$$

Fubini's Theorem guarantees $Kx \in L^2([0,1])$, and using the Cauchy-Schwarz inequality and Fubini's Theorem again it follows that

$$\|K\|_{\mathcal{L}(X,X)} \leq \|k\|_{L^2([0,1])} \,,$$

[Draft of October 6, 2021. Not for dissemination.]

which furthermore implies that K is a bounded operator from $L^2([0,1])$ to $L^2([0,1])$.

Since the kernel function $k \in L^2([0,1]^2)$ is measurable, there exists a sequence $(k_n) \subset L^2([0,1]^2)$ of simple piecewise constant functions such that $k_n \to k$ in $L^2([0,1]^2)$. Let $E_1, \ldots, E_n$ be a finite disjoint partitioning of $[0,1]$ with $|E_i| \leq cn^{-1}$, $i = 1, \ldots, n$. Then we could choose for example

$$k_n(s,t) = \sum_{i,j}^{n} a_{ij} \xi_{E_i}(s) \xi_{E_j}(t)\,,$$

with ξ_E the characteristic function of $E \subset [0,1]$. If K_n denotes the integral operator with kernel k_n instead of k, then it follows from the linearity of the integral that

$$\|K_n - K\|_{\mathcal{L}(X,X)} \leq \|k_n - k\|_{L^2([0,1]^2)} \to 0\,, \quad \text{as} \;\; n \to \infty,$$

i.e., $K_n \to K$ in $\mathcal{L}(X,X)$. On the other hand,

$$[K_n x](t) = \int_0^1 k_n(s,t) x(s) \mathrm{d}s = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij} \int_{E_i} x(s) \mathrm{d}s \right) \xi_{E_j}(t)$$

and thus $K_n x$ is a linear combination of the n functions ξ_{E_i}, $i = 1, \ldots, n$. As the limit of a sequence (K_n) of linear operators with finite dimensional range, the operator K is compact.

For the adjoint operator $K^* \in \mathcal{L}(X,X)$ we have

$$[K^* y](s) = \int_0^1 k(s,t) y(t) \mathrm{d}t \qquad \text{for almost all } s \in [0,1]\,.$$

Hence, an integral operator is self-adjoint iff the kernel function is symmetric, i.e., $k(s,t) = k(t,s)$ for almost all $s, t \in [0,1]$.

The analogy between compact operators and matrices is primarily related to the fact that compact linear operators have only countably many eigenvalues. (For bounded linear operators that is not necessarily the case!) We even have the following extension of the Schur decomposition for matrices.

Theorem A.2.6 (Spectral Theorem)**.** *Let X be a Hilbert space let $K \in \mathcal{K}(X,X)$ be self-adjoint. Then there exists an orthonormal system $(u_n) \subset X$ and a null sequence $(\lambda_n) \subset \mathbb{R} \setminus \{0\}$ with*

$$Kx = \sum_{n=1}^{\infty} \lambda_n \langle x, u_n \rangle_X u_n \qquad \textit{for all } x \in X\,.$$

The sequence (u_n) forms an ONB for $\overline{\mathcal{R}(K)}$.

Letting $x = u_n$, we can see that u_n is an eigenvector corresponding to the eigenvalue λ_n with $Ku_n = \lambda_n u_n$. Typically, the eigenvalues and the corresponding eigenvectors are ordered such that

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq 0\,.$$

It follows that $\|K\|_{\mathcal{L}(X,X)} = |\lambda_1|$.

[Draft of October 6, 2021. Not for dissemination.]