Technische Universität München

Fakultät für Mathematik
Professur für Wissenschaftliches Rechnen (Prof. Dr. Elisabeth Ullmann)

Exploring and exploiting hierarchies in Bayesian inverse problems

Jonas Latz, M.Sc.

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Christian Kühn

Prüfer der Dissertation:

1. Prof. Dr. Elisabeth Ullmann
2. Prof. Dr. Robert Scheichl (Ruprecht-Karls-Universität Heidelberg)
3. Prof. Dr. Andrew M. Stuart (California Institute of Technology, USA)
   nur schriftliche Beurteilung

Die Dissertation wurde am 13.08.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 20.11.2019 angenommen.

Titel in deutscher Sprache:
Hierarchische Methoden und Modelle in der Bayes'schen Inversion

Zusammenfassung.
In dieser Arbeit untersuchen wir Aspekte der Bayes'schen Inversion: Wohlgestelltheit, Diskretisierung und Algorithmen. Der erste Forschungsbeitrag dieser Arbeit ist die Einführung eines neuen Wohlgestelltheitsbegriffs Bayes'scher inverser Probleme. Dieses leicht abgeschwächte Konzept erlaubt uns die Wohlgestelltheit Bayes'scher inverser Probleme mit sehr allgemeinen Modellen zu zeigen. Bei endlichdimensionalem, nicht-degenerierten Gauß'schem Fehler reicht zum Beispiel schon die Messbarkeit des Modells.
Dann untersuchen wir die Diskretisierung Bayes'scher inverser Probleme. Genauer betrachten wir hierarchische Bayes'sche inverse Probleme, in denen das A-priori-Zufallsfeld parametrisiert ist. Standardmethoden, wie Spektralentwicklungen des Kovarianzoperators, können hier oft nicht verwendet werden, da mehrere Zufallsfelddiskretisierungen notwendig sind und diese eine übermäßig hohe Rechenzeit in Anspruch nehmen würden. Wir lösen dieses Problem mit einer Reduzierte-Basis-Methode für die schnelle Diskretisierung von parametrisierten Zufallsfeldern. Die Reduzierte-Basis-Methode in diesem Setup ist der zweite Forschungsbeitrag dieser Arbeit.
Die Lösung eines Bayes'schen inversen Problems ist die A-posteriori-Verteilung eines unsicheren Parameters oder Prozesses. Das Generieren von Zufallsvariablen bezüglich dieser Verteilung mit Markov Chain Monte Carlo (MCMC) oder Importance Sampling benötigt ebenfalls eine übermäßig hohe Rechenzeit, wenn das zugrundeliegende Modell einen hohen Rechenaufwand verursacht; also zum Beispiel eine partielle Differentialgleichung ist. Effizienter sind hier Sequentielle Monte-Carlo-Methoden, die auf Hierarchien von Modelldiskretisierungen und temperierten Likelihoods aufbauen. Der dritte Forschungsbeitrag dieser Arbeit ist eine adaptive Strategie, welche diese Hierarchien kombiniert: die Multilevel-Sequential^2-Monte-Carlo-Methode. Wir leiten diesen Algorithmus her und vergleichen ihn in Experimenten mit anderen Sequentiellen Monte-Carlo-Verfahren.
Wir interpretieren die Sequentielle Monte-Carlo-Methode als Markov-Kette zufälliger Maße. Der vierte Forschungsbeitrag dieser Arbeit ist eine Diskussion des asymptotischen Verhalten dieser Markov-Ketten und der Konvergenz der Sequential-Monte-Carlo-Methode. Wir beschließen die Arbeit mit einem Ausblick auf zukünftige Forschungsthemen.

Abstract. We make several novel contributions to aspects of the Bayesian approach to inverse problems: well-posedness, discretisation, and algorithms. The first main contribution of this work is a new concept of well-posedness of Bayesian inverse problems. In contrast to the existing concepts, our slightly simplified concept allows us to make well-posedness statements with respect to general mathematical models. Under e.g. finite-dimensional, non-degenerate Gaussian noise assumptions, we only need to show measurability of the underlying model.
Next, we move on to the discretisation of Bayesian inverse problems. We consider hierarchical Bayesian inverse problems in which the prior random field is parameterised. Typically, random fields are discretised by truncated spectral expansions, such as the Karhunen–Loève expansion. Such discretisation strategies may adhoc be not suitable in hierarchical settings, since the parameterisation may require a large number of random field discretisations. This is computationally infeasible. The second main contribution of this thesis is a reduced basis method allowing for a computationally cheap, parameterised discretisation.
The solution of a Bayesian inverse problem is the posterior distribution. Sampling from the posterior distribution, with e.g. Markov chain Monte Carlo (MCMC) or Importance Sampling, may be unsuitable if the Bayesian inverse problems are constrained by a computationally tasking mathematical model, e.g. by a partial differential equation (PDE). More suitable are Sequential Monte Carlo strategies that use hierarchies of model discretisations and tempered likelihoods. An adaptive combination of these hierarchies leads to the third main contribution of this work: the highly efficient Multilevel Sequential^2 Monte Carlo algorithm. We derive this method and compare it numerically with standard Sequential Monte Carlo methods. Moreover, we interpret Sequential Monte Carlo in a framework where it is a Markov chain of random measures. In this setting, we discuss the long-time behaviour of these Markov chains and, thus, the convergence of Sequential Monte Carlo. This is the fourth main contribution of this work. We conclude by pointing the reader to directions for future research.

Acknowledgements

I express my gratitude and deep appreciation for the help and support of my advisor Elisabeth Ullmann. My work and research would not have been the same without her knowledge, experience, and patience. I also want to thank her for the freedom she gave me to pursue my own ideas and own projects.

During my doctoral research, I had the opportunity to collaborate with various, brilliant researchers from Technical University of Munich (TUM) and beyond (place in parantheses, if beyond): Wolfgang Betz, Matthieu Bulté, Hans-Joachim Bungartz, Marvin Eisenberger, Ionuţ-Gabriel Farcaş, Jan Hasenauer (Bonn), Christian Kahle, Andrew Kei Fong Lam (Hongkong), Daniel Kressner (Lausanne), Juan Pablo Madrigal Cianci (Lausanne), Stefano Massei (Lausanne), Tobias Neckel, Fabio Nobile (Lausanne), Iason Papaioannou, Sebastian Potthoff (Bremen), Daniel Schaden, Claudia Schillings (Mannheim), Björn Sprungk (Göttingen), Daniel Straub, Raul F. Tempone (Aachen and Thuwal), Felipe Uribe, and Fabian Wagner. Outside of immediate collaborations, many brilliant researchers were happy to discuss ideas and problems with me and to give me their insights: Elizabeth Bismut, Melina Freitag (Bath), Thomas Horger, Kody J. H. Law (Manchester), Steven A. Mattis, Benjamin Peherstorfer (New York), Catherine E. Powell (Manchester), Laura Scarabosio, Tim J. Sullivan (Berlin), Piotr Swierczynski, Barbara Wohlmuth, and Linus Wunderlich (London). This manuscript profited from comments and remarks made by Vanja Nikolić and Daniel Walter (Linz), who carefully reviewed it. I would like to thank all these people for their openness, their trust, their time, and for arising friendships.

My doctoral research was mainly funded by Deutsche Forschungsgemeinschaft and TUM through the International Graduate School of Science and Engineering. École polytechnique fédérale de Lausanne partially supported my research stay in its mathematics department. I obtained travel grants from Mathematisches Forschungsinstitut Oberwolfach, the Society for Industrial and Applied Mathematics (Philadelphia), and the Isaac Newton Institute (Cambridge). I thank these institutions for their generous financial support.

During the last years, I have spent hardly any time with family and friends. When I did, I received support, comprehension (for my enthusiam, not for the mathematics), and love. I will do my best to return these goods at a higher rate in the next years.

Publications by the author

Parts of this thesis are excerpts from articles that were published or submitted for publication as a part of Jonas Latz's doctoral research. Jonas Latz is single or main author of these articles, which are also listed below.

[154] J Latz. On the well-posedness of Bayesian inverse problems. *Under review*, 2019. (available as *arXiv e-print* 1902.10257)

[155] J Latz, M Eisenberger, and E Ullmann. Fast sampling of parameterised Gaussian random fields. *Comput. Methods in Appl. Mech. Engrg.*, 348:978–1012, 2019.

[156] J Latz, I Papaioannou, and E Ullmann. Multilevel Sequential^2 Monte Carlo for Bayesian inverse problems. *J. Comput. Phys.*, 368:154–178, 2018.

Excerpts of [154] appear in §1 and §2; excerpts of [155] appear in §1, §3, and §4; excerpts of [156] appear in §1, §5, and §6. Short excerpts of [154, 155, 156] may also appear in the chapters not noted here.

Jonas Latz worked on further articles during his doctoral research. We list those articles below, even though they are not contained in this thesis.

[31] M Bulté, J Latz, and E Ullmann. A practical example for the non-linear Bayesian filtering of model parameters. *Accepted for publication in Springer QUIET 2017 Special Volume.* (available as *arXiv e-print*, 1807.08713)

[87] I-G Farcaş, J Latz, E Ullmann, T Neckel, and H-J Bungartz. Multilevel adaptive sparse Leja approximations for Bayesian inverse problems. *Under review*, 2019. (available as *arXiv e-prints*, 1904.12204)

[134] C Kahle, K F Lam, J Latz, and E Ullmann. Bayesian parameter identification in Cahn–Hilliard models for biological growth. *SIAM/ASA J. Uncertain. Quantif.*, 7(2):526–552, 2019.

[248] F Uribe, I Papaioannou, J Latz, W Betz, E Ullmann, and D Straub. Bayesian inference with subset simulation in spaces of varying dimension. *Under review*, 2019. (available at `https://www.bgu.tum.de/fileadmin/w00blj/era/Papers/2019_uribe_et_al_JCP.pdf` *[accessed August 10th 2019]*)

August 12th, 2019

Contents

Introduction

> The probability that two subsequent events will both happen is a ratio compounded of the probability of the 1st, and the probability of the 2[n]d on supposition the 1st happens.
>
> Thomas Bayes [13, Proposition 3]

Machine learning and artificial intelligence have brought mathematical models into everyday life. Smartphones and smart homes have voice assistants that use models for voice recognition. Digital photo albums use image recognition models to sort photos with respect to the person the photos are showing. Self-driving cars are not a metaphor for scientific and technical progress, but exist and are the result of good models for image and video segmentation.

In science and engineering, mathematical models have been present for a much longer time. Being able to describe a system of interest mathematically has immense advantages in terms of understanding processes, predicting behaviour, and making decisions. In this thesis, we consider the process that is called *training* in machine learning and *calibration*, *parameter identification*, or *inverse problem* in the mathematics and engineering community. Here, a basic model is fitted to a particular system of interest, using observations from said system.

Bayes [13] and Laplace [152] considered parameter estimation already in the 18^{th} century and introduced the basic concept of – what we call nowadays – *Bayesian inference* to approach such problems; see e.g. [54]. The least squares approach, as used by Gauss (see e.g. Teets and Whitehead [245]), has been of more use in practical problems. This may be due to the non-affordable computational expenses coupled to the Bayesian approach. In recent years, ever since the seminal works by Tarantola [242] and Kaipio and Somersalo [135], the Bayesian approach has gained popularity in inverse problems. Due to the technical progress, the computational expenses have become affordable.

An argument in favour of the Bayesian approach to inverse problems is its stability with respect to perturbations in the data. This has been discussed extensively by Stuart [236] and others, e.g., [84, 126, 127, 239, 240]. Since inverse problems are often discussed in terms of Hadamard's [110] *well-posedness*, this term was also introduced to discuss existence, uniqueness, and stability properties of solutions to Bayesian inverse problems.

In this thesis, we extend the theory on well-posedness of Bayesian inverse problems. In principle, we weaken the so far popular concept of - what we will call - Lipschitz well-posedness. Here, stability is quantified in terms of Lipschitz continuity of the

solution to the Bayesian inverse problem with respect to perturbations in the data. In our proposed concept, stability is quantified only in terms of continuity. This will allow us to extend the class of well-posed Bayesian inverse problems significantly, while still fitting to Hadamard's concept. Moreover, we will be able to discuss stability in cases where we cannot analyse the underlying mathematical models.

Hierarchies

In Applied Mathematics, hierarchies occur in various settings in which they have different meanings. In the following, we list the two kinds of hierarchies that we consider throughout this thesis.

In numerical analysis, hierarchies refer to different accuracies with which a mathematical model is approximated. A partial differential equation (PDE) can be solved, for instance, on a mesh that is refined throughout the process. The different refinement steps give the levels in a hierarchy of meshes. We show such a hierarchy in Figure 0.1, where the coarse solution on the left-hand side becomes more accurate when approaching the right-hand side. We will refer to a method as hierarchical if it uses different resolutions of the same mathematical model. A classical example for such hierarchical methods would be *multigrid solvers* for linear PDEs; see e.g., [109, 246]. In uncertainty quantification, so-called *multilevel methods* have recently gained attention; see [18, 19, 45, 87, 88, 102, 103, 243, 244, 247], to name a few. The

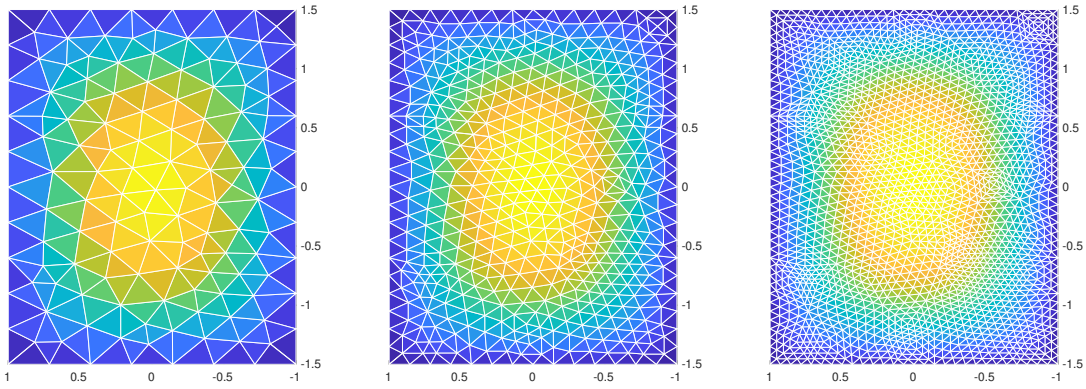


Figure 0.1. An example for a hierarchy of resolutions. We show numerical solutions of the PDE $-\triangle u = 1$ on $(-1, 1) \times (-1.5, 1.5)$ with $u = 0$ on the boundary. We start with a coarse resolution in the left figure, which is refined when going to the right. The PDE solution is given by the colourmap; the mesh is indicated by the white grid. We used the MATLAB PARTIAL DIFFERENTIAL EQUATION TOOLBOX 3.0 to generate the mesh, the mesh refinements, and the PDE solutions.

levels in multilevel refer again to different mesh sizes. Hence, multilevel methods are hierarchical as well. They *exploit* the hierarchy to speed-up an estimation process. In this thesis, we contribute to the field of multilevel methods by proposing a novel, highly efficient approach to PDE-constraint Bayesian inverse problems; *the Multilevel Sequential*^{2} *Monte Carlo algorithm*. This method is based on the Sequential Monte Carlo framework. This is a general framework to approximate sequences –

or maybe, hierarchies – of probability measures. We will interpret Sequential Monte Carlo in terms of a Markov chain of random measures. This setting will allow us to consider the long-time behaviour of Sequential Monte Carlo methods with finite sets of particles.

In statistics and probability theory, hierarchies usually describe compositions of probability measures and Markov kernels. They are used if a basic probability measure cannot describe the complexity of the random experiment of interest; basic probability measures being, e.g. Gaussian, Cauchy, Poisson, or binomial distributions. A way to construct such a hierarchy is to take a basic probability measure, parameterise it, and consider its parameters as a random variable. This random variable is then distributed according to another probability measure. This process can be extended recursively by parameterising the second probability measure as well, considering its parameter to be random, and so on. In Figure 0.2, we show such a process that allows us to describe the probability measure of a random variable $\boldsymbol{\theta}$. We see the probability measure of $\boldsymbol{\theta}$ on the right-hand side as a measure depending on three parameters. These parameters are random as well and their probability distributions are also parameterised. This process goes on until a fourth layer. Visually and also conceptually, such *hierarchical measures* are very similar to

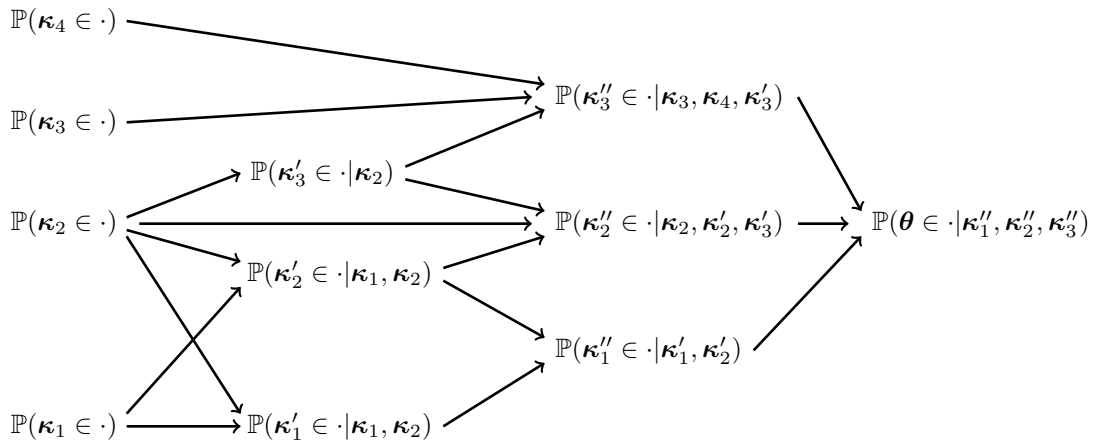


Figure 0.2. A hierarchical measure shown as a network of conditional measures.

deep neural networks that are popular in the machine learning community. While a hierarchical measure is constructed as successively composed Markov kernels and measures, deep neural networks consist of a sequence of composed activation functions, weight matrices, and bias vectors.
In uncertainty quantification, hierarchical measures and models have recently gained attention, especially in problems with random fields; see, e.g. [28, 76, 77, 78, 214, 258]. However, from a computational point of view, working with parameterised (or *hierarchical*) random fields in high-dimensional setting has so far received little attention. Here, parameterised random fields require hundreds of thousands of standard random field discretisations, i.e. with finite elements or spectral expansions. This is often computationally infeasible. To enable us to *explore* such hierarchical random fields, we propose the *fast sampling* framework. Here, we use the classical reduced basis setting to construct a surrogate for the parametric Karhunen–Loève eigenproblem. This reduces the cost of a random field discretisation significantly.

It allows us to perform hierarchical forward uncertainty quantification and solve hierarchical Bayesian inverse problems. Moreover, we study theoretical properties of both these problems; in particular the well-definedness of both problems and well-posedness of the hierarchical Bayesian inverse problem.

Organisation of the thesis

In this section, we lay out the organisation of the thesis. In Figure 0.3, we give a overview of dependencies of the chapters with respect to each other.

In §1 we build the mathematical foundation for this thesis. We start by discussing mathematical models that are either physics-based or data-driven. Parameters in

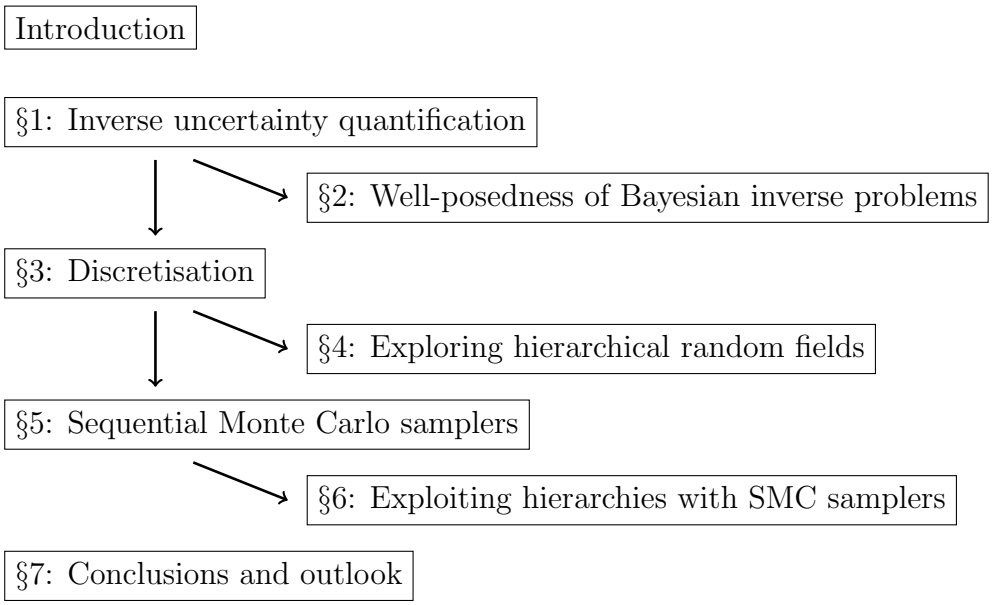


Figure 0.3. Dependencies among the chapters in this thesis. $x \to y$ indicates that y depends on material from x. Moreover, $\to$ is transitive; hence $x \to y$ and $y \to z$ imply $x \to z$.

these models are to be considered uncertain. We will represent these uncertainties with randomness and, hence, introduce the necessary background in probability theory. Moreover, we introduce inverse problems as problems of statistical inference and explain the Bayesian approach to inverse problems. In addition, we review Stuart's [236] concept of well-posedness of Bayesian inverse problems and show an implication on posterior mean estimation. Finally, we introduce hierarchical measures as a way of representing uncertainties. Here, we show the well-definedness of hierarchical forward and Bayesian inverse problems.

In §2 we introduce a new concept of well-posedness of Bayesian inverse problems. We consider stability only in terms of continuity, rather than Lipschitz continuity. We explain advantages of this approach and give weak conditions under which we obtain well-posedness. These conditions will always be satisfied if, e.g., the noise is finite-dimensional and non-degenerate Gaussian. This has also implications on the well-posedness of hierarchical Bayesian inverse problems. Then we extend the

stability results to probability metrics other than the Hellinger distance, which is by now standard in the literature. Finally, we illustrate our results with a couple of numerical examples.

In §3 we discuss the numerical approximation of Bayesian inverse problems. This consideration includes the discretisation of mathematical models, the discretisation of random fields, and the discretisation of probability measures. Here, we consider the Galerkin approach for mathematical models and random fields. Moreover, we consider the Karhunen–Loève expansion, a spectral approach for the discretisation of random fields. For the approximation of probability measures, which arise in forward and inverse uncertainty quantification, we discuss Monte Carlo, Importance Sampling, and Markov chain Monte Carlo. Finally, we discuss strategies that allow for a faster discretisation of forward and Bayesian inverse problems. Those include surrogate models and variance reduction via multilevel methods.

In §4 we explain the discretisation of random fields in hierarchical settings. The computational cost in these settings can quickly exceed any computational budget. To overcome this issue, we introduce a reduced basis surrogate for the parameterised Karhunen–Loève eigenproblem. Furthermore, we propose a linearly separable approximation to Matérn covariance operators that allow for a fast offline-online decomposition. We verify our approach in numerical experiments and show significant speed-ups. In addition, we present results of high-dimensional hierarchical Bayesian inverse problems that illustrate the applicability of our method.

In §5 we consider Sequential Monte Carlo samplers. We first introduce the problem framework, which is a sequence of measures that should be approximated. Then we discuss Sequential Importance Sampling and Sequential Monte Carlo, which are particle filters that allow for the discretisation of such a sequence. We introduce both methods as ways to construct sequences of random measures. In the following, we prove that this concept is well-defined; and that the methods produce measure-valued Markov chains. Starting from the latter point of view, we then use concepts from MCMC analysis to increase our understanding of Sequential Monte Carlo samplers with a finite number of particles. This is a novel approach towards the investigation of particle filters. Finally, we discuss how the particle filters can be used in the discretisation of Bayesian inverse problems, e.g. by tempering.

In §6 we introduce a novel multilevel approach that is based on a Sequential Monte Carlo sampler. After introducing the Multilevel Bridging algorithm and the Multilevel Sequential Monte Carlo sampler, we discuss problems of these algorithms that may be caused by the large discrepancy of posterior measures with different discretisation levels in the underlying model. Our *Multilevel Sequential*^{2} *Monte Carlo* method solves this problem with a fully adaptive strategy that combines a tempering of the likelihood and Multilevel Bridging. In numerical experiments, we show that the proposed method can outperform Multilevel Bridging and single-level Sequential Monte Carlo.

In §7 we conclude the thesis and propose directions for future research.

Chapter 1

Inverse uncertainty quantification

> [...] en nommant p la probabilité *à priori* de la cause que nous venons de considérer; on aura $E = Hp$; et en suivant le raisonnement précédent, on trouvera
>
> $$P = \frac{Hp}{S.Hp};$$
>
> Pierre-Simon Laplace [152, p. 182]

In the following sections, we discuss the mathematical background of inverse uncertainty quantification. We commence by introducing mathematical models that are used in science and engineering in §1.1. In §1.2, we consider parts of these mathematical models as uncertain and model the uncertainty probabilistically. The main feature of inverse – as opposed to forward – uncertainty quantification is the incorporation of data into uncertain models. We consider the statistical modelling of data in the context of inverse problems in §1.3. Additionally, we discuss the framework of Bayesian inference that will be used throughout this work for the inversion process. In §1.4, we discuss hierarchical models in uncertainty quantification: we motivate the use of such models, and introduce hyperpriors, as well as Bayesian model selection.

1.1 Mathematical models

In his iconic book [2], Ackoff commences Chapter 4 with the following summary:

> "The word *model* is used as a noun, adjective, and verb, and in each instance has a slightly different connotation. As a noun 'model' is a *representation* in the sense in which an architect constructs a small-scale model of a building or a physicist a large-scale model of an atom. As an adjective 'model' implies a degree of perfection or idealization, as in reference to a model home, a model student, or a model husband. As a verb 'to model' means to demonstrate, to reveal, to show what thing is like. Scientific models have all these connotations."

Hence, we define models as an *idealised representation* used to *demonstrate* the behaviour of a system of interest.

In the following, we discuss models that can be represented mathematically; those are also called *symbolic models* or *mathematical models*. A mathematical model is a function $u : D \to \mathbb{R}^K$. The domain $D \subseteq \mathbb{R}^d$ typically represents spatial positions in or at the system of interest, time points, or both. Given some $x \in D$, the vector $u(x) \in \mathbb{R}^K$ describes the behaviour at said spatial position or time point.
We distinguish two classes of models:

1. *Physics-based models* that arise from scientific laws, such as mass conservation and gravity in classical mechanics, Heisenberg's uncertainty principle in quantum mechanics, or reaction equilibria in chemistry.

2. *Data-driven models* (or sometimes *non-physics-based models*) are not defined by underlying scientific laws, but give flexible frameworks that have to be calibrated with data. Examples are models in machine learning, like deep neural networks and Gaussian processes.

1.1.1 Physics-based models

Physics-based models can be defined implicitly as the unique solution of a certain equation. Such an equation is e.g. an ordinary differential equation (ODE) or a partial differential equation. Let H be a space of functions mapping from D to $\mathbb{R}^K$ or equivalence classes of such mappings and H' be some further space. H is called *solution space*, H' is called *test space*. Moreover, let $\mathcal{E} : H \times H' \to \mathbb{R}$ be an operator, where

$$\mathcal{E}(u^*, v) = 0 \text{ for a unique } u^* \in H \text{ and all } v \in H'. \tag{1.1}$$

The solution u^* is then the mathematical model defined by $\mathcal{E}$. Throughout this chapter and the remainder of this work, we consider *elliptic (partial differential) equations* as a recurring example. As a fundamental resource for such PDEs in weak and strong form, we cite Gilbarg and Trudinger [101] and, as a historical resource, Fick [91]. First of all, we consider the *Poisson equation*.

Example 1.1 (Poisson equation; strong form)**.** Let $D \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, be an open, bounded and connected set, with sufficiently smooth boundary. We aim to construct a function $u : D \to \mathbb{R}$ that models the distribution of temperature in the domain D, subject to the following assumptions:

- D has a homogeneous (i.e. constant) thermal conductivity $a > 0$,
- heating and cooling within D is modelled by a continuous function $f : \overline{D} \to \mathbb{R}$,
- energy is conserved,
- the environment around the structure is cooled down to temperature zero,
- the temperature does not change over time.

Subject to temperature u and thermal conductivity a, one can compute the heat flux density $q : D \to \mathbb{R}^d$ in the domain D by Fourier's law:

$$q(x) = a\nabla u(x) \qquad (x \in D) \qquad \text{(Fourier's law)}$$

Moreover, conservation of energy implies that changes in energy are only controlled by the source term f:

$$-\nabla \cdot q(x) = f(x) \qquad (x \in D). \tag{conservation}$$

The temperature being set to zero at the boundaries implies Dirichlet boundary conditions. Combining Fourier's law, conservation of energy, and boundary conditions, we obtain the *strong form of the Poisson equation*

$$\begin{aligned} -\triangle u(x) &= a^{-1} f(x) && (x \in D) \\ u(x) &= 0 && (x \in \partial D). \end{aligned} \tag{1.2}$$

◇

We introduced the Poisson equation as a model for thermal conduction. Our derivation can be found in many textbooks on thermodynamics, e.g. Astarita [9, §7.1]. Note that we show approximate solution of a Poisson equation in Figure 0.1.
Next, we consider the weak form of the Poisson equation. It can be obtained by multiplying the strong Poisson equation with test functions, integrating on both sides, and then applying integration by parts. The weak formulation gives us an example of a model in terms of the operator $\mathcal{E}$, solution space H, and test space H', given in (1.1). Furthermore, the weak formulation is the basis for finite element discretisations, which we review in §3.1.

Example 1.2 (Poisson equation; weak form)**.** We consider the set-up from Example 1.1. Let $H := H' := \mathbf{H}_0^1(D)$ be the Sobolev space of square integrable functions, with zero boundary values, and square integrable first order weak derivatives. The function $u^* \in H$ solves the *weak form of the Poisson equation*, if

$$\mathcal{E}(u^*, v) := \int_D \langle \nabla u^*(x), \nabla v(x) \rangle_D \mathrm{d}x - \int_D a^{-1} f(x) v(x) \mathrm{d}x = 0 \text{ for all } v \in H'. \tag{1.3}$$

The *Lax–Milgram theorem*, see e.g. [238], implies the unique solvability of the weak formulation of the Poisson equation. ◇

In this work, we consider parameterised mathematical models. Hence, we define classes of mathematical models. Such a class is a subset of H and determines e.g. an underlying physical law that is not sufficient to determine the system's behaviour. In a class, a variety of systems of interest can be represented. Note that rather than defining a set of models, we define classes by parametrisations. Let X be the so-called *parameter space*, which we assume to be a Borel measurable subset of some separable Banach space. Note that X may be infinite-dimensional. A *parameterised model* is then a function $G : X \to H$ that maps a parameter in X to a model in H. The class represented by G is the image of G. We denote implictly defined parameterised models by $\mathcal{E} : H \times H' \times X \to \mathbb{R}$. In this case, the function $G : X \to H$ maps $X \ni \theta \mapsto u^* \in H$, where

$$\mathcal{E}(u^*, v; \theta) = 0 \text{ for a unique } u^* \in H \text{ and all } v \in H'. \tag{1.4}$$

A straightforward example for a parameterised model in the context of Example 1.1 and 1.2 is the class of Poisson equations with continuous source terms. In that

case, the parameter space is the Banach space $X := \mathbf{C}^0(\overline{D})$ of continuous functions mapping from the closure of D to $\mathbb{R}$ and the function G maps the parameter $f \in X$ to $u \in H$ given in Equation (1.3).
We now move from the Poisson equation to the more general *stationary diffusion equation*, which is also an elliptic equation. Since we do not consider non-stationary diffusion equations in this thesis, we sometimes drop the term *stationary* in the following. The diffusion equation appears in the mathematical modelling of fluid flow in a porous medium. Moreover, it is commonly used as an academic example in uncertainty quantification, see our pointers to literature in §1.2.5. We introduce the diffusion equation with a brief explanation of the governing physical laws in the following Example 1.3. These explanations are very similar to the discussion in Example 1.1. Similarly to Example 1.2, we then give the weak formulation of the diffusion equation. We close this section by discussing the parametrisation that we consider throughout this thesis.

Example 1.3 (Stationary diffusion equation; strong form)**.** Let $D \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, be an open, bounded and connected set, with sufficiently smooth boundary, that models a groundwater reservoir in 1D, 2D, or 3D. We aim to construct a function $u : D \to \mathbb{R}$ that models the pressure head of the water in D, subject to the following assumptions:

- the permeability of the reservoir is modelled by a continuously differentiable function $a : \overline{D} \to (0, \infty)$,
- in- and outflow of water is modelled by a continuous function $f : \overline{D} \to \mathbb{R}$,
- the pressure outside of the reservoir is equal to zero,
- the pressure does not change over time,
- the water is incompressible.

From pressure head u and permeability a, one can compute the flux $q : D \to \mathbb{R}^d$ in D by Darcy's law:

$$q(x) = -a(x)\nabla u(x) \qquad (x \in D) \qquad \text{(Darcy's law)}$$

Moreover, the incompressibility of the water implies that changes in mass and volume are only controlled by the source term f:

$$\nabla \cdot q(x) = f(x) \qquad (x \in D). \qquad \text{(incompressibility)}$$

Since the pressure head is zero outside of the reservoir, we consider Dirichlet boundary conditions. Combining Darcy's law, incompressibility, and boundary conditions, we obtain the *strong form of the diffusion equation*

$$\begin{aligned} -\nabla \cdot a(x)\nabla u(x) &= f(x) && (x \in D) \\ u(x) &= 0 && (x \in \partial D). \end{aligned}$$

In the context of elliptic equations, a is called *diffusion coefficient*. ◇

Our derivation of the diffusion equation is based on textbooks on hydrology, e.g. Hölting and Coldewey [122, §4], and (computational) fluid dynamics, e.g. Petrila and Trif [201]. For a theoretical derivation of Darcy's law, we refer to Whitaker [256]. We typically parameterise the diffusion equation by parameterising the diffusion coefficient a. In particular, we define $a : \overline{D} \times X \to (0, \infty)$. This gives us a class of models representing the flow in a reservoir with respect to the particular permeability structure $a(\cdot, \theta)$, specified by $\theta \in X$. Next, we give the weak formulation of the diffusion equation and give an example for the parameterisation.

Example 1.4 (Stationary diffusion equation; weak form)**.** We consider the setting from Example 1.3. Let the parameter space $X := \mathbf{C}^0(\overline{D})$ be the Banach space of continuous functions from $\overline{D}$ to $\mathbb{R}$. The permeability is given by $a(\theta, x) := \exp(\theta(x))$. Moreover, let again $H := H' := \mathbf{H}^1_0(D)$. The parameterised model $G : X \to H$ maps $\theta \in X$ to $u^* \in H$, where u^* solves the *weak form of the diffusion equation*

$$\mathcal{E}(u^*, v; \theta) := \int_D a(\theta, x) \langle \nabla u^*(x), \nabla v(x) \rangle_D \mathrm{d}x - \int_D f(x) v(x) \mathrm{d}x = 0 \text{ for all } v \in H'.$$

The existence and uniqueness of $u^* \in H$ is again guaranteed by the Lax–Milgram theorem for any $\theta \in X$. Note that Lax–Milgram holds, since $\overline{D}$ is compact, θ is continuous, and $a = \exp(\theta)$, thus, bounded away from 0 and ∞. ◇

In the following, we concentrate only on the weak formulations of the PDEs. We drop the term *weak*, when referring to weak Poisson or diffusion equations. Considering only weak formulations comes at some cost. This is the subject of the following remark.

Remark 1.5. One can show that strong solutions of Poisson and diffusion equations also satisfy the corresponding weak equations. Due to missing regularity of D and a, strong solutions may not exist, even if weak solutions are available. Note for instance that we have only assumed $a \in \mathbf{C}^0(\overline{D})$ for the weak formulation, but $a \in \mathbf{C}^1(\overline{D})$ for the strong formulation. Hence, the converse is in general not true: weak solutions do not necessarily satisfy the strong equations.
We note that even weaker assumptions on the functions a and f can be made in the examples above. For instance, one can consider the function $a(\cdot)$ to be measurable, strictly positive, and bounded above and $f \in \mathbf{L}^2(\overline{D})$. This still implies the existence of weak solutions, as discussed by e.g. [101, Theorem 8.3]. ◇

Throughout this thesis, we always assume that mathematical models are parameterised. Therefore, we usually drop the term *parameterised* and refer already to G as a *model*.

1.1.2 Data-driven models

Data-driven models are employed, if the underlying physical process is unknown or too complicated to be described. Those models require data that reflects the underlying process. Data-driven models are particularly popular in machine learning applications, such as pattern recognition problems. Here, we refer to Murphy [188]; see also the following example.

Example 1.6 (Image recognition)**.** Literate humans are typically able to recognise handwritten digits. However, we do not know a mathematical model or physical process describing the map from $\{0, 1, \dots, 9\}$ to the handwritten versions of each of these digits. Data-driven approaches to model this map (or its inverse) are very popular. They have for instance been considered in the works of LeCun and co-authors [159, 160, 161]. ◇

Data-driven models are again given by a function $G : X \to H$ mapping from a parameter space X to a solution space H. Physics-based models have parameters to adjust the given laws to a particular process the underlying physics of which is known. The parameters in data-driven models adjust and control all underlying processes and relationships.
We define data-driven models on the space $H := \mathbf{M}(D; \mathbb{R}^K)$ of measurable functions from $D := \mathbb{R}^d$ to $\mathbb{R}^K$. Typical examples include linear models and artificial neural networks (ANNs):

Example 1.7 (Basic models)**.** Let the parameter space $X := \mathbb{R}^{K\times d} \times \mathbb{R}^K$ contain pairs of matrices and vectors. The matrices represent linear maps from D to $\mathbb{R}^K$. The vectors are in $\mathbb{R}^K$.

- A *linear model* G maps

$$X \ni (W, b) \mapsto \left(D \ni x \mapsto Wx + b \in \mathbb{R}^K\right) \in H.$$

- Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an *activation function*. An *artificial neural network* G maps

$$X \ni (W, b) \mapsto \left(D \ni x \mapsto \sigma\left(Wx + b\right) \in \mathbb{R}^K\right) \in H,$$

  where σ is applied component-wise to vectors. In artificial neural networks, W is called *weight*, b is called *bias*.

◇

Linear models are used if the underlying process can be assumed to be affine linear. Note that the denomination *linear model* is used in the statistics literature even if the classes of models consist of affine linear functions. Linear models in statistics are discussed thoroughly in Rencher and Schaalje [209]. An artificial neural network models K neurons in the human brain. The ANN represents the state of the each of the K neurons when perceiving x. Originally, the activation function σ is given by the *Heaviside function* $\mathbf{1}_{[0,\infty)}$. The state of each of the neurons is then either active (= 1) or inactive (= 0). Other activation functions lead to more general approximability properties of the ANN. As historical resources concerning artificial neural networks, we refer to [142, 182].
More complex models can for instance be constructed by a combination of the basic models. Products and sums of products of linear models lead to polynomial models. In the last decades, *deep models* have especially gained popularity. Deep models arise when composing a hierarchy of a number $N_{\text{dep}} \in \mathbb{N}$ of basic models. Here, the basic models are called *layers*. By composing ANNs, we obtain a deep neural network (DNN). Those we discuss in the following example.

Example 1.8 (Deep neural network)**.** Let $k \in \mathbb{N}^{N_{\text{dep}}+1}$ be a vector configurating the size of the layers of the deep neural network. Here, $k_0 = d$, which is the dimension of the input space D and $k_{N_{\text{dep}}} = K$ is the dimension of the output space. The parameter space X contains N_{dep} pairs of weight and bias matrices:

$$X := \prod_{n=1}^{N_{\text{dep}}} \mathbb{R}^{k_n \times k_{n-1}} \times \mathbb{R}^{k_n}$$

The deep neural network $\mathcal{G}$ maps $(W^{(n)}, b^{(n)})_{n=1}^{N_{\text{dep}}} \in X$ to $u : D \to \mathbb{R}^K$, where

$$\begin{aligned} u(x) &:= x_{N_{\text{dep}}}, \\ x_n &:= \sigma\left(W^{(n)} x_{n-1} + b^{(n)}\right), && (n = 1, \dots, N_{\text{dep}}) \\ x_0 &:= x, && (x \in D). \end{aligned}$$

Here, σ is again an activation function, which may vary throughout the layers. $\diamond$

Recent reviews on Deep Neural Networks are given by Higham and Higham [120], and LeCun et al. [159]. Deep models can also be based on compositions of linear models and ANNs. So-called convolutional neural networks (CNNs) appear in image recognition problems and can contain linear models and ANNs. CNNs use many localised neural networks that capture subsets of the images. The information captured from these subimages are combined in so-called pooling layers. These pooling layers can be linear functions. For a thorough explanation of image recognition with neural networks, we refer to Bishop [24].
In the following chapters, we focus on physics-based models. Nonetheless, when discussing hierarchical models, we come back to the notion of deep models. Also, the problems that we consider for physics-based models can be easily transferred to the data-driven setting.

1.2 Uncertainty and probability

Ackoff [2, Chapter 4] views scientific models as idealised. A model contains a feature of the reality, if and only if it is relevant. In practice, relevant features may be unknown or uncertain in the modelling process. In this section, we discuss such *uncertainties* that appear in mathematical models.
Uncertainties are particularly problematic if they influence a decision making process. Such decisions can for instance be of *engineering* or *medical nature*, or *purely scientific*. Examples are: '*How many pillars should Tower Bridge have?*', '*How should the glioblastoma of Patient A be treated?*', or '*Does a Higgs boson exist?*', respectively. Uncertainties in the model influence such decisions. For instance, the *Advanced Learner's Dictionary of Current English* [125] defines *uncertain* as something

> "[...] 1. changeable; not reliable [...] 2. not certainly knowing or known [...]"

Hence, in common language, uncertainties make a model unreliable. In the decision-theoretic context, uncertainty is often considered as opposed to *risk*. Here, a *decision* is called

1. decision under *certainty*, if the outcome of any possible decision is known,

2. decision under *risk*, if the probability distribution over the possible outcomes of any possible decision is known,

3. decision under *uncertainty*, if said probability distributions are unknown.

A model containing uncertainties leads to decisions under uncertainties, unless the probability distribution of the outcomes is determined. We define *uncertainty quantification* (UQ) as the process of determining said probability distribution. As a reference concerning decision-making under certainty, risk, and uncertainty, we mention Hansson [114].
In this work, we use probability measures and random variables to represent uncertain parameters. We briefly summarise concepts of probability theory that form the foundation of this work in §1.2.1. Those concepts are probability measures, random variables, moments, convergence of random variables, and ways to characterise probability measures. Then, we introduce metric and topological spaces of probability measures in §1.2.2. Those allow us to compare and analyse probability measures in the following sections. Having defined random variables in §1.2.1, we introduce function-valued random variables, so-called *random fields* in §1.2.3. Those play an important role in the modelling of uncertainties in partial differential equations. In §1.2.4, we discuss the concepts of conditional probability and Markov kernels. These – closely related – concepts allow for the modelling of hierarchies, such as random processes. Finally, we investigate mathematical models that contain uncertainties, as well as the quantification of said uncertainties in §1.2.5. As an example, we consider the diffusion equation with random diffusion coefficient.

1.2.1 Representation of uncertainty by probability

In this work, we represent uncertainty via randomness; that is via the outcome of a random experiment. Those random experiments are modelled using probability theory. The approach to model uncertainties with randomness has been advocated by Cox [50].

Random experiments and probability spaces. Let Ω be a set from which an element ω shall be drawn randomly. We specify which element is drawn by defining a probability measure.

Definition 1.9. Let $\mathcal{A} \subseteq 2^\Omega$ be a σ-algebra. A *probability measure* or *probability distribution* $\mathbb{P}$ is a map from $\mathcal{A}$ to $[0,1]$ that satisfies the following axioms:

(i) $\mathbb{P}(\emptyset) = 0$,

(ii) Let $I \subseteq \mathbb{N}$ be a countable index set. Moreover, let $(A_i)_{i\in I} \in \mathcal{A}^I$ be a family of pairwise disjoint events. Then, $\sum_{i\in I} \mathbb{P}(A_i) = \mathbb{P}\left(\bigcup_{i\in I} A_i\right)$,

(iii) $\mathbb{P}(\Omega) = 1$.

If $\mathbb{P}$ satisfies only (i)-(ii), it is called *measure*. ◇

An element $A \in \mathcal{A}$ is called *event* and the assigned value $\mathbb{P}(A)$ is the *probability* of the event. We formalise the random experiment of drawing $\omega \in \Omega$ according to a probability measure $\mathbb{P}$ by writing down the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The axioms in Definition 1.9 are equivalent to those first stated by Kolmogorov in his book *Grundbegriffe* [144]. Note that while being axiomatised in this book, probability theory has been considered for hundreds of years, e.g. by Laplace in 1812 [152]. Recent introductions to probability theory are given by [8, 23, 136, 143]. The following paragraphs are mostly reviews of known results from the books above.

Random variables. In the following, we intend to represent randomness in variables and parameters. Each of the so-called random variables represents a single random experiment, given by some probability space. To couple all of these random experiments, we proceed as follows: We first define an *underlying* probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Moreover, we define Z to be some space and $\mathcal{Z} \subseteq 2^Z$ to be a σ-algebra on Z.

Definition 1.10. A function $\boldsymbol{\xi} : \Omega \to Z$ is called *measurable* or *(Z-valued) random variable* if $\boldsymbol{\xi}^{-1}(Z') := \{\boldsymbol{\xi} \in Z'\} := \{\omega \in \Omega : \boldsymbol{\xi}(\omega) \in Z'\} \in \mathcal{A} \quad (Z' \in \mathcal{Z})$. ◇

As before, we denote the set of measurable functions by

$$\mathbf{M}(\Omega; Z) := \{f : \Omega \to Z : f \text{ is measurable}\}.$$

The random variable $\boldsymbol{\xi}$ corresponds to a random experiment on the probability space $(Z, \mathcal{Z}, \mathbb{P}(\boldsymbol{\xi} \in \cdot))$. The probability measure $\mathbb{P}(\boldsymbol{\xi} \in \cdot)$ is called *(probability) distribution of* $\boldsymbol{\xi}$ or *pushforward measure of* $\boldsymbol{\xi}$. If $\mu = \mathbb{P}(\boldsymbol{\xi} \in \cdot)$, we also write $\boldsymbol{\xi} \sim \mu$. Throughout this thesis, we use normal letters to denote elements of sets and the corresponding boldface letters to denote random variables taking values in the according set. Hence, if $\xi \in Z$, $\boldsymbol{\xi}$ is a Z-valued random variable.
Since all random variables are defined on $(\Omega, \mathcal{A}, \mathbb{P})$, any random experiment that is performed in the following is represented by drawing a *sample* $\omega \in \Omega$ according to the probability measure $\mathbb{P}$. The outcome of the random experiment on $(Z, \mathcal{Z}, \mathbb{P}(\boldsymbol{\xi} \in \cdot))$ is then given by $\boldsymbol{\xi}(\omega)$. The value $\xi' := \boldsymbol{\xi}(\omega') \in Z$ for an outcome ω' of a random experiment on $(\Omega, \mathcal{A}, \mathbb{P})$ is called *realisation* of $\boldsymbol{\xi}$. As usual in probability theory, we neither define the underlying space $(\Omega, \mathcal{A}, \mathbb{P})$, nor the mappings from Ω to the respective space (like Z). Instead, we define random variables via the probability space they represent, i.e. $(Z, \mathcal{Z}, \mathbb{P}(\boldsymbol{\xi} \in \cdot))$. Doing so, it is vital to specify and clarify the interdepence of random variables. Consider two random variables $\boldsymbol{\xi}_i$ on probability spaces $(Z_i, \mathcal{Z}_i, \mu_i)$ $(i = 1, 2)$. The so-called *marginal distributions* μ_1 and μ_2 do not specify the interdependence of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. Their interdependence is specified by the *joint distribution* $\mu := \mathbb{P}((\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \cdot)$. This is a probability measure on the product space $(Z_1 \times Z_2, \mathcal{Z}_1 \otimes \mathcal{Z}_2)$. Finally, we call $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ *(stochastically) independent*, if their joint distribution equals the product of their marginals: $\mathbb{P}((\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \cdot) = \mu_1 \otimes \mu_2$.

Moments. Let Z now be a separable Banach space equipped with the *Borel-σ-algebra* $\mathcal{B}Z$, and let $\boldsymbol{\xi} \sim \mu$ be a random variable on Z. Let $k \in \mathbb{N}$. We define the *k-th moment of* $\boldsymbol{\xi}$ by

$$\mathbb{E}[\boldsymbol{\xi}^k] := \int_\Omega \boldsymbol{\xi}(\omega)^k \mathrm{d}\mathbb{P}(\omega) := \int_Z \xi^k \mathrm{d}\mu(\xi),$$

if the (Bochner-)integral on the right-hand side exists and is finite. The first moment ($k = 1$) is called *expected value*, or *mean*. Moreover, the second, centralised moment is called *variance*

$$\mathrm{Var}(\boldsymbol{\xi}) := \mathbb{E}\left[(\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}])^2\right] = \mathbb{E}\left[\boldsymbol{\xi}^2\right] - \mathbb{E}[\boldsymbol{\xi}]^2$$

and the square-root of the variance is called *standard deviation*

$$\mathrm{StD}(\boldsymbol{\xi}) := \sqrt{\mathrm{Var}(\boldsymbol{\xi})}.$$

If Z is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_Z$, we define the *covariance operator* of $\boldsymbol{\xi}$ by

$$\mathrm{Cov}(\boldsymbol{\xi}) : Z \to Z, \qquad z \mapsto \mathbb{E}\left[\langle \cdot, \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}] \rangle_Z \langle \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}], z \rangle_Z\right].$$

Note, that the Riesz representation theorem, see [219, Theorem 4.12], implies the well-definedness of this map. We sometimes consider moments not with respect to $\mathbb{P}$, but with respect to other measures. Let Z_1 be another separable Banach space that is equipped with the Borel-σ-algebra $\mathcal{B}Z_1$. Furthermore, let $\varphi : Z \to Z_1$ be a measurable function, and $\boldsymbol{\xi} \sim \mu$. We abbreviate,

$$\mathbb{E}_\mu[\varphi] := \mathbb{E}[\varphi(\boldsymbol{\xi})], \qquad \mathrm{Var}_\mu(\varphi) := \mathrm{Var}(\varphi(\boldsymbol{\xi})), \qquad \mathrm{StD}_\mu(\varphi) := \mathrm{StD}(\varphi(\boldsymbol{\xi})),$$

if the corresponding integrals exist.

Convergence of random variables. Moments of random variables can be approximated by so-called *Monte Carlo* methods. We discuss those in §3.3. The theoretical foundation of these methods are the laws of large numbers. Given a set of independent and identically distributed (i.i.d.) random variables $(\boldsymbol{\xi}_i)_{i \in \mathbb{N}}$, the laws of large numbers treat the behaviour of the sample mean $\overline{\boldsymbol{\xi}}_n := \sum_{i=1}^n \boldsymbol{\xi}_i / n$ as $n \to \infty$.
Having the laws of large numbers in mind, we first discuss different kinds of convergence of random variables.

Definition 1.11. Let Z be a separable Banach space that is equipped with $\mathcal{B}Z$. Moreover, let $\boldsymbol{\xi}_i : \Omega \to Z$ be a Z-valued random variable for $i \in \mathbb{N}$; and let $\boldsymbol{\xi} : \Omega \to Z$ be another random variable. We say,

(i) $\boldsymbol{\xi}_i \to \boldsymbol{\xi}$ converges *in probability* as $i \to \infty$, if for all $\varepsilon > 0$ it holds

$$\lim_{i \to \infty} \mathbb{P}(\|\boldsymbol{\xi}_i - \boldsymbol{\xi}\|_Z > \varepsilon) = 0;$$

(ii) $\boldsymbol{\xi}_i \to \boldsymbol{\xi}$ converges *almost surely* (a.s.) as $i \to \infty$, if

$$\mathbb{P}(\lim_{i \to \infty} \|\boldsymbol{\xi}_i - \boldsymbol{\xi}\|_Z = 0) = 1.$$

$\diamond$

One can show that almost sure convergence of a sequence of random variables implies convergence in probability, see [143, Remark 6.4]. This justifies the nomenclature of the following *Weak* and *Strong Laws of Large Numbers*:

Theorem 1.12 (Weak Law of Large Numbers)**.** Let $(\boldsymbol{\xi}_i)_{i=1}^{\infty}$ be a sequence of i.i.d. random variables. The sample mean of the first $i \in \mathbb{N}$ random variables is given by $\overline{\boldsymbol{\xi}}_i = \frac{1}{n}\sum_{n=1}^{i} \boldsymbol{\xi}_n$. If $\mathbb{E}[\|\boldsymbol{\xi}_1\|_Z^2] < \infty$, we have

$$\mathbb{P}(\|\overline{\boldsymbol{\xi}}_n - \mathbb{E}[\boldsymbol{\xi}_1]\|_Z > \varepsilon) \leq \frac{\mathrm{Var}(\|\boldsymbol{\xi}_1\|_Z)}{n \cdot \varepsilon^2} \tag{1.5}$$

for every $\varepsilon > 0$. Hence, $\overline{\boldsymbol{\xi}}_n \to \mathbb{E}[\boldsymbol{\xi}_1]$ in probability, as $n \to \infty$.

Proof. (1.5) above follows from Chebyshev's inequality [8, §2.4.9]. □

The upper bound (1.5) gives information about the speed of convergence of the approximation. However, this information is rather limited, since it depends on the choice of ε. Instead, one should consider the *root mean square error* (RMSE) of sample mean and first moment. In the setting of Theorem 1.12, the RMSE reads

$$\sqrt{\mathbb{E}[\|\overline{\boldsymbol{\xi}}_n - \mathbb{E}[\boldsymbol{\xi}_1]\|_Z^2]} = \mathrm{StD}(\|\boldsymbol{\xi}_1\|_Z)/\sqrt{n}. \tag{1.6}$$

Hence, (1.6) indicates a rate of convergence of $O(1/\sqrt{n}; n \to \infty)$.
We state the Strong Law of Large Numbers also in the setting of random variables with finite second moment. More general situations are possible, we refer to [8, Chapter 6] and [121] for details.

Theorem 1.13 (Strong Law of Large Numbers; [121, Theorem 2.1])**.** Let Z be a separable Hilbert space. Let $(\boldsymbol{\xi}_i)_{i=1}^{\infty}$ be a sequence of i.i.d. random variables. If $\mathbb{E}[\|\boldsymbol{\xi}_1\|_Z^2]$ is finite, the sample mean $\overline{\boldsymbol{\xi}}_n$ converges almost surely to the mean, as $n \to \infty$. In other words,

$$\mathbb{P}\left(\lim_{n\to\infty} \|\overline{\boldsymbol{\xi}}_n - \mathbb{E}[\boldsymbol{\xi}_1]\|_Z = 0\right) = 1.$$

◇

Characterising probability measures. Finally, we discuss ways to represent and characterise probability measures. To this end, we use functions that are defined on the space Z or its topological dual Z^*. First, we consider *probability density functions* to characterise probability measures. Here, Z is some set that is equipped with a generic σ-algebra $\mathcal{Z}$.
Let $(Z, \mathcal{Z}, \nu_Z)$ be a σ-finite measure space, and $f : Z \to [0, \infty)$ be a measurable function that integrates to 1 on said measure space, i.e.

$$\int_Z f \mathrm{d}\nu_Z = 1.$$

Then, f defines a probability measure μ on $(Z, \mathcal{Z})$, by

$$\mu(A) := \int_A f \mathrm{d}\nu_Z \qquad (A \in \mathcal{Z}).$$

The function f is called *ν_Z-(probability) density (function)* (*ν_Z-pdf*) of μ. We call it just *(probability) density (function)* (*pdf*), if ν_Z is the Lebesgue measure on Z, or if no particular ν_Z is defined. Note that the Radon–Nikodym theorem [8, §2.2.1]

implies that if f exists as above, μ is absolutely continuous with respect to ν_Z. By the Radon–Nikodym theorem, f is ν_Z-almost everywhere (a.e.) uniquely defined. Hence, we can represent μ by f and f ν_Z-a.e. by μ. Thus, probability density functions give us a practical way to characterise probability measures.
Unfortunately, probability density functions require the existence of a dominating measure ν_Z on $(Z, \mathcal{Z})$. Natural dominating measures on Z are the counting measure (if Z is countable) or the Lebesgue measure (if Z is finite dimensional). Especially in the infinite dimensional setting, there is no measure that can easily replace these measures to define basic probability measures.
Alternatively, if Z is a separable Banach space, we can consider the *characteristic function* of a probability measure.

Definition 1.14. Let Z be a separable Banach space. Let Z^* be the *topological dual space* of Z and $(\cdot,\cdot)_{Z^*,Z} : Z^* \times Z \to \mathbb{R}$ be the *duality pairing* of Z and Z^*. The *characteristic function* or *Fourier transform* of a probability measure μ on $(Z, \mathcal{B}Z)$ is given by

$$\varphi_\mu : Z^* \to \mathbb{C}, \qquad t \mapsto \int_Z \exp(\mathrm{i}(t,\theta)_{Z^*,Z})\mathrm{d}\mu(\theta).$$

If Z is a Hilbert space, we employ its self-duality and define φ_μ on Z. ◇

One can show that every probability measure is uniquely defined by its characteristic function; see [27, A.3.18].

1.2.2 Spaces of probability measures

In §1.2.1, we have discussed types of convergence of sequences of random variables. Such sequences arise e.g. when approximating moments of probability measures with samples of those. Closely related to sequences of random variables are sequences of probability measures, and the convergence of the latter. They arise when approximating probability measures, in central limit theorems, and when considering marginal perturbations in probability measures. To discuss such sequences, we now discuss spaces of probability measures.
Let Z be again some separable Banach space that we always associate with $\mathcal{B}Z$ in this section. We define the space of probability measures on $(Z, \mathcal{B}Z)$ by

$$\mathrm{Prob}(Z) := \mathrm{Prob}(Z, \mathcal{B}Z) := \{\mu : \mathcal{B}Z \to [0,1] : \mu \text{ is probability measure}\}.$$

Moreover, for some σ-finite measure ν_Z on $(Z, \mathcal{B}Z)$, we define the space of probability measures that are absolutely continuous with respect to ν_Z by

$$\mathrm{Prob}(Z, \nu_Z) := \mathrm{Prob}(Z, \mathcal{B}Z, \nu_Z) := \{\mu \in \mathrm{Prob}(Z) : \mu \ll \nu_Z\}.$$

Finally, we define the space of probability measures with finite absolute p-th moment

$$\mathrm{Prob}_p(Z) := \mathrm{Prob}_p(Z, \mathcal{B}Z) := \left\{\mu \in \mathrm{Prob}(Z) : \int_Z \|\xi\|_Z^p \mathrm{d}\mu(\xi) < \infty\right\},$$

for $p \in [1, \infty)$. Note, that Jensen's inequality implies that

$$\mathrm{Prob}_p(Z) \subseteq \mathrm{Prob}_q(Z) \qquad (p \geq q \geq 1).$$

In the following, we consider topologies and metrics on each of these spaces. Those allow us to discuss the convergence of sequences of probability measures. Moreover, the metrics can be used to measure the distance between probability measures. Note that a thorough review concerning metrics on probability spaces is given by Gibbs and Su [99]. We start with the weak topology on $\mathrm{Prob}(Z)$.

The weak topology. Let $(\mu_i)_{i\in\mathbb{N}} \in \mathrm{Prob}(Z)^{\mathbb{N}}$ and $\mu \in \mathrm{Prob}(Z)$ be a further measure. We say $\mu_i \to \mu$ converges *weakly*, as $i \to \infty$, if

$$\lim_{i\to\infty} \int_Z f(\xi)\mu_i(\mathrm{d}\xi) = \int_Z f(\xi)\mu(\mathrm{d}\xi), \tag{1.7}$$

for any bounded and continuous function $f : Z \to \mathbb{R}$. We abbreviate this with $\mu_i \xrightarrow{\mathrm{w}} \mu$ $(i \to \infty)$.
There is a multitude of equivalent criteria for weak convergence. We mention that weak convergence is implied by pointwise convergence of the associated characteristic functions. For other criteria for weak convergence, we refer to the *Portmanteau Theorem*, see [8, Theorem 2.8.1]. Importantly, note that weak convergence does not imply that probabilities of events need to converge, i.e.

$$\mu_i \xrightarrow{\mathrm{w}} \mu \quad (i \to \infty) \qquad \not\Rightarrow \qquad \lim_{i\to\infty} \mu_i(B) = \mu(B), \tag{1.8}$$

for general $B \in \mathcal{B}Z$. We illustrate this statement with the following example.

Example 1.15. Let $Z := \mathbb{R}$ and $\mu_i := \delta(\cdot - 1/i)$ be the *Dirac measure* concentrated in $1/i$, for $i \in \mathbb{N}$. Then, $\mu_i \xrightarrow{w} \mu := \delta(\cdot - 0)$, as $i \to \infty$. Let now $B := (0, 1]$. Then, we have $\mu_i(B) = 1$, $i \in \mathbb{N}$, but $\mu(B) = 0$. Hence, $\lim_{i\to\infty} \mu_i(B) = 1 \neq 0 = \mu(B)$. $\diamond$

Weak convergence can be equivalently represented – *metrised* – by the *(Lévy-) Prokhorov distance* on the space $\mathrm{Prob}(Z)$:

$$\mathrm{d}_{\mathrm{Prok}}(\mu, \mu') := \inf\left\{\varepsilon > 0 : \mu(B) \leq \mu'(B^\varepsilon) + \varepsilon, B \in \mathcal{B}X\right\},$$

where $B^\varepsilon := \{b \in B : \exists b' \in Z : \|b - b'\|_Z < \varepsilon\}$ is the *open generalised ε-ball* around $B \in \mathcal{B}Z$.

Proposition 1.16 (Prokhorov; [205, Theorem 1.11])**.** $\mathrm{d}_{\mathrm{Prok}}$ is a metric on $\mathrm{Prob}(Z)$ and metrises the weak convergence, i.e. $\mu_i \xrightarrow{\mathrm{w}} \mu$ $(i \to \infty)$, if and only if

$$\lim_{i\to\infty} \mathrm{d}_{\mathrm{Prok}}(\mu, \mu_i) = 0.$$

$\diamond$

Motivated by Proposition 1.16, we define the *weak topology* on $\mathrm{Prob}(Z)$ as the topology induced by the Prokhorov metric $\mathrm{d}_{\mathrm{Prok}}$, i.e.

$$\mathcal{O}_{\mathrm{Prok}} := \{B \subseteq \mathrm{Prob}(Z) : \forall(\mu \in B)\exists(\varepsilon > 0)\{\mu' : \mathrm{d}_{\mathrm{Prok}}(\mu, \mu') < \varepsilon\} \subseteq B\};$$

see, e.g. [145, §4.2 (1)]. In practice, it is rather difficult to compute and to interpret the Prokhorov metric between two measures. The *Wasserstein distance* has gained attention as a metric closely connected to the weak topology. [70] is the first article mentioning the *Wasserstein distance* between probability measures. It is given as

the cost of an optimal transport from one probability measure to another. This connection is thoroughly discussed in Villani [252]. We now define the Wasserstein distance and discuss its connection to the weak topology. Let $p \in (0, \infty)$ and let

$$\mathrm{Coup}(\mu, \mu') := \{\Lambda' \in \mathrm{Prob}(Z^2) : \mu(B) = \Lambda'(B \times Z), \mu'(B) = \Lambda'(Z \times B), B \in \mathcal{B}Z\}$$

be the set of couplings of two measures $\mu, \mu' \in \mathrm{Prob}(Z)$. The couplings are the probability measures on $(Z^2, \mathcal{B}Z^2)$ that have μ and μ' as marginals. The Wasserstein(p)-metric is given by

$$\mathrm{d}_{\mathrm{Was(p)}}(\mu, \mu') := \left(\inf_{\Lambda \in \mathrm{Coup}(\mu, \mu')} \int_{Z^2} \|\xi - \xi'\|_Z^p \mathrm{d}\Lambda(\xi, \xi') \right)^{1/p},$$

if the integral on the right-hand side exists. The integral on the right-hand side exists if $\mu, \mu' \in \mathrm{Prob}_p(Z)$. On this space, convergence in the Wasserstein(p) metric is stronger than weak convergence

Proposition 1.17 ([252, Theorem 6.9])**.** Let $p \in [1, \infty)$. $\mathrm{d}_{\mathrm{Was(p)}}$ is a metric on $\mathrm{Prob}_p(Z)$ and for $\mu, \mu_i \in \mathrm{Prob}_p(Z)$ $(i \in \mathbb{N})$, we have $\lim_{i \to \infty} \mathrm{d}_{\mathrm{Was(p)}}(\mu_i, \mu) = 0$, if and only if

$$\mu_i \xrightarrow{\mathrm{w}} \mu \ (i \to \infty) \text{ and } \lim_{i \to \infty} \int_Z \|\theta\|_Z^p \mu_i(\mathrm{d}\theta) = \int_Z \|\theta\|_Z^p \mu(\mathrm{d}\theta).$$

Hence, convergence in $\mathrm{d}_{\mathrm{Was(p)}}$ implies convergence in the weak topology. ◇

Total variation and Hellinger distance. In (1.8), we mention that weak convergence does not necessarily imply convergence of probabilities of events. Convergence in total variation is a stronger concept. Here, indeed probabilities of all events converge. The total variation distance is defined as the smallest uniform bound over all differences between probabilities. In particular, the *total variation* (tv) distance of $\mu, \mu' \in \mathrm{Prob}(Z)$ is

$$\mathrm{d}_{\mathrm{tv}}(\mu, \mu') := \sup_{A \in \mathcal{B}Z} |\mu(A) - \mu'(A)|.$$

Let now $\mu, \mu' \in \mathrm{Prob}(Z, \nu_Z)$, where ν_Z is a σ-finite measure. Then, the total variation distance can be expressed as an $\mathbf{L}^1$-distance between the Radon–Nikodym derivatives of μ, μ' with respect to ν_Z:

$$\mathrm{d}_{\mathrm{tv}}(\mu, \mu') = \frac{1}{2} \int_Z \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu_Z}(\xi) - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu_Z}(\xi) \right| \nu_Z(\mathrm{d}\xi) = \frac{1}{2} \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu_Z} - \frac{\mathrm{d}\mu'}{\mathrm{d}\nu_Z} \right\|_{\mathbf{L}^1(Z, \nu_Z)}.$$

The proof of this statement is based on the Hahn–Jordan decomposition of the signed measure $\mu - \mu'$, see [219, Theorem 6.13].
Finally, we discuss the *Hellinger distance*. The Hellinger distance on $\mathrm{Prob}(Z, \nu_Z)$ is given by

$$\mathrm{d}_{\mathrm{Hel}}(\mu, \mu') := \sqrt{\frac{1}{2} \int_Z \left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu_Z}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu_Z}} \right)^2 \mathrm{d}\nu_Z} = \sqrt{\frac{1}{2}} \cdot \left\| \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu_Z}} - \sqrt{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu_Z}} \right\|_{\mathbf{L}^2(Z, \nu_Z)}.$$

The Hellinger distance is based on the work of Hellinger [118]. It can be understood as an $\mathbf{L}^2$-distance of the square-root of the densities. Note that it is well-defined, since

$$f \in \mathbf{L}^1(D;[0,\infty)) \Longrightarrow f^{1/2} \in \mathbf{L}^2(D;[0,\infty)).$$

The Hellinger distance is a metric that is *topologically equivalent* to the total variation distance on $\mathrm{Prob}(Z,\nu_Z)$, see Lemma 2.14. It is sometimes preferred over the total variation distance, as it scales like the distance in expected values with respect to the measures.

Lemma 1.18. Let $C := 2\int_Z \|\theta\|_X^2 \mu(\mathrm{d}\theta) + 2\int \|\theta\|_X^2 \mu'(\mathrm{d}\theta) < \infty$. Then,

$$\left\| \int_Z \theta\mu(\mathrm{d}\theta) - \int_Z \theta\mu'(\mathrm{d}\theta) \right\|_X \leq C\mathrm{d}_{\mathrm{Hel}}(\mu,\mu').$$

Proof. This Lemma generalises [157, Lemma 1.30]. The proof proceeds analoguously. □

We use this result in Theorem 1.51 to compare posterior means (i.e. conditional means) under perturbation of the data.

1.2.3 Random fields

We aim to model uncertain parameters in mathematical models with random variables. Parameters in mathematical models sometimes take values in function spaces. As an example, we recall the function-valued permeability in the diffusion equation in Example 1.4. Function-valued random variables are called *random fields*. The theory of random fields as function-valued random variables has been discussed by [27, 139, 136, 169, 236, 238]. We now particularly discuss Gaussian random fields that yield continuous realisations.
We start by defining Gaussian measures on separable Banach spaces. First, we recall the notion of a real-valued Gaussian random variable which induces a Gaussian measure on $\mathbb{R}$.

Definition 1.19. The random variable $\boldsymbol{\xi} : \Omega \to \mathbb{R}$ follows a *non-degenerate Gaussian measure*, if

$$\mathbb{P}(\boldsymbol{\xi} \leq \xi) := \mathrm{N}(a,b^2)((-\infty,\xi]) := \int_{-\infty}^{\xi} \frac{1}{(2\pi)^{1/2}b} \exp\left(-\frac{(\xi'-a)^2}{b^2}\right) \mathrm{d}\xi' \quad (\xi \in \mathbb{R}),$$

for some $a \in \mathbb{R}$ and $b > 0$. The Gaussian measure is *degenerate* if $b = 0$. In this case, we define $\mathrm{N}(a,0) := \delta(\cdot - a)$, the Dirac measure concentrated in a. ◇

Let Z denote a separable Banach space with Borel-σ-algebra $\mathcal{B}Z$. We now introduce Gaussian measures on Z.

Definition 1.20. The Z-valued random variable $\boldsymbol{\theta} : \Omega \to Z$ has a Gaussian measure, if $(T,\boldsymbol{\theta})_{Z^*,Z}$ follows a Gaussian measure for any $T \in Z^*$ in the topological dual space of Z. ◇

In the following, we assume that Z is a separable Hilbert space. This simplifies the following discussion, even though many of the concepts introduced below can be generalised to the separable Banach space setting; or even more general spaces.
In Definition 1.20, we distinguish two cases. If Z is finite-dimensional, we call $\boldsymbol{\theta}$ a *(multivariate) Gaussian random variable* with mean vector $m := \mathbb{E}[\boldsymbol{\theta}]$ and covariance matrix $\mathcal{C} := \mathrm{Cov}(\boldsymbol{\theta})$. If Z is infinite-dimensional, then $\boldsymbol{\theta}$ is called *Gaussian random field* with mean function $m := \mathbb{E}[\boldsymbol{\theta}]$ and covariance operator $\mathcal{C} := \mathrm{Cov}(\boldsymbol{\theta})$.
In the next section, we discuss the construction of Gaussian measures. First, we state required properties of mean and covariance. While any $m \in Z$ can be used as a mean function, the covariance operator $\mathcal{C} : Z \to Z$ has to be linear, nuclear, positive semidefinite, and self-adjoint. This is a consequence of the following theorem.

Theorem 1.21 ([27, Theorem 2.3.1])**.** Let μ be a measure on $(Z, \mathcal{B}Z)$. μ is Gaussian, if and only if there is an $m' \in Z$ and a linear operator $\mathcal{C}' : Z \to Z$, that is nuclear, positive semidefinite, and self-adjoint, such that the characteristic function φ_μ of μ satisfies

$$\varphi_\mu(t) = \exp\left(\mathrm{i}\langle m', t\rangle_Z - \frac{1}{2}\langle \mathcal{C}'t, t\rangle_Z\right) \qquad (t \in Z). \tag{1.9}$$

In this case, $\mu = \mathrm{N}(m', \mathcal{C}')$.
Moreover, all functions of the form (1.9) are valid characteristic functions. Hence, for all $m', \mathcal{C}'$ as specified above, the Gaussian measure $\mathrm{N}(m', \mathcal{C}')$ is well-defined. ◇

Note that an operator is *nuclear* (or *trace-class*) if it is compact and the sum of the norms of its eigenvalues is finite. We denote the set of valid covariance operators on Z by $\mathrm{CO}(Z)$.

Construction of Gaussian measures. We move on to the construction of Gaussian measures. If $\dim Z < \infty$, we can identify a Gaussian measure on Z in terms of a probability density function w.r.t. the Lebesgue measure.

Proposition 1.22. Let $N_{\mathrm{sto}} \in \mathbb{N}$, $Z := \mathbb{R}^{N_{\mathrm{sto}}}$, $m \in Z$, and $\mathcal{C} \in \mathrm{CO}(Z)$ with full rank. Then, the Gaussian measure can be written as

$$\mathrm{N}(m, \mathcal{C})(B) = \int_B \mathrm{n}(\theta; m, \mathcal{C})\mathrm{d}\theta \quad (B \in \mathcal{B}Z), \tag{1.10}$$

where

$$\mathrm{n}(\theta; m, \mathcal{C}) = \det(2\pi\mathcal{C})^{-1/2} \exp\left(-(1/2)\langle \theta - m, \mathcal{C}^{-1}(\theta - m)\rangle\right) \tag{1.11}$$

is the associated probability density function. ◇

If $\dim Z = \infty$, there is no Lebesgue measure with respect to which we can define a densitiy to obtain a Gaussian measure. However, it is possible to define a Gaussian measure with respect to another Gaussian measure, if one arises from the other by a translation of the mean. This is a result of the *Cameron–Martin theorem*:

Theorem 1.23 (Cameron–Martin; [27, Corollary 2.4.3])**.** Let $\mu := \mathrm{N}(m, \mathcal{C})$ be a Gaussian measure on Z. Let $\mathrm{CM}(\mu)$ be the *Cameron–Martin* space of μ, i.e. $\mathrm{CM}(\mu)$ is the Hilbert space defined as the completion of $\mathrm{img}(\mathcal{C}^{1/2})$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathrm{CM}(\mu)} := \langle C^{-1/2}\cdot, C^{-1/2}\cdot\rangle_Z$. Then,

$$\mathrm{N}(m + h, \mathcal{C}) \ll \mu \text{ and } \frac{\mathrm{dN}(m + h, \mathcal{C})}{\mathrm{d}\mu}(\theta) = \exp\left(\langle h, \theta\rangle_{\mathrm{CM}(\mu)} - \frac{\|h\|^2_{\mathrm{CM}(\mu)}}{2}\right) \quad (\mu\text{-a.s.}),$$

if and only if $h \in \mathrm{CM}(\mu)$. ◇

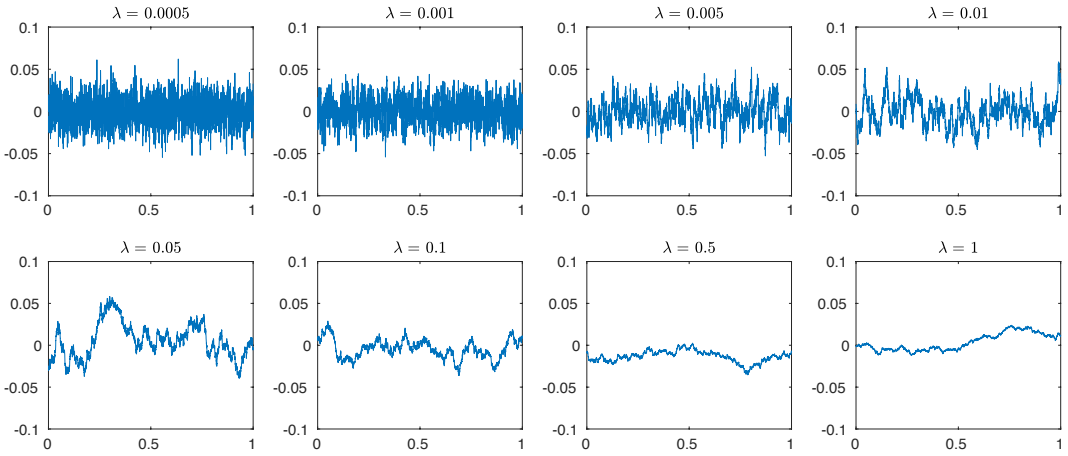


Figure 1.1. Samples of mean-zero Gaussian random fields in 1D with exponential covariance, with $\sigma := 1$ and $\lambda \in \{0.0005, 0.001, \ldots, 1\}$, where $\|\cdot\|_{\overline{D}} := |\cdot|$ is the standard modulus. The samples are discretised with 4000 piecewise constant finite elements, distributed regularly over $D = [0, 1]$.

The Cameron–Martin Theorem tells us how to construct one Gaussian measure using another. However, we do not learn how to commence the process with the first Gaussian measure. Here, one option it to employ the *Karhunen–Loève (KL) expansion*.

Definition 1.24. Let $\dim Z = \infty$ and let $(\alpha_i, \psi_i)_{i=1}^{\infty}$ denote the eigenpairs of $\mathcal{C}$, where $(\psi_i)_{i=1}^{\infty}$ forms an orthonormal basis of Z. Let $\boldsymbol{\xi} : \Omega \to \mathbb{R}^{\mathbb{N}}$ be a measurable function. Furthermore, let the components of $\boldsymbol{\xi}$ form a sequence $(\boldsymbol{\xi}_i)_{i=1}^{\infty}$ of i.i.d. random variables, where $\boldsymbol{\xi}_1 \sim \mathrm{N}(0, 1)$. Then, the expansion

$$\boldsymbol{\theta}_{\mathrm{KL}} := m + \sum_{i=1}^{\infty} \sqrt{\alpha_i} \boldsymbol{\xi}_i \psi_i$$

is called KL expansion. ◇

One can easily verify the following proposition.

Proposition 1.25. $\boldsymbol{\theta}_{\mathrm{KL}}$ is distributed according to $\mathrm{N}(m, \mathcal{C})$.

Proof. Apply Theorem 1.21 to $\boldsymbol{\theta}_{\mathrm{KL}}$. □

In the remainder of this work we assume that eigenpairs are ordered descendantly with respect to the absolute value of the associated eigenvalues.
For illustration purposes we give a example for a Gaussian measure on an infinite-dimensional, separable Hilbert space.

Example 1.26. Let $Z := \mathbf{L}^2(\overline{D}; \mathbb{R})$. We define the *exponential covariance operator*,

$$\begin{aligned} \mathcal{C}_{\mathrm{exp}}^{(\lambda,\sigma)} &: Z \to Z, \ \varphi \mapsto \int_{\overline{D}} c_{\mathrm{exp}}^{(\lambda,\sigma)}(x, \cdot)\varphi(x)\mathrm{d}x, \\ c_{\mathrm{exp}}^{(\lambda,\sigma)} &: \overline{D} \times \overline{D} \to \mathbb{R}, \ (x, x') \mapsto \sigma^2 \exp\left(-\lambda^{-1}\|x - x'\|_D\right), \end{aligned} \tag{1.12}$$

where $\lambda > 0$ is called *correlation length* and $\sigma > 0$ is called *standard deviation*. Then, $\mathcal{C}_{\text{exp}}^{(\lambda,\sigma)} \in \text{CO}(Z)$ and $\text{N}(0, \mathcal{C}_{\text{exp}}^{(\lambda,\sigma)})$ is a well-defined Gaussian measure. We show samples of $\text{N}(0, \mathcal{C}_{\text{exp}}^{(\lambda,\sigma)})$ in 1D and 2D in Figures 1.1–1.2. ◇

Functions like $c_{\text{exp}}^{(\lambda,\sigma)}$ in Example 1.26 are called *covariance functions* or *covariance kernels*. In the following, we discuss spatial properties of random fields, such as continuity of realisations. These properties can often be derived from properties of their covariance kernels.

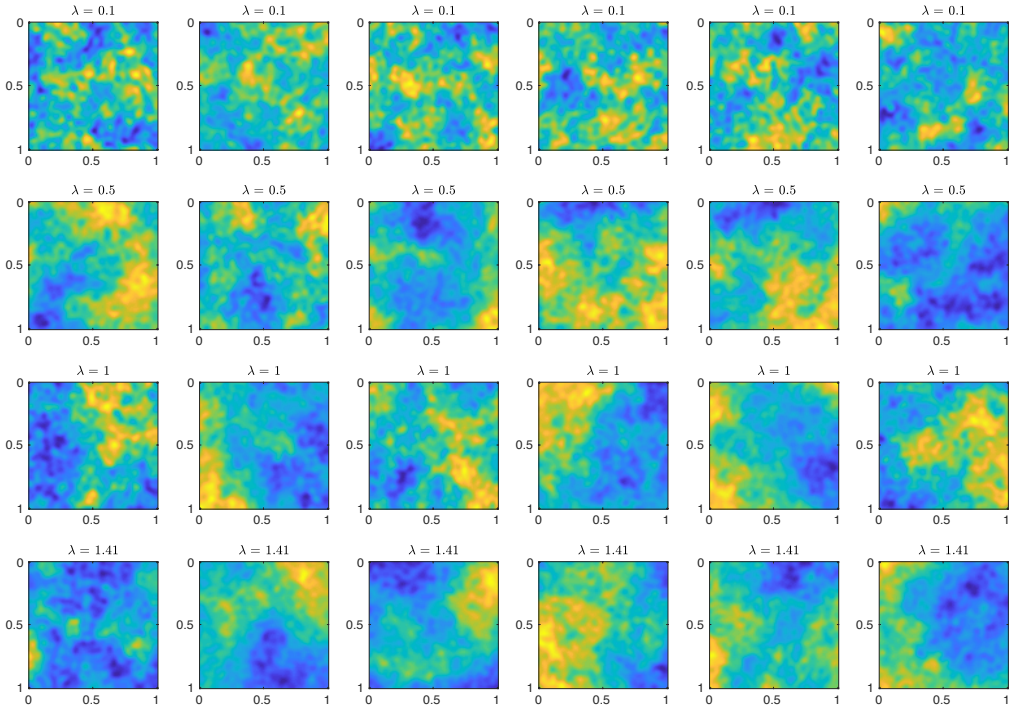


Figure 1.2. Samples of mean-zero Gaussian random fields in 2D with exponential covariance, with $\sigma := 1$ and $\lambda \in \{0.1, 0.5, 1.0, 1.41\}$, where $\|\cdot\|_{\overline{D}} := \|\cdot\|_2$ is the 2-norm. The samples were generated with a Karhunen–Loève expansion, which was truncated after $N_{\text{sto}} = 800$ terms; retaining a minimum of 90.1% of the random field's variances. The expansions were computed using 100×100 square piecewise constant finite elements on a regular grid.

Covariance kernels and continuity. We come back to the diffusion equation. Here, the parameter is a continuous function $\theta : \overline{D} \to (0, \infty)$. We aim to model this parameter with a random field. Hence, we need to obtain continuous realisations from the random field.

Indeed, when looking at Figures 1.1–1.2, we see that at least the discretised samples following a Gaussian measure with exponential covariance operators are continuous functions.

We now investigate this observation more rigorously and discuss the regularity of random fields. This plays a central role in the books by Adler [3, 4]. Recent results were also discussed by Potthoff [203].

First we define covariance kernels, show that they imply valid covariance operators, and then discuss the continuity of random fields.

Definition 1.27. Let $c : \overline{D} \times \overline{D} \to \mathbb{R}$ be a function satisfying

(i) c is symmetric, i.e. for all $x, y \in \overline{D}$, it holds $c(x, y) = c(y, x)$,

(ii) c is positive semidefinite, i.e. for all $n \in \mathbb{N}, x_1, \ldots, x_n \in \overline{D}$, and $a_1, \ldots, a_n \in \mathbb{R}$, it holds $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j c(x_i, x_j) \geq 0$,

(iii) c is continuous.

Then, c is a *covariance kernel*. ◇

Note that the assumptions in Definition 1.27 also imply that c is a Mercer kernel, see [238, Definition 11.2]. There are various ways to construct covariance kernels, we refer to Duvenaud [79, §2] for a review of those. Moreover, note that a huge class of covariance kernels is the (Whittle–)Matérn class, see Matérn [178]. We will discuss those in §4.2.3.
A covariance kernel can indeed be used to define a covariance operator.

Proposition 1.28. Let c be a covariance kernel and $\overline{D}$ be compact. Moreover, let $Z := \mathbf{L}^2(D)$. Then, $\mathcal{C} : Z \to Z$, defined by

$$Z \ni \varphi \mapsto \mathcal{C}\varphi := \int_{\overline{D}} \varphi(x) c(x, \cdot) \mathrm{d}x \in Z$$

defines a covariance operator on Z.

Proof. Since $\overline{D}$ is compact, and c is a Mercer kernel, we can apply Mercer's Theorem [238, Theorem 11.3]. □

Some properties of a random field can be investigated by studying the covariance kernel. Let $c : D \times D \to [0, \infty)$ be a covariance kernel. If there is a function $R : \{x - y : x, y \in D\} \to [0, \infty)$, such that $c(x, y) = R(x - y)$ $(x, y \in D)$, the covariance kernel is called *stationary*. Stationarity implies that a translation of the random field does not change its probability distribution. If there is a function $R : [0, \infty) \to [0, \infty)$, such that $c(x, y) = R(\|x - y\|)$ $(x, y \in D)$, the covariance kernel is called *isotropic*. Isotropy implies stationarity.
So far, we have defined a Gaussian random field as a random variable taking values in some separable Hilbert space Z of functions from $\overline{D}$ to $\mathbb{R}$. Now, we discuss continuity of samples of a measure $\mathrm{N}(0, \mathcal{C})$, i.e., we aim to show

$$\mathbb{P}\left(\lim_{y \to x} |\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y)| = 0, \quad x \in \overline{D} \right) = 1.$$

This implies that a sample from $\mathrm{N}(0, \mathcal{C})$ is continuous with probability one, or *almost surely continuous*.
We often cannot discuss continuity of $\theta \in Z$, since point-evaluations of θ may be not defined in Z. Alternatively, we need to define a Gaussian random field on a function space, in which functions have point evaluations. Note that a Gaussian measure can already be defined on a topological space Z, which is locally convex. We equip this space with the *cylindrical σ-algebra* $\mathcal{F}$ defined by

$$\sigma\left\{ \{\theta \in Z : ((T_1, \theta)_{Z^*,Z}, \ldots, (T_n, \theta)_{Z^*,Z}) \in B\} : n \in \mathbb{N},\ B \in \mathcal{B}\mathbb{R}^n,\ T_1, \ldots, T_n \in Z^* \right\}$$

and define the Gaussian measure on $(Z, \mathcal{F})$, as in Definition 1.20; see Bogachev [27, Definition 2.2.1(ii)]. If point-evaluations in Z are defined, those are linear functionals. Hence, we can access the *finite-dimensional distributions*

$$\mathbb{P}((\boldsymbol{\theta}(x_1, \cdot), \dots, \boldsymbol{\theta}(x_k, \cdot)) \in \cdot),$$

where $\boldsymbol{\theta} : \Omega \to Z$ is a measurable map, $k \in \mathbb{N}$, and $(x_1, \dots, x_k) \in \overline{D}^k$. In this case, we have

$$\mathbb{P}((\boldsymbol{\theta}(x_1), \dots, \boldsymbol{\theta}(x_k)) \in \cdot) = \mathrm{N}\left(0, \begin{pmatrix} c(x_1, x_1) & \cdots & c(x_1, x_k) \\ \vdots & \ddots & \vdots \\ c(x_k, x_1) & \cdots & c(x_k, x_k) \end{pmatrix}\right),$$

for some covariance kernel c. However, sets of the form

$$\left\{\theta \in Z : \lim_{y \to x} |\theta(x) - \theta(y)| = 0, \quad x \in \overline{D}\right\} \notin \mathcal{F}$$

are not measurable, since the described event contains an infinite number of point-evaluations. Sets in $\mathcal{F}$ however contain only finitely many point-evaluations at a time. We can discuss continuity of random fields $\boldsymbol{\theta}$ subject to *modifications*. A random field $\boldsymbol{\theta}'$ is a modification of $\boldsymbol{\theta}$, if

$$\mathbb{P}(\boldsymbol{\theta}(x) = \boldsymbol{\theta}'(x)) = 1 \qquad (x \in \overline{D}).$$

Note that from a probabilistic point of view, we cannot distinguish $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Finally, we move on to the continuity of a modification of a random field. We state a standard result on the continuity of random fields, linking continuity of a random field with continuity of the covariance kernel. Results of this kind are known as the *Kolmogorov(-Chentsov) continuity theorem*.

Theorem 1.29. Let D be a bounded, open set and let $Z := \mathbb{R}^{\overline{D}} := \{f : \overline{D} \to \mathbb{R}\}$ be the set of functions from $\overline{D}$ to $\mathbb{R}$. Furthermore, let $\boldsymbol{\theta} \sim \mathrm{N}(0, \mathcal{C})$ be a random field and $\mathcal{C}$ be given by a covariance kernel $c : \overline{D} \times \overline{D} \to \mathbb{R}$. If some $K, \varepsilon \in (0, \infty)$ exist, such that

$$\mathbb{E}\left[(\boldsymbol{\theta}(x) - \boldsymbol{\theta}(y))^2\right] \leq \frac{K}{|\log \|x - y\||^{1+\varepsilon}} \qquad (x, y \in \overline{D}),$$

then, there is a modification $\boldsymbol{\theta}'$ of $\boldsymbol{\theta}$, such that $\boldsymbol{\theta}'$ is almost surely continuous.

Proof. The space Z is locally convex, see the discussion in [27, §2.2]. See [3, Theorem 3.4.1] for a proof of this theorem. □

In the following, we will not distinguish between a random field and its modifications. A continuous random field will be a random field which has a modification that is almost surely continuous. Sometimes, we are slightly imprecise in the following, and discuss continuous Gaussian random fields on e.g. $(Z, \mathcal{B}Z) := (\mathbf{L}^2(\overline{D}), \mathcal{B}\mathbf{L}^2(\overline{D}))$. This means, we can define the random field on a subset or superset Z', such that there is a modification of the random field that is almost surely continuous. Indeed, we always assume in the following that a Gaussian measure is defined on a separable Hilbert space.

Gaussian white noise. In the last subsection, we have discussed regular random fields that have a continuous modification. However, sometimes rather irregular functions shall be modelled, e.g., by random fields without correlation structure. Gaussian white noise is an example for these. Let $Z := \mathbf{L}^2(D)$, and let $\mathrm{id}_Z : Z \to Z$ be the identity operator on Z. *Gaussian white noise* shall be understood as the random field corresponding to $\mathrm{N}(0, \mathrm{id}_Z)$. Equivalently, we can characterise $\boldsymbol{\theta} \sim \mathrm{N}(0, \mathrm{id}_Z)$, as the random field satisfying $\langle z, \boldsymbol{\theta} \rangle_Z \sim \mathrm{N}(0, \|z\|_Z^2)$, for any $z \in Z$.
However, a Gaussian random field with covariance operator id_Z does not exist on Z. Indeed, id_Z is not nuclear in Z: Let $(\varphi_n)_{n\in\mathbb{N}} \in Z^{\mathbb{N}}$ be an orthonormal basis of Z. Then, since all eigenvalues of id_Z are equal to 1, we have

$$\mathrm{trace}(\mathrm{id}_Z) = \sum_{n\in\mathbb{N}} \langle (\mathrm{id}_Z^* \mathrm{id}_Z)^{1/2} \varphi_n, \varphi_n \rangle = \sum_{n\in\mathbb{N}} \|\varphi_n\|_Z^2 = \sum_{n\in\mathbb{N}} 1 = \infty;$$

see also [27, Corollary 2.3.2]. Therefore, according to Theorem 1.21, $\mathrm{N}(0, \mathrm{id}_Z)$ is not defined. Nonetheless, it is possible to define a *generalised Gaussian random field* resembling Gaussian white noise. We will first introduce the principal idea in a functional theoretical setting, and then give an example on how the spaces may be chosen appropriately.
The principal idea consists in defining a Banach space $W \supseteq Z$ such that Z can be compactly embedded into W and such that $\mathrm{N}(0, \mathrm{Id}_Z)$ is well-defined on $(W, \mathcal{F})$. $\mathcal{F}$ is now the cylindrical σ-algebra on W. The tuple (Z, W) is called *abstract Wiener space*. Moreover, let W^* be the dual space of W with respect to the inner product of Z. That means, $W^* := \{T \in Z : \langle T, w \rangle_Z \in \mathbb{R}, w \in W\}$. Since Z is a Hilbert space, it is self-dual, i.e., $Z = Z^*$, and $W^* \subseteq Z$ is dense. The inclusion relation $W^* \subseteq Z \subseteq W$ is called *Gelfand triple*.
Gaussian white noise on (Z, W) is then defined as the probability measure μ on W with the characteristic function

$$\int_W \exp(\mathrm{i}\langle T, w \rangle_Z) \mathrm{d}\mu(w) = \exp\left(-\frac{1}{2}\|T\|_Z^2\right) \qquad (T \in W^*).$$

Hence, we define Gaussian white noise on a larger set than Z, but test it with functions in a subset of Z. Coming back to the desired property of Gaussian white noise: if $\boldsymbol{\theta} \sim \mu$, the random variable $\langle T, \boldsymbol{\theta} \rangle_Z \sim \mathrm{N}(0, \|T\|_Z^2)$. This is now true for any $T \in W^*$. As W^* is dense in Z, we can extend this for $z \in Z$: Let $(T_n)_{n\in\mathbb{N}} \in (W^*)^{\mathbb{N}}$, such that $\lim_{n\to\infty} T_n = z$. Then, $\mathbb{P}(\langle T_n, \boldsymbol{\theta} \rangle_Z \in \cdot) \xrightarrow{\mathrm{w}} \mathrm{N}(0, \|z\|_Z^2)$, as $n \to \infty$.
Finally, we give the standard example for the choice of W and W^*. Let D be the whole space $\mathbb{R}$ and $Z := \mathbf{L}^2(D)$. Let $W^* := \mathcal{S}(\mathbb{R})$ be the space of *rapidly decreasing functions*, i.e.,

$$\mathcal{S}(\mathbb{R}) := \left\{ f \in \mathbf{C}^\infty(\mathbb{R}; \mathbb{C}) : \forall i, j \in \mathbb{N}_0 : \lim_{|x|\to\infty} \left| x^i \frac{\mathrm{d}^j f}{\mathrm{d}x^j}(x) \right| = 0 \right\}.$$

One can show that indeed $W^* \subseteq Z$ and that W^* is dense in Z. We obtain the space W as the dual of W^* with respect to $\langle \cdot, \cdot \rangle_Z$. Here, $W := \mathcal{S}'(\mathbb{R})$ is the space of *tempered distributions*. Intuitively, we sample white noise in the space of tempered distributions. Those are generalised functions and rougher than $\mathbf{L}^2$-functions. We test the white noise with rapidly decreasing functions. Those are much smoother than standard $\mathbf{L}^2$-functions.

For this definition and construction of Gaussian white noise, we refer to the book by Kuo [151] and also to Kahle et al. [134]. Note that Kuo also explains others approaches to introduce white noise; e.g. as a weak derivative of the Brownian motion.

1.2.4 Markov kernels and conditional probability

In this section, we consider Markov kernels and conditional probabilities. These structures allow to represent hierarchies in random experiments. Such hierarchies may for instance be of temporal nature: Two random experiments are performed one after another, and the second experiment depends on the outcome of the first experiment. Markov kernels are parameterised measures. While the first random experiment is represented by a probability measure, the second experiment is given by a Markov kernel, where the parameter represents the outcome of the first experiment.
Conditional probability measures are Markov kernels that describe the probability distribution of the outcome of an experiment if the outcome of another experiment is known.
Markov kernels and conditional probability are the basis for Bayesian inference and parametric statistical models. Additionally, we use Markov kernels to construct probabilistic deep models.
For the here-presented results, we refer to Ash [8] and Dobrushin [70].

Markov kernels. We start with the definition of a Markov kernel.

Definition 1.30. Let $(Z_1, \mathcal{Z}_1)$ and $(Z_2, \mathcal{Z}_2)$ be two measurable spaces. A map $M : Z_1 \times \mathcal{Z}_2 \to [0,1]$ a *Markov kernel (mapping) from* $(Z_1, \mathcal{Z}_1)$ *to* $(Z_2, \mathcal{Z}_2)$, if it is a measurable function in the first component and a probability measure in the second component, i.e.

$$\text{(i) } M(\xi_1, \cdot) \in \mathrm{Prob}(Z_2) \quad (\xi_1 \in Z_1), \qquad \text{(ii) } M(\cdot, B) \in \mathbf{M}(Z_2; [0,1]) \quad (B \in \mathcal{Z}_2),$$

where $[0,1]$ is associated with its Borel-σ-algebra $\mathcal{B}[0,1]$. Let $\mu \in \mathrm{Prob}(Z_1)$ be a probability measure. The *composition* of μ and M is a probability measure on Z_2 and is defined by $\mu M(B) := \int_{Z_1} M(\xi_1, B) \mathrm{d}\mu(\xi_1) \quad (B \in \mathcal{Z}_2)$. $\diamond$

The measure μM describes the following process: If $\boldsymbol{\xi}_1 \sim \mu$ and $\boldsymbol{\xi}_2 \sim M(\boldsymbol{\xi}_1, \cdot)$, then $\boldsymbol{\xi}_2 \sim \mu M$. The state of $\boldsymbol{\xi}_1$ is *hidden* in the composition μM. We can also consider the joint measure of $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$, in which $\boldsymbol{\xi}_1$ is not hidden. Here, we define the *semidirect product* of μ, M, which is a probability measure on $(Z_1 \times Z_2, \mathcal{Z}_1 \otimes \mathcal{Z}_2)$. We denote it by $\mu \odot M$. It is defined to be the measure satisfying

$$(\mu \odot M)(B_1 \times B_2) = \int_{B_1} M(\xi_1, B_2) \mathrm{d}\mu(\xi_1) \quad (B_1 \in \mathcal{Z}_1, B_2 \in \mathcal{Z}_2).$$

Finally, we note that we use the convention to denote Markov kernels by $M(\cdot\,|*) := M(*, \cdot)$.

Conditional probability. We move on to conditional probability. First, we introduce the *elementary definition* of conditional probabilities. Let $A, B \in \mathcal{A}$ be two events, where $\mathbb{P}(B) > 0$. Then, the *conditional probability* of A given B is defined by

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.13}$$

The conditional probability describes the probability of A, if it is already known, that the event B occurs. This elementary definition is rather restrictive since we typically cannot expect that $\mathbb{P}(B) > 0$. This is for instance the case, if B is an event of the form

$$B = \{\boldsymbol{\xi} = \xi\},$$

where $\boldsymbol{\xi} : \Omega \to Z$ is a random variable with continuous distribution, and $\xi \in Z$ is a realisation of $\boldsymbol{\xi}$. The concept of conditional probability can be generalised to allow conditioning with respect to such events. This is based on the following theorem.

Theorem 1.31 ([8, Theorem 5.3.1])**.** Let $A \in \mathcal{A}$, and $\boldsymbol{\xi} : \Omega \to Z$ be a random variable on a generic measurable space $(Z, \mathcal{Z})$. Then, there is a measurable function $f : Z \to [0, 1]$, such that

$$\mathbb{P}(A \cap \{\boldsymbol{\xi} \in C\}) = \int_{\{\boldsymbol{\xi} \in C\}} (f \circ \boldsymbol{\xi})(\omega)\mathbb{P}(\mathrm{d}\omega) \qquad (C \in \mathcal{Z}).$$

The function f is $\mathbb{P}(\boldsymbol{\xi} \in \cdot)$-a.s. unique. ◇

The function f evaluated in $\xi \in Z$ in Theorem 1.31 represents the *conditional probability* of the event A, given that $B = \{\boldsymbol{\xi} = \xi\}$. We define $f(\xi) =: \mathbb{P}(A|\boldsymbol{\xi} = \xi)$, for $\xi \in Z$, $\mathbb{P}(\boldsymbol{\xi} \in \cdot)$-almost surely.

Remark 1.32. This definition of a conditional probability reminds us of the weak formulation of a PDE, as in Examples 1.2 and 1.4. However, the conditional probability f is not tested with test functions from a Hilbert space H', but with indicator functions of type $\mathbf{1}_{\{\boldsymbol{\xi} \in C\}} : \Omega \to [0, 1]$, for any $C \in \mathcal{Z}$. ◇

Note that this definition of a conditional probability is consistent with the definition given in Equation (1.13). Hence, when considering a random variable $\boldsymbol{\xi}$, such that $\mathbb{P}(\boldsymbol{\xi} = \xi) > 0$ for some $\xi \in Z$, we have

$$f(\xi) = \frac{\mathbb{P}(A \cap \{\boldsymbol{\xi} = \xi\})}{\mathbb{P}(\boldsymbol{\xi} = \xi)}.$$

We have given a formulation of the conditional probability of a certain event $A \in \mathcal{A}$, given that $\boldsymbol{\xi} = \xi$. This formulation is similar to a weak formulation. The conditional probability is a function of $\xi \in X$. For fixed A this function is measurable. We now rephrase a theorem giving conditions under which conditional probabilities indeed represent a probability measure with respect to A for fixed $\xi \in Z$. This makes the conditional probability a Markov kernel from $(Z, \mathcal{Z})$ to $(\Omega, \mathcal{A})$. This Markov kernel is called *regular conditional probability*.

Theorem 1.33 ([158, Theorem 3.1])**.** Let Ω be a Radon space, i.e. a separable, complete metric space, and $\boldsymbol{\xi} : \Omega \to Z$ be a random variable. Then, there is a Markov kernel $M : Z \times \mathcal{A} \to [0,1]$, such that

$$\mathbb{P}(A \cap \{\boldsymbol{\xi} \in C\}) = \int_{\{\boldsymbol{\xi} \in C\}} M(\boldsymbol{\xi}(\omega), A)\mathbb{P}(\mathrm{d}\omega) \qquad (C \in \mathcal{B}Z, A \in \mathcal{A}).$$

The Markov kernel is $\mathbb{P}(\boldsymbol{\xi} \in \cdot)$-a.s. unique. ◇

Throughout this work, we assume that the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is Radon. In this case, the assumptions of Theorem 1.33 are always satisfied. Henceforth, a regular conditional probability always exists.

Remark 1.34. We briefly argue, why we can assume that Ω is Radon. Throughout this work, we consider at most countably many random variables, taking values in separable Banach spaces. Let $I \subseteq \mathbb{N}$. We define the spaces by $(Z_i)_{i\in I}$, and the random variables $\boldsymbol{\xi}_i \in \mathbf{M}(\Omega; Z_i)$, for $i \in I$. Then, we can model $\Omega := \prod_{i\in I} Z_i$, and $\boldsymbol{\xi}_i : \Omega \to Z_i$ to be the canonical projection mapping $\Omega \ni (\xi_j)_{j\in I} \mapsto \xi_i \in Z_i$, for $i \in I$. Now, Ω is the product of at most countably many separable Banach spaces. Therefore, Ω is also a separable Banach space; hence a Radon space; e.g. take Ω as the ℓ^p direct sum of $(Z_i)_{i\in I}$. ◇

Construction of conditional probabilities. The definitions of (regular) conditional probabilities arise from Theorem 1.31 and Theorem 1.33. These definitions are not constructive. In this section, we discuss how to model conditional probabilities in practical situations. Moreover, we show how to compute conditional probabilities using probability density functions.
In the last section, we showed that regular conditional probabilities are Markov kernels. On the other hand, when given a Markov kernel, we can always construct a set of random variables, such that the Markov kernel represents a conditional probability with respect to these random variables. This is independent of the distribution of the random variables.

Theorem 1.35. Let M be a Markov kernel from $(Z_1, \mathcal{Z}_1)$ to $(Z_2, \mathcal{Z}_2)$. Then, random variables $\boldsymbol{\xi}_1 : \Omega \to Z_1$ and $\boldsymbol{\xi}_2 : \Omega \to Z_2$ exists, such that $M(B|\xi_1) = \mathbb{P}(\boldsymbol{\xi}_2 \in B|\boldsymbol{\xi}_1 = \xi_1)$, for $B \in \mathcal{Z}_2, \xi_1 \in Z_1, \mathbb{P}(\boldsymbol{\xi}_1 \in \cdot)$-a.s.. ◇

Proof. Without loss of generality, we assume that $(\Omega, \mathcal{A}) := (Z_1 \times Z_2, \mathcal{Z}_1 \otimes \mathcal{Z}_2)$. Moreover, let $\mu' \in \mathrm{Prob}(Z_1)$ be some arbitrary probability measure. We define $\boldsymbol{\xi}_i$ to be the i-th canonical projection from Ω onto Z_i, $i = 1, 2$. Moreover, we define $\mathbb{P} \in \mathrm{Prob}(\Omega)$ to be the measure satisfying

$$\mathbb{P}(B_1 \times B_2) := \int_{B_1} M(B_2|\xi_1)\mu'(\mathrm{d}\xi_1) \qquad (B_1 \in \mathcal{Z}_1, \quad B_2 \in \mathcal{Z}_2).$$

Note that $\mathbb{P}$ is uniquely defined by Carathéodory's Extension Theorem [143, Theorem 1.41]. Then, $\boldsymbol{\xi}_1 \sim \mu'$ and $\boldsymbol{\xi}_2 \sim \mu' M$. Moreover, $\{B_1 \times B_2\} = \{\{\boldsymbol{\xi}_1 \in B_1\} \cap \{\boldsymbol{\xi}_2 \in B_2\}\}$. Therefore, the following equation is satisfied,

$$\begin{aligned}\mathbb{P}(\{\boldsymbol{\xi}_1 \in B_1\} \cap \{\boldsymbol{\xi}_2 \in B_2\}) &= \int_{B_1} M(B_2|\xi_1)\mu'(\mathrm{d}\xi_1) \\ &= \int_{\{\boldsymbol{\xi}_1 \in B_1\}} M(B_2|\boldsymbol{\xi}_1(\omega))\mathbb{P}(\mathrm{d}\omega) \qquad (B_1 \in \mathcal{Z}_1, \quad B_2 \in \mathcal{Z}_2).\end{aligned}$$

The last statement is a result about integration with respect to pushforward measures, see [143, Theorem 4.10]. According to Theorem 1.33,

$$M(B|\xi_1) = \mathbb{P}(\boldsymbol{\xi}_2 \in B|\boldsymbol{\xi}_1 = \xi_1) \qquad (B \in \mathcal{Z}_2, \quad \xi_1 \in Z_1, \mathbb{P}(\boldsymbol{\xi}_1 \in \cdot)\text{-a.s.}),$$

concluding this proof. □

Therefore, we can construct conditional probability measures by modelling Markov kernels. This also implies the notational convention $M(\cdot\,|*) := M(*, \cdot)$.
Having discussed how to model a conditional probability measure, we now discuss how to derive it from a set of random variables. In particular, let $\boldsymbol{\xi}_1 : \Omega \to Z_1$, $\boldsymbol{\xi}_2 : \Omega \to Z_2$ be random variables. Starting from the joint probability distributions of these random variables, we would like to compute $\mathbb{P}(\boldsymbol{\xi}_2 \in \cdot\,|\boldsymbol{\xi}_1 = \cdot)$. In §1.2.1, we have discussed the representation of probability measures with probability density functions. We now use the probability density function of $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ to compute the *conditional density* representing the conditional probability measure.

Theorem 1.36 (Conditional densities)**.** Let Z_1, Z_2 be separable Banach spaces, and $\boldsymbol{\xi}_1 : \Omega \to Z_1$, $\boldsymbol{\xi}_2 : \Omega \to Z_2$ be random variables on those. Moreover, let ν_1, ν_2 be σ-finite measures on $(Z_1, \mathcal{Z}_1)$, $(Z_2, \mathcal{Z}_2)$, respectively, with

$$\mathbb{P}((\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \cdot) \ll \nu_1 \otimes \nu_2 \quad \text{and} \quad f := \frac{\mathrm{d}\mathbb{P}((\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \cdot)}{\mathrm{d}\nu_1 \otimes \nu_2}, \quad \nu_1 \otimes \nu_2\text{-a.e..}$$

Then, $\mathbb{P}(\boldsymbol{\xi}_1 \in \cdot) \ll \nu_1$ and $\mathbb{P}(\boldsymbol{\xi}_2 \in \cdot) \ll \nu_2$, where

$$\frac{\mathrm{d}\mathbb{P}(\boldsymbol{\xi}_1 \in \cdot)}{\mathrm{d}\nu_1} = \int_{Z_2} f(\cdot, \xi_2)\nu_2(\mathrm{d}\xi_2) =: g_1 \quad (\mathrm{d}\nu_1\text{-a.e.}),$$

$$\frac{\mathrm{d}\mathbb{P}(\boldsymbol{\xi}_2 \in \cdot)}{\mathrm{d}\nu_2} = \int_{Z_1} f(\xi_1, \cdot)\nu_1(\mathrm{d}\xi_1) =: g_2 \quad (\mathrm{d}\nu_2\text{-a.e.}).$$

Moreover, we have

$$\mathbb{P}(\boldsymbol{\xi}_2 \in \cdot\,|\boldsymbol{\xi}_1 = \xi_1) \ll \nu_2 \quad \text{and} \quad \frac{\mathrm{d}\mathbb{P}(\boldsymbol{\xi}_2 \in \cdot\,|\boldsymbol{\xi}_1 = \xi_1)}{\mathrm{d}\nu_2} = \begin{cases} \frac{f(\xi_1, \cdot)}{g_1(\xi_1)}, & \text{if } g_1(\xi_1) \neq 0, \\ 0, & \text{otherwise} \end{cases} \tag{1.14}$$

$$(\xi_1 \in Z_1, \mathbb{P}(\boldsymbol{\xi}_1 \in \cdot)\text{-a.s.}; \nu_2\text{-a.e.}),$$

$$\mathbb{P}(\boldsymbol{\xi}_1 \in \cdot\,|\boldsymbol{\xi}_2 = \xi_2) \ll \nu_1 \quad \text{and} \quad \frac{\mathrm{d}\mathbb{P}(\boldsymbol{\xi}_1 \in \cdot\,|\boldsymbol{\xi}_2 = \xi_2)}{\mathrm{d}\nu_1} = \begin{cases} \frac{f(\cdot, \xi_2)}{g_2(\xi_2)}, & \text{if } g_2(\xi_2) \neq 0, \\ 0, & \text{otherwise} \end{cases}$$

$$(\xi_2 \in Z_2, \mathbb{P}(\boldsymbol{\xi}_2 \in \cdot)\text{-a.s.}; \nu_1\text{-a.e.}).$$

◇

Proof. The first part of the theorem follows from the fact that for $i = 1, 2$, and $j := 3 - i$ we have

$$\mathbb{P}(\boldsymbol{\xi}_i \in A) = \mathbb{P}(\boldsymbol{\xi}_i \in A, \boldsymbol{\xi}_j \in Z_j) = \int_A \int_{Z_j} f(\xi_i, \xi_j)\nu_j(\mathrm{d}\xi_j)\nu_i(\mathrm{d}\xi_i), \qquad (A \in \mathcal{Z}_i)$$

by Tonelli's theorem. As the probability on the left-hand side exists, the integrals on the right-hand side are well-defined.
To show that the conditional densities are correct, we need to test

$$M_i(\cdot\,|\xi_j) = \int_\bullet \mathbf{1}(g_j(\xi_j) \neq 0)\frac{f(\xi_i,\xi_j)}{g_j(\xi_j)}\nu_i(\mathrm{d}\xi_i)$$

in Theorem 1.33, which is well-defined, since Z_1, Z_2 are separable Banach spaces. Let $A \in \mathcal{Z}_i$, and $C \in \mathcal{Z}_j$. Then,

$$\begin{aligned}\int_{\{\boldsymbol{\xi}_j \in C\}} M_i(A|\boldsymbol{\xi}_j(\omega))\mathbb{P}(\mathrm{d}\omega) &= \int_C M_i(A|\xi_j)g_j(\xi_j)\nu_j(\mathrm{d}\xi_j)\\ &= \int_C\int_A \mathbf{1}(g_j(\xi_j)\neq 0)\frac{f(\xi_i,\xi_j)}{g_j(\xi_j)}\nu_i(\mathrm{d}\xi_i)g_j(\xi_j)\nu_j(\mathrm{d}\xi_j)\\ &= \int_C\int_A \mathbf{1}(g_j(\xi_j)\neq 0)f(\xi_i,\xi_j)\nu_i(\mathrm{d}\xi_i)\nu_j(\mathrm{d}\xi_j)\\ &= \mathbb{P}(\{\boldsymbol{\xi}_i \in A\}\cap\{\boldsymbol{\xi}_j \in C\}).\end{aligned}$$

The last equality is true, since $\mathbb{P}(g_j(\boldsymbol{\xi}_j)=0) = \int_{\{g_j(\boldsymbol{\xi}_j)=0\}} g_j(\xi_j)\nu_j(\mathrm{d}\xi_j) = 0$. □

The formulae (1.14) remind us of the formula for elementary conditional probabilities (1.13). Indeed, (1.13) is a special case of (1.14), where $\boldsymbol{\xi}_1 \equiv \mathbf{1}_A$ and $\boldsymbol{\xi}_2 \equiv \mathbf{1}_B$ are binary random variables on $(Z_1,\mathcal{Z}_1) := (Z_2,\mathcal{Z}_2) := (\{0,1\}, 2^{\{0,1\}})$. Here, the measures $\nu_1 := \nu_2 := \#$ are identical to the *counting measure* on $(\{0,1\}, 2^{\{0,1\}})$.

1.2.5 Uncertainties in mathematical models

In the following, we assume to be in a decision making process. We consider a mathematical model $G : X \to H$ and some parameter $\theta \in X$. This model shall be used for the decision making process. Indeed, we assume that the outcome of the decision is completely determined by $G(\theta)$.
Let now the parameter $\theta \in X$ be uncertain. Therefore, the model $G(\theta)$ is uncertain as well and the decisions making process is *under uncertainties*. As mentioned in §1.2.1, we represent the uncertainty in θ by a probability distribution, i.e. $\boldsymbol{\theta} \sim \mu$ is a random variable with a given distribution. In this case, $G(\boldsymbol{\theta})$ is an H-valued random variable. This H-valued random variable is now a *model under uncertainties*. Note that if we know $\mu(G \in \cdot) := \mathbb{P}(G(\boldsymbol{\theta}) \in \cdot)$, we also know the probability distribution over the outcome of the decision problem. In this case, the associated decision problem is a decision *under risk*. A decision under risk is favourable compared to a decision under uncertainty.
Given a mathematical model G and an uncertain parameter $\boldsymbol{\theta} \sim \mu$, the task of determining the push-forward measure $\mu(G \in \cdot)$ is called *forward uncertainty quantification*, *forward propagation of uncertainty*, or *forward problem*. With any of these terms, we also refer to computing moments of this push-forward measure.
We conclude this section by discussing the diffusion equation from Example 1.4 with uncertain parameter. This problem has become a popular academic example in forward uncertainty quantification. It has been considered by [45, 83, 96, 111, 112, 150, 170, 232, 238, 243, 244], to name a few.

Example 1.37 (Pathwise stochastic diffusion equation)**.** We consider the setting from Examples 1.3-1.4. Hence, we have a model G given by the weak formulation on $H := H' := \mathbf{H}_0^1(D)$:

$$\mathcal{E}(u^*, v; \theta) := \int_D a(\theta, x)\langle \nabla u^*(x), \nabla v(x)\rangle_D \mathrm{d}x - \int_D f(x)v(x)\mathrm{d}x = 0 \text{ for all } v \in H'.$$

The permeability is given by $a(\theta, \cdot) := \exp(\theta)$. We now replace θ by a random variable $\boldsymbol{\theta} \sim \mathrm{N}(0, \mathcal{C})$, which is a random field with continuous realisations on $\overline{D}$. Hence, for $\mathbb{P}$-a.e. $\omega \in \Omega$, the random field realisation $\boldsymbol{\theta}(\cdot, \omega)$ is continuous or can be replaced by a continuous modification. Thus, the function $G \circ \boldsymbol{\theta} : \Omega \to H$, $\omega \mapsto G(\theta(\omega)) = u^*$ is well-defined with probability one. Hence, we can choose a σ-algebra on H, such that $G \circ \boldsymbol{\theta}$ is well-defined as a random variable. This is the so-called *pathwise formulation of the stochastic diffusion equation.* ◇

1.3 Statistical inverse problems

In the previous section, we have discussed decisions under risk or uncertainty. In practice, such decisions are influenced by mathematical models and probabilistic assumptions, but also by data. In this section, we discuss how data can be used to construct and enhance mathematical models.
The task of constructing or enhancing a mathematical model with data is called *inverse problem*. We define inverse problems in §1.3.1. In practice, data is often noisy and sparse, leading to an ill-posedness of the associated inverse problem. This indicates a statistical consideration of the inverse problem, which we pursue in the following subsections. First, we introduce statistical models representing inverse problems in §1.3.2. We aim to solve the according inference problems with the Bayesian approach. We introduce this approach and discuss its Lipschitz well-posedness in §1.3.3.

1.3.1 Inverse problems

Let $G : X \to H$ be a mathematical model. We assume that G models the true physical behaviour of the system of interest for a certain $\theta^\dagger \in X$. We call $\theta^\dagger$ the *true (underlying) parameter*.
We now discuss observations from the physical system. Let Y be a separable Banach space, called *data space*. We obtain a data set $y^\dagger \in Y$ when seeing the observable part of $G(\theta^\dagger) \in H$. The observable part of the system is defined via an *observation operator* $\mathcal{O} : H \to Y$. Moreover, we define the parameter-to-observation map by $\mathcal{G} := \mathcal{O} \circ G$ and refer to it by *forward response operator*. Hence, the data is defined by

$$y^\dagger := \mathcal{G}(\theta^\dagger).$$

We typically assume that the data is polluted by *observational noise* $\eta^\dagger \in Y$. The data is then defined by

$$y^\dagger := \mathcal{G}(\theta^\dagger) + \eta^\dagger. \tag{1.15}$$

In Equation (1.15), we have implictly assumed that the observational noise is additive. This is not a strict requirement. Without loss of generality, we make this assumption throughout most of this thesis.

An *inverse problem* (IP) consists in the identification of the true parameter $\theta^\dagger$, once the data $y^\dagger$ has been observed. In particular, we consider the following problem.

$$\text{Find } \theta^\dagger \in X : y^\dagger := \mathcal{G}(\theta^\dagger) + \eta^\dagger. \tag{1.16}$$

Here, we defined an inverse problem as a noisy equation that shall be solved with respect to a parameter. This concept is similar to the basic definition of Stuart [236, §2] and Sullivan [238, §6.1]. If no observational noise has been added to the data, we call the inverse problem *noise-free*.
We now discuss the solvability of inverse problems; or more particularly, their *well-posedness*. Hadamard [110] established this term considering the solvability of PDEs. Today it also refers to the solvability of an inverse problem.

Definition 1.38. An inverse problem (1.16) is *well-posed*, if

(i) a solution $\theta^\dagger \in X$ exists,

(ii) the solution is unique,

(iii) the solution is stable, i.e. the map $Y \ni y^\dagger \mapsto \theta^\dagger \in X$ is continuous.

Otherwise, the inverse problem is *ill-posed*. ◇

In the noisy case, we can typically not hope for well-posedness. Indeed, since the observational noise is unknown, we effectively need to identify both $\theta^\dagger$ and $\eta^\dagger$.

Proposition 1.39. Let X contain at least two elements and let $\eta^\dagger \in Y$ be unknown. Then, the inverse problem (1.16) is ill-posed.

Proof. Let $\theta_1, \theta_2 \in X$ be not identical. Let $i = 1, 2$. We set $\eta_i := y^\dagger - \mathcal{G}(\theta_i)$. Then, $\mathcal{G}(\theta_i) + \eta_i = y^\dagger$. Hence, both (θ_1, η_1) and (θ_2, η_2) are a solution to the inverse problem (1.16). Since $\theta_1 \neq \theta_2$, the solution is not unique. □

Remark 1.40 (Least squares)**.** To overcome the ill-posedness, inverse problems are classically defined and/or solved via an optimisation problem minimising the distance between data and model. If this distance is derived from an $\mathbf{L}^2$-norm this is the so-called *least squares approach*. Such approaches are for instance discussed in Chavent [36] for nonlinear models and in Groetsch [107] for linear models. As an original example for an inverse problem solved via the least squares approach, we mention Gauss' calculation of the orbit of the dwarf planet *Ceres*, see Teets and Whitehead [245] for an overview. We revisit the least squares approach briefly in §1.3.3, but do not consider it throughout the rest of this work. ◇

In Proposition 1.39, we discussed a noisy inverse problem. A noise-free inverse problem is for instance well-posed, if the operator $\mathcal{G}$ is *homeomorphic*. In this case, the operator $\mathcal{G}$ is continuous, bijective, and its inverse $\mathcal{G}^{-1}$ is also continuous. Since the function $\mathcal{G}$ is bijective, we can find $\theta^\dagger = \mathcal{G}^{-1}(y^\dagger)$. The continuity of $\mathcal{G}^{-1}$ gives us stability. This is a rather strong assumption. In fact, we typically face the problem that the data space is of a much lower dimension than the parameter space. Hence, if e.g. $\mathcal{G}$ is affine linear, there is no inverse of $\mathcal{G}$, if the parameter space is higher dimensional than the data space.
We have an infinite-dimensional parameter space and a finite-dimensional data space in the elliptic inverse problem that we discuss in the following. Note however that $\mathcal{G}$ is non-linear in this case.

Example 1.41 (Elliptic inverse problem)**.** We consider the setting from Examples 1.3-1.4. Hence, we have a model G given by the weak formulation on $H := H' := \mathbf{H}_0^1(D)$:

$$\mathcal{E}(u^*, v; \theta) := \int_D a(\theta, x)\langle \nabla u^*(x), \nabla v(x)\rangle_D \mathrm{d}x - \int_D f(x)v(x)\mathrm{d}x = 0 \text{ for all } v \in H'.$$

The model G maps $X \ni \theta \mapsto u^*$ which is the unique solution to the weak equation above. The diffusion coefficient is given by $a(\theta, \cdot) := \exp(\theta)$, where $\theta \in \mathbf{C}^0(D)$ is a continuous function. We now observe $u^\dagger = G(\theta^\dagger)$ that is based on $\theta^\dagger \in X$. It is observed in $N_{\mathrm{obs}} \in \mathbb{N}$ locations $(o_i)_{i=1}^{N_{\mathrm{obs}}} \in D^{N_{\mathrm{obs}}}$. Hence, $Y := \mathbb{R}^{N_{\mathrm{obs}}}$. If point-evaluations are defined in H, we model the observation operator by $\mathcal{O} : H \to Y$, where

$$\mathcal{O}(u) := (u(o_i))_{i=1}^{N_{\mathrm{obs}}}. \tag{1.17}$$

If point-evaluations are not defined in H, we observe the average pressure in a small ball around the respective o_i, $i = 1, \ldots, N_{\mathrm{obs}}$. In particular, let $\varepsilon > 0$ and $B(o_i, \varepsilon) \subseteq D$ be the open ε-ball around o_i contained in D, for $i = 1, \ldots, N_{\mathrm{obs}}$. The observation operator $\mathcal{O} : H \to Y$ is given by

$$\mathcal{O}(u) := \left(\frac{\int_{B(o_i,\varepsilon)} u(x)\mathrm{d}x}{\int_{B(o_i,\varepsilon)} \mathrm{d}x}\right)_{i=1}^{N_{\mathrm{obs}}} = \left(\frac{\Gamma(1+d/2)}{\pi^{d/2}\varepsilon^d}\int_{B(o_i,\varepsilon)} u(x)\mathrm{d}x\right)_{i=1}^{N_{\mathrm{obs}}}, \tag{1.18}$$

where the second "=" is implied by basic results about the volume of d-dimensional balls, see e.g. [171]. *Lebesgue's differentiation theorem* [219, Theorem 7.10] implies that the observation operator (1.18) approximates (1.17), for $\varepsilon > 0$ small. We typically do not distinguish the two observation operators in (1.17)-(1.18), but just refer by *point evaluations* in any situation to the appropriate one. ◇

In more physical terms, Example 1.41 models the inverse problem built on the following experimental setup: Given is a groundwater reservoir D with a given log-permeability $\theta^\dagger : D \to \mathbb{R}$. To identify $\theta^\dagger$, sensors are placed at positions $(o_i)_{i=1}^{N_{\mathrm{obs}}}$. The sensors are used to measure the pressure in the positions. The inverse problem consists in the identification of the permeability given the measurements.
Elliptic inverse problems have been discussed as early as e.g. 1981 by Richter [210]. As for the forward problem, the elliptic inverse problem has become a popular academic example. We refer to e.g. [20, 57, 71, 87, 177, 226, 236].

1.3.2 Statistical models

In §1.3.1, we defined an inverse problem as an equation that shall be solved with respect to the true parameter. This equation typically contains observational noise. We showed in Proposition 1.39 that this noise makes the inverse problem hard or impossible to solve. We now discuss noisy inverse problems from a statistical point of view. The noise is then modelled as a random variable, data is a realisation of a random variable, and the identification of the true parameter can be stated as a statistical inference problem. In the following, we consider statistical models and likelihoods. Then, we show how an inverse problems can be fit into a statistical model; and how the associated likelihoods are constructed.

Let Y be the data space and let $\mathcal{P} \subseteq \mathrm{Prob}(Y)$ be a set of probability measures on Y. The tuple $(Y, \mathcal{P})$ is called *statistical model*. The statistical model $(Y, \mathcal{P})$ is called *parametric*, if the statistical model can be represented by a parameterised measure $\mu_L : X \times \mathcal{B}Y \to [0, 1]$, where

$$\mathcal{P} := \{\mu_L(\cdot\,|\theta) : \theta \in X\}\,.$$

In more rigorous terms, the statistical model is the image of some Markov kernel μ_L that is defined from the parameter space $(X, \mathcal{B}X)$ to the data space $(Y, \mathcal{B}Y)$. X is still a measurable subset of a separable Banach space.
The *statistical inference* problem corresponding to the statistical model $(Y, \mathcal{P})$ is the following. Let $\mu^\dagger \in \mathcal{P}$ and $y^\dagger$ be a realisation of the random variable $\boldsymbol{y} \sim \mu^\dagger$. Here, $\mu^\dagger$ is called *data generating measure*, and $y^\dagger$ is called *data*. *Observing data* $y^\dagger$ can be interpreted as 'the event $\{\boldsymbol{y} = y^\dagger\}$ occurs'. The corresponding statistical inference problem consists in identifying $\mu^\dagger$ given the data $y^\dagger$. If the statistical model is parametric, we typically represent the data-generating measure $\mu^\dagger$ by the parameter $\theta^\dagger \in X$. Here, $\mu^\dagger = \mu_L(\cdot\,|\theta^\dagger)$. In this case, the statistical inference problem consists in identifying $\theta^\dagger \in X$.
Rather than defining $\mathcal{P}$ or μ_L, it is simpler and sometimes advantageous to define a statistical model via a *(data) likelihood*. This is possible, if a parametric statistical model is – what we call – *dominated*. This holds, if for any $\theta \in X$: $\mu_L(\cdot\,|\theta) \ll \nu_Y$, where ν_Y is a σ-finite measure on $(Y, \mathcal{B}Y)$. Hence, μ_L is uniformly dominated by ν_Y. The *likelihood* L is defined by the probability density function

$$L(y|\theta) = \frac{\mathrm{d}\mu_L(y|\theta)}{\mathrm{d}\nu_Y} \qquad (\theta \in X; y \in Y, \nu_Y\text{-a.e.}).$$

In §1.2.4, we have proven that Markov kernels can always be interpreted as conditional probability measures. Therefore, we can interpret $\mu_L(\cdot\,|\theta)$ as conditional probability measure of y given that the parameter equals $\theta \in X$. In this set-up, the likelihood L is a conditional density.
For more information on statistical models, we refer to McCullagh [181], or alternatively to various textbooks on statistical inference, e.g. van der Vaart [250]. We note that as opposed to our definition, statistical models with parameters in function spaces are often considered *non-parametric*, see e.g. [55, 97]. There, parametric statistical models have a finite or finite-dimensional parameter space.

Statistical modelling of inverse problems. We consider the noisy inverse problem (1.15). Now, the observational noise $\eta^\dagger$ is modelled as the realisation of a random variable $\boldsymbol{\eta} : \Omega \to Y$. The probability distribution of $\boldsymbol{\eta}$ is defined by μ_{noise}. The data $y^\dagger$ is a realisation of $\boldsymbol{y} := \mathcal{G}(\theta^\dagger) + \boldsymbol{\eta}$, where $\theta^\dagger$ is the true parameter that shall be identified. The inverse problem is represented by the parametric statistical model

$$(Y, \mathcal{P}) := (Y, \{\mu_{\text{noise}}(\cdot - \mathcal{G}(\theta)) : \theta \in X\})\,.$$

Moreover, the data generating measure is defined by $\mu^\dagger := \mu_{\text{noise}}(\cdot - \mathcal{G}(\theta^\dagger))$.

Gaussian noise. In practice, observational noise is often modelled by a Gaussian random variable. In the next paragraphs, we show in which cases the according parametric statistical model is dominated and define the likelihood function. We consider three different cases: finite dimensional data space, infinite dimensional data space, and infinite dimensional white noise.

Finite dimensions, non-degenerate. Let the data space $Y := \mathbb{R}^{N_{\text{obs}}}$ be finite-dimensional. Moreover, let $\Gamma \in \mathbb{R}^{N_{\text{obs}} \times N_{\text{obs}}}$ be positive definite and $\mu_{\text{noise}} = \mathrm{N}(0, \Gamma)$. The statistical model $(Y, \mathcal{P})$ is then parametric given by $\mu_L(\cdot\,|\theta) := \mathrm{N}(\mathcal{G}(\theta), \Gamma)$. This Markov kernel is uniformly dominated by the Lebesgue measure $\mathrm{Leb}(N_{\text{obs}})$ on $\mathbb{R}^{N_{\text{obs}}}$:

$$\mu_L(\cdot\,|\theta) \ll \mathrm{Leb}(N_{\text{obs}}) \qquad (\theta \in X).$$

Hence, $(Y, \mathcal{P})$ is parametric and dominated and we can define the likelihood function by

$$L(y^\dagger|\theta) := \mathrm{n}(y^\dagger|\mathcal{G}(\theta), \Gamma) := \det(2\pi\Gamma)^{1/2} \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(y^\dagger - \mathcal{G}(\theta)\|_Y^2\right),$$

where $y^\dagger \in Y$, $\mathrm{Leb}(N_{\text{obs}})$-a.e., and $\theta \in X$.

Infinite dimensions. Let now Y be an infinite-dimensional separable Hilbert space, and let $\Gamma \in \mathrm{CO}(Y)$ be a valid covariance operator on Y. We have again a parametric statistical model $(Y, \mathcal{P})$, where $\mu_L(\cdot\,|\theta) := \mathrm{N}(\mathcal{G}(\theta), \Gamma)$, $\theta \in X$. In the finite-dimensional case, this Markov kernel was uniformly dominated by a Lebesgue measure; and $(Y, \mathcal{P})$ therefore parametric and dominated. Since there is no Lebesgue measure in infinite dimensions, we cannot show dominatedness in this way. Nonetheless, we can sometimes show that the Markov kernel is dominated by another measure. As suggested by Stuart [236, Remark 3.8], this may be some other Gaussian measure.

Corollary 1.42. Assume there is some $m \in Y$, such that $\mathcal{G}(\theta) - m$ is in the Cameron–Martin space $\mathrm{CM}(\mathrm{N}(m, \Gamma))$, for $\theta \in X$. Then, $\mu_L(\cdot\,|\theta) \ll \mathrm{N}(m, \Gamma)$ and

$$\frac{\mathrm{d}\mu_L(\cdot\,|\theta)}{\mathrm{dN}(m, \Gamma)}(y^\dagger) = \exp\left(\langle \mathcal{G}(\theta) - m, y^\dagger\rangle_{\mathrm{CM}(\mathrm{N}(m,\Gamma))} - \frac{1}{2}\|\mathcal{G}(\theta) - m\|^2_{\mathrm{CM}(\mathrm{N}(m,\Gamma))}\right),$$

for $\theta \in X$, and $y^\dagger \in Y$, $\mathrm{N}(m, \Gamma)$-a.s.

Proof. This is a result of Theorem 1.23 (Cameron–Martin), where $\mu = \mathrm{N}(m, \mathcal{C})$ and $h := \mathcal{G}(\theta) - m$. □

Hence, if the assumptions of Corollary 1.42 are satisfied, we can construct a likelihood as a conditional density with respect to the Gaussian measure $\mathrm{N}(m, \Gamma)$. The likelihood is then given by $L(\theta|\cdot) := \frac{\mathrm{d}\mu_L(\cdot\,|\theta)}{\mathrm{dN}(m,\Gamma)}$, for $\theta \in X$.

Infinite-dimensional white noise. Another modelling choice for the noise on an infinite dimensional data space is Gaussian white noise. This has been discussed e.g. by Kahle et al. [134], who also derived a likelihood for this modelling choice. We review their result in the following. Additionally, we give the underlying statistical model.

Let $Y := \mathbf{L}^2(D)$ be again an infinite-dimensional separable Hilbert space. We remind the reader that Gaussian white noise is a generalised random field on Y, corresponding to the Gaussian measure with covariance operator Id_Y. Let $W \supseteq Y$ be a Banach space such that the tuple (Y, W) forms an abstract Wiener space, and let $\mu = \mathrm{N}(0, \mathrm{Id}_Y)$ be Gaussian white noise on (Y, W); see our thorough discussion in

§1.2.3. Note that μ is now a measure on the Banach space W. Hence, we replace the data space Y by W and obtain a statistical model $(W, \mathcal{P})$. This statistical model is parametric and given by the Markov kernel $\mu_L(\cdot\,|\theta) := \mu(\cdot - \mathcal{G}(\theta))$, $\theta \in X$.
In this setting, we can show that the white noise model will always lead to a parametric and dominated statistical model: We proceed as in the infinite-dimensional Gaussian case. The Hilbert space Y can be understood as the Cameron–Martin space of μ. Since $\mathrm{img}(\mathcal{G}) \subseteq Y$, we can apply a Cameron–Martin-type formula for Gaussian white noise; see [151, p. 8]. Then, we obtain a probability density function with respect to μ:

$$\frac{\mathrm{d}\mu_L(\cdot\,|\theta)}{\mathrm{d}\mu}(w) = \exp\left(\langle \mathcal{G}(\theta), w\rangle_Y - \frac{1}{2}\|\mathcal{G}(\theta)\|_Y^2\right) \quad (\theta \in X, w \in W, \mu\text{-a.s.}). \tag{1.19}$$

Hence, $(W, \mathcal{P})$ is parametric and dominated, and the likelihood L is given by

$$\frac{\mathrm{d}\mu_L(\cdot\,|\cdot)}{\mathrm{d}\mu}.$$

Remark 1.43. Let $\theta \in X$. Note that $\langle \mathcal{G}(\theta), w\rangle_Y$ is not well-defined, if $w \in W\backslash Y$ and $\mathcal{G}(\theta) \in Y\backslash W^*$. In this case, the density in (1.19) appears to be not well-defined. However, by the discussion in §1.2.3, the space W^* is dense in Y. Hence, we define in this case

$$\langle \mathcal{G}(\theta), w\rangle_Y := \lim_{n\to\infty} \langle y_n, w\rangle_Y,$$

where $(y_n)_{n=1}^\infty \in (W^*)^{\mathbb{N}}$ and $\lim_{n\to\infty} y_n = \mathcal{G}(\theta)$. ◇

1.3.3 Bayesian statistics

In the last section, we have discussed dominated statistical models extensively. Now, we introduce Bayesian inference as a method to solve the parameter identification problem associated to a dominated statistical model $(Y, \mathcal{P})$. In the following, we will be interested in the Bayesian inference of a statistical model that has been derived from a PDE-based inverse problem, e.g. the one given in Example 1.41.
For an introduction of Bayesian inference, we refer to [97, 179, 195, 211]. As a historic resource containing Bayes' Theorem we cite Laplace [152], and for a historical overview [54, 212]. Using Bayesian inference to solve inverse problems has been suggested in [135, 242]
To identify the true parameter $\theta^\dagger \in X$, the Bayesian approach proceeds as follows: First, we model the parameter as a random variable $\boldsymbol{\theta} : \Omega \to X$, which is distributed according to $\mu_{\mathrm{prior}} \in \mathrm{Prob}(X)$. μ_{prior} is the so-called *prior (measure)*. The random variable $\boldsymbol{\theta}$ reflects the knowledge and uncertainty in the parameter, as discussed in §1.2.5. Rather than computing a point estimator $\widehat{\theta} \in X$, we aim to use the data $y^\dagger$ to *update* our knowledge by the data. We mentioned already in §1.3.2 that observing data shall be understood as an event $\{\boldsymbol{y} = y^\dagger\} \in \mathcal{A}$. To express that this event has occurred, we condition the parameter with respect to this event and obtain the so-called *posterior (measure)*:

$$\mu_{\mathrm{post}}^\dagger := \mathbb{P}(\boldsymbol{\theta} \in \cdot\,|\boldsymbol{y} = y^\dagger). \tag{1.20}$$

The posterior measure reflects our knowledge and uncertainty concerning the parameter $\boldsymbol{\theta}$, given that we have observed the data $y^\dagger$.

In the following, we discuss *Bayes' Theorem*, which gives us a way to obtain the posterior measure. Indeed, Bayes' Theorem gives a connection of the likelihood L, prior μ_{prior}, and posterior $\mu^\dagger_{\text{post}}$, in terms of their probability density functions. We assume that $\mu_{\text{prior}} \ll \nu_X$, where ν_X is some σ-finite measure on $(X, \mathcal{B}X)$, and define

$$\pi_{\text{prior}} = \frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}\nu_X}.$$

Note that $\nu_X := \mu_{\text{prior}}$ and $\pi_{\text{prior}} \equiv 1$ is a feasible choice. In Bayes' Theorem, the posterior measure is given in terms of a probability density function with respect to the same measure ν_X. Below, we briefly discuss a measure-theoretic sublety we encounter with conditional probabilities and their densities. Then, we move on to our formulation of Bayes' Theorem.

Remark 1.44. As stated in Theorem 1.33, conditional probabilities like $\mu^\dagger_{\text{post}} = \mathbb{P}(\boldsymbol{\theta} \in \cdot\,|\boldsymbol{y} = y^\dagger)$ are only defined for $\mathbb{P}(\boldsymbol{y} \in \cdot)$-a.s. every $y^\dagger \in Y$. This implies that if $\mathbb{P}(\boldsymbol{y} \in \cdot)$ has a continuous distribution, point evaluations in Y of the function $\mathbb{P}(\boldsymbol{\theta} \in A|\boldsymbol{y} = \cdot)$ may not be well-defined, for $A \in \mathcal{B}X$. In this case, one would not be able to compute the posterior measure for any single data set in Y. Also, the statement in (1.20), should be understood only for $\mathbb{P}(\boldsymbol{y} \in \cdot)$-a.s. every $y^\dagger \in Y$. ◇

Theorem 1.45 (Bayes)**.** Let $y^\dagger \in Y$ be $\mathbb{P}(\boldsymbol{y} \in \cdot)$-almost surely defined. Moreover, let $L(y^\dagger|\cdot)$ be in $\mathbf{L}^1(X, \mu_{\text{prior}})$ and strictly positive. Then,

$$Z(y^\dagger) := \int_X L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta) \in (0, \infty).$$

Moreover, the posterior measure $\mu^\dagger_{\text{post}} \ll \nu_X$ exists, it is unique, and it has the ν_X-density

$$\pi^\dagger_{\text{post}}(\theta') = \frac{L(y^\dagger|\theta')\pi_{\text{prior}}(\theta')}{Z(y^\dagger)} \qquad (\theta' \in X, \nu_X\text{-a.e.}). \tag{1.21}$$

◇

Proof. The following statements hold $\mathbb{P}(\boldsymbol{y} \in \cdot)$-a.s. for $y^\dagger \in Y$.
We first show that $Z(y^\dagger) > 0$. Since we assume that $L(y^\dagger|\cdot)$ is μ_{prior}-a.s. strictly positive, we can write:

$$Z(y^\dagger) = \int_X L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta) = \int_{\{L(y^\dagger|\cdot)>0\}} L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta). \tag{1.22}$$

Now let $n \in \mathbb{N}$. As the integrand in (1.22) is positive, Chebyshev's inequality, [8, Theorem 2.4.9], implies that

$$n \cdot \int_{\{L(y^\dagger|\cdot)>0\}} L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta) \geq \mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}). \tag{1.23}$$

We aim to show that the probability on the right-hand side of this equation converges to 1 as $n \to \infty$. Knowing this, we can conclude that preasymptotically the right-hand side is strictly positive for all $n \geq N$, for some $N \in \mathbb{N}$.
Note that measures are continuous with respect to increasing sequences of sets. We define the set

$$B_n := \{L(y^\dagger|\cdot) > n^{-1}\}$$

and observe that $(B_n)_{n=1}^{\infty}$ is indeed an increasing sequence. Moreover, note that

$$B_\infty = \bigcup_{m=1}^{\infty} B_m = \{L(y^\dagger|\cdot) > 0\},$$

and that $\mu_{\text{prior}}(B_\infty) = 1$. Hence, we have

$$\lim_{n\to\infty} \mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) = \mu_{\text{prior}}(L(y^\dagger|\cdot) > 0) = 1.$$

As mentioned earlier, we now deduce that for some $\varepsilon \in (0,1)$, there is an index $N \in \mathbb{N}$ such that

$$|\mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) - 1| \leq \varepsilon < 1 \qquad (n \geq N)$$

and thus $\mu_{\text{prior}}(L(y^\dagger|\cdot) > n^{-1}) > 0$, for $n \geq N$. Plugged into Equation (1.23), this gives us $Z(y^\dagger) > 0$. We have also $Z(y^\dagger) < \infty$, since $L(y^\dagger|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$, Thus, the posterior density (1.21) is well-defined.
We now apply [58, Theorem 3.4] (or alternatively Theorem 1.36) and obtain

$$\frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu_{\text{prior}}}(\theta') = \frac{L(y^\dagger|\theta')}{Z(y^\dagger)} \qquad (\theta' \in X, \mu_{\text{prior}}\text{-a.s.}).$$

This implies

$$\pi^\dagger_{\text{post}}(\theta') = \frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\nu_X}(\theta') = \frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu_{\text{prior}}}(\theta')\frac{\mathrm{d}\mu_{\text{prior}}}{\mathrm{d}\nu_X}(\theta') = \frac{L(y^\dagger|\theta')\pi_{\text{prior}}(\theta')}{Z(y^\dagger)},$$

where $\theta' \in X, \nu_X$-a.e., by application of standard results concerning Radon–Nikodym derivatives. This concludes the proof. □

The quantity in the denominator of Bayes' formula

$$Z(y^\dagger) := \int_X L(y^\dagger|\theta)\mathrm{d}\mu_{\text{prior}}(\theta)$$

is the ν_Y-density of $\mathbb{P}(\boldsymbol{y} \in \cdot)$ and is called *(model) evidence*; see Theorem 1.36. We made various assumptions in Theorem 1.45. We comment on those in §2.1.2. Our version of Bayes' theorem is mainly built on [58, Theorem 3.4].
In Remark 1.44, we have mentioned that the posterior measure is only $\mathbb{P}(y \in \cdot)$-a.s. uniquely defined. Hence, the map $y^\dagger \mapsto \mu^\dagger_{\text{post}}$ is not well-defined. We resolve this issue by fixing the definition of the likelihood $L(y^\dagger|\theta')$ for every $y^\dagger \in Y$ and μ_{prior}-a.s. every $\theta' \in X$. According to Theorem 1.45, we then obtain indeed a unique posterior measure for any data set $y^\dagger \in Y$. We define the *Bayesian inverse problem* (BIP) with prior μ_{prior} and likelihood L by

$$\text{Find } \mu^\dagger_{\text{post}} \in \text{Prob}(X, \mu_{\text{prior}}) \text{ with } \nu_X\text{-density } \pi_{\text{post}}(\theta|y^\dagger) = \frac{L(y^\dagger|\theta)\pi_{\text{prior}}(\theta)}{Z(y^\dagger)}. \quad (1.24)$$

Throughout the thesis when refering to a BIP (1.24), we always assume to have a likelihood $L : X \times Y \to \mathbb{R}$ and a prior measure μ_{prior} satisfying the assumptions in Theorem 1.45. Hence, a BIP is always well-defined.

Lipschitz well-posedness. In Proposition 1.39, we have shown that noisy inverse problems are often ill-posed. Stuart showed in his work [236] that the Bayesian approach to a noisy inverse problem may be well-posed. He gave assumptions under which Bayesian inverse problems satisfy, what we call *Lipschitz well-posedness*. Similarly to the well-posedness definition of the classical problem (1.16), we consider an existence, a uniqueness and a stability condition. Stability is quantified in terms of the *Hellinger distance*. We now formalise the concept of Lipschitz well-posedness for Bayesian inverse problems.

Definition 1.46 (Lipschitz well-posedness)**.** The problem (1.24) is *Lipschitz well-posed*, if

(i) $\mu^\dagger_{\text{post}} \in \text{Prob}(X, \mu_{\text{prior}})$ exists *(existence)*,

(ii) $\mu^\dagger_{\text{post}}$ is unique in $\text{Prob}(X, \mu_{\text{prior}})$ *(uniqueness)*, and

(iii) $(Y, \|\cdot\|_Y) \ni y^\dagger \mapsto \mu^\dagger_{\text{post}} \in (\text{Prob}(X, \mu_{\text{prior}}), \text{d}_{\text{Hel}})$ is locally Lipschitz continuous *(stability)*.

◇

We finish this section by reviewing the well-posedness result of Stuart [236]. First, we give two sets of assumptions; on the prior and on the likelihood. Then, we formulate the result in [236].

Assumption 1.47 (Prior)**.** Let μ_{prior} fulfill the following assumptions

(i) μ_{prior} is *light-tailed*; i.e. some $\varepsilon > 0$ exists, such that $\int_X \exp(\varepsilon\|\theta\|_X^2)\mu_{\text{prior}}(\text{d}\theta) < \infty$,

(ii) centred, open R-balls $B(0,R) := \{\theta \in X : \|\theta\|_X < R\} \subseteq X$ have positive probability, i.e. $\mu_{\text{prior}}(\{\theta \in X : \|\theta\| < R\}) > 0$, for any radius $R > 0$.

◇

This set of assumptions is not explicitly stated in [236], but required when the prior is not Gaussian. This has been mentioned for the second of those assumptions in [134, Theorem 3.1].
By assumption in (1.24), the likelihood L is strictly positive. Therefore, we sometimes define L in terms of its *potential*, which is the *negative log-likelihood*:

$$\Phi(\theta; y) := -\log L(y|\theta) \qquad (y \in Y, \theta \in X).$$

Note that if Y is finite dimensional and the noise is non-degenerate Gaussian, Φ is an $\mathbf{L}^2$ distance of model $\mathcal{G}$ and data $y^\dagger$. Therefore, we sometimes call Φ *(data) misfit*. The second set of assumptions is given in terms of the potential.

Assumption 1.48 (Potential, [236, Assumption 2.6])**.** Let the potential $\Phi : X \times Y \to \mathbb{R}$ satisfy the following conditions:

(i) for every $\varepsilon, r > 0$ there is an $M(\varepsilon, r) \in \mathbb{R}$ such that

$$\Phi(\theta; y) \geq M(\varepsilon, r) - \varepsilon\|\theta\|_X^2 \quad (\theta \in X, y \in Y, \text{ where } \|y\|_Y < r);$$

(ii) for every $r > 0$ there is a $K(r) > 0$ such that

$$\Phi(\theta; y) \leq K(r) \quad (\theta \in X, y \in Y, \text{ where } \max\{\|\theta\|_X, \|y\|_Y\} < r);$$

(iii) for every $r > 0$ there is an $L(r) > 0$ such that

$$|\Phi(\theta_1; y) - \Phi(\theta_2; y)| < L(r)\|\theta_1 - \theta_2\|_X \quad (\theta_1, \theta_2 \in X, y \in Y, \text{ where } \max\{\|\theta_1\|_X, \|\theta_2\|_X, \|y\|_Y\} < r);$$

(iv) for every $\varepsilon, r > 0$ there is a $C(\varepsilon, r) \in \mathbb{R}$ such that

$$|\Phi(\theta; y_1) - \Phi(\theta; y_2)| \leq \exp(\varepsilon\|\theta\|_X^2 + C(\varepsilon, r))\|y_1 - y_2\|_Y \quad (\theta \in X, y_1, y_2 \in Y, \text{ where } \max\{\|y_1\|_Y, \|y_2\|_Y\} < r)$$

◇

Now, we state the theorem.

Theorem 1.49 (Lipschitz well-posedness; [236, Theorems 4.1, 4.2])**.** Let μ_{prior} fulfill Assumption 1.47 and let Φ fulfill Assumption 1.48. Then, the Bayesian inverse problem (1.24) with prior μ_{prior} and likelihood $L = \exp(-\Phi)$ is Lipschitz well-posed. ◇

Since Stuart [236] has introduced the concept of Lipschitz well-posedness, his concept has been discussed and generalised by various researchers. Dashti and Stuart [58] simplified Assumption 1.48. Various PDE-based BIPs have been investigated with respect to Lipschitz well-posedness, e.g. the elliptic partial differential equation [57, 130], level-set inversion [129], Helmholtz source identification with Dirac sources [82], a Cahn-Hilliard model for tumour growth [134]. Moreover, BIPs with more general prior models have been discussed, e.g. stable priors in quasi-Banach spaces [239, 240], convex and heavy-tailed priors [126, 127]. We mention Ernst et al. [84], who have considered uniform and Hölder continuity of posterior measures with respect to data, and given sufficient assumptions in this setting. We refer to these as *Hölder* and *uniform well-posedness*, respectively. Finally, we refer to [233] for a discussion of Lipschitz well-posedness, where stability is measured in different metrics, such as the Wasserstein distance.

Bayesian point estimators. The solution of the Bayesian inverse problem is a probability measure. Sometimes, it is necessary to obtain a point estimate for $\theta^\dagger$; rather than a probability distribution over possible values. A point estimator is a measurable function $\widehat{\theta} : Y \to X$ mapping the observed data into the parameter space. The point estimator shall often represent the best fit for the data. Here, *best* needs to be specified.

Maximum likelihood estimators form a popular class of point estimators in frequentist statistics. They are defined as the global maximum of the log-likelihood function, after inserting the observed data. Hence, it is the parameter that explains the data output optimally. In the inverse problem framework, we can often not expect that the log-likelihood is concave. In this case, there may be no global maximum.

Furthermore, if there is a global maximum, finding it computationally is a hard problem.
Bayesian point estimators can sometimes overcome these problems. In the following, we will discuss two popular examples of this class: the *posterior mean*, also known as the *Bayes estimator*; and the *Maximum-A-Posteriori (MAP)* estimator.
For the following discussion of Bayesian point estimators, we refer to e.g. van der Vaart [250, §10] for the Bayes estimator, and to Dashti et al. [56] for the MAP.

Bayes estimator.

Definition 1.50. The *Bayes estimator* or *posterior mean* is the expected value of the posterior measure

$$\widehat{\theta}(y^\dagger) := \int_X \theta \mu^\dagger_{\text{post}}(\mathrm{d}\theta),$$

if well-defined. ◇

If the posterior measure has a finite second moment, one can show that the Bayes estimator is the $\mathbf{L}^2$-optimal estimator of $\theta^\dagger$, see [153, Theorem 2.11]. If in addition the associated BIP is Lipschitz well-posed, we can show a Lipschitz well-posedness result for the Bayes estimator.

Theorem 1.51. Consider a Lipschitz well-posed BIP and let the posterior measure $\mu^\dagger_{\text{post}}$ have a finite second moment for all $y^\dagger \in Y$. Then,

(i) the Bayes estimator exists,

(ii) the Bayes estimator is unique,

(iii) the map $y^\dagger \to \widehat{\theta}(y^\dagger)$ is locally Lipschitz continuous in X.

Hence, the problem of computing the posterior mean is well-posed.

Proof. The Bayes estimator exists and is unique, since the posterior measure has a finite first moment. Let $y^\dagger, y^\ddagger \in Y$. Since $\mu^\dagger_{\text{post}}$, $\mu^\ddagger_{\text{post}}$ have finite second moment, $C := 2\int_X \|\theta\|_X^2 \mu^\dagger_{\text{post}}(\mathrm{d}\theta) + 2\int_X \|\theta\|_X^2 \mu^\ddagger_{\text{post}}(\mathrm{d}\theta) < \infty$. Therefore, by Lemma 1.18

$$\left\| \int_X \theta \mu^\dagger_{\text{post}}(\mathrm{d}\theta) - \int_X \theta \mu^\ddagger_{\text{post}}(\mathrm{d}\theta) \right\|_X \leq C \mathrm{d}_{\text{Hel}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}})$$

Moreover, since the is Lipschitz well-posed, we can find a $K > 0$, such that

$$\mathrm{d}_{\text{Hel}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) \leq K \|y^\dagger - y^\ddagger\|_Y.$$

□

Hence, the Bayes estimator inherits the well-posedness property from the posterior measure. This is not necessarily true for the MAP estimator, which we discuss next.

Maximum-A-Posteriori estimator. The *Maximum-A-Posteriori estimator* is typically defined as the maximum of the posterior density with respect to the Lebesgue measure. This definition does not cover posterior measures that do not have a Lebesgue density - i.e. if X is infinite dimensional. Instead, the MAP is characterised as a *(weak) mode* of the posterior measure.

Definition 1.52. A *mode* $\theta^* \in X$ of $\mu \in \mathrm{Prob}(X)$ satisfies $\lim_{R \downarrow 0} \frac{\sup_{\theta \in X} \mu(B(\theta,R))}{\mu(B(\theta^*,R))} = 1$. ◇

This means that translating the ball $B(\theta^*, R)$ by $\theta - \theta^*$ does not increase its probability, even if the ball is very small. For details on modes of probability measures, we refer to [162].
We discuss the MAP here, since it often corresponds to a *regularised least squares approach*. The regularised least squares approach is a popular deterministic approach to an inverse problem; see also Remark 1.40. Given an inverse problem (1.16), it proceeds by solving the following optimisation problem

$$\widehat{\theta}(y^\dagger) \in \mathrm{argmin}_{\theta \in X} \frac{1}{2} \|\Gamma^{-1/2}(\mathcal{G}(\theta) - y^\dagger)\|_Y^2 + \lambda \mathrm{Reg}(\theta).$$

The function $\mathrm{Reg} : X \to \mathbb{R}$ is called *regulariser* and $\lambda > 0$ determines the strength of the regularisation. A popular class of regularisers is the *Tikhonov class*. It consists of regularisers of type $\mathrm{Reg} := \frac{1}{2}\|\mathcal{D}(\cdot - m)\|_X^2$, for some $m \in X$ and $\mathcal{D} : X \to X$ linear and bounded. Using a Tikhonov regulariser, one can show a certain correspondence between MAP estimation and the regularised least squares approach. We explain this correspondence now intuitively and refer to [56, 58] for a detailed and rigorous discussion.
Consider a BIP with a finite-dimensional data space Y and parameter space X. We assume to have a Gaussian prior $\mu_{\mathrm{prior}} = \mathrm{N}(m, \mathcal{C})$ and a likelihood that is induced by an additive Gaussian noise model

$$L(y^\dagger|\cdot) := \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(\mathcal{G} - y^\dagger)\|_Y^2\right),$$

where Γ is symmetric positive definite. Under further assumptions, one can show that

$$\widehat{\theta}(y^\dagger) \in \mathrm{argmin}_{\theta \in X} \frac{1}{2}\|\Gamma^{-1/2}(\mathcal{G}(\theta) - y^\dagger)\|_Y^2 + \frac{1}{2}\|\mathcal{C}^{-1/2}(\theta - m)\|_X^2$$

is a mode of $\mu_{\mathrm{post}}^\dagger$. Hence, a Tikhonov-regularised inverse problem gives the MAP estimator of the associated BIP.
Note that the MAP estimator is conceptually similar to the maximum likelihood estimator. We mentioned before that the maximum likelihood estimator tends to be computationally inaccessible since the log-likelihood is not concave. The prior in the Bayesian setting acts as a regularisation and may give rise to a convex minimisation problem, which is computationally much easier to solve.

1.4 Hierarchical uncertainty quantification

In §1.1.2, we have considered data-driven mathematical models. More specifically, we have discussed the use of *deep models*. Those are built as a composition of simple

models, such as linear models and artifical neural networks. Deep models allow to represent more complex behaviour than simple models.
Similarly, we want to construct more complex probability measures to represent uncertain parameters. We do so by constructing a hierarchy of Markov kernels and a probability measure, representing the *layers*. The *hierarchical measure* is then given as the composition along this hierarchy. For the sake of simplicity, we assume throughout this thesis that we have at most two layers, i.e. a depth of one. We discuss the case of more layers in Remark 1.54.

Definition 1.53 (Hierarchical measure)**.** Let R be a measurable subset of a separable Banach space. We call R *hyperparameter space* and equip it with the Borel σ-algebra $\mathcal{B}R$. Moreover, let $K : R \times \mathcal{B}X \to [0,1]$ be a Markov kernel, and $\mu' \in \mathrm{Prob}(R)$. We define the *hierarchical measure* μ'' as the composition of μ' and K, i.e.

$$\mu'' := \mu' K := \int_R K(\cdot\,|\kappa)\mu'(\mathrm{d}\kappa).$$

Sometimes, we consider the joint measure of both layers $\mu \in \mathrm{Prob}(X \times R)$, given by

$$\mu(A \times A_1) := (\mu \odot K)(A \times A_1) := \int_{A_1} K(A|\kappa)\mu'(\mathrm{d}\kappa) \qquad (A \in \mathcal{B}X, A_1 \in \mathcal{B}R).$$

◇

Remark 1.54 (Deep hierarchical measures)**.** Let now $N_{\mathrm{dep}} \in \mathbb{N}$ be the *depth* of the hierarchical measure. We define $R_1, \dots, R_{N_{\mathrm{dep}}}$ to be measurable subsets of separable Banach spaces. Let $\mu^{(N_{\mathrm{dep}})} \in \mathrm{Prob}(R_{N_{\mathrm{dep}}})$ be the probability measure representing the deepest layer in the hierarchy. Moreover, let $K^{(k)}$ be a Markov kernel from $(R_{k+1}, \mathcal{B}R_{k+1})$ to $(R_k, \mathcal{B}R_k)$, $k = 0, \dots, N_{\mathrm{dep}}$. Here, we set $(R_0, \mathcal{B}R_0) := (X, \mathcal{B}X)$. The associated hierarchical measure μ'' on $(X, \mathcal{B}X)$ is

$$\begin{aligned}\mu'' &:= \mu^{N_{\mathrm{dep}}} K^{(N_{\mathrm{dep}}-1)} \cdots K^{(1)} K^{(0)} \\ &:= \int_{R_{N_{\mathrm{dep}}}} \cdots \int_{R_1} K^{(0)}(\cdot\,|\kappa^1) K^{(1)}(\mathrm{d}\kappa^1|\kappa^2) \cdots \\ &\qquad\qquad \cdots K^{(N_{\mathrm{dep}}-1)}(\mathrm{d}\kappa^{N_{\mathrm{dep}}-1}|\kappa^{N_{\mathrm{dep}}}) \mu^{N_{\mathrm{dep}}}(\mathrm{d}\kappa^{N_{\mathrm{dep}}}).\end{aligned}$$

This joint measure $\mu \in \mathrm{Prob}(X \times R_1 \times \cdots \times R_{N_{\mathrm{dep}}})$ of all layers is again defined by the semidirect product

$$\begin{aligned}&\mu(A \times A_1 \times \cdots \times A_{N_{\mathrm{dep}}}) \\ &\quad := \mu^{N_{\mathrm{dep}}} \odot K^{(N_{\mathrm{dep}}-1)} \odot \cdots \odot K^{(1)} \odot K^{(0)} \\ &\quad := \int_{A_{N_{\mathrm{dep}}}} \cdots \int_{A_1} K^{(0)}(A|\kappa^1) K^{(1)}(\mathrm{d}\kappa^1|\kappa^2) \cdots \\ &\qquad\qquad \cdots K^{(N_{\mathrm{dep}}-1)}(\mathrm{d}\kappa^{N_{\mathrm{dep}}-1}|\kappa^{N_{\mathrm{dep}}}) \mu^{N_{\mathrm{dep}}}(\mathrm{d}\kappa^{N_{\mathrm{dep}}}),\end{aligned}$$

where $A \in \mathcal{B}X, A_1 \in \mathcal{B}R_1, \dots, A_{N_{\mathrm{dep}}} \in \mathcal{B}R_{N_{\mathrm{dep}}}$. As mentioned above, we only consider hierarchical measures with depths of up to $N_{\mathrm{dep}} = 1$. Generalisations for $N_{\mathrm{dep}} \geq 2$ can be obtained recursively. From the setting in this remark, we obtain the two layer setting by defining $R_1 =: R$, $\kappa^1 =: \kappa$, $K^{(1)} =: K$, and $\mu^1 =: \mu'$. ◇

Hierarchical measures are supposed to capture more complexity than a classical probability measure. In §1.4.1, we motivate the use of hierarchical measures in a particular Bayesian inverse problem. We proceed in §1.4.2 by considering the forward propagation of uncertainties and Bayesian inversion with hierarchical measures. In the forward case, we show existence of the desired push-forward measure. In the inverse case, we show that the posterior measure is well-defined.
In Bayesian inverse problems, it is possible to have not only a hierarchical prior, but also a hierarchical likelihood. Hierarchical likelihoods appear, for instance, when not only the parameter θ should be estimated, but also the correct mathematical model should be found in a finite set of potential models. This process is termed *Bayesian model selection* and will be briefly addressed in §1.4.3.

1.4.1 Motivation

Gaussian random fields are completely characterised by the mean function and covariance operator, and are thus simple models of spatially varying functions. They are also flexible; depending on the regularity of the covariance operator it is possible to generate realisations with different degrees of smoothness. However, in some practical situations the full information on the covariance operator might not be available, e.g. the correlation length, smoothness, and point-wise variance of the random field are not known. Of course, these model parameters can be fixed a priori. However, the posterior measure of a Bayesian inverse problem is often very sensitive to prior information. We illustrate this in the following simple example.

Example 1.55. We consider a Gaussian random field with exponential covariance operator on the unit square $\overline{D} = [0,1]^2$ and correlation length $\lambda = 0.5$; see Example 1.26. The goal is to estimate the statistics of this field given 9 noisy point observations within the framework of Bayesian inversion. The noise is centered Gaussian with a noise level about 1%. In Figure 1.3 we plot a realisation of the true field together with the posterior mean and variance associated with four prior fields with a different (fixed) correlation length each. The posterior mean and variance have been computed analytically. Note that this is possible since λ is fixed, the prior and noise are Gaussian, and the forward response operator is linear. ◇

We make two observations in Figure 1.3. First, we see that it is not possible to identify the true random field perfectly in this experiment. This is due to the sparsity of the data; it is not a defect of the Bayesian inversion. The posterior measure is well-defined, and has been computed analytically without a sample error in this experiment.
The second observation is at the same time the motivation for hierarchical Bayesian inversion. We clearly see in Figure 1.3 that the posterior measure depends crucially on the underlying prior measure and associated correlation length. If the assumed correlation length is too small compared to the truth, then the posterior mean estimate is only accurate close to the observation locations. If, on the other hand, the assumed correlation length is too large, we obtain an overconfident posterior mean estimate. Inaccurate, fixed prior random field parameters can substantially deteriorate the estimation result in Bayesian inverse problems. We treat this problem by modelling unknown *hyperparameters* as random variables.

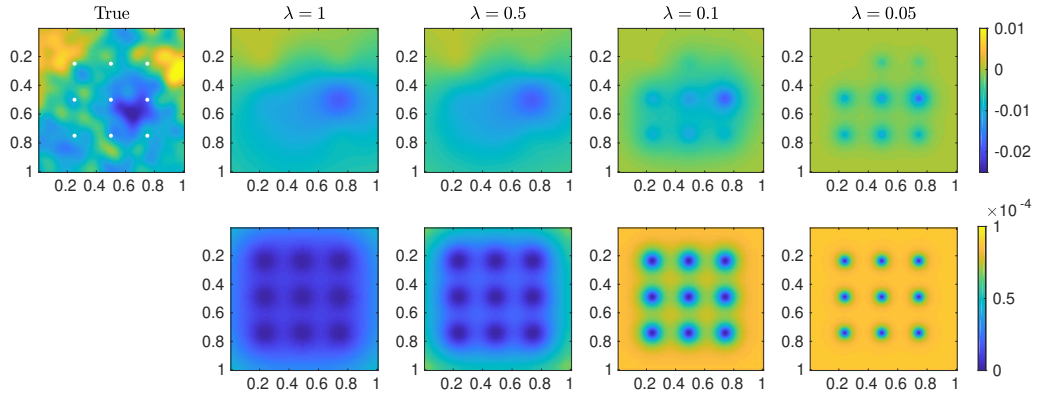


Figure 1.3. Estimation of a Gaussian random field. The top-left figure shows a realisation of the true random field. The task is to estimate this field given 9 noisy point evaluations (white dots). The four top-right (bottom-right) figures show the posterior mean (pointwise posterior variance) for mean-zero Gaussian prior random fields with exponential covariance operator, standard deviation $\sigma = 1$, and correlation lengths $\lambda = 1, 0.5, 0.1, 0.05$.

In this case, the prior is formed by a hierarchical measure, in which the inner layer represents the unknown hyperparameters, i.e. $\kappa = (\lambda, \sigma)$, and the outer layer represent a Gaussian random field subject to the unknown hyperparameters.

1.4.2 Forward and inverse problem

In this section, we describe a general setting for forward uncertainty propagation and an associated Bayesian inverse problem, with hierarchical measures. Importantly, we investigate the well-definedness of these problems if the uncertain elements are modelled by hierarchical measures. This is a necessary extension of the by now well-established solution theory for e.g. Gaussian random inputs.
Hierarchical measures in forward and inverse UQ have been discussed by Wikle [258]. We also refer to Robert [211, §10] for general hierarchical Bayesian analyses. Hierarchical Bayesian inverse problems have been considered in [28, 78, 76, 77, 214] from a mathematical/statistical perspective and in [133, 234, 241] from a computational perspective. Hierarchical models are also considered in the frequentist approach to inference, see e.g. [116, 186] for random field models and spatial statistics.
Let $G : X \to H$ be a mathematical model. Moreover, let the hyperparameter space R, Markov kernel $K : R \times \mathcal{B}X \to [0, 1]$ and measure $\mu' \in \mathrm{Prob}(R)$ form a hierarchical measure $\mu'' := \mu' K \in \mathrm{Prob}(X)$. Moreover, we define $(\boldsymbol{\kappa}, \boldsymbol{\theta}) \sim \mu := \mu' \odot K$.

Forward uncertainty propagation. Let μ'' represent the uncertainty in the parameter of the model G. To quantify this uncertainty, we aim to determine the push-forward measure $\mu''(G \in \cdot)$. The well-definedness of pushforward measures has been discussed extensively for a list of underlying measures. In the following, we discuss the existence of $\mu''(G \in \cdot)$. Hence, we add hierarchical measures to this list. Moreover, we also study the existence of moments of $\mu''(G \in \cdot)$. To this end we make the following assumptions.

Assumption 1.56. (i) $K(G \in \cdot\,|\kappa) := \mathbb{P}(G(\boldsymbol{\theta}) \in \cdot\,|\boldsymbol{\kappa} = \kappa)$ is well-defined for μ'-a.e. $\kappa \in R$.

(ii) For some $k \in \mathbb{N}$ it holds

$$m_k(\kappa) := \int G(\theta)^k K(\mathrm{d}\theta|\kappa) < \infty$$

for μ'-a.e. $\kappa \in R$ and $\int m_k(\kappa)\mu'(\mathrm{d}\kappa) < \infty$.

◇

Let, for instance, $K(\cdot\,|\kappa)$ be a Gaussian random field with continuous realisations for $\kappa \in R$, μ'-a.s. Moreover, let $G : X \to H$ be the elliptic PDE with unknown log-diffusion coefficient from Example 1.37. Then, Assumption 1.56 hold, see [35, 83]. We now show that in such a case, also the hierarchical forward problem is well-posed.

Theorem 1.57. Let Assumption 1.56 hold. Then, the measure $\mu''(G \in \cdot)$ is well-defined. Moreover, $\int G(\theta)^k \mu''(\mathrm{d}\theta) < \infty$, where $k \in \mathbb{N}$ is as in Assumption 1.56(ii). ◇

Proof. By Assumption 1.56(i), $K(G \in \cdot\,|\kappa)$ is well-defined and a probability measure for μ'-a.e. $\kappa \in R$. Hence,

$$\mu''(G \in \cdot) = \int_R K(G \in \cdot\,|\kappa)\mu'(\mathrm{d}\kappa)$$

is well-defined and a probability measure. The finiteness of the moments can be shown analogously. □

Bayesian inverse problem. Now we consider a Bayesian inverse problem, in which the prior is the hierarchical measure μ''. Here, the measure μ' of the hyperparameter is called *hyperprior*. We now actually distinguish two different cases: $\mu_{\text{prior}} := \mu'' := \mu' K$ and $\mu_{\text{prior}} := \mu := \mu' \odot K$. In the first case, we are only interested in the posterior distribution of the unknown parameter $\boldsymbol{\theta}$:

$$\mu''^{\dagger}_{\text{post}} = \mathbb{P}(\boldsymbol{\theta} \in \cdot\,|y = y^\dagger) \in \mathrm{Prob}(X).$$

In the second case, we aim to compute the joint posterior of $\boldsymbol{\theta}$ and the hyperparameters $\boldsymbol{\kappa}$:

$$\mu^{\dagger}_{\text{post}} = \mathbb{P}((\boldsymbol{\theta}, \boldsymbol{\kappa}) \in \cdot\,|y = y^\dagger) \in \mathrm{Prob}(X \times R).$$

Determining either of these posterior measures will be called *hierarchical Bayesian inverse problem.*

We ignore the first case $\mu_{\text{prior}} := \mu''$ since we can obtain $\mu''^{\dagger}_{\text{post}}$ by marginialising $\mu^{\dagger}_{\text{post}}$. Then, we get

$$\mu''^{\dagger}_{\text{post}} = \mu^{\dagger}_{\text{post}}(\cdot \times R).$$

Hence, we consider the second case, where $\mu_{\text{prior}} := \mu := \mu' \odot K$, and discuss Bayes' formulae for $\mu^{\dagger}_{\text{post}}$. Then, we give properties under which the hierarchical Bayesian inverse problem is well-defined.

To determine $\mu^\dagger_{\text{post}}$, we can apply Bayes' formula from Theorem 1.45, where we set $\nu_{X\times R} := \mu_{\text{prior}}$ and $\pi_{\text{prior}} \equiv 1$. We obtain

$$\frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu_{\text{prior}}}(\theta,\kappa) = \frac{L(y^\dagger|\theta)}{Z(y^\dagger)}, \tag{1.25}$$

where

$$Z(y^\dagger) := \iint_{R\times X} L(\theta|y^\dagger)K(\mathrm{d}\theta|\kappa)\mu'(\mathrm{d}\kappa).$$

The right-hand side of Equation (1.25) is constant in κ. Indeed, it seems as if observing the data $y^\dagger$ has no influence on the distribution of $\boldsymbol{\kappa} \sim \mu'$. Actually, $\boldsymbol{\theta} \sim \mu''$ and $\boldsymbol{\kappa}$ are completely coupled through μ_{prior}. By considering the Radon–Nikodym derivative with respect to μ_{prior}, we do not see the dependence.
We now assume that the coupling can be seen through a conditional density. We assume that ν_X is a σ-finite measure on X dominating $K(\cdot\,|\kappa)$ for μ'-a.e. $\kappa \in R$. We define the conditional density

$$\pi^K(\theta|\kappa) := \frac{\mathrm{d}K(\cdot\,|\kappa)}{\mathrm{d}\nu_X}(\theta) \qquad ((\theta,\kappa) \in X\times R, \nu_X\otimes\mu'\text{-a.e.}).$$

Moreover, let ν_R be a σ-finite measure on R dominating μ', with

$$\pi'(\kappa) := \frac{\mathrm{d}\mu'}{\mathrm{d}\nu_R}(\kappa) \qquad (\kappa \in R, \nu_R\text{-a.e.}).$$

Then, Bayes formula reads:

$$\frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}(\nu_X\otimes\nu_R)}(\theta,\kappa) = \frac{L(y^\dagger|\theta)\pi^K(\theta|\kappa)\pi'(\kappa)}{Z(y^\dagger)}.$$

Now, one can easily see the coupling through π^K. Note that from either of these Bayes' formulae, we can obtain a formula for the posterior $\mu''^\dagger_{\text{post}}$. Since $\mu''^\dagger_{\text{post}} = \mu^\dagger_{\text{post}}(\cdot\times R)$, we just need to integrate over R respect to κ. In particular, we have

$$\frac{\mathrm{d}\mu''^\dagger_{\text{post}}}{\mathrm{d}\nu_X}(\theta) = \int_R \frac{L(y^\dagger|\theta)\pi^K(\theta|\kappa)}{Z(y^\dagger)}\mu'(\mathrm{d}\kappa).$$

We will now show that the posterior measures $\mu^\dagger_{\text{post}}$ and $\mu''^\dagger_{\text{post}}$ based on the hierarchical priors μ and μ'' are well-defined.

Theorem 1.58. Let K be a Markov kernel from $(R,\mathcal{B}R)$ to $(X,\mathcal{B}X)$. Moreover, let $\mu^{\kappa,\dagger}_{\text{post}}$ be the posterior measure of the BIP with prior $K(\cdot\,|\kappa)$, where $\kappa \in R$ is fixed and likelihood $L : X\times Y \to \mathbb{R}$. In particular, let

$$\mu^{\kappa,\dagger}_{\text{post}}(B) := Z(y^\dagger,\kappa)^{-1}\int_B L(\theta|y^\dagger)K(\mathrm{d}\theta|\kappa), \quad B \in \mathcal{B}X,$$
$$Z(y^\dagger,\kappa) := \int_X L(\theta|y^\dagger)K(\mathrm{d}\theta|\kappa).$$

Moreover, let $\mu' \in \mathrm{Prob}(R)$, and $\mu'' := \mu'K$ be the hierarchical measure and μ be the joint measure of μ' and K. Finally, we assume $(\kappa \mapsto Z(y^\dagger,\kappa)) \in \mathbf{L}^1(R,\mathcal{B}R,\mu';\mathbb{R})$. Then, the BIPs

(i) with prior μ'' and likelihood L, and

(ii) with prior μ and likelihood L

are well-defined.

Proof. We only need to discuss (ii). (i) follows from (ii), since we can derive $\mu''^{\dagger}_{\text{post}} = \mu^{\dagger}_{\text{post}}(\cdot \times R)$. By Theorem 1.45, the posterior measure $\mu^{\dagger}_{\text{post}} = \iint_{(\cdot)} \mu^{\kappa,\dagger}_{\text{post}}(\mathrm{d}\theta)\mu'(\mathrm{d}\kappa)$ is well-defined and unique if the normalising constant $Z(y^{\dagger})$ is positive and finite. By assumption it holds $Z(y,\cdot) \in (0,\infty)$. Hence, $Z(y^{\dagger}) = \int_R Z(y^{\dagger},\kappa)\mu'(\mathrm{d}\kappa) > 0$. Furthermore, also by assumption, we have $Z(y^{\dagger}) = \int_R Z(y^{\dagger},\kappa)\mu'(\mathrm{d}\kappa) = \|Z(y^{\dagger},\cdot)\|_1 < \infty$. □

1.4.3 Bayesian model selection

In *Bayesian model selection*, not only a model parameter is identified, but also the correct model. Note that Bayesian model selection is just another example of a hierarchical Bayesian inverse problem. In this case, the hierarchy does not only influence the prior, but also parameter space and likelihood. We refer to Robert [211, §7] and Wassermann [255] for an introduction to Bayesian model selection.
Let $N_{\text{ms}} \in \mathbb{N}$ be the number of considered models. The collection of models is given by $\{\mathcal{G}_1, \dots, \mathcal{G}_{N_{\text{ms}}}\}$. Each of the models may have an individual noise model, an individual set of parameters, and an individual prior model. Hence, we define for each model index $i = 1, \dots, N_{\text{ms}}$ a parameter space X_i, a likelihood $L_i : X_i \times Y \to \mathbb{R}$, and a prior $\mu^{(i)}_{\text{prior}} \in \text{Prob}(X_i)$.
Let $I = \{1, \dots, N_{\text{ms}}\}$ be the index set of the models. In Bayesian model selection, we define a prior $\mu^{\text{ms}}_{\text{prior}} \in \text{Prob}(I)$. Let $\boldsymbol{i} \sim \mu^{\text{ms}}_{\text{prior}}$. We want to use data $y^{\dagger}$ to update our belief concerning the choice of the model

$$\mu^{\dagger,\text{ms}}_{\text{post}} := \mathbb{P}(\boldsymbol{i} \in \cdot\, | \boldsymbol{y_i} = y^{\dagger}),$$

where $\boldsymbol{y}_i \sim L_i(\cdot\,|\theta)$ is the random variable representing the noisy model output, given that the model index is $i \in I$ and the model parameter is $\theta \in X_i$. We obtain a single parameter space representing all possible model configurations by

$$X := \bigcup_{i \in I} \{i\} \times X_i.$$

Moreover, we construct a single parametric model $L : X \times Y \to \mathbb{R}$, where

$$L(y|(i,\theta)) := L_i(y|\theta) \qquad ((i,\theta) \in X, y \in Y).$$

To apply Bayes' theorem to compute $\mu^{\dagger}_{\text{post}}$, we need to find the likelihood of y given i, i.e. the likelihood $L'(y|i)$. Let $i \in I$. We need to represent the joint distribution of $(\boldsymbol{y}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim \mu^{(i)}_{\text{prior}}$ and $\boldsymbol{y}_i \sim L_i(\cdot\,|\boldsymbol{\theta})$. We obtain

$$\mathbb{P}((\boldsymbol{y}_i, \boldsymbol{\theta}) \in \cdot\, | \boldsymbol{i} = i) := \iint_{(\cdot)} L(y|\theta)\mu^{(i)}_{\text{prior}}(\mathrm{d}\theta)\nu_Y(\mathrm{d}y),$$

where ν_Y is the σ-finite measure with respect to which the likelihood is defined as a conditional density. We integrate with respect to θ and obtain as a likelihood of $y \in Y$ given the model $i \in I$:

$$L'(y|i) = \int_{X_i} L(y|(i,\theta))\mu_{\text{prior}}^{(i)}(\mathrm{d}\theta) > 0 \qquad (i \in I).$$

The positivity of $L'(y|i)$ is implied by the positivity of $L(y|(i,\theta))$. Note that the so-called *marginal* likelihood $L'(y|i)$ is identical to the model evidence when computing the posterior for model i. We obtain the posterior using Bayes' formula

$$\mu_{\text{post}}^{\dagger,\text{ms}}(\{i\}) = \frac{L'(y|i)\mu_{\text{prior}}^{\text{ms}}(\{i\})}{\sum_{i' \in I} L'(y|i')\mu_{\text{prior}}^{\text{ms}}(\{i'\})}. \tag{1.26}$$

(1.26) is again a result of Theorem 1.45. The theorem holds since $L' > 0$ and since the sum in the denominator of (1.26) is finite.
Finally, we mention that it is also possible to compute the joint posterior of $\boldsymbol{\theta}$ and $\boldsymbol{i}$, i.e.

$$\mu_{\text{post}}^{\dagger} := \mathbb{P}((\boldsymbol{i},\boldsymbol{\theta}) \in \cdot\, |\boldsymbol{y_i} = y^{\dagger}).$$

The posterior $\mu_{\text{post}}^{\dagger}$ is then a measure on the joint space X.
In inverse problems, various model selection problems have been discussed. A Bayesian inverse problem with model selection has been approached in the articles of Lima and co-authors [165, 166]. Here, the correct model is identified to represent the growth of a brain tumour in mice. [187, 248] consider model selection in hydrogeological inverse problems. Here, the number of terms in the truncated KL expansion was supposed to be identified. Finally, we mention Mark et al. [176]. They consider Bayesian model selection in various complex dynamical systems: the policy assessment in coal-mining safety, the invasiveness of tumour cells, and stock market fluctuations.

Chapter 2

Well-posedness of Bayesian inverse problems

> [...] ces problémes [...] bien posé, je veux dire comme possible et déterminé.
>
> Jacques Hadamard [110, p. 49]

In this chapter, we try to identify general settings in which we can show the well-posedness of BIPs, using no or very limited assumptions on the underlying mathematical model or the forward response operator. This is in contrast to the rather restrictive Assumptions 1.47 and 1.48, which are required to show Lipschitz well-posedness in Theorem 1.49.
In particular, we aim to find assumptions on the likelihood $L(y^\dagger|\theta')$ (or rather the statistical model) that imply well-posedness and that are independent of the underlying forward response operator

$$\mathcal{G} \in \mathbf{M}(X;Y) := \{f : X \to Y \text{ measurable}\}.$$

Existence and *uniqueness* are results of Theorem 1.45, which has already rather mild assumptions. The crux is the *local Lipschitz continuity* condition, reflecting stability. In §2.1, we suggest to substitute Lipschitz continuity by continuity and argue that this generalisation is sensible. The main result of this chapter is a less restrictive set of assumptions, implying existence, uniqueness, and continuity of the posterior measure. The new set of assumptions allows us to show in §2.2 that BIPs with finite-dimensional, non-degenerate Gaussian noise are always well-posed; independent of the choice of $\mathcal{G} \in \mathbf{M}(X;Y)$. In §2.3, we extend the theory in a different direction: throughout this work, we have considered stability of the posterior in the Hellinger distance. Now, we extend the theory to weak convergence, total variation, Wasserstein distance, and Kullback–Leibler divergence. Finally, we illustrate the theoretical results in §2.4.

2.1 Redefining well-posedness

2.1.1 Relaxing Lipschitz

In the following, we give examples for well-defined BIPs in which local Lipschitz continuity does not hold in the posterior measure or is hard to verify by using results in the literature. In all of these cases, we show that the posterior measures are continuous in the data. Given that Hadamard's classical formulation of well-posedness

does not require local Lipschitz continuity and that local Lipschitz continuity may be too strong for general statements, we use these examples to advocate a relaxation of the local Lipschitz continuity condition.
Ill-posedness in the Lipschitz sense can, for instance, occur when data has been transformed by a non-Lipschitz continuous function. As an example, we consider a Bayesian inverse problem that is linear and Gaussian, however, the data is transformed by the cubic root function.

Example 2.1 (Cubic inverse problem)**.** Let $X := Y := \mathbb{R}$. We consider the Bayesian approach to the inverse problem

$$y^\dagger = (\theta + \eta)^3,$$

where θ is the unknown parameter and η is observational noise; both are independent. The probability measure of parameter and noise are given by $\mu_{\text{prior}} := \mu_{\text{noise}} := \mathrm{N}(0, 1^2)$. The likelihood of the BIP is

$$L(y^\dagger|\theta) \propto \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\theta - \sqrt[3]{y^\dagger}\|^2\right).$$

Since prior and noise are Gaussian, and the forward model is linear (the identity operator), we can compute the posterior measure analytically, see [6, §3]. We obtain $\mu^\dagger_{\text{post}} := \mathrm{N}\left(\sqrt[3]{y^\dagger}/2, (1/\sqrt{2})^2\right)$. Moreover, one can show that

$$\mathrm{d}_{\text{Hel}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) = \sqrt{1 - \exp\left(-\frac{1}{8}\left(\sqrt[3]{y^\dagger} - \sqrt[3]{y^\ddagger}\right)^2\right)}, \tag{2.1}$$

where $\mu^\ddagger_{\text{post}}$ is the posterior measure based on a second data set $y^\ddagger \in Y$. One can show analytically that this Hellinger distance in (2.1) is not locally Lipschitz as $|y^\dagger - y^\ddagger| \to 0$. It is however continuous. We plot the Hellinger distance in Figure 2.1 on the left-hand side, where we set $y^\ddagger := 0$ and vary only $y^\dagger \in (-1, 1)$. We observe indeed that the Hellinger distance is continuous, but not Lipschitz continuous. In the plot on the right-hand side, we show the Hellinger distance, when considering $\sqrt[3]{y^\dagger}$ as the data set, rather than $y^\dagger$. In this case, the Hellinger distance is locally Lipschitz in the data. ◇

The Bayesian inverse problem in Example 2.1 is ill-posed in the sense of Definition 1.46 since the posterior is only continuous, but not Lipschitz in the data. However, we can heal this ill-posedness by transforming $y^\dagger \mapsto \sqrt[3]{y^\dagger}$. Hence, the Lipschitz well-posedness property reduces to a continuous data transformation problem.
Other examples may be Lipschitz well-posed, but this may be difficult to verify in practice or for general forward response operators. We recall Assumption 1.48. According to Theorem 1.49, those assumptions are sufficient, but not necessary to prove well-posedness. In Assumption 1.48(iv), we consider local Lipschitz continuity in the log-likelihood $\log L$ with respect to the data. Here, the Lipschitz constant is supposed to be a positive function that is monotonically non-decreasing in $\|\theta\|_X$. This assumption is not satisfied in the following example.

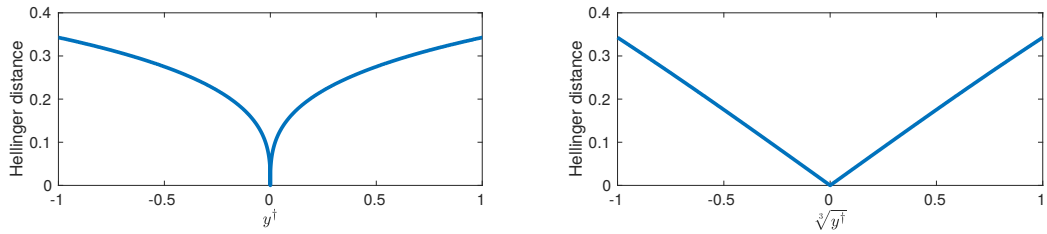


Figure 2.1. Hellinger distances between posterior measures in Example 2.1. The posterior measures are based on two data sets: $y^\dagger$ that varies in $(-1, 1)$ and $y^\ddagger := 0$. In the left figure, we show the relationship between data and Hellinger distance. In the right figure, we replace the data by $y^\dagger := \sqrt[3]{y^\dagger}$, $y^\ddagger := \sqrt[3]{y^\ddagger}$. In both plots, we observe a continuous relationship between Hellinger distance and data, which is also Lipschitz continuous in the right figure, but not in the left figure.

Example 2.2. Let $X := (0, 1)$ and $Y := \mathbb{R}$. We consider the Bayesian approach to the inverse problem

$$y^\dagger = \theta^{-1} + \eta$$

where θ is the unknown parameter and η is observational noise. Neglecting linear prefactors, this inverse problem can be thought of as the recovery of a wavelength θ from a noisy frequecy measurement $y^\dagger$.
The prior measure of θ is given by $\mu_{\text{prior}} = \text{Unif}(0, 1)$. The noise is distributed according to $\mu_{\text{noise}} = \text{N}(0, 1^2)$. Moreover, note that parameter and noise are independent random variables. The likelihood of the BIP is

$$L(y^\dagger|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\theta^{-1} - y^\dagger\|^2\right).$$

For fixed $\theta \in X$, the logarithm of the likelihood in this setting is Lipschitz continuous in the data. However, as $\theta \downarrow 0$, the Lipschitz constant explodes. Hence, the likelihood does not fulfil Assumption 1.48(iv). ◇

Hence, we cannot use the theory discussed in §1.3.3 to show Lipschitz well-posedness of the Bayesian inverse problem in Example 2.2. We expect a similar problem for other forward response operators that are not locally bounded. In Corollary 2.6, we revisit Example 2.2 and show that the posterior measure is continuous with respect to the data.
Up to now we presented rather academic examples in this section. A practically more relevant problem is the Bayesian elliptic inverse problem. That is the Bayesian approach to the inverse problem that we have discussed in Example 1.41. When discussing this example, we allow only continuous diffusion coefficients. In practical applications, this may be too restrictive. Iglesias et al. [130] consider more realistic geometric priors measures. In [130, Theorem 3.5], the authors show local Lipschitz continuity for some of those prior measures, but only Hölder continuity with coefficient $\gamma = 0.5$ for others. This is another example where Lipschitz well-posedness in the sense of Definition 1.46 has not been shown, but continuity in the posterior measure is satisfied.
In the next section, we weaken the *Lipschitz well-posedness* by replacing Lipschitz continuity with continuity as a stability condition. Looking back at the discussion here, we consider this weakening tolerable for practical problems.

2.1.2 Definition and main result

Definition 2.3 (Well-posedness)**.** The problem (1.24) is *(Hellinger) well-posed*, if

(i) $\mu^\dagger_{\text{post}} \in \text{Prob}(X, \mu_{\text{prior}})$ exists *(existence)*,

(ii) $\mu^\dagger_{\text{post}}$ is unique in $\text{Prob}(X, \mu_{\text{prior}})$ *(uniqueness)*, and

(iii) $(Y, \|\cdot\|_Y) \ni y^\dagger \mapsto \mu^\dagger_{\text{post}} \in (\text{Prob}(X, \mu_{\text{prior}}), \text{d}_{\text{Hel}})$ is a continuous function. *(stability)*

◇

Lipschitz continuity implies continuity. Hence, Lipschitz well-posedness in Definition 1.46 is a stronger property than well-posedness in Definition 2.3. In Example 2.1, we have investigated a BIP that is not Lipschitz well-posed, but well-posed. Hence, we also know that Lipschitz well-posedness is a strictly stronger statement. We now give assumptions, under which a Bayesian inverse problem can be shown to be well-posed.

Assumption 2.4. Consider a BIP. Let the following assumptions hold for μ_{prior}-a.e. $\theta' \in X$ and every $y^\dagger \in Y$:

(A1) $L(\cdot\,|\theta')$ is a strictly positive probability density function,

(A2) $L(y^\dagger|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$,

(A3) $g \in \mathbf{L}^1(X, \mu_{\text{prior}})$ exists such that $L(y^\ddagger|\cdot) \leq g$ for all $y^\ddagger \in Y$.

(A4) $L(\cdot\,|\theta')$ is continuous.

◇

(A1) means that any data set $y^\dagger \in Y$ has a positive likelihood under any parameter $\theta' \in X$. We conservatively assume that no combination of parameter and data values is impossible, but some may be unlikely. This can usually be satisfied by continuously transforming the forward response operator and/or by choosing a noise distribution that is concentrated on all of Y. Note that the assumption that $L(y^\dagger|\theta')$ being a probability density function can be relaxed to $c \cdot L(y^\dagger|\theta')$ is a probability density function, where $c > 0$ does depend neither on $y^\dagger$, nor θ'. (A2)-(A3) imply that the likelihood is integrable with respect to the prior and that it is bounded from above uniformly in the data by an integrable function. These assumptions are, for instance, satisfied when the likelihood is bounded from above by a constant. Noise models with bounded probability density function on Y should generally imply a bounded likelihood. Note that (A3) implies (A2). We have stated those assumptions separately, since (A1) and (A2) alone imply Theorem 1.45. (A4) requires the continuity of the likelihood with respect to the data. Continuity in the data is for instance, given when considering noise models with continuous probability density functions and a continuous connection of noise and model. We give examples in §2.4 showing that we can not neglect the continuity in the data. We continue with the main result of this chapter.

Theorem 2.5. Let (A1)-(A4) hold for a BIP. Then the BIP is well-posed.

Proof. Note that existence and uniqueness of the measure $\mu^{\dagger}_{\text{post}}$ are results of Theorem 1.45 that holds since (A1)-(A2) are satisfied. We proceed as follows: we show that the likelihood is continuous as a function from Y to $\mathbf{L}^1(X, \mu_{\text{prior}})$ and that at the same time $y^{\dagger} \mapsto Z(y^{\dagger})$ is continuous. This implies that $y^{\dagger} \mapsto L(y^{\dagger}|\cdot)^{1/2} \in \mathbf{L}^2(X, \mu_{\text{prior}})$ is continuous as well. Then, we collect all of this information and show the continuity in the Hellinger distance, which is the desired result.
1. We now show continuity in $y^{\dagger} \in Y$ when integrating $L(y^{\dagger}|\cdot)$ with respect to μ_{prior}. This is a standard application of *Lebesgue's Dominated Convergence Theorem* (DCT; [219, p. 26]): Let $(y_n)_{n=1}^{\infty} \in Y^{\mathbb{N}}$ be a sequence converging to $y^{\dagger}$, as $n \to \infty$. (A4) implies that $\lim_{n\to\infty} L(y_n|\cdot) = L(y^{\dagger}|\cdot)$ pointwise in X. We obtain by the DCT

$$\lim_{n\to\infty} \int_X L(y_n|\cdot)\mathrm{d}\mu_{\text{prior}} = \int \lim_{n\to\infty} L(y_n|\cdot)\mathrm{d}\mu_{\text{prior}} = \int_X L(y^{\dagger}|\cdot)\mathrm{d}\mu_{\text{prior}},$$

since the sequence $(L(y_n|\cdot))_{n=1}^{\infty}$ is bounded from above by $g \in \mathbf{L}^1(X, \mu_{\text{prior}})$ and bounded from below by 0, see (A1) and (A3). Hence, the functions

$$Y \ni y^{\dagger} \mapsto \int_X L(y^{\dagger}|\cdot)\mathrm{d}\mu_{\text{prior}} = Z(y^{\dagger}) \in \mathbb{R}, \qquad Y \ni y^{\dagger} \mapsto L(y^{\dagger}|\cdot) \in \mathbf{L}^1(X, \mu_{\text{prior}})$$

are continuous. Moreover, note that Theorem 1.45 implies that $Z(y^{\dagger})$ is finite and strictly larger than 0.
2. The continuity in $\mathbf{L}^1(X, \mu_{\text{prior}})$ implies that for every $y^{\dagger} \in Y$, we have for $\varepsilon_1 > 0$ some $\delta_1(\varepsilon_1) > 0$, such that

$$\|L(y^{\dagger}|\cdot) - L(y^{\ddagger}|\cdot)\|_{\mathbf{L}^1(X,\mu_{\text{prior}})} \leq \varepsilon_1 \quad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta_1(\varepsilon_1)).$$

Using this, we can show that $y^{\dagger} \mapsto L(y^{\dagger}|\cdot)^{1/2}$ is continuous in $\mathbf{L}^2(X, \mu_{\text{prior}})$. Let $y^{\dagger} \in Y$ and $\varepsilon_1, \delta_1(\varepsilon_1), y^{\ddagger}$ be chosen as above. We have

$$\begin{aligned}
&\|L(y^{\dagger}|\cdot)^{1/2} - L(y^{\ddagger}|\cdot)^{1/2}\|^2_{\mathbf{L}^2(X,\mu_{\text{prior}})} \\
&\qquad = \int_X \left|L(y^{\dagger}|\cdot)^{1/2} - L(y^{\ddagger}|\cdot)^{1/2}\right|^2 \mathrm{d}\mu_{\text{prior}} \\
&\qquad \leq \int_X \left|L(y^{\dagger}|\cdot)^{1/2} - L(y^{\ddagger}|\cdot)^{1/2}\right| \times \left|L(y^{\dagger}|\cdot)^{1/2} + L(y^{\ddagger}|\cdot)^{1/2}\right| \mathrm{d}\mu_{\text{prior}} \\
&\qquad = \int_X \left|L(y^{\dagger}|\cdot) - L(y^{\ddagger}|\cdot)\right| \mathrm{d}\mu_{\text{prior}} \leq \varepsilon_1.
\end{aligned}$$

Now, we take the square-root on each side of this inequality. For every $\varepsilon_2 > 0$, choose $\delta_2(\varepsilon_2) := \delta_1(\varepsilon_2^{1/2}) > 0$. Then,

$$\|L(y^{\dagger}|\cdot)^{1/2} - L(y^{\ddagger}|\cdot)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\text{prior}})} \leq \varepsilon_2 \quad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta_2(\varepsilon_2))$$

gives us the desired continuity result.
3. Using the continuity result in 1. and the composition of continuous functions, we also know that $y^{\dagger} \mapsto Z(y^{\dagger})^{-1/2} \in (0, \infty)$ is continuous. Hence, we have for every $y^{\dagger} \in Y$ and every $\varepsilon_3 > 0$ a $\delta_3(\varepsilon_3) > 0$ with

$$|Z(y^{\dagger})^{-1/2} - Z(y^{\ddagger})^{-1/2}| \leq \varepsilon_3 \quad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta_3(\varepsilon_3)).$$

Given this and all the previous results, we now employ a technique that is typically used to prove the continuity of the product of two continuous functions. Let $y^\dagger \in Y$, $\varepsilon_2, \varepsilon_3 > 0$, $\delta_4 = \min\{\delta_2(\varepsilon_2), \delta_3(\varepsilon_3)\}$ and $y^\dagger \in Y$: $\|y^\dagger - y^\ddagger\|_Y \leq \delta_4$. We arrive at

$$\begin{aligned}
\mathrm{d_{Hel}}(\mu^\dagger_{\mathrm{post}}, \mu^\ddagger_{\mathrm{post}}) &= \|Z(y^\dagger)^{-1/2} L(y^\dagger|\theta)^{1/2} - Z(y^\ddagger)^{-1/2} L(y^\ddagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\mathrm{prior}})} \\
&\leq |Z(y^\ddagger)^{-1/2}| \times \|L(y^\ddagger|\theta)^{1/2} - L(y^\dagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\mathrm{prior}})} \\
&\quad + \|L(y^\dagger|\theta)^{1/2}\|_{\mathbf{L}^2(X,\mu_{\mathrm{prior}})} |Z(y^\ddagger)^{-1/2} - Z(y^\dagger)^{-1/2}| \\
&\leq Z(y^\ddagger)^{-1/2}\varepsilon_2 + Z(y^\dagger)^{1/2}\varepsilon_3 \\
&\leq ((Z(y^\dagger)^{-1/2} + \varepsilon_3)\varepsilon_2 + Z(y^\dagger)^{1/2}\varepsilon_3,
\end{aligned}$$

where we have used in the last step that $|Z(y^\dagger)^{-1/2} - Z(y^\ddagger)^{-1/2}| \leq \varepsilon_3$. We now choose some $\varepsilon_4 > 0$ and set $\delta_4 = \min\{\delta_2(\varepsilon_2'), \delta_3(\varepsilon_3')\}$, where

$$\varepsilon_2' := \frac{\varepsilon_4 Z(y^\dagger)^{1/2}}{\varepsilon_4 + 2}, \quad \varepsilon_3' := \frac{\varepsilon_4}{2Z(y^\dagger)^{1/2}}.$$

Then, we obtain that $\mathrm{d_{Hel}}(\mu^\dagger_{\mathrm{post}}, \mu^\ddagger_{\mathrm{post}}) \leq \varepsilon_4$ for any $y^\ddagger \in Y$, such that $\|y^\dagger - y^\ddagger\|_Y \leq \delta_4$. This implies the continuity of the posterior measure in Hellinger distance. □

2.2 The additive Gaussian noise case

In practice, the measurement data space is typically finite dimensional and the measurement error is often modelled by additive non-degenerate Gaussian noise. In this case, one can verify assumptions (A1)-(A4) independently of prior μ_{prior} and forward response operator $\mathcal{G}$. Hence, this very popular setting leads to a well-posed Bayesian inverse problem in the sense of Definition 2.3.

Corollary 2.6. Let $Y := \mathbb{R}^k$ and $\Gamma \in \mathbb{R}^{k\times k}$ be symmetric positive definite. Let $\mathcal{G} \in \mathbf{M}(X; Y)$ be a measurable function. A Bayesian inverse problem with additive non-degenerate Gaussian noise $\boldsymbol{\eta} \sim \mathrm{N}(0, \Gamma)$ is given by the following likelihood:

$$L(y^\dagger|\theta) = \det(2\pi\Gamma)^{-1/2} \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(\mathcal{G}(\theta) - y^\dagger)\|_Y^2\right).$$

The BIP corresponding to any prior probability measure μ_{prior} on $(X, \mathcal{B}X)$ and likelihood L is well-posed.

Proof. We verify (A1)-(A4). (A1): By definition, the likelihood is a strictly positive probability density function for any $\theta' \in X$. (A2)-(A3): The likelihood is bounded above uniformly by $g \equiv \det(2\pi\Gamma)^{-1/2}$ which is integrable with respect to any probability measure on $(X, \mathcal{B}X)$. (A4): The likelihood is continuous in $y^\dagger$ for any $\theta' \in X$. □

Remark 2.7. Let X contain at least two elements. The non-Bayesian inverse problem (1.16) corresponding to the additive Gaussian noise setting in Corollary 2.6 is ill-posed. We have shown this in Proposition 1.39. Hence, in case of Gaussian noise, the Bayesian approach using any prior measure always gives a well-posed Bayesian inverse problem, in contrast to the always ill-posed inverse problem (1.16). ◇

The fact that we can show well-posedness under *any* prior measure and *any* forward response operator has relatively strong implications for practical problems. We now comment on the deterministic discretisation of posterior measures, hierarchical models, Bayesian model selection, and the Bayesian elliptic inverse problem.

Remark 2.8 (Deterministic discretisation)**.** Bayesian inverse problems can be discretised with deterministic quadrature rules; such are quasi-Monte Carlo [68], sparse grids [229], or Gaussian quadrature. Those are then used to approximate the model evidence and to integrate with respect to the posterior. Deterministic quadrature rules often behave like discrete approximations of the prior measure. If this discrete approximation is a probability measure as well, we can apply Corollary 2.6 and show that the BIP based on the discretised prior is well-posed. ◇

Remark 2.9 (Hierarchical prior)**.** We have discussed hierarchical Bayesian inverse problems in §1.4.2. Let L be a likelihood as in Corollary 2.6 on the parameter space X. Moreover, we define a hyperprior K, which is a Markov kernel from $(R, \mathcal{B}R)$ to $(X, \mathcal{B}X)$ and a prior on R, that is μ'. We can now consider either of two different BIP: in the first case, we set $\mu_{\text{prior}} := \mu' K$, in the second case, we set $\mu_{\text{prior}} := \mu' \odot K$. In the first case, the uncertain parameter is only $\boldsymbol{\theta} : \Omega \to X$. In the second case, we consider also $\boldsymbol{\kappa} : \Omega \to R$ to be uncertain, where $\boldsymbol{\kappa} \sim \mu'$. Recall that in the first case, the posterior measure is given by $\mathbb{P}(\boldsymbol{\theta} \in \cdot\,|\boldsymbol{y} = y^\dagger) \in \text{Prob}(X)$, whereas in the second case, it is given by $\mathbb{P}((\boldsymbol{\theta}, \boldsymbol{\kappa}) \in \cdot\,|\boldsymbol{y} = y^\dagger) \in \text{Prob}(X \times R)$. In either case, we obtain a well-posed BIP. The first case follows straight from Corollary 2.6. The second case follows from the same corollary. Here, we need to extend the parameter space to $X \times R$. The likelihood is still strictly positive and bounded on this extended parameter space, and constant on with respect to R. Moreover, the likelihood is still continuous in the data. Hence, Assumption 2.4 is satisfied and the BIP is indeed well-posed. ◇

Remark 2.10 (Model selection)**.** As in Remark 2.9, we can use Corollary 2.6 to show well-posedness of the Bayesian model selection problem that we have discussed in §1.4.3. Here, we only need to assume that the likelihood for any model $\mathcal{G} \in \mathbf{M}'$ is of the form given in Corollary 2.6. ◇

Remark 2.11 (Bayesian elliptic inverse problem)**.** We consider the Bayesian approach to the elliptic inverse problem discussed in Example 1.41. In this inverse problem, we aim to identify the diffusion coefficient of an elliptic PDE, given finitely many observations from the PDE solution. Recall that the Lipschitz and Hölder well-posedness of the Bayesian elliptic inverse problem has been discussed by [57, 130]. This already implies the well-posedness in our setting. However, we can also apply Corollary 2.6, as we do in the following.
We assume that X is a separable Hilbert space and $P : X \to \mathbf{C}^0(\overline{D})$ is a map from X to the space of continuous functions. We define the model $G : X \to H$ mapping $\theta \in X$ to u satisfying

$$\int_D (\exp \circ P(\theta))(x) \langle \nabla u^*(x), \nabla v(x) \rangle_D \mathrm{d}x - \int_D f(x) v(x) \mathrm{d}x = 0 \text{ for all } v \in H',$$

where H is an appropriate solution space and H' is an appropriate test space. Moreover, we define the observation operator $\mathcal{O}$ as in Example 1.41. We assume that

$\mathcal{G} := \mathcal{O} \circ G$ is a well-defined, measurable map from X to Y. Moreover, we assume that the noise is non-degenerate, finite-dimensional Gaussian. Then, according to Corollary 2.6, the Bayesian elliptic inverse problem is well-posed.
Finally, we comment on the choice of X and P. We will later discuss the discretisation of a random field using a truncated KL expansion; see §3.2.1. In this case, we have, e.g. $X := \mathbb{R}^{N_{\text{sto}}}$ and $P(\theta) := m(x) + \sum_{i=1}^{N_{\text{sto}}} \psi_i(x)\theta_i$, where m and ψ_i $(i = 1, \dots, N_{\text{sto}})$ are continuous functions. Alternatively, we may choose $X := \mathbb{R}^{\overline{D}}$ and choose a prior measure whose samples have realisations with continuous modifications. Note that X is not a separable Hilbert space, it may be not even Radon. Hence on this space, we cannot hope for the existence of regular conditional measures; see Theorem 1.33. However, for a fixed data set, the Bayesian inverse problem given in (1.24) may be still solvable. ◇

In this subsection, we have discussed finite-dimensional data and additive non-degenerate Gaussian noise. These results cannot trivially be extended to the infinite-dimensional data case or to the degenerate noise case. The infinite-dimensional data requires a likelihood definition via the Cameron-Martin Theorem, which requires conditions on the forward response operator and noise covariance. For a discussion of infinite-dimensional data spaces, we refer to §1.3.2, or also to [236, Remark 3.8] for compact covariance operators and [134, §2.1] specifically for Gaussian white noise generalised random fields. Degenerate Gaussian likelihoods do not satisfy (A1) and can lead to degenerate posterior measures. We discuss concepts of well-posedness that can handle degenerate posteriors in §2.3.

2.3 Perturbed posteriors in other metrics

In Remark 2.7, we consider a setting, where the BIP is always well-posed, but the inverse problem (1.16) is always ill-posed. Incidentally, we can give an example where a converse statement holds:

Example 2.12 (Noise-free inverse problem)**.** Consider a parameter space X and data space Y. Let $\mathcal{G} : X \to Y$ be a homeomorphism, i.e. it is bijective, continuous, and its inverse $\mathcal{G}^{-1} : Y \to X$ is continuous as well. Let $y^\dagger := \mathcal{G}(\theta^\dagger)$ be a data set observed *noise-free* from the forward response operator $\mathcal{G}$ based on the true parameter $\theta^\dagger \in X$. Then, the inverse problem (1.16) is well-posed. ◇

We now apply *a* Bayesian approach to the inverse problem in Example 2.12. Let μ_{prior} be a prior measure that is concentrated on all of X. The likelihood is

$$L(y^\dagger|\theta) := \begin{cases} 1, & \text{if } \mathcal{G}(\theta) = y^\dagger, \\ 0, & \text{otherwise.} \end{cases}$$

The posterior measure in this setting is the Dirac measure concentrated in the true value:

$$\mu^\dagger_{\text{post}} = \mathbb{P}(\boldsymbol{\theta} \in \cdot\, | \mathcal{G}(\boldsymbol{\theta}) = y^\dagger) = \delta(\cdot - \mathcal{G}^{-1}(y^\dagger)) = \delta(\cdot - \theta^\dagger).$$

Hence, after seeing the data, we have identified the true parameter $\theta^\dagger$ and we are certain about it. Note that this posterior measure is not computed using Theorem 1.45, which would not hold in this setting. Instead, the Bayesian inversion is

defined via the disintegration theorem, a topic thoroughly discussed by Cockayne et al. [47]. Also $\mu^\dagger_{\text{post}} \notin \text{Prob}(X, \mu_{\text{prior}})$. Hence, we cannot compute the Hellinger distance between posteriors $\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}$, where $\mu^\ddagger_{\text{post}}$ is based on $y^\ddagger \neq y^\dagger$. Instead, we consider the closely related total variation distance and obtain

$$\mathrm{d}_{\text{tv}}(\mu^\dagger_{\text{post}}, \mu^\ddagger_{\text{post}}) := \sup_{B \in \mathcal{B}X} \left| \mu^\dagger_{\text{post}}(B) - \mu^\ddagger_{\text{post}}(B) \right| = 1.$$

Hence $\mu^\ddagger_{\text{post}} \not\to \mu^\dagger_{\text{post}}$ in total variation as $y^\ddagger \to y^\dagger$. Thus, indeed the inverse problem is well-posed, while the associated Bayesian inverse problem is ill-posed in the total variation distance.
However, we have $\mu^\ddagger_{\text{post}} \xrightarrow{\text{w}} \mu^\dagger_{\text{post}}$ as $y^\ddagger \to y^\dagger$. Hence we observe continuity in the weak topology on the space $\text{Prob}(X)$.
Summarising this discussion, we have seen:

- In settings, where Theorem 1.45 does not hold, there may still be a solution to the Bayesian inverse problem. However, the Hellinger distance with respect to the prior may then be not defined. *Can we anyway discuss well-posedness?*
- Different metrics on $\text{Prob}(X)$ give different well-posedness results. *Are there connections between those?*

Motivated by these questions, we now investigate in more detail the choice of metrics on probability spaces when discussing the well-posedness of Bayesian inverse problems.

Definition 2.13 (d-Well-posedness)**.** Let $\mathrm{P} \subseteq \text{Prob}(X)$ be a space of probability measures and $d : \mathrm{P}^2 \to [0, \infty)$ be a metric on P. A Bayesian inverse problem is *d-well-posed* if

(i) $\mu^\dagger_{\text{post}} \in \mathrm{P}$ exists, *(existence)*

(ii) $\mu^\dagger_{\text{post}}$ is unique in P, *(uniqueness)*

(iii) $(Y, \|\cdot\|_Y) \ni y^\dagger \mapsto \mu^\dagger_{\text{post}} \in (\mathrm{P}, d)$ is a continuous function. *(stability)*

◇

We consider the following concepts of well-posedness: *weak* well-posedness (if $d = \mathrm{d}_{\text{Prok}}$ is the *Prokhorov metric*), *total variation* well-posedness (if $d = \mathrm{d}_{\text{tv}}$ is the *total variation distance*), and *Wasserstein(p)* well-posedness (if $d = \mathrm{d}_{\text{Was(p)}}$ is the *Wasserstein(p) metric*).
Two different notions of well-posedness can be compared in terms of the topologies in which the continuity is discussed. A coarser topology contains more continuous functions. Hence, the well-posedness results obtained on some topology can easily be extended to a coarser topology. The converse is in general not true. We only consider metric spaces of measures in this section. This simplifies the topological discussion to the following:

Lemma 2.14. Let A, B be two sets and let (A, d_A), (B, d_1) and (B, d_2) be metric spaces. Let $f : (A, d_A) \to (B, d_2)$ be a continuous function. Moreover, let $t : [0, \infty) \to [0, \infty)$ be continuous in 0, with $t(0) = 0$. Finally, let

$$d_1(b, b') \leq t(d_2(b, b')) \qquad (b, b' \in B).$$

Then, $f : (A, d_A) \to (B, d_1)$ is continuous as well.

Proof. For every $a \in A$ and $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$, with

$$d_2(f(a), f(a')) \leq \varepsilon \qquad (a' \in A : d_A(a, a') \leq \delta(\varepsilon)).$$

Hence, for the same a, a', ε and δ, we have

$$d_1(f(a), f(a')) \leq t(d_2(f(a), f(a'))) \leq t(\varepsilon)$$

Since t is continuous in 0, we find for every $\varepsilon' > 0$ some $\delta'(\varepsilon') > 0$, such that $|t(x)| \leq \varepsilon'$ for $x \in [0, \infty) : |x| \leq \delta'(\varepsilon')$. Now, we choose for every $a \in A$ and $\varepsilon'' > 0$: $\delta''(\varepsilon'') := \delta(\delta'(\varepsilon''))$. Then,

$$d_1(f(a), f(a')) \leq t(d_2(f(a), f(a'))) \leq t(\delta'(\varepsilon'')) \leq \varepsilon'' \qquad (a' \in A : d_A(a, a') \leq \delta''(\varepsilon'')),$$

which results in continuity in (B, d_1). □

In the setting of Lemma 2.14, we call d_1 *coarser* than d_2 and d_2 *finer* than d_1, respectively. We can now compare the well-posedness in different metrics. The proposition below follows immediately from Lemma 2.14.

Proposition 2.15. Let d_1, d_2 be metrics on $\mathrm{P} \subseteq \mathrm{Prob}(X)$ and let d_1 be coarser than d_2. Then, a Bayesian inverse problem that is d_2-wellposed, is also d_1-wellposed. ◇

We now assume that $\mathrm{P} \subseteq \mathrm{Prob}(X)$ is a space of probability measures such that the previously defined metrics are all well-defined on P. Applying Proposition 2.15 to the previously mentioned concepts of well-posedness, we obtain the relations shown in Figure 2.2. Here, we refer again to [99] for the appropriate bounds between the metrics, which imply the following statements:

- Total variation well-posedness and (Hellinger) well-posedness are equivalent. Hence in settings, where the Hellinger distance is not defined, the total variation distance shall be the metric of choice.
- Weak well-posedness is indeed the weakest of the considered concepts.
- According Proposition 1.17, we know that the convergence in the $\mathrm{d}_{\mathrm{Was(p)}}$ is stronger than weak convergence on $\mathrm{P}_p(X)$. Hence, Wasserstein(p) well-posedness implies weak well-posedness.

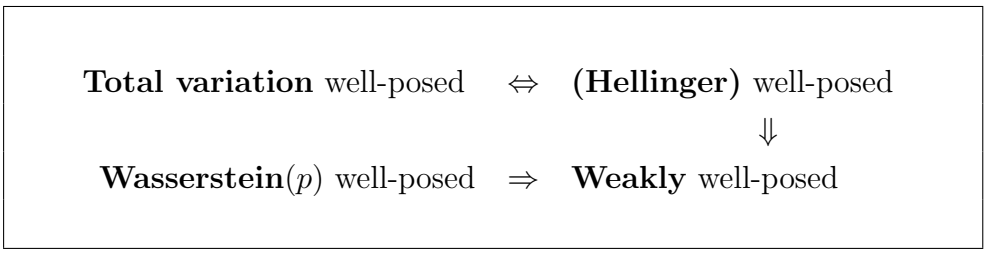


Figure 2.2. Relations between concepts of well-posedness

Which concept of well-posedness should we consider in practice? Weak well-posedness implies the continuity of posterior expectations of bounded, continuous quantities of interest. If this is the task of interest, weak well-posedness should be sufficient.

Hellinger and tv distance imply the convergence of any (existing) posterior expectation. Hence, if discontinuous functions are integrated, or probabilities computed, those distances should be chosen. Wasserstein(p) distances have gained popularity in the convergence and stability theory of Markov chain Monte Carlo (MCMC) algorithms; see e.g. [98, 220]. Hence, Wasserstein(p) well-posedness may be the right tool when discussing the well-posedness of solving a Bayesian inverse problem via MCMC.

2.3.1 Kullback–Leibler divergence

The *Kullback–Leibler divergence* (KLD), *relative entropy*, or *directed divergence* is a popular 'metric' in information theory and machine learning. It is used to describe the *information gain* when going from $\mu \in \mathrm{Prob}(X)$ to another measure $\mu' \in \mathrm{Prob}(X, \mu)$. If well-defined, it is given by

$$\mathrm{D}_{\mathrm{KL}}(\mu' \| \mu) := \int_X \log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right) \mathrm{d}\mu'.$$

Note that this is not actually a metric, since it is neither symmetric, nor does it fulfil the triangle inequality. However, we can describe continuity in the KLD, which also induces a topology, see [14]. This allows us to consider the Kullback–Leibler well-posedness of Bayesian inverse problems. This concept bridges information theory and Bayesian inverse problems; and allows statements about the loss of information in the posterior measure when the data is perturbed. In particular, we define this *loss of information* by the information gain when going from the posterior $\mu^{\ddagger}_{\mathrm{post}}$ with perturbed data $y^{\ddagger}$ to the posterior $\mu^{\dagger}_{\mathrm{post}}$ with unperturbed data $y^{\dagger}$. Hence, the loss of information is equal to $\mathrm{D}_{\mathrm{KL}}(\mu^{\dagger}_{\mathrm{post}} \| \mu^{\ddagger}_{\mathrm{post}})$. A Bayesian inverse problem is Kullback–Leibler well-posed if the posterior measure exists, if it is unique, and if the information loss is continuous with respect to the data.

Definition 2.16 (Kullback–Leibler well-posed)**.** A BIP is *Kullback–Leibler well-posed* if

(i) $\mu^{\dagger}_{\mathrm{post}} \in \mathrm{Prob}(X, \mu_{\mathrm{prior}})$ exists *(existence)*,

(ii) $\mu^{\dagger}_{\mathrm{post}}$ is unique in $\mathrm{Prob}(X, \mu_{\mathrm{prior}})$ *(uniqueness)*, and

(iii) for all $y^{\dagger} \in Y$ and $\varepsilon > 0$, there is $\delta(\varepsilon) > 0$, such that

$$\mathrm{D}_{\mathrm{KL}}(\mu^{\dagger}_{\mathrm{post}} \| \mu^{\ddagger}_{\mathrm{post}}) \leq \varepsilon \quad (y^{\ddagger} \in Y : \|y^{\dagger} - y^{\ddagger}\|_Y \leq \delta(\varepsilon)) \qquad \textit{(stability)}.$$

◇

In the setting of Theorem 1.45, Assumption 2.4 is not sufficient to show Kullback–Leibler well-posedness; indeed, the Kullback–Leibler divergence may be not even well-defined. We require the following additional assumption on the log-likelihood.

Assumption 2.17. Consider a BIP. Let the following assumption hold for μ_{prior}-a.e. $\theta' \in X$ and every $y^{\dagger} \in Y$.

(A5) there is a $\delta > 0$ and a function $h(\cdot, y^\dagger) \in \mathbf{L}^1(X, \mu^\dagger_{\text{post}})$ such that

$$|\log L(y^\ddagger|\cdot)| \leq h(\cdot, y^\dagger) \qquad (y^\ddagger \in Y : \|y^\dagger - y^\ddagger\|_Y \leq \delta).$$

◇

Assumption (A5) seems much stronger than (A1)-(A4). Indeed, we now require some integrability condition on the forward response operator. That condition may be hard to verify when the posterior measure has heavy tails, the model is unbounded, or when the model cannot be analysed.

Theorem 2.18. Let (A1)-(A5) hold for a BIP. Then, BIP is Kullback–Leibler well-posed.

Proof. First note that (A1)-(A4) imply the existence and the uniqueness of the posterior measure, as well as the continuity of $y^\dagger \mapsto Z(y^\dagger)$. Let $y^\dagger \in Y$ and $y^\ddagger \in Y$, with $\|y^\dagger - y^\ddagger\|_Y \leq \delta$. $\delta > 0$ is chosen as in (A5). We have

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mu^\dagger_{\text{post}} \| \mu^\ddagger_{\text{post}}) &= \int_X \log\left(\frac{\mathrm{d}\mu^\dagger_{\text{post}}}{\mathrm{d}\mu^\ddagger_{\text{post}}}\right) \mathrm{d}\mu^\dagger_{\text{post}} \\ &= \int_X \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu^\dagger_{\text{post}} + \left(\log Z(y^\ddagger) - \log Z(y^\dagger)\right),\end{aligned}$$

where the right-hand side of this equation is well-defined since $Z(y^\dagger), Z(y^\ddagger) \in (0, \infty)$ by Theorem 2.5 and since (A5) holds. Moreover, the continuity in the model evidence implies that $\left(\log Z(y^\ddagger) - \log Z(y^\dagger)\right) \to 0$, as $y^\ddagger \to y^\dagger$. Also, note that $\log L(\cdot\,|\theta')$ is continuous by (A4), which implies

$$\lim_{y^\ddagger \to y^\dagger} \int_X \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu^\dagger_{\text{post}} = \int_X \lim_{y^\ddagger \to y^\dagger} \log L(y^\dagger|\cdot) - \log L(y^\ddagger|\cdot) \mathrm{d}\mu^\dagger_{\text{post}} = 0,$$

where we applied the DCT with $2h(\cdot\,|y^\dagger)$ as a dominating function. □

2.4 Numerical illustrations

We illustrate some of the results shown in the previous sections with numerical examples. Firstly, we consider some simple one-dimensional examples complementing the examples we have considered throughout the chapter. Those include Bayesian inverse problems with likelihoods that are discontinuous in parameter or data. Secondly, we consider an inverse problem that is high-dimensional in terms of data and parameters. The high-dimensional inverse problem is concerned with the reconstruction of an image by Gaussian process regression.

2.4.1 Discontinuities in the likelihood

In some previous works, the Lipschitz continuity of the log-likelihood in the data and (at least) continuity in the parameter has been assumed; see [236]. In this thesis, we prove results that do not require continuity in the parameter, however, we still require continuity in the data. We now illustrate these results with simple numerical experiments. Indeed, we show that Assumption (A4) is crucial by comparing BIP posteriors with likelihoods that are continuous and discontinuous in the data.

Example 2.19 (Continuity of $y \mapsto L(y|\cdot)$)**.** We define data and parameter space by $Y := \mathbb{R}$ and $X := [0,1]$. We consider the BIPs with prior measure $\mu_{\text{prior}} := \text{Unif}(0,1)$ on X and one of the following likelihoods

(a) $L(y^\dagger|\theta) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\|y^\dagger - \theta\|_Y^2)$,

(b) $L(y^\dagger|\theta) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\|\lfloor y^\dagger \rfloor - \theta\|_Y^2)$.

Moreover, we assume that the parameter $\theta \sim \mu_{\text{prior}} := \text{Unif}(0,1)$ follows a uniform prior distribution. ◇

We solve the inverse problems in Example 2.19 with numerical quadrature. In particular, we compute the model evidences for $y^\dagger \in \{-5, -4.999, -4.998, \dots 5\}$ and the Hellinger distances between $\mu_{\text{post}}^\dagger$ and $\mu_{\text{post}}^\ddagger$, where $y^\ddagger = 1$. In Figure 2.3, we plot the likelihood functions at $\theta = 0$, the logarithms of the posterior densities, and the Hellinger distances. The top row in the figure refers to (a), the bottom row refers to (b). In the continuous setting (a), we see continuity with respect to $y^\dagger$ in all images. Indeed, the BIP in (a) fulfills (A1)-(A4). The inverse problem in (b) satisfies (A1)-(A3), but not (A4). Also, we see discontinuities with respect to the data in all figures referring to (b). Especially, the figure of the Hellinger distances is discontinuous which leads to the conclusion that this inverse problem is not well-posed. Hence, (A4) is indeed crucial to obtain well-posedness of a Bayesian inverse problem.

Remark 2.20. A likelihood as in Example 2.19(b) can arise, when considering cumulative or categorial data, rather than real-valued continuous data as in (a). Categorial data arises in classification problems. ◇

While continuity in the data is important, we now illustrate that continuity in the forward response operator is not necessary to obtain continuity in the data to posterior map. We give an example that can be understood as learning the bias in an artificial neural network. Recall Example 1.7 for the definition of ANNs.

Example 2.21 (Continuity in $\theta \mapsto L(\cdot\,|\theta)$)**.** We define data and parameter space by $Y := \mathbb{R}$ and $X := [0,1]$. Let $w \in [1, \infty]$ be a known weight parameter. We define the forward response operator with weight w by

$$\mathcal{G}_w : X \to Y, \qquad \theta \mapsto \frac{1}{1 + \exp(-w(0.5 - \theta))}.$$

If $w < \infty$, the forward response operator resembles a single layer artificial neural network with sigmoid activation function evaluated at 0.5. This ANN has known weight w and uncertain bias θ, see Example 1.7. Moreover, note that in the limiting setting $w = \infty$, the sigmoid function is there replaced by the Heaviside function with step at θ, evaluated also at $x = 0.5$. That is,

$$\mathcal{G}_\infty : X \to Y, \qquad \theta \mapsto \begin{cases} 1, & \text{if } 0.5 \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

We consider the BIP of estimating the true bias $\theta^\dagger$, given an observation $y_w^\dagger := \mathcal{G}_w(\theta^\dagger) + \eta^\dagger$. Here, we consider the noise $\eta^\dagger$ to be a realisation of $\eta \sim \text{N}(0, 1^2)$. Moreover, we assume that the parameter $\theta \sim \mu_{\text{prior}} = \text{Unif}(0,1)$ follows a uniform prior distribution. ◇

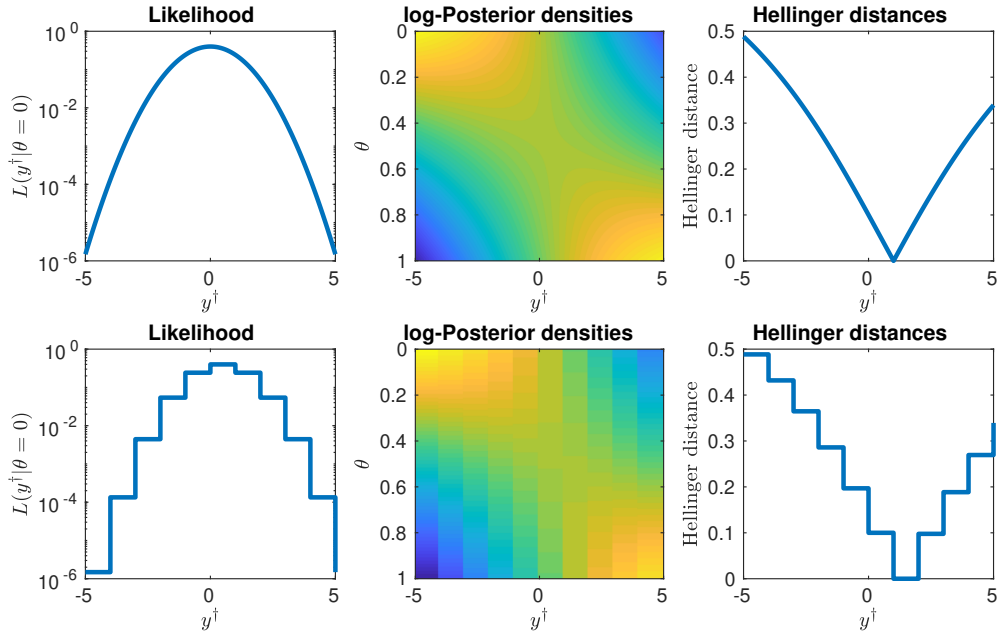


Figure 2.3. Posterior measures with likelihoods that are continuous and discontinuous in the data. Top row: Example 2.19(a), bottom row: Example 2.19(b). Left: Likelihood at $\theta = 0$. Centre: Log-posterior densities corresponding to the Bayesian inference problems. The colormaps show a descent in the posterior density, when going from yellow (high) to dark blue (low). Right: Hellinger distance between the posterior $\mu^{\ddagger}_{\text{post}}$ with $y^{\ddagger} = 1$ and posterior $\mu^{\dagger}_{\text{post}}$ with $y^{\dagger}$ varying between -5 and 5.

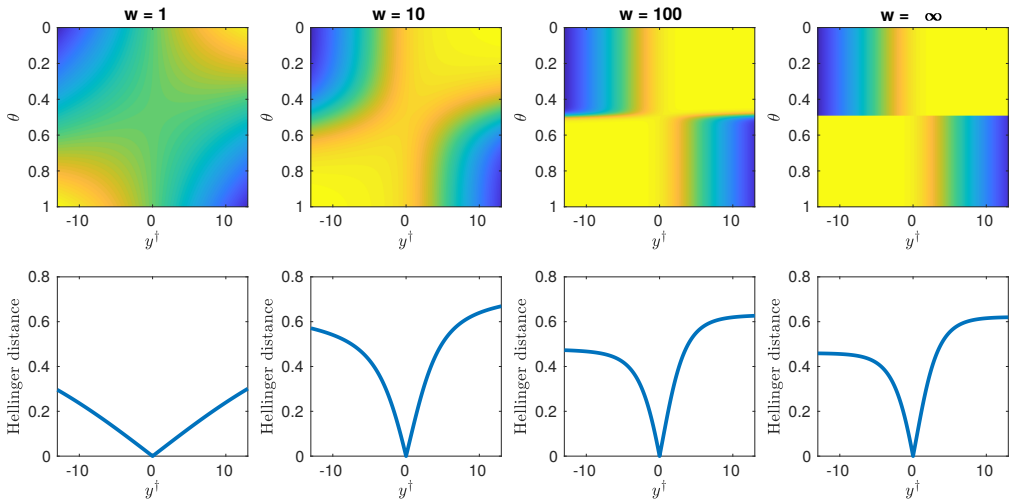


Figure 2.4. Posterior measures with likelihoods that are continuous and discontinuous in the parameter. From left to right: Example 2.21 given $w \in \{1, 10, 100, \infty\}$. Top row: Log-posterior densities corresponding to the Bayesian inference problems. The colormaps show a descent in posterior density, when going from yellow (high) to dark blue (low). Bottom row: Hellinger distance between the posterior $\mu^{\ddagger}_{\text{post}}$ with $y^{\ddagger} = 1$ and posterior $\mu^{\dagger}_{\text{post}}$ with $y^{\dagger}$ varying between -13 and 13.

We solve the BIPs in Example 2.21 with weights $w \in \{1, 10, 100, \infty\}$ again with numerical quadrature for $y^\dagger \in \{-13, -12.99, -12.98, \dots, 13\}$. We compute the Hellinger distance between $\mu^\dagger_{\text{post}}$ and $\mu^\ddagger_{\text{post}}$, where $y^\ddagger = 0$. We plot the logarithms of the posterior densities obtained in Example 2.21 in Figure 2.4, along with the Hellinger distances. We observe that all of the posteriors are continuous with respect to the data. This includes the posterior that is based on the discontinuous forward response operator $\mathcal{G}_\infty$. It is discontinuous in the parameter, but continuous in the data. The BIP considered here satisfy again Assumptions (A1)-(A4). Hence, also these numerical experiments verify the statement of Theorem 2.5.

Remark 2.22. In *deep learning*, sigmoid functions $\mathcal{G}_w$ ($w < \infty$) are considered as smooth approximations to the Heaviside function $\mathcal{G}_\infty$, which shall be used as an activation function. The smooth sigmoid functions allow to train the deep neural network with a gradient based optimisation algorithm. When training the neural network with a Bayesian approach, rather than an optimisation approach, we see that we can use Heaviside functions in place of smooth approximations and obtain a well-posed Bayesian inverse problem.
It is unclear, whether Heaviside activation functions are actually benificial. The Universal Approximation Theorem (see e.g. [53, Theorem 1]) discusses the approximability of continuous functions with neural networks and requires continuous activation functions. ◇

2.4.2 A high-dimensional inverse problem

We now consider an inverse problem that is high-dimensional in parameter and data space. In particular, we observe single, noisy pixels of a grayscale photograph. The inverse problem consists in the reconstruction of the image, for which we use Gaussian process regression. We then perturb the data by adding white noise to the image and investigate changes in the posterior, as we rescale the noise.

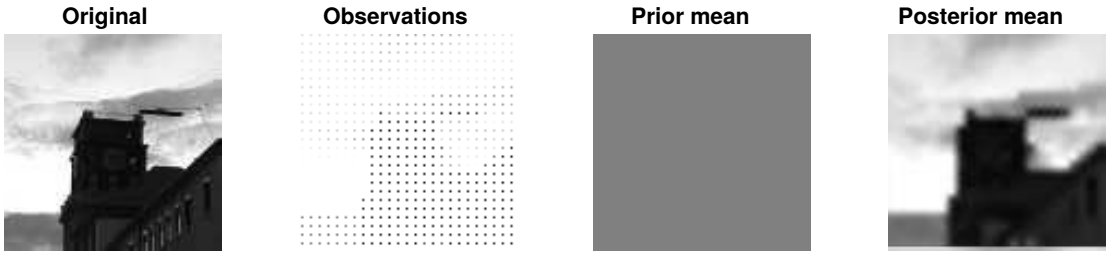


Figure 2.5. Reconstruction of an image with Gaussian process regression. From left to right: original image, observational data (white parts are unobserved), prior mean, and posterior mean.

Example 2.23. Let the parameter space $X := \mathbb{R}^{100\times 100}$ contain grayscale images of 100×100 pixels. The data space $Y := \mathbb{R}^{25\times 25}$ consists of 25×25 pixels that are observed in a single picture. Returning those 25×25 pixels from a 100×100 pixels image is modelled by the function $\mathcal{G} : X \to Y$. Let $\theta^\dagger \in X$ be a full image. Given

$$y^\dagger = \mathcal{G}(\theta^\dagger) + \eta,$$

we shall recover the full image $\theta^\dagger$. Here, $\eta \sim \mathrm{N}(0, 5^2 I)$ is normally distributed noise, with a noise level of about $5/\max(y) = 2\%$. We assume a Gaussian prior on X:

$$\mu_{\text{prior}} = \mathrm{N}\left(\begin{pmatrix} 128 & \cdots & 128 \\ \vdots & \ddots & \vdots \\ 128 & \cdots & 128 \end{pmatrix}, \mathcal{C}\right),$$

where $\mathcal{C} \in \mathbb{R}^{100\times 4}$ is a covariance tensor assigning the following covariances:

$$\mathrm{Cov}(\theta_{i,j}, \theta_{\ell,k}) = 10000 \cdot \exp\left(-\frac{\sqrt{(i-\ell)^2 + (j-k)^2}}{15}\right).$$

Note that this is a discretised version of an exponential covariance kernel for a Gaussian process in 2D space; see Example 1.26. ◇

The Bayesian inverse problem in Example 2.23 can be solved analytically since $\mathcal{G}$ is linear, and prior and noise are Gaussian. We obtain the posterior measure by Gaussian process regression. In Figure 2.5, we present the original image, observations, prior mean image, and posterior mean image. The reconstruction is rather coarse, which is not surprising given that we observe only 6.25E2 of 1E4 pixels of the image. We now investigate how the posterior measure changes under marginal changes in the data. To do so, we perturb the image additively with scaled white noise. In particular, we add $\mathrm{N}(0, \sigma^2)$-distributed, independent random variables to each pixel. In Figure 2.6, we show images and associated observations, where the standard deviations (StD) of the noise is $\sigma \in \{1, 10, 100\}$. Using Gaussian process regression,

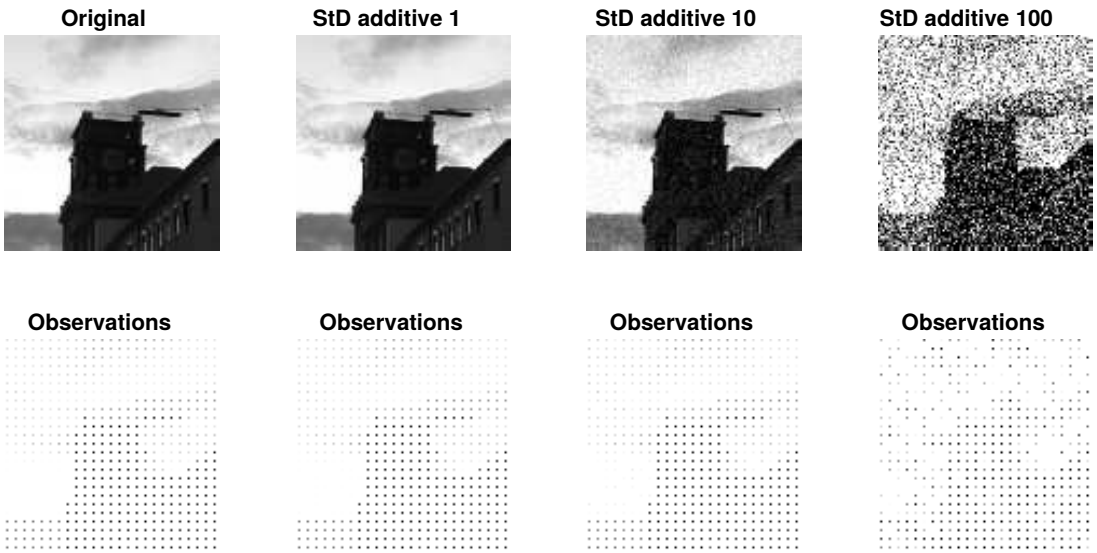


Figure 2.6. Original and perturbed images and data. Top row: Original image and images perturbed with scaled white noise, given $\sigma \in \{1, 10, 100\}$. Bottom row: Observations obtained from the perturbed image.

we compute the posteriors after perturbing the images with scaled white noise $\sigma \in \{1\mathrm{E}-17, 1\mathrm{E}-16, \ldots, 1\mathrm{E}2\}$. Between the original posterior with no perturbation in

the data and all others, we compute the Hellinger distance and the relative Frobenius distance between the (matrix-valued) posterior means, defined as

$$\text{Relative Frobenius distance} := \frac{\left\| \int_X \theta \mathrm{d}\mu^{\ddagger}_{\text{post}}(\theta) - \int_X \theta \mathrm{d}\mu^{\dagger}_{\text{post}}(\theta) \right\|_{\mathrm{F}}}{\left\| \int_X \theta \mathrm{d}\mu^{\ddagger}_{\text{post}}(\theta) \right\|_{\mathrm{F}}},$$

where $\mu^{\dagger}_{\text{post}}$ (respectively $\mu^{\ddagger}_{\text{post}}$) is the posterior referring to the perturbed data $y^{\dagger}$ (respectively non-perturbed data $y^{\ddagger}$). Since the perturbation is random, we perform

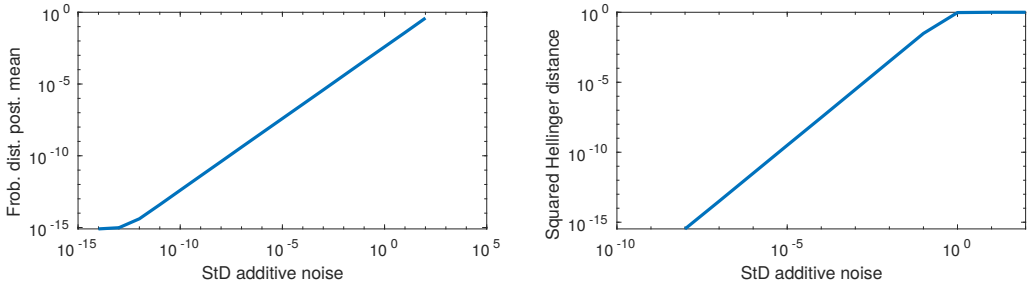


Figure 2.7. Mean relative Frobenius distances and mean squared Hellinger distances computed between the posteriors $\mu^{\dagger}_{\text{post}}$ and posteriors $\mu^{\ddagger}_{\text{post}}$ in which the underlying image was perturbed with white noise that has been scaled by StD $\sigma = 0, 1\text{E-}17, 1\text{E-}16, \ldots, 1\text{E}2$. 'Mean' refers to the fact that the perturbations are random and the distances have been computed for 20 random perturbations and then averaged. When $\sigma \to 0$, the distances go to 0. For $\sigma < 10^{-8}$ (respectively $\sigma < 10^{-15}$) the distances are equal to zero up to machine precision and are not shown in the plot on the right-hand side (respectively left-hand side).

this process 20 times and compute the mean of these distances. The standard deviation in these metrics is negligibly small. We plot the results in Figure 2.7, where we see indeed continuity w.r.t. the standard deviation of the noise.
In light of Corollary 2.6, this is what we expect: First, note that the Bayesian inverse problem falls in the category *additive finite-dimensional Gaussian noise* and is therefore well-posed. Hence, also in this high-dimensional setting, we are able to verify our analytical results concerning well-posedness.

Chapter 3

Discretisation

> **Problema 36.** Formulae integralis cujuscunque $y = \int X \mathrm{d}x$ valorem vero proxime indagare.
>
> Leonhard Euler [85, p. 178]

Up to now, we have discussed mathematical models, the representation of uncertainties, statistical estimation, well-posedness of inverse problems, and Bayesian inverse problems. All of these discussions have taken place on a metalevel. Aside from *toy examples* in §2.4, we did not actually approach a Bayesian inverse problem computationally. In this chapter, we change this and consider computational strategies for BIPs.
A computational solution of a Bayesian inverse problem contains various subtasks: First, we need to approximate the mathematical model G respectively $\mathcal{E}$. Second, if the uncertain parameter is represented by a random field, this random field has to be discretised as well. Finally, the posterior measure, which is the solution of the Bayesian inverse problem, has to be discretised. In §§3.1-3.3, we discuss these three topics respectively. Then, in §3.4, we review techniques used in forward and inverse uncertainty quantification to increase the efficiency of the strategies discussed in the sections before.

3.1 Discretisation of mathematical models

In this section, we focus on the discretisation of linear elliptic PDEs. Doing so, we give brief explanations of the Galerkin approach. For a more thorough discussion, we refer to classical textbooks, such as [30, 206]. Concerning the discretisation of models that are not based on linear elliptic PDEs, we refer to, e.g. Stoer and Bulirsch [235] (ODEs), Bartels [12] (nonlinear PDEs), and Lord et al. [170] (stochastic differential equations and stochastic partial differential equations).
Recall our setting of non-parametric models, given by an implicit formula; see (1.1). Hence, a model is a function $u^* \in H$, where H is a Hilbert space of functions mapping from D to $\mathbb{R}$. The model is defined implicitly as the unique solution $u^* \in H$, such that

$$\mathcal{E}(u^*, v) = 0 \text{ for all } v \in H',$$

where H' is another Hilbert space, and $\mathcal{E} : H \times H' \to \mathbb{R}$ is a continuous and bounded operator. In the following, we assume that $\mathcal{E}$ is affine linear in the first component

and linear in the second component. Finding $u^* \in H$ analytically is often impossible. Moreover, since, H, H' are often infinite-dimensional, numerical algorithms cannot easily approach this problem either.
For these reasons we consider a Galerkin approach to approximate the solution u^*. Let H_ℓ and H'_ℓ be two finite dimensional Hilbert spaces, with

$$H_\ell \subseteq H, \qquad H'_\ell \subseteq H'.$$

The subscript ℓ denotes the accuracy of H_ℓ and H'_ℓ, respectively. We typically see the spaces H_ℓ and H'_ℓ as part of an increasing sequence of spaces. Hence, increasing ℓ leads to an at least as accurate representation of functions in H respectively H' by functions in H_ℓ respectively H'_ℓ.
The Galerkin approximation $u^*_\ell \in H_\ell$ satisfies

$$\mathcal{E}(u^*_\ell, v_\ell) = 0 \text{ for all } v_\ell \in H'_\ell.$$

Hence, we replace the model and the test space by finite-dimensional spaces. Moreover, we assume that such a u^*_ℓ exists and that it is unique. The problem of finding u^*_ℓ simplifies to a finite-dimensional linear system of equations: Let $N_\ell := \dim(H_\ell)$ and $N'_\ell := \dim(H'_\ell)$. Moreover, let $B_\ell := (\varphi_1, \ldots, \varphi_{N_\ell})$ be a basis of H_ℓ and $B'_\ell := (\varphi'_1, \ldots, \varphi'_{N'_\ell})$ be a basis of H'_ℓ respectively. We can now represent elements $u_\ell \in H_\ell$ by

$$u_\ell := \sum_{i=1}^{N_\ell} c_i \varphi_i,$$

where $c \in \mathbb{R}^{N_\ell}$ is a vector. In the same fashion, we can represent $v_\ell \in H'_\ell$ with elements in B'_ℓ. Using this representation, we obtain

$$\begin{aligned}\mathcal{E}(u_\ell, v_\ell) &= \mathcal{E}\left(\sum_{i=1}^{N_\ell} c_i \varphi_i, \sum_{j=1}^{N'_\ell} c'_j \varphi'_j\right) \\ &= \sum_{j=1}^{N'_\ell} c'_j \left(\mathcal{E}\left(\sum_{i=1}^{N_\ell} c_i \varphi_i, \varphi'_j\right) - \mathcal{E}(0, \varphi'_j) + \mathcal{E}(0, \varphi'_j)\right).\end{aligned}$$

Note that $u_\ell \mapsto \mathcal{E}(u_\ell, v_\ell) - \mathcal{E}(0, v_\ell)$ is linear. Hence,

$$\begin{aligned}\mathcal{E}(u_\ell, v_\ell) = 0 &\Leftrightarrow \left(\sum_{i=1}^{N_\ell} c_i \left(\mathcal{E}\left(\varphi_i, \varphi'_j\right) - \mathcal{E}(0, \varphi'_j)\right)\right) = -\mathcal{E}(0, \varphi'_j) \quad (j = 1, \ldots, N'_\ell) \\ &\Leftrightarrow \mathcal{E}_\ell c = b_\ell,\end{aligned}$$

where $\mathcal{E}_\ell \in \mathbb{R}^{N'_\ell \times N_\ell}$ and $b_\ell \in \mathbb{R}^{N'_\ell}$, are given by

$$(\mathcal{E}_\ell)_{j,i} = \mathcal{E}(\varphi_i, \varphi'_j) - \mathcal{E}(0, \varphi'_j), \quad (b_\ell)_j = -\mathcal{E}(0, \varphi'_j) \qquad (i = 1, \ldots, N_\ell, j = 1, \ldots, N'_\ell).$$

Hence, we obtain $u^*_\ell \in H_\ell$ by solving $\mathcal{E}_\ell c = b_\ell$ numerically with respect to c and then set $u^*_\ell := B_\ell c$. The matrix $\mathcal{E}_\ell$ is called *system matrix*.

Elliptic equations. The presented Galerkin approach is very natural for elliptic equations, like the Poisson and the diffusion equation. These we have discussed in Examples 1.2 and 1.4. There we set $H := H' := \mathbf{H}_0^1(D)$, and

$$\mathcal{E}(u, v; \theta) := \int_D \exp(\theta(x))\langle \nabla u(x), \nabla v(x)\rangle_D \mathrm{d}x - \int_D f(x)v(x)\mathrm{d}x \qquad (u, v \in H)$$

for the diffusion equation and

$$\mathcal{E}(u, v) := \mathcal{E}(u, v; 0) \qquad (u, v \in H)$$

for the Poisson equation. The operators $\mathcal{E}$ in these examples are affine linear in the first and linear in the second component.

When introducing the weak formulations, we mentioned that one can apply the Lax-Milgram theorem to show that the weak formulations have a unique solution. If indeed $H_\ell = H'_\ell \subseteq H$, we can apply again the Lax–Milgram theorem and obtain the unique solvability also for the discretised problem.

Finite elements. *How do we choose B_ℓ and B'_ℓ?* Note that properties of the system matrix $\mathcal{E}_\ell$ depend heavily on the choice of bases B_ℓ and B'_ℓ. A popular class of basis functions are the so-called *finite elements*. Finite elements are piecewise polynomials supported on small bounded sets. The union of these supports covers the space D. The locally supported functions in combination with a local differential operator often lead to a sparse system matrix. Linear systems with sparse system matrices can be solved computationally fast. Examples for finite elements are listed in, e.g. [30, II §5].

3.2 Discretisation of random fields

Consider a Gaussian random field $\mathrm{N}(m, \mathcal{C})$ on $(X, \mathcal{B}X)$, where X is now a separable Hilbert space. For practical computations the possibly infinite-dimensional parameter space X must be discretised. First, in §3.2.1, we discuss random field discretisation via a truncation of the Karhunen–Loève expansion. For details on the KL expansion and its truncation, we refer to [96, 139, 169, 238]. Moreover, see e.g. the works [22, 48, 140, 204, 222, 230, 260] for the computational strategies to obtain the KL expansion. One of these computational strategies is a Galerkin approach to approximate the eigenfunctions, which proceeds similarly as the strategy introduced in §3.1.

Instead of solving the KL eigenproblem with a Galerkin approach, we can also immediately use Galerkin to approximate a random field; see, e.g. [168]. Here, we consider the covariance operator instead of an elliptic operator and we aim to compute a Cholesky or spectral decomposition of the operator, rather than solving a linear equation. Given, e.g. a Galerkin discretisation, we can identify the random field with a finite-dimensional Gaussian random vector. In §3.2.2, we discuss the finite element discretisation and the sampling of finite-dimensional Gaussian random vectors.

3.2.1 Karhunen–Loève expansion

We have introduced the Karhunen–Loève expansion in Definition 1.24 as a way to construct a Gaussian measure in infinite dimensions. We can use this framework also to discretise a Gaussian random field by truncating it after N_{sto} terms. We define

$$\boldsymbol{\theta}_{\text{KL}} := m + \sum_{i=1}^{\infty} \sqrt{\alpha_i}\boldsymbol{\xi}_i\psi_i$$

where $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots \sim \text{N}(0,1)$ are i.i.d. standard Gaussian random variables, and (α_i, ψ_i) are eigenpairs of $\mathcal{C}$ ordered descendantly in α_i. The familiy $(\psi_i)_{i=1}^{\infty}$ forms an orthonormal basis of X. Let now $N_{\text{sto}} \in \mathbb{N}$ and let

$$\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}} := m + \sum_{i=1}^{N_{\text{sto}}} \sqrt{\alpha_i}\boldsymbol{\xi}_i\psi_i,$$

be the random field obtained after truncating the expansion above. The sum of the remaining eigenvalues in the truncated KL expansion gives the following $\mathbf{L}^2$-error bound:

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{\text{KL}} - \boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}\|_X^2\right] = \sum_{i=N_{\text{sto}}+1}^{\infty} \alpha_i; \tag{3.1}$$

see e.g. the proof of [238, Theorem 11.4]. Furthermore, the truncated KL expansion $\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}$ solves the minimisation problem

$$\min_{\widehat{\theta} \in \mathbf{Lin}(\mathbb{R}^{N_{\text{sto}}}; X)} \mathbb{E}\left[\|\boldsymbol{\theta}_{\text{KL}} - \widehat{\theta}(\boldsymbol{\xi})\|_X^2\right], \tag{3.2}$$

for any given $N_{\text{sto}} \in \mathbb{N}$. Hence, the truncated KL expansion $\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}$ is the optimal N_{sto}-dimensional affine linear function which approximates $\boldsymbol{\theta}_{\text{KL}}$ using i.i.d. standard Gaussian random variables, see [96, §2.3.2].
Unfortunately, the eigenpairs $(\alpha_i, \psi_i)_{i=1}^{\infty}$ of the covariance operator $\mathcal{C}$ are often not analytically accessible. However, it may be possible to compute the eigenpairs with a Galerkin approach. This is one of the options which we discuss in the following subsection.

3.2.2 Finite element discretisation

We now discuss the approximation of a random field with finite elements. Let $\text{B}_\ell := (\varphi_i : i = 1, \ldots, N_\ell) \in X^{N_\ell}$ denote an N_ℓ-dimensional basis of a finite element space. We approximate X by $X_\ell := \text{span}(\text{B}_\ell)$. Note that we can identify $X_\ell \cong \mathbb{R}^{N_\ell}$. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product on $\mathbb{R}^{N_\ell}$. Note that $\mathbb{R}^{N_\ell}$ is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{M_\ell} = \langle \cdot, M_\ell \cdot \rangle$, where

$$M_\ell = (\langle \varphi_i, \varphi_j \rangle_X : i, j = 1, ..., N_\ell)$$

is the Gramian matrix associated with the finite element basis B_ℓ. The Gaussian measure $\text{N}(m, \mathcal{C})$ on $(X, \mathcal{B}X)$ can then be approximated on $\mathbb{R}^{N_\ell}$ by the measure $\text{N}(m_\ell, \mathcal{C}_\ell)$ with mean vector $m_\ell := (\langle \varphi_i, m \rangle_X : i = 1, ..., N_\ell)$ and covariance matrix

$$\mathcal{C}_\ell := (\langle \varphi_i, \mathcal{C}\varphi_j \rangle_X : i, j = 1, ..., N_\ell)\,.$$

In the following, we set $X_\ell := \mathbb{R}^{N_\ell}$ with inner product $\langle \cdot, \cdot \rangle_{M_\ell}$. Moreover, we assume that the Gaussian measure $\mathrm{N}(m_\ell, \mathcal{C}_\ell) \in \mathrm{Prob}(X_\ell)$ has a mean equal to zero, that is $m_\ell \equiv 0$. The covariance operator $\mathcal{C}_\ell \in \mathbb{R}^{N_\ell \times N_\ell}$ is now a matrix. We next discuss sampling from $\mathrm{N}(0, \mathcal{C}_\ell)$.
A simple sampling strategy uses the Cholesky decomposition $LL^* = \mathcal{C}_\ell$. Let $\boldsymbol{\xi} \sim \mathrm{N}(0, \mathrm{Id}_{X_\ell})$. Then, it is easy to see that $L\boldsymbol{\xi} \sim \mathrm{N}(0, \mathcal{C}_\ell)$. The computational cost of a Cholesky decomposition is $O(N_\ell^3; N_\ell \to \infty)$, see [104, Algorithm 4.2.1].
Alternatively, we can use a spectral decomposition of $\mathcal{C}_\ell$, which corresponds to a discrete KL expansion. We obtain

$$\sum_{i=1}^{N_\ell} \sqrt{\alpha_i} \boldsymbol{\xi}_i \psi_i \sim \mathrm{N}(0, \mathcal{C}),$$

where $(\alpha_i, \psi_i)_{i=1}^{N_\ell}$ are generalised eigenpairs satisfying

$$\mathcal{C}_\ell \psi_i = \alpha_i M_\ell \psi_i \qquad (i = 1, \ldots, N_\ell). \tag{3.3}$$

Note that the eigenvectors form an orthonormal basis of the space $(X_\ell, \langle \cdot, \cdot \rangle_\ell)$. Note also that if $\mathcal{C}_\ell$ was obtained by a Galerkin approximation of $\mathcal{C}$, this corresponds to a Galerkin approximation of the KL expansion.
Computing the spectral decomposition of a symmetric matrix is typically more expensive compared to the Cholesky decomposition, even though both problems have a cubic cost in N_ℓ. Compare, e.g. the cost of the Cholesky decomposition [104, Algorithm 4.2.1] ($N_\ell^3/3$ flops) and the cost of computing a Schur decomposition [104, Algorithm 8.3.3] ($4N_\ell^3/3$ flops). From the Schur decomposition, one can compute the spectral decomposition in less than cubic time.
However, the spectral decomposition is of further use. In §3.2.1, we discuss a truncated KL expansion as a technique to reduce the dimension of X from infinity to a finite number. Similarly, we can use the KL for dimension reduction from a high-dimensional to a low-dimensional stochastic space. Let $N_{\mathrm{sto}} \in \mathbb{N}, N_{\mathrm{sto}} < N_\ell$ and consider the truncated KL expansion

$$\boldsymbol{\theta}_{\mathrm{KL}}^{N_{\mathrm{sto}}} := \sum_{i=1}^{N_{\mathrm{sto}}} \sqrt{\alpha_i} \boldsymbol{\xi}_i \psi_i.$$

As in (3.1), the truncation error can again be represented by the sum of the remaining eigenvalues in the sequence:

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{\mathrm{KL}} - \boldsymbol{\theta}_{\mathrm{KL}}^{N_{\mathrm{sto}}}\|_X^2 \right] = \sum_{i=N_{\mathrm{sto}}+1}^{N_\ell} \alpha_i.$$

We have mentioned that the spectral decomposition of a matrix is typically more expensive than a Cholesky decomposition. However, when using a truncated KL expansion on the finite dimensional space, we only need to compute the leading N_{sto} eigenpairs. This reduces the computational cost to $O(N_\ell^2 N_{\mathrm{sto}}; N_\ell \to \infty)$ when using an *implictly restarted Lanczos method*, see [33].

Remark 3.1. If $N_\ell \in \mathbb{N}$ is large, computational costs of $O(N_{\mathrm{sto}} N_\ell^2)$ and $O(N_\ell^3)$ may be prohibitively expensive. There are more efficient methods to sample high-dimensional Gaussian random vectors.

The *circulant embedding method* may be viable for sampling in $O(N_\ell \log N_\ell; N_\ell \to \infty)$ if the covariance is discretised on a regular grid in 1D, 2D, or 3D, and given by a stationary covariance function. It uses the *Fast Fourier Transform* and has been discussed by [11, 34, 69, 105]. Other than that, covariance matrices can be represented as *hierarchical matrices* or in other low-rank formats to speed-up computations down to a cost of $O(N_\ell \log N_\ell; N_\ell \to \infty)$, see, e.g. [15, 37, 65, 90, 140].
More recently, the so called *stochastic-partial-differential-equation-based sampling* has been developed in the works [28, 29, 141, 193, 194, 215]. The major idea is to generate samples of Gaussian fields with Matérn covariance operators by solving a certain discretised fractional PDE with white noise forcing. If the fractional PDE can be represented by a sparse system matrix, this can lead to a computational cost $O(N_\ell; N_\ell \to \infty)$. Last we mention the *adapative cross approximation* or *pivoted Cholesky decomposition*, which leads to an incomplete Cholesky decomposition. It has been studied in [115, 224]. ◇

3.3 Monte Carlo methods

We now discuss the approximation of probability measures with so-called *Monte Carlo methods*. These methods will later be used to approximate posterior measures of Bayesian inverse problems. As main resources concerning Monte Carlo methods, we refer to [167, 213, 217].
Monte Carlo methods are fairly old. They date back to, e.g. Metropolis and Ulam in 1949 [184], as well as to von Neumann in 1951 [254].
We start with *standard Monte Carlo* in §3.3.1, where we assume that we can sample independently from the probability measure that is approximated. This is one of the most basic strategies in forward uncertainty quantification. In Bayesian inverse problems however, independent posterior samples are typically not accessible. Alternatively, we discuss *Importance Sampling* in §3.3.2 and *Markov chain Monte Carlo* (MCMC) in §3.3.3. In Importance Sampling, we sample from a different measure and reweigh the samples to obtain an approximation of the posterior. In MCMC, we sample *dependently*.

3.3.1 Standard Monte Carlo

Let $\mu \in \mathrm{Prob}(X)$ and $Q : X \to \mathbb{R}$ be a function in $\mathbf{L}^2(X, \mu)$. Fundamentally, Monte Carlo methods aim to approximate integrals of the form $\int_X Q(\theta)\mu(\mathrm{d}\theta)$. To this end, the Weak or Strong Law of Large Numbers shall be applied; see Theorem 1.12 and Theorem 1.13:
We produce $N_{\mathrm{smp}} \in \mathbb{N}$ samples $Q(\boldsymbol{\theta}_1), \dots, Q(\boldsymbol{\theta}_{N_{\mathrm{smp}}})$, where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu$ i.i.d. Then, we approximate the integral by the sample mean:

$$\overline{Q(\boldsymbol{\theta})}_{N_{\mathrm{smp}}} := \frac{1}{N_{\mathrm{smp}}} \sum_{n=1}^{N_{\mathrm{smp}}} Q(\boldsymbol{\theta}_n) \approx \int_X Q(\theta)\mu(\mathrm{d}\theta).$$

This is indeed justified, since Theorem 1.13 implies that

$$\overline{Q(\boldsymbol{\theta})}_{N_{\mathrm{smp}}} \to \int_X Q(\theta)\mu(\mathrm{d}\theta) \qquad (\mathbb{P}\text{-a.s.}, N_{\mathrm{smp}} \to \infty). \tag{3.4}$$

(3.4) implies that $\overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}}$ is a *consistent estimator*. Moreover, by integrating over $\overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}}$ w.r.t. $\mathbb{P}$, one can show that the estimator is *unbiased*. To determine the quality of the Monte Carlo estimation, we can use the formula for the root mean square error in (1.6). It reads

$$\mathbb{E}\left[\left(\overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}} - \int_X Q(\theta)\mu(\mathrm{d}\theta)\right)^2\right]^{1/2} = \frac{\mathrm{StD}(Q(\boldsymbol{\theta}_1))}{\sqrt{N_{\text{smp}}}}. \tag{3.5}$$

So far, we have used Monte Carlo to approximate integrals. Actually, we aim to approximate *measures*. To this end, note that we can interpret $\overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}}$ as an integral with respect to a certain discrete measure. This discrete measure is given by the following *empirical measure*

$$\widehat{\boldsymbol{\mu}} := \frac{1}{N_{\text{smp}}} \sum_{n=1}^{N_{\text{smp}}} \delta(\cdot - \boldsymbol{\theta}_n). \tag{3.6}$$

Note that this measure is a random measure, since $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\text{smp}}}$ are random variables. $\widehat{\boldsymbol{\mu}}$ is indeed well-defined as a random measure, we show this in Proposition 5.7. Applying the Strong Law of Large Numbers – Theorem 1.13 – we can show that $\widehat{\boldsymbol{\mu}}$ converges to μ, as $N_{\text{smp}} \to \infty$. We specify this result in the following proposition.

Proposition 3.2. Let $\mu \in \mathrm{Prob}(X)$ and $\widehat{\boldsymbol{\mu}}$ be the measure in (3.6), based on $N_{\text{smp}} \in \mathbb{N}$ samples from μ. Then,

$$\mathbb{P}\left(\widehat{\boldsymbol{\mu}} \xrightarrow{w} \mu, \text{ as } N_{\text{smp}} \to \infty\right) = 1.$$

Proof. Let $Q \in \mathbf{C}^0(X)$ and bounded. Then, $Q \in \mathbf{L}^2(X, \mu)$. Theorem 1.13 implies that

$$\mathbb{P}\left(\lim_{N_{\text{smp}}\to\infty} \int_X Q(\theta)\widehat{\boldsymbol{\mu}}(d\theta) = \int_X Q(\theta)\mu(d\theta)\right) = 1,$$

since $\int_X Q(\theta)\widehat{\boldsymbol{\mu}}(d\theta) = \overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}}$. Moreover, since this result is independent of the choice of Q, $\lim_{N_{\text{smp}}\to\infty} \int_X Q(\theta)\widehat{\boldsymbol{\mu}}(d\theta) = \int_X Q(\theta)\mu(d\theta)$ implies weak convergence of $\widehat{\boldsymbol{\mu}} \to \mu$. □

Proposition 3.2 refers to a convergence with probability one of a sequence of random variables to a deterministic value. The random variables take values in a space of probability measures and converge to a probability measure. The probability measures in this sequence converge in the weak topology.
Next, we consider the distance between μ and $\widehat{\boldsymbol{\mu}}$. This will allow us to estimate the error between the original and the approximate measure. Note that the distances in §1.2.2 are not applicable, since $\widehat{\boldsymbol{\mu}}$ is a random measure. Therefore, we introduce the *expected total variation distance* between two random measures $\boldsymbol{\mu}, \boldsymbol{\mu}' : \Omega \to \mathrm{Prob}(X)$. It is defined by

$$\mathrm{d}_{\text{etv}}(\boldsymbol{\mu}, \boldsymbol{\mu}') := \sup_{A \in \mathcal{B}X} \mathbb{E}[|\boldsymbol{\mu}(A) - \boldsymbol{\mu}'(A)|].$$

Alternatively, we could use the RMSE estimate in (3.5) to measure the distance between μ and $\widehat{\boldsymbol{\mu}}$, and then take the supremum over bounded functions $Q : X \to \mathbb{R}$. This has been done, for instance, in [5, 20]. We take the expected total variation distance to be consistent with the discussion in §1.2.2 and §2.3. For $\mu, \widehat{\boldsymbol{\mu}}$, we obtain the following error estimate in d_{etv}.

Proposition 3.3. Let $\mu \in \mathrm{Prob}(X)$ and $\widehat{\boldsymbol{\mu}}$ be the random measure in (3.6), based on $N_{\mathrm{smp}} \in \mathbb{N}$ samples from μ. Then,

$$\mathrm{d}_{\mathrm{etv}}(\widehat{\boldsymbol{\mu}}, \mu) \leq \frac{2}{\sqrt{N_{\mathrm{smp}}}}.$$

Proof. This is a result of the more general Proposition 3.5, where we set $\gamma \equiv 1$. □

3.3.2 Importance Sampling

Let $\mu_0 \in \mathrm{Prob}(X)$. We aim to approximate a measure $\mu_1 \in \mathrm{Prob}(X, \mu_0)$, where we have access to the density

$$\gamma(\theta) \propto \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}(\theta) \qquad (\theta \in X, \mu_0\text{-a.s.})$$

and to independent samples

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_0.$$

Note that we assume to know the μ_0-pdf γ only up to a normalising constant. μ_0 is called *importance measure*. Now, we aim to use the samples to approximate μ_1. We consider the following identity:

$$\int_X Q(\theta)\mu_1(\mathrm{d}\theta) = \frac{\int_X \gamma(\theta)Q(\theta)\mu_0(\mathrm{d}\theta)}{\int_X \gamma(\theta)\mu_0(\mathrm{d}\theta)}. \tag{3.7}$$

It implies that we can represent integrals with respect to μ_1 by a ratio of integrals with respect to μ_0. Hence, using Monte Carlo, we can indeed use μ_0-distributed samples to approximate integrals with respect to μ_1. This idea has been mentioned for instance by Metropolis et al. [183]. For a more recent review, we refer to Agapiou et al. [5].

Remark 3.4. We briefly note how to understand the setting above in the Bayesian framework. In Bayesian inverse problems, we apply Importance Sampling to approximate the posterior measure $\mu_1 := \mu^\dagger_{\mathrm{post}}$ using samples from the prior $\mu_0 := \mu_{\mathrm{prior}}$. The density $\gamma := L(y^\dagger|\cdot)$ is given by the likelihood. The relationship

$$\frac{\mathrm{d}\mu^\dagger_{\mathrm{post}}}{\mathrm{d}\mu_{\mathrm{prior}}} \propto L(y^\dagger|\cdot)$$

is a result of Theorem 1.45. ◇

We move on to the Importance Sampling approximation. We assume that $\gamma \in \mathbf{L}^2(X, \mu_0)$. Moreover, let $Q : X \to \mathbb{R}$ be a quantity of interest, given such that $(Q \cdot \gamma) \in \mathbf{L}^2(X, \mu_0)$. Then,

$$\frac{\frac{1}{N_{\mathrm{smp}}}\sum_{n=1}^{N_{\mathrm{smp}}} \gamma(\boldsymbol{\theta}_n)Q(\boldsymbol{\theta}_n)}{\frac{1}{N_{\mathrm{smp}}}\sum_{n=1}^{N_{\mathrm{smp}}} \gamma(\boldsymbol{\theta}_n)} = \frac{\sum_{n=1}^{N_{\mathrm{smp}}} \gamma(\boldsymbol{\theta}_n)Q(\boldsymbol{\theta}_n)}{\sum_{n=1}^{N_{\mathrm{smp}}} \gamma(\boldsymbol{\theta}_n)} \approx \int_X Q(\theta)\mu_1(\mathrm{d}\theta).$$

This is again a result of the Strong Law of Large Numbers, which implies that denominator and numerator converge to the respective integrals in (3.7). Note that

this again implies that the estimator is *consistent*. However, as opposed to the standard Monte Carlo estimator, the Importance Sampling estimator is *biased*, see, e.g. [5, Theorem 2.1].
Similarly to (3.6), we can use Importance Sampling to directly approximate the probability measure μ_1. We have

$$\widehat{\boldsymbol{\mu}}_1 := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{\gamma(\boldsymbol{\theta}_n)}{\sum_{m=1}^{N_{\mathrm{smp}}} \gamma(\boldsymbol{\theta}_m)} \delta(\cdot - \boldsymbol{\theta}_n),$$

which is again a random measure, see Proposition 5.7. We can again compute an error bound in terms of the expected total variation distance $\mathrm{d}_{\mathrm{etv}}$.

Proposition 3.5. Let $\mu_0 \in \mathrm{Prob}(X)$, $\mu_1 \in \mathrm{Prob}(X, \mu_0)$, and $\gamma \propto \mathrm{d}\mu_1/\mathrm{d}\mu_0$, with $\gamma \in \mathbf{L}^2(X, \mu_0)$. Moreover, let $\widehat{\boldsymbol{\mu}}_1$ be the Importance Sampling estimate given above. Then,

$$\mathrm{d}_{\mathrm{etv}}(\mu_1, \widehat{\boldsymbol{\mu}}_1) \leq 2 \cdot \sqrt{\frac{\rho}{N_{\mathrm{smp}}}},$$

where $\rho := \int_X \gamma^2(\theta)\mu_0(\mathrm{d}\theta) / \left(\int_X \gamma(\theta)\mu_0(\mathrm{d}\theta)\right)^2$.

Proof. Let $A \in \mathcal{B}X$. Then, by Jensen's inequality and [5, Theorem 2.1], we have

$$\left(\mathbb{E}\left[|\mu_1(A) - \widehat{\boldsymbol{\mu}}_1(A)|\right]\right)^2 \leq \mathbb{E}\left[(\mu_1(A) - \widehat{\boldsymbol{\mu}}_1(A))^2\right] \leq \frac{4\rho}{N_{\mathrm{smp}}}.$$

By taking the square root on both sides and taking the supremum over $A \in \mathcal{B}X$, we obtain the desired result. □

To assess the quality of an Importance Sampling estimate, we often consider the *effective sample size* (ESS). Consider some Importance Sampling estimate $\widehat{\boldsymbol{\mu}}_1$. The effective sample size

$$\mathrm{ESS} := N_{\mathrm{smp}}/\rho \tag{3.8}$$

is used to compare the Importance Sampling estimate $\widehat{\boldsymbol{\mu}}_1$ of μ_1 with a standard Monte Carlo estimate of μ_1. In particular, note that according to Proposition 3.5 the Importance Sampling error $\mathrm{d}_{\mathrm{etv}}(\widehat{\boldsymbol{\mu}}_1, \mu_1)$ is bounded above by $2/\sqrt{\mathrm{ESS}}$. This is the same error bound we obtain in Proposition 3.3 when sampling $N'_{\mathrm{smp}} := \mathrm{ESS}$ times independently from μ_1 and using these samples to approximate μ_1. Hence, the effective sample size is the number of independent samples we would need in a standard Monte Carlo estimation to obtain the same error as in the Importance Sampling estimate $\widehat{\boldsymbol{\mu}}_1$.

3.3.3 Markov chain Monte Carlo

In Monte Carlo, we sample independently from the correct probability measure. In Importance Sampling, we sample independently from an auxiliary probability measure and correct the samples with weights. Now we consider Markov chain Monte Carlo (MCMC). In MCMC, we drop the independence and sample dependently from – eventually – the correct probability measure. Those dependent samples form a Markov chain. In the next paragraphs, we discuss basic properties of Markov chains. For more details, we refer to the textbooks [185, 213].

Definition 3.6 (Markov chain)**.** Let $\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_t)_{t=1}^{\infty}$ be a sequence of random variables, each of which takes values in X. $\underline{\boldsymbol{\theta}}$ is a *Markov chain* if for all $t \in \mathbb{N}, t \geq 2$, and $\theta_1, \ldots, \theta_{t-1} \in X$, we have: $\mathbb{P}(\boldsymbol{\theta}_t \in \cdot \,|\boldsymbol{\theta}_{t-1} = \theta_{t-1}, \boldsymbol{\theta}_{t-2} = \theta_{t-2}, \ldots, \boldsymbol{\theta}_1 = \theta_1) = \mathbb{P}(\boldsymbol{\theta}_t \in \cdot \,|\boldsymbol{\theta}_{t-1} = \theta_{t-1})$. ◇

This means, the conditional probability distribution of the state $\boldsymbol{\theta}_t$ given all the other states is identical to the conditional probability distribution of $\boldsymbol{\theta}_t$ given only the last state at $t-1$. Hence, the distribution of the chain is independent of what happened before $t-1$. This property is called *Markov property*. While being semantically similar, note that this property does not imply that $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_s$ are independent, for $|s-t| \geq 2$.
The Markov property implies that if some of the past states of the Markov chain are known, we only need the last of those to describe the conditional probability measure: for $t \in \mathbb{N}, t \geq 2$, $U \subseteq \{1, \ldots, t-1\}$, and $(\theta_s)_{s \in U} \in X^U$, we have

$$\mathbb{P}\left(\boldsymbol{\theta}_t \in \cdot \,|\boldsymbol{\theta}_s = \theta_s, s \in U\right) = \mathbb{P}(\boldsymbol{\theta}_t \in \cdot \,|\boldsymbol{\theta}_{(\max U)} = \theta_{(\max U)}).$$

Markov chains whose conditional probability distributions are invariant with respect to shifts in the index t are called *(time-)homogeneous*. Otherwise, the Markov chain is *inhomogeneous*. More particularly, those satisfy

$$\mathbb{P}(\boldsymbol{\theta}_2 \in \cdot \,|\boldsymbol{\theta}_1 = \theta) = \mathbb{P}(\boldsymbol{\theta}_t \in \cdot \,|\boldsymbol{\theta}_{t-1} = \theta) \qquad (t \geq 2)$$

for $\theta \in X$. A homogeneous Markov chain can be completely described by a Markov kernel K mapping from $(X, \mathcal{B}X)$ into itself, where

$$K(A|\theta) = \mathbb{P}(\boldsymbol{\theta}_t \in A|\boldsymbol{\theta}_{t-1} = \theta) \qquad (\theta \in X, A \in \mathcal{B}X, t \in \mathbb{N}, t \geq 2).$$

For an initial measure $\mu' \in \mathrm{Prob}(X)$, the Markov chain corresponding to K can be defined uniquely by

$$\begin{aligned} &\boldsymbol{\theta}_1 \sim \mu', \\ &\boldsymbol{\theta}_t \sim K(\cdot \,|\boldsymbol{\theta}_{t-1}) \qquad (t \in \mathbb{N}, t \geq 2). \end{aligned}$$

The Markov kernel K is called *transition kernel* of $\underline{\boldsymbol{\theta}}$. In the following, we always assume that $\underline{\boldsymbol{\theta}}$ is homogeneous and that it has a transition kernel K.
As a dynamical system, a Markov chain can have fixed points on the space of probability measures. Indeed, a probability measure $\mu \in \mathrm{Prob}(X)$ is *stationary* with respect to $\underline{\boldsymbol{\theta}}$ if

$$\boldsymbol{\theta}_{t-1} \sim \mu \Rightarrow \boldsymbol{\theta}_t \sim \mu,$$

for some $t \in \mathbb{N}, t \geq 2$. We can write this equivalently as

$$\mu = \mu K.$$

Next, we define the reversibility of a Markov chain. Intuitively, $\underline{\boldsymbol{\theta}}$ is called *reversible* if a step from $\theta \in X$ to $\theta' \in X$ is as likely as a step from θ' to θ. Rigorously, $\underline{\boldsymbol{\theta}}$ is reversible if for $A, B \in \mathcal{B}X$ it holds

$$(\mu \odot K)(A \times B) = (\mu \odot K)(B \times A) \tag{3.9}$$

where μ is the stationary measure of $\underline{\boldsymbol{\theta}}$. (3.9) is called *detailed balance condition* and can be equivalently written as

$$\int_A K(B|\theta)\mu(\mathrm{d}\theta) = \int_B K(A|\theta)\mu(\mathrm{d}\theta). \tag{3.10}$$

If the detailed balance condition holds for an arbitrary probability measure $\mu' \in \mathrm{Prob}(X)$, the Markov chain $\underline{\boldsymbol{\theta}}$ is stationary with respect to $\mu' = \mu$. Indeed, by setting in (3.9) $A := X$ and choosing $B \in \mathcal{B}X$ arbitrarily we obtain:

$$\mu K(B) = \mu \odot K(X \times B) = \mu \odot K(B \times X) = \int_B K(X|\theta)\mu(\mathrm{d}\theta) = \int_B 1\mu(\mathrm{d}\theta) = \mu(B).$$

Hence, we can prove stationarity of a probability measure by showing the detailed balance condition.
We justified Monte Carlo and Importance Sampling methods applying the Strong Law of Large Numbers; see Theorem 1.13. We can show a similar result for a Markov chain. Under assumptions that we specify below, one can show that the following holds

$$\frac{1}{T}\sum_{t=1}^{T} Q(\boldsymbol{\theta}_t) \to \int_X Q(\theta)\mu(\mathrm{d}\theta) \qquad (t \to \infty; \mathbb{P}\text{-a.s.}), \tag{3.11}$$

where $\underline{\boldsymbol{\theta}}$ is a Markov chain that is stationary with respect to μ and $Q : X \to \mathbb{R}$ is a bounded function. Hence, if we can construct a Markov chain that is stationary w.r.t. a measure of interest, e.g. a posterior measure, we can use the sample path of the Markov chain to approximate integrals with respect to this measure. This is how MCMC methods proceed.
The assumption that has to be satisfied for (3.11) to hold is *Harris recurrence*. Before rigorously formulating this statement in Theorem 3.9, we introduce Harris recurrence in the following section.

Harris recurrence of Markov chains. We now introduce the concept of Harris recurrence of a Markov chain that is required in (3.11). Moreover, note that we have implictly assumed that stationary measures exist and that they are unique. In this section, we investigate this more rigorously.
We start by introducing recurrence and Harris recurrence of events. Let $A \in \mathcal{B}X$ be an event in X. A is *recurrent*, if a Markov chain starting in A is expected to return to A infinitely often. More precisely, let

$$\mathbf{N}_A := \sum_{t=1}^{\infty} \mathbf{1}_A(\boldsymbol{\theta}_t)$$

be the number of visits to A. The set A is *recurrent* if $\mathbb{E}\left[\mathbf{N}_A|\boldsymbol{\theta}_1 = \theta\right] = \infty$ for any initial state $\theta \in A$.
Harris recurrence of a set is a stronger concept. It applies if a Markov chain intialised in any $\theta \in A$ visits the set A infinitely often with probability one:

$$\mathbb{P}\left(\mathbf{N}_A = \infty|\boldsymbol{\theta}_1 = \theta\right) = 1 \qquad (\theta \in A).$$

Starting from the concept of (Harris) recurrence of a single set, we now aim to discuss the (Harris) recurrence of a Markov chain. The Markov chain is (Harris)

recurrent, if it is (Harris) recurrent for all sets that have a sufficient mass, and if it mixes throughout all of these sets. The mass is specified by a measure ν_X on $(X, \mathcal{B}X)$. The mixing of a Markov chain is specified by its irreducibility, which we define next.
The Markov chain $\underline{\boldsymbol{\theta}}$ is *ν_X-irreducible*, if for every $A \in \{\nu_X > 0\} := \{B \in \mathcal{B}X : \nu_X(B) > 0\}$, some $t \in \mathbb{N}$ exists, such that

$$K^t(A|\theta) > 0 \qquad (\theta \in X).$$

$K^t := K \cdots K$ refers to the t-fold composition of K. Hence, ν_X-irreducibility implies that within a finite number t of steps the chain reaches any set A that has sufficient measure, i.e. $\nu_X(A) > 0$, starting at any initial state $\theta \in X$.

Definition 3.7. A Markov chain $\boldsymbol{\theta}$ is *(Harris) recurrent* if it is ν_X-irreducible, for some measure ν_X on $(X, \mathcal{B}X)$, and if every set $A \in \{\nu_X > 0\}$ is *(Harris) recurrent*. ◇

In MCMC methods, it is important that the Markov chain indeed converges to the correct stationary measure. Fortunately, a recurrent Markov chain has a unique stationary measure. We have the following result.

Theorem 3.8 ([185, Theorem 10.0.1])**.** Let $\underline{\boldsymbol{\theta}}$ be a recurrent Markov chain. Then, a σ-finite measure μ exists in $(X, \mathcal{B}X)$, such that

$$\mu K = \mu.$$

Moreover, the measure μ is unique up to a normalising constant. ◇

Note that μ in Theorem 3.8 is not necessarily a probability measure; therefore we do not call it a *stationary measure*. However, if we know a stationary measure $\mu \in \text{Prob}(X)$ of the Markov chain $\underline{\boldsymbol{\theta}}$, Theorem 3.8 implies that it is unique. In this case, where $\underline{\boldsymbol{\theta}}$ is recurrent and μ is a probability measure, the chain $\underline{\boldsymbol{\theta}}$ is called *positive*. Since Harris recurrence implies recurrence, Harris recurrence also implies the statement of Theorem 3.8. A Harris recurrent and positive chain is called *Harris positive*. Harris positivity implies the statement in (3.11) which we now formalise in Theorem 3.9.

Theorem 3.9 (Strong ergodicity)**.** Let $\underline{\boldsymbol{\theta}}$ be a Harris positive Markov chain in X that is stationary with respect to $\mu \in \text{Prob}(X)$. Moreover, let $Q : X \to \mathbb{R}$ be some bounded function. Then,

$$\frac{1}{T}\sum_{t=1}^{T} Q(\boldsymbol{\theta}_t) \to \int_X Q(\theta)\mu(\mathrm{d}\theta) \qquad (t \to \infty; \mathbb{P}\text{-a.s.})$$

for every initial value $\theta_1 \in X$.

Proof. See [185, Theorem 17.3.2], where μ is a probability measure, $f := Q$ and $g := 1$. □

Remark 3.10 (How about ν_X?)**.** We note that recurrence and Harris recurrence as introduced in Definition 3.7 seem to depend on the choice of ν_X, which is the so-called *irreducibility measure*. However, ν_X appears neither in Theorem 3.8 nor in

Theorem 3.9. Indeed, it is not needed. The existence of *some* σ-finite measure ν_X, such that $\underline{\boldsymbol{\theta}}$ is (Harris) recurrent with respect to it, is sufficient.
There is, however, a connection between ν_X and μ. One can show that any irreducibility measure ν_X of a Markov chain $\boldsymbol{\theta}$ is dominated by the stationary measure μ; i.e. $\nu_X \ll \mu$. For a proof of this claim, we refer to [185, Proposition 4.2.2, Theorem 10.4.9].
Hence, when testing reducibility we can automatically ignore sets in $\{\mu = 0\}$. $\diamond$

MCMC algorithms. In Theorem 3.9 we show that we can approximate integrals with respect to a probability measure using samples from a Markov chain that is stationary with respect to this measure. Now, we aim to use this procedure to compute integrals with respect to, e.g. the posterior measure in a Bayesian inverse problem. In this section, we focus on the algorithmic construction of such a Markov chain. That is, we explain how the transition kernel K has to be chosen, such that the stationary measure is the measure of interest (e.g. the posterior measure). We start with the *Metropolis–Hastings (MH) framework*. Then, we introduce the *Gibbs sampler*. Finally, we combine both of these methods and obtain the *Metropolis-within-Gibbs method*, which we will use to sample from hierarchical posteriors.

Metropolis–Hastings. Let ν_X be some σ-finite measure on $(X, \mathcal{B}X)$ and $\mu \in \mathrm{Prob}(X, \nu_X)$ be the probability measure that we aim to approximate. Similarly to Importance Sampling, we assume in the following to have only access to a function $\gamma : X \to \mathbb{R}$, which satisfies

$$\gamma \propto \frac{\mathrm{d}\mu}{\mathrm{d}\nu_X} \qquad (\nu_X\text{-a.e.}).$$

We describe how the Metropolis–Hastings MCMC method constructs a Markov kernel K_{MH}. Let $\theta \in X$ be the current state. By using some auxilliary Markov kernel that is based on θ, a value $\theta' \in X$ is proposed. Then, a Bernoulli experiment with probability $a(\theta, \theta')$ is performed. $a(\theta, \theta')$ is called *acceptance probability*. If the Bernoulli experiment succeeds, i.e. if the output is 1, the output of the Metropolis–Hastings Markov kernel K_{MH} is the proposed state θ'. Otherwise, if the output is 0, the kernel K_{MH} returns the current state θ.
The procedure of sampling the new value θ' is called *proposal step*. The Bernoulli experiment is termed *accept-reject step*. Let $Q : X \times \mathcal{B}X \to [0, 1]$ be an auxilliary Markov kernel mapping from $(X, \mathcal{B}X)$ into itself. Q represents the proposal step. Moreover, let $Q(\cdot\,|\theta) \ll \nu_X$ $(\theta \in X)$ and let $q : X \times X \to [0, \infty)$ be a conditional density:

$$q(\theta'|\theta) = \frac{\mathrm{d}Q(\cdot\,|\theta)}{\mathrm{d}\nu_X}(\theta') \qquad (\theta \in X; \theta' \in X, \nu_X\text{-a.s.}).$$

For the accept-reject step, we define the acceptance probability

$$a(\theta, \theta') = \min\left\{1, \frac{\gamma(\theta')q(\theta|\theta')}{\gamma(\theta)q(\theta'|\theta)}\right\} \qquad (\theta, \theta' \in X).$$

Combining Q and a, we obtain the *Metropolis–Hastings kernel* $K_{\mathrm{MH}} : X \times \mathcal{B}X \to [0, 1]$, where

$$K_{\mathrm{MH}}(A|\theta) := \int_X (1 - a(\theta, \theta'))\delta(A - \theta) + a(\theta, \theta')\delta(A - \theta')Q(\mathrm{d}\theta'|\theta) \quad (A \in \mathcal{B}X, \theta \in X).$$

It can be proved that K_{MH} satisfies the detailed balance condition (3.9). Hence, the Metropolis–Hastings kernel induces a reversible Markov chain that is stationary with respect to μ. Moreover, one can show that μ-irreducibility of the MH chain already implies the Harris positivity of the chain. According to Theorem 3.9, Harris positivity implies convergence of the sample mean of the chain to the correct mean of the measure of interest. We summarise these statements below.

Theorem 3.11. Let $\underline{\boldsymbol{\theta}}$ be a Markov chain in X with transition kernel K_{MH}. Then, $\underline{\boldsymbol{\theta}}$ is stationary with respect to μ and reversible. Moreover, if in addition $\underline{\boldsymbol{\theta}}$ is μ-irreducible, $\underline{\boldsymbol{\theta}}$ is Harris positive.

Proof. [213, Theorem 7.2] shows the detailed balance condition. [213, Lemma 7.3] shows Harris positivity. □

The critical point in Metropolis–Hastings algorithms is the construction of the proposal distribution Q. In the following, we give two examples for potential proposal distributions. We commence with the *Random Walk Metropolis* (RWM) proposal.

Example 3.12 (Random Walk Metropolis)**.** Let $X := \mathbb{R}^{N_{\mathrm{sto}}}$ be a finite dimensional space. Moreover, let $\mu \in \mathrm{Prob}(X)$ be the measure of interest. A *random walk Metropolis* algorithm is given, if the density $q : X \times X \to \mathbb{R}$ is symmetric, i.e. $q(\theta|\theta') = q(\theta'|\theta)$. Notably, the acceptance probability simplifies to

$$a(\theta, \theta') := \min\{1, \gamma(\theta')/\gamma(\theta)\} \qquad (\theta, \theta' \in X)$$

in this case. The term *Random Walk Metropolis* is often used to describe the special case of a Gaussian random walk proposal:

$$Q(\cdot\,|\theta) := \mathrm{N}(\theta, \Sigma) \qquad (\theta \in X),$$

where $\Sigma \in \mathrm{CO}(X)$ is a covariance matrix. In this case, the proposal consists in adding a Gaussian random variable to the current state. Hence, the proposed value $\theta' \in X$ is a realisation of $\theta + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \Sigma)$. ◇

Metropolis et al. [183] have proposed the random walk Metropolis algorithm in 1953. The Metropolis–Hastings framework is younger; it has been introduced by Hastings [117] in 1970 as a generalisation of RWM, allowing for non-symmetric proposals. Non-symmetric proposals are particularly useful in high- and infinite dimensional parameter space settings, as we will see below.
We aim to apply MCMC algorithms to approximate posterior measures in Bayesian inverse problems in high- and infinite-dimensional parameter spaces X. Unfortunately, RWM algorithms are often inefficient in high and infinite dimensions, see [49, §4.2]. In Bayesian inverse problems with Gaussian priors, we can instead apply the *preconditioned Crank–Nicolson* (pCN) proposal.

Example 3.13 (Preconditioned Crank–Nicolson MCMC)**.** Let X be a separable Hilbert space. We consider a BIP with $\mu_{\mathrm{prior}} = \mathrm{N}(0, \mathcal{C}) \in \mathrm{Prob}(X)$ and likelihood $L(y^\dagger|\cdot) : X \to \mathbb{R}$. Let $\beta \in (0, 1]$. The *preconditioned Crank-Nicolson proposal* is given by

$$Q(\cdot\,|\theta) := \mathrm{N}\left((1-\beta^2)^{1/2}\theta, \beta\mathcal{C}\right) \qquad (\theta \in X).$$

The associated acceptance probability is

$$a(\theta, \theta') := \min\left\{1, L(y^\dagger|\theta')/L(y^\dagger|\theta)\right\} \qquad (\theta, \theta' \in X).$$

◇

In the special case $\beta := 1$, we propose a value by sampling independently from the prior measure. Therefore, the associated MCMC algorithm is called *independence sampler*. Note that this does not mean that the associated MCMC algorithm generates independent samples: the coupling is given through the accept-reject step. We can apply the pCN sampler also in the case, where the prior is not mean-zero. Here, we translate prior and likelihood. Let $\mu_{\text{prior}} := \mathrm{N}(m, \mathcal{C})$, then we redefine $\mu_{\text{prior}} := \mathrm{N}(0, \mathcal{C})$ and $L(y^\dagger|\cdot) := L(y^\dagger|m + \cdot)$.
The pCN proposal is based on the work by [21, 49], to which we refer for the derivation of the acceptance probability. It is derived from a Crank–Nicolson discretisation of a certain Langevin dynamic. This explains the origin of its name. pCN MCMC has recently been generalised by Rudolf and Sprungk [221].

Gibbs sampler. Let $\boldsymbol{\theta} : \Omega \to X$ and $\boldsymbol{\kappa} : \Omega \to R$ be two random variables and let $\mu \in \mathrm{Prob}(R \times X)$ be the joint measure of $(\boldsymbol{\kappa}, \boldsymbol{\theta})$, i.e. $\mu := \mathbb{P}((\boldsymbol{\kappa}, \boldsymbol{\theta}) \in \cdot)$. μ could be, for instance, a posterior measure in a hierarchical Bayesian inverse problem. We now discuss MCMC algorithms used to approximate integrals with respect to μ.
The most basic of these methods is the *Gibbs sampler*. It is a method using samples from the *full conditionals*. Those are

$$M'(\cdot\,|\theta) = \mathbb{P}(\boldsymbol{\kappa} \in \cdot\,|\boldsymbol{\theta} = \theta) \quad (\theta \in X), \qquad M''(\cdot\,|\kappa) = \mathbb{P}(\boldsymbol{\theta} \in \cdot\,|\boldsymbol{\kappa} = \kappa) \quad (\kappa \in R).$$

It proceeds in the following way: Let (κ, θ) be the current state of the Markov chain. The next state is given by (κ', θ), where κ' is a realisation of $\boldsymbol{\kappa} \sim M'(\cdot\,|\theta)$. Then the chain proceeds with (κ', θ'), where θ' is a realisation of $\boldsymbol{\theta} \sim M''(\cdot\,|\kappa')$. Hence, one coordinate is always fixed and the other coordinate is sampled with respect to the fixed state. We can define the Gibbs sampler formally by the transition kernel

$$K_{\text{Gibbs}} = K'K'' \tag{3.12}$$

from $(R \times X, \mathcal{B}R \otimes \mathcal{B}X)$ into itself, where

$$\begin{aligned} K'(\cdot \times \cdot\,|\kappa, \theta) &:= M'(\cdot\,|\theta) \otimes \delta(\cdot - \theta), \\ K''(\cdot \times \cdot\,|\kappa, \theta) &:= \delta(\cdot - \kappa) \otimes M''(\cdot\,|\kappa) \qquad (\theta \in X, \kappa \in R). \end{aligned}$$

Various ways to analyse Gibbs samplers have been discussed in the past. One can for instance use the Metropolis–Hastings framework to show that μ is stationary with respect to both K' and K''; as well as the reversibility of either kernel: We choose $Q \in \{K', K''\}$ as a proposal kernel, and observe that $a \equiv 1$. Hence, every proposed sample is immediately accepted. Therefore, K' and K'' are Metropolis–Hastings kernels and they satisfy the first part of Theorem 3.11. However, neither K', nor K'' is μ-irreducible on its own: Note that K' moves only in R, and K'' moves only in X. Hence, getting from any state (κ, θ) to the state (κ', θ') for which $\theta' \neq \theta$ (respectively $\kappa' \neq \kappa$) is impossible when applying K' (respectively K''). The composition $K_{\text{Gibbs}} = K'K''$ may however be μ-irreducible, see [213, Theorem 10.8], and also Harris positive, see [213, Lemma 10.9].
An interesting fact is that RWM was originally introduced to sample from a system of energy states, not a posterior measure. Such a system is for instance given in the Ising model [131]. Simulating states in the Ising model, on the other hand, has later become the motivation to introduce the Gibbs sampler, see [94].

Metropolis-within-Gibbs. In practice, especially when considering BIPs, we may be not able to sample from the full conditionals M' and/or M''. In such a case, the *Metropolis-within-Gibbs* (MWG) framework can be applied. It replaces the sampling from the full conditionals by Metropolis–Hastings kernels that are stationary with respect to the full conditionals: Let K'_{MH} be an MH Markov kernel from $(R \times X, \mathcal{B}R \otimes \mathcal{B}X)$ to $(R, \mathcal{B}R)$ and K''_{MH} be a MH Markov kernel from $(R \times X, \mathcal{B}R \otimes \mathcal{B}X)$ to $(X, \mathcal{B}X)$. Moreover, we assume that $M'(\cdot\,|\theta)$ is stationary with respect to $K'_{\mathrm{MH}}(\cdot\,|\cdot, \theta)$, $\theta \in X$, and $M''(\cdot\,|\kappa)$ is stationary with respect to $K''_{\mathrm{MH}}(\cdot\,|\kappa, \cdot)$, $\kappa \in R$. The MWG sampler is defined by the following kernel:

$$K_{\mathrm{MWG}} = K'K'',$$

where

$$\begin{aligned} K'(\cdot \times * \,|\kappa, \theta) &:= K'_{\mathrm{MH}}(\cdot\,|\kappa, \theta) \otimes \delta(* - \theta), \\ K''(\cdot \times * \,|\kappa, \theta) &:= \delta(\cdot - \kappa) \otimes K''_{\mathrm{MH}}(*\,|\kappa, \theta) \qquad (\theta \in X, \kappa \in R). \end{aligned}$$

As opposed to the 'pure' Gibbs sampler (3.12), the MWG has been applied in hierarchical Bayesian inverse problems by various researchers. We refer to, e.g. [41, 49, 76, 78, 248].

3.4 Improving efficiency

Consider a BIP that shall be solved with Importance Sampling or MCMC. Furthermore, we assume that the likelihood contains a forward response operator $\mathcal{G}$ that we need to discretise. Thus, we replace $\mathcal{G}$ by $\mathcal{G}_\ell$. Note that in Importance Sampling and most MCMC methods, we need to evaluate $\mathcal{G}_\ell$ whenever we produce a sample. Those evaluations are required to compute the weights in Importance Sampling or to compute the acceptance probability in MCMC. If evaluations of $\mathcal{G}_\ell$ are computationally demanding, applying either algorithm may be prohibitively expensive.
In the following, we briefly discuss two methods to reduce the computational cost of sampling methods: either we replace $\mathcal{G}_\ell$ by a *surrogate*, or we *reduce the variance* of the estimator using, e.g. a multilevel algorithm.

3.4.1 Surrogate models

Let $\mathcal{G}$ be a forward response operator in a Bayesian inverse problem, which is approximated by $\mathcal{G}_\ell$. We assume that evaluations of $\mathcal{G}_\ell$ are too expensive to be evaluated $N_{\mathrm{smp}} \in \mathbb{N}$ times - as may be required by MCMC or Importance Sampling.
Surrogate methods aim to add another layer of approximation that reduces the cost of $\mathcal{G}_\ell$ evaluations significantly. Indeed, a function $\widehat{\mathcal{G}} : X \to Y$ is constructed that is computationally cheap to evaluate and where $\mathcal{G}_\ell \approx \widehat{\mathcal{G}}$. The function $\widehat{\mathcal{G}}$ is called *surrogate* or sometimes *emulator*.
Surrogate methods typically proceed in two phases: *the offline and the online phase*. The offline phase happens prior to the sampling algorithm and is used for the construction of the function $\widehat{\mathcal{G}}$. The online phase contains the sampling algorithm that is now based on the surrogate $\widehat{\mathcal{G}}$. Importantly, the online phase should not depend on evaluations of the full model $\mathcal{G}_\ell$.

Reduced basis method. Reduced basis (RB) methods are a class of surrogates that are used to speed-up the evaluation of parameterised PDEs. Let $G : X \to H$ be a mathematical model given implicitly by $\theta \mapsto u^*$, where u^* solves

$$\mathcal{E}(u^*, v; \theta) = 0 \qquad (v \in H').$$

In §3.1, we have discussed Galerkin approximations using finite element bases to discretise H, and H'. In the following, we assume $H := H'$. A finite element space H_ℓ gives a good approximation of H – and thus eventually u^* – , but may be too high-dimensional to allow for many evaluations of G_ℓ.
To approximate G in particular, we actually do not need a good approximation of H, but rather of $\mathrm{img}(G) \subseteq H$. In *reduced basis methods*, we assume that we can approximate $\mathrm{img}(G)$ well with a very low-dimensional linear space H_{RB}. We can then replace H_ℓ by H_{RB} in the discretised formulation of the model and reduce the computational cost significantly. The *reduced space* H_{RB} is defined as the span of the *reduced basis* $B_{\mathrm{RB}} = (\varphi_1^{\mathrm{RB}}, \dots, \varphi_{N_{\mathrm{RB}}}^{\mathrm{RB}}) \in H^{N_{\mathrm{RB}}}$.
Reduced basis methods were introduced in [192], and are typically used to solve PDEs for a large number of parameter configurations, see, e.g. [119, 207]. In uncertainty quantification, reduced basis methods have been applied and analysed by, e.g. [38, 40, 75, 81] for forward problems and by, e.g. [39, 164, 173, 218] for Bayesian inverse problems.

Proper orthogonal decomposition. The *proper orthogonal decomposition* (POD) is a computational approach to constructing the reduced basis B_{RB}, for a particular mathematical model G given by $\mathcal{E}$. Here, we also refer to [207, §6].
Let G_ℓ be the discretised version of G on an N_ℓ-dimensional finite element space. The POD strategy starts in the following way: A family of N_{snap} elements of X is chosen:

$$\underline{\theta}_{\mathrm{snap}} = (\theta_{\mathrm{snap}}^{(1)}, \dots, \theta_{\mathrm{snap}}^{(N_{\mathrm{snap}})}) \in X^{N_{\mathrm{snap}}}.$$

In the following, we will approximate $\mathrm{img}(G)$ by $\mathrm{span}\{G_\ell(\theta_{\mathrm{snap}}^{(1)}), \dots, G(\theta_{\mathrm{snap}}^{(N_{\mathrm{snap}})})\}$. Various ways to find the parameter vector $\underline{\theta}_{\mathrm{snap}}$ have been discussed in the literature. Such are randomised and deterministic strategies, as well as adaptive greedy approaches. In any of these approaches, it is vital that the resulting $N_{\mathrm{snap}} \ll \min\{N_{\mathrm{smp}}, N_\ell\}$.
As indicated above, the model G_ℓ is evaluated for each of the components of $\underline{\theta}_{\mathrm{snap}}$. These evaluations

$$W_{\mathrm{snap}} := \underline{G_\ell(\theta_{\mathrm{snap}})} := \big(G_\ell(\theta_{\mathrm{snap}}^{(1)}), \dots, G_\ell(\theta_{\mathrm{snap}}^{(N_{\mathrm{snap}})})\big) \in H_\ell^{N_{\mathrm{snap}}}$$

are called *snapshots*. In the following, we identify H_ℓ with $\mathbb{R}^{N_\ell}$. Then, $W_{\mathrm{snap}} \in \mathbb{R}^{N_\ell \times N_{\mathrm{snap}}}$. Now, we construct the reduced basis from W_{snap}. The POD proceeds by orthonormalising W_{snap} and by removing vectors that are redundant or insignificant. The POD is obtained by computing the reduced singular value decomposition of W_{snap}:

$$U \cdot \Sigma \cdot V^* = W_{\mathrm{snap}},$$

where $U \in \mathbb{R}^{N_\ell \times N_{\mathrm{snap}}}$, and $V \in \mathbb{R}^{N_{\mathrm{snap}} \times N_{\mathrm{snap}}}$ are orthonormal matrices, and

$$\Sigma = \mathrm{diag}(\sigma_s : s = 1, \dots, N_{\mathrm{snap}})$$

is the diagonal matrix containing the singular values of W_{snap}. The reduced basis consists of columns of U. Let $s = 1, \dots, N_{\text{snap}}$ and let $u_s \in \mathbb{R}^{N_\ell}$ be the s-th column vector of U. If the singular value $\sigma_s = 0$, the vector u_s does not contribute to the reduced basis. Hence, we eliminate those and define

$$W := (u_s : s = 1, \dots, N_{\text{snap}}, \sigma_s > 0).$$

For further dimension reduction, we can also eliminate those vectors with singular values smaller than some threshold $\underline{\sigma} > 0$. In this case, we obtain a smaller reduced basis

$$W := (u_s : s = 1, \dots, N_{\text{snap}}, \sigma_s > \underline{\sigma}).$$

Such a further dimension reduction is very critical for the speed-up in a reduced basis method. While clearly $N_{\text{RB}} < N_\ell$, this relation may be not sufficient for a speed-up. We finish this section with an example, in which we clarify that a significant dimension reduction is necessary for speed-up.

Example 3.14. Let G be the parameter-to-solution map for a parameterised elliptic PDE on an interval, a square, or a cube. We fix the parameter and discretise the PDE with linear finite elements on a regular mesh and obtain an N_ℓ-dimensional FEM space. Due to the choice of finite elements, the $N_\ell \times N_\ell$ system matrix is sparse. Using a multigrid method, the discretised equation can be solved in $O(N_\ell; N_\ell \to \infty)$; see, e.g. Hackbusch [109, §4.3] who discusses the Poisson equation and Trottenberg et al. [246, §4.7] who gives conditions under which PDEs with more general linear operators can be solved by multigrid.
Now, we aim to solve the same PDE with a reduced basis. We have constructed the reduced basis from snapshots – not from locally supported finite element functions. While having a much lower dimension, the system equation on the reduced basis is dense. Hence, the computational cost in the online phase increases to asymptotically $O(N_{\text{RB}}^2; N_{\text{RB}} \to \infty)$. Therefore, we only obtain a speed-up, if $N_{\text{RB}}^2 < N_\ell$. ◇

Other surrogates. Reduced basis methods are a popular surrogate technique that is based on Galerkin projection. This is not the only option for surrogates. Other authors have been using *sparse grid interpolation* [87, 89, 225, 229], *polynomials* [52, 177, 234], *piecewise polynomials* [180], and *Gaussian processes* [163, 237].

3.4.2 Variance reduction

Variance reduction techniques follow a different paradigm. Here, not the cost of a single expensive model evaluation shall be reduced, but their total quantity. In the following, we only refer to the simple Monte Carlo integration, with independent samples from the correct measure. Similar techniques can also be applied within MCMC and Importance Sampling to solve Bayesian inverse problems. We mention those further below, when reviewing previous works.
Let $\mu \in \text{Prob}(X)$ be a measure from which we can draw independent samples, and $Q \in \mathbf{L}^2(X, \mu)$ be a scalar-valued function. In §3.3.1, we have discussed the standard Monte Carlo algorithm to approximate integrals of the form $\int_X Q(\theta)\mu(\mathrm{d}\theta)$. Recall (3.5) where the root mean square error of the Monte Carlo estimator with $N_{\text{smp}} \in \mathbb{N}$

samples of this integral is given by

$$\mathbb{E}\left[\left(\overline{Q(\boldsymbol{\theta})}_{N_{\text{smp}}} - \int_X Q(\theta)\mu(\mathrm{d}\theta)\right)^2\right]^{1/2} = \sqrt{\frac{\mathrm{Var}_\mu(Q)}{N_{\text{smp}}}}.$$

Hence, to obtain a small error, we need either a small variance $\mathrm{Var}_\mu(Q)$, or a large number of samples N_{smp}.
Let now Q be based on a mathematical model that has to be discretised. Hence, we have $Q_\ell \approx Q$ and a potentially high computational cost to evaluate Q_ℓ. In the previous section, we have discussed surrogate methods that reduce the computational cost of evaluations of Q_ℓ. Hence, surrogates allow us to increase the size of N_{smp}. *Variance reduction* techniques aim at finding an estimator that has a variance that is smaller than $\mathrm{Var}_\mu(Q_\ell)$. Given a smaller variance, we can perform Monte Carlo estimations with fewer samples. Below, we briefly introduce variance reduction techniques and refer to the book by Rubinstein and Kroese [217, Chapter 5] for details.
Variance reduction techniques proceed as follows. We define an estimator $V : X \to \mathbb{R}$, with

$$\left|\int_X V(\theta) - Q_\ell(\theta)\mu(\mathrm{d}\theta)\right| = 0 \qquad \text{and} \qquad \mathrm{Var}_\mu(V) < \mathrm{Var}_\mu(Q_\ell).$$

Thus, a Monte Carlo estimation of $\int_X V(\theta)\mu(\mathrm{d}\theta)$ leads asymptotically to the same value, but converges faster.
Popular variance reduction techniques are *antithetic variables* and *control variates*. *Importance sampling* can also be used for variance reduction. In the context of PDE-based uncertainty quantification, *Multilevel Monte Carlo* (MLMC) methods constitute a further class of variance reduction methods. Here, we consider a sequence of functions $(Q_k)_{k=1}^\ell$ that approximate Q with different accuracies $k = 1, \dots, \ell$. Here, $k = 1$ refers to the coarsest discretisation level and $k = \ell$ to the finest discretisation level. Moreover, we define the estimator V by the following telescopic sum

$$\begin{aligned} V &:= Q_\ell + (Q_{\ell-1} - Q_{\ell-1}) + \cdots + (Q_1 - Q_1) \\ &= (Q_\ell - Q_{\ell-1}) + (Q_{\ell-1} - Q_{\ell-2}) + \cdots + (Q_2 - Q_1) + Q_1. \end{aligned}$$

Looking at the first line of the displayed equation, one can easily see that V and Q_ℓ have identical expected values. In the estimation process, the bracketed terms in the second row above are coupled probabilistically. If this leads to (Q_k, Q_{k-1}) being positively correlated for $k = 2, \dots, \ell$, the estimator R has indeed a smaller variance than Q_ℓ. Additionally, if the computational cost of the coarse discretisations $k = 1, \dots, \ell$ is sufficiently small, the overall cost is indeed reduced.
Multilevel Monte Carlo has been first introduced for stochastic differential equation (SDE) path simulations by Giles [102]. Later, the method has been applied to the stochastic elliptic PDE in Example 1.37 by, e.g. [45, 112, 244]. Other problems in forward UQ have been approached with Multilevel Monte Carlo in [7, 113, 148]. For an overview on Multilevel Monte Carlo, we refer to [103]. Bayesian inverse problems have been approached with Multilevel Markov chain Monte Carlo [71, 72] and Multilevel Sequential Monte Carlo [18, 19]. In addition, the Multilevel Monte Carlo idea has been extended to deterministic quadrature, such as stochastic

collocation [111, 243], sparse grid quadrature [87, 88] and quasi-Monte Carlo [67, 149, 150]. Finally, we mention multifidelity methods, which use a hierarchy of surrogate models, rather than a hierarchy of FEM/time-stepping discretisation levels; see [199, 200] for multifidelity Monte Carlo.

Chapter 4

Exploring hierarchical random fields

> Yet, such *deep learning* may require considerable amount of work.
>
> Rina Dechter [59, p. 180]

In §1.4, we motivated and discussed hierarchical models in uncertainty quantification. Next, we consider especially *parameterised Gaussian random fields*. Those are,

$$K(\cdot\,|\kappa) = \mathrm{N}(m(\kappa), \mathcal{C}(\kappa)) \qquad (\kappa \in R).$$

Notably, the resulting random field $\mu'' = \mu' K$ is not necessarily Gaussian and allows us to model a larger class of spatial variations. However, the greater flexibility of parameterised Gaussian fields brings new computational challenges, as we explain next.

Assume that we discretise a Gaussian random field by a KL expansion. The basis functions in this expansion are the eigenfunctions of the covariance operator. For fixed, deterministic hyperparameters it is sufficient to compute the KL eigenpairs only a single time since the covariance operator is fixed. However, changing the hyperparameters changes the covariance operator. This may require to re-compute the KL eigenpairs. Dunlop et al. [78], for instance, consider parameterised covariance operators with eigenpairs that are analytically accessible. Other than that, it is typically necessary to re-compute the eigenpairs. The associated cost and memory requirement of these re-computations scales at least quadratically in the number of spatial unknowns. Here, it is often practically impossible to use uncertain hyperparameters in a (Gaussian) random field model in 2D or 3D physical space. We do not consider other sampling strategies, such as those discussed in Remark 3.1, as they can be limited in terms of applicability.

To overcome this limitation we now suggest and study a reduced basis surrogate for the efficient computation of parameter dependent KL expansions. This approach is similar to that discussed in §3.4.1. In contrast, our reduced basis surrogate approximates the KL eigenpairs rather than a PDE solution. Reduced basis approaches to solve parameterised eigenproblems have been discussed in [92, 123, 124, 172, 231, 249]. Those authors have focused on eigenproblems with differential operators. Particularly for the KL eigenproblem, reduced basis ideas have been discussed by Contreras et al. [48]. However, they do not consider parameterised Gaussian random fields, but use the reduced basis for domain decomposition in very high-dimensional KL eigenproblems. Moreover, we mention Sraj et al. [234], who consider a polynomial chaos surrogate to speed-up a hierarchical Bayesian inverse problem.

In §4.1, we discuss parameterised Gaussian measures, and their representation via a KL expansion. Furthermore, we collect various tasks that we need to perform with Gaussian random fields and note their computational cost. Next, in §4.2, we consider the low-rank approximation of covariance operators using the reduced basis and POD approach. Moreover, we discuss how to approximate Matérn covariance functions to make them suitable for the POD approach. In §4.3, we discuss the *reduced basis sampling* algorithm that explains the online phase of the algorithm. We also derive the overall computational cost of the method. Finally, in §4.4, we verify and illustrate the method using numerical experiments.

4.1 Parameterised Gaussian measures

Let $R \subseteq \mathbb{R}^{\mathrm{N}_R}$ be again the space in which the hyperparameters live; assume that R is non-empty and finite-dimensional. It forms a measurable space together with its Borel-σ-algebra $\mathcal{B}R$. As mentioned above, we consider Markov kernels where $K(\cdot\,|\kappa)$ is a Gaussian measure for all $\kappa \in R$. Particularly, we define

$$R \ni \kappa \mapsto K(\cdot\,|\kappa) := \mathrm{N}(m(\kappa), \mathcal{C}(\kappa)) \in \mathrm{Prob}(X), \tag{4.1}$$

where

$$m : R \to X, \quad \mathcal{C} : R \to \mathrm{CO}(X)$$

are measurable functions, and X is a separable Hilbert space. The distribution of the parameter $\boldsymbol{\kappa} : \Omega \to R$ is again called μ'. Moreover, the distribution of the composition and the joint distribution are again defined by

$$\mu'' = \mu K \in \mathrm{Prob}(X), \qquad \mu = \mu' \odot K \in \mathrm{Prob}(R \times X),$$

respectively.

Remark 4.1. We point out that even if $K(\cdot\,|\kappa)$ is a Gaussian measure for any $\kappa \in R$, the composition μ'' is not necessarily a Gaussian measure. We give two examples.

(a) Let $R := X := \mathbb{R}$, let $\mu' := \mathrm{N}(m_0, \sigma_0^2)$ be a Gaussian measure and $K(\cdot\,|\kappa) := \mathrm{N}(\kappa, \sigma^2)$. This construction models a family of Gaussian random variables where the mean value is another Gaussian random variable. Here, we have $\mu'' = \mathrm{N}(m_0, \sigma_0^2 + \sigma^2)$ is a Gaussian measure.

(b) Let R be a finite set. Then μ'' is called *Gaussian mixture* and is often not Gaussian. See §1.1 and §2.1 in [86].

◇

In Example 1.26, we have discussed a Gaussian random field with exponential covariance operator. We now revisit this example and construct a parameterised Gaussian measure, in which correlation length and standard deviation are unknown.

Example 4.2. We consider again the exponential covariance operator in (1.12). Let $\underline{\lambda} > 0$ and $\overline{\sigma} > \underline{\sigma} > 0$. For any $\lambda \in [\underline{\lambda}, \mathrm{diam}(D)]$ and $\sigma \in [\underline{\sigma}, \overline{\sigma}]$, one can show that $\mathcal{C}_{\mathrm{exp}}^{(\lambda,\sigma)} \in \mathrm{CO}(X)$ is a valid covariance operator. The parameters $\boldsymbol{\kappa} = (\boldsymbol{\lambda}, \boldsymbol{\sigma})$ are

random variables on a non-empty set $R := [\underline{\lambda}, \mathrm{diam}(D)] \times [\underline{\sigma}, \overline{\sigma}]$. The associated probability measure μ' is given by

$$\mu' := \mu'_\lambda \otimes \mu'_\sigma.$$

Here, μ'_λ is given such that $\boldsymbol{\lambda}^{-1} \sim \mathrm{Unif}[\mathrm{diam}(D)^{-1}, \underline{\lambda}^{-1}]$; we show the pdf of μ'_λ in Figure 4.9. μ'_σ is a Gaussian measure that is truncated outside of $[\underline{\sigma}, \overline{\sigma}]$. $\boldsymbol{\sigma}$ models the standard deviation of $\theta(x)$, for any $x \in D$. The measure μ' and the Markov kernel $K(\cdot\,|\lambda, \sigma) = \mathrm{N}(0, \mathcal{C}_{\exp}^{(\lambda,\sigma)})$ induce a joint measure μ. This can now be understood as follows:

1. Sample $\boldsymbol{\kappa}$ from μ':
   (a) Sample the correlation length $\boldsymbol{\lambda} \sim \mu'_\lambda$,
   (b) Sample the standard deviation $\boldsymbol{\sigma} \sim \mu'_\sigma$.
2. Sample the random field $\theta \sim K(\cdot\,|\boldsymbol{\lambda}, \boldsymbol{\sigma})$ with exponential covariance operator, standard deviation $\boldsymbol{\sigma}$ and correlation length $\boldsymbol{\lambda}$.

Hence, we modelled a Gaussian random field with exponential covariance, where the correlation length and standard deviation are unknown. ◇

4.1.1 Discretisation of parameterised random fields

We have discussed the discretisation of Gaussian random fields in §3.2. We next extend this to parameterised Gaussian measures.
In the following, we will be interested in solving hierarchical forward and hierarchical Bayesian inverse problems, as discussed in §1.4.2. To this end, we require samples $(\boldsymbol{\kappa}_1, \boldsymbol{\theta}_1), \ldots, (\boldsymbol{\kappa}_{N_{\mathrm{smp}}}, \boldsymbol{\theta}_{N_{\mathrm{smp}}}) \sim \mu$. We assume that sampling from μ' is possible and inexpensive. However, for each sample $\boldsymbol{\kappa}_n \sim \mu'$ we also need to sample $\boldsymbol{\theta}_n \sim K(\cdot\,|\boldsymbol{\kappa}_n) := \mathrm{N}(m(\boldsymbol{\kappa}_n), \mathcal{C}(\boldsymbol{\kappa}_n))$, for which we want to use a truncated KL expansion. This requires the assembly of the (dense) covariance matrix $\mathcal{C}(\boldsymbol{\kappa}_n)$, and the computation of its leading N_{sto} eigenpairs. We abbreviate this process by the function $\mathrm{eigs}(\mathcal{C}(\boldsymbol{\kappa}_n), N_{\mathrm{sto}})$ which returns $\Psi_n := (\alpha_i^{1/2}(\boldsymbol{\kappa}_n)\psi_i(\boldsymbol{\kappa}_n))_{i=1}^{N_{\mathrm{sto}}}$. The complete sampling procedure is given in Algorithm 1.

Algorithm 1: Sampling from a parameterised Gaussian measure μ

```
for n ∈ {1, ..., N_smp} do
    Sample κ_n ~ μ'
    Ψ_n ← eigs(C(κ_n), N_sto)
    Sample ξ ~ N(0, Id_{N_sto})
    θ_n ← m(κ_n) + Ψ_n ξ
end
```

Parameterised Karhunen–Loève expansion. In §3.2 we study the truncation error of the Karhunen–Loève expansion of a Gaussian random field. We now extend this study to parameterised Gaussian random fields. In this case, we compute a parameterised KL expansion; hence the KL expansion is Gaussian and depends

on the parameters. In particular, we discuss the choice of N_{sto} for parameterised Gaussian random fields. Throughout the rest of this chapter, we assume that μ'' has a finite second moment, i.e.

Assumption 4.3. $\int_X \|\theta\|_X^2 \mu''(\mathrm{d}\theta) = \iint_{X\times R} \|\theta\|_X^2 K(\mathrm{d}\theta|\kappa)\mu'(\mathrm{d}\kappa) < \infty$. ◇

This assumption is satisfied for Example 4.2, since K has a finite second moment, and $\boldsymbol{\kappa}$ lives in R – a compact space.
We move on to the parameterised KL expansion. Let

$$\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}} := \sum_{i=1}^{N_{\text{sto}}} \sqrt{\alpha_i(\boldsymbol{\kappa})}\boldsymbol{\xi}_i\psi_i(\boldsymbol{\kappa}), \tag{4.2}$$

where $\boldsymbol{\kappa} \sim \mu'$ and $\boldsymbol{\xi} \sim \mathrm{N}(0, \mathrm{Id}_{N_{\text{sto}}})$. Note that $\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}$ is an approximation to the parameterised Gaussian random field $\boldsymbol{\theta} \sim \mu''$. The mean square error of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}$ can be computed as follows:

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}}\|_X^2\right] &= \iint_{R\times\mathbb{R}^{N_{\text{sto}}}} \left(\sum_{i=N_{\text{sto}}+1}^{\infty} \sqrt{\alpha_i(\boldsymbol{\kappa})}\boldsymbol{\xi}_i\psi_i(\boldsymbol{\kappa})\right)^2 \mathrm{N}(0, \mathrm{Id}_{N_{\text{sto}}})(\mathrm{d}\boldsymbol{\xi})\mu'(\mathrm{d}\boldsymbol{\kappa}) \\ &= \int_R \sum_{i=N_{\text{sto}}+1}^{\infty} \alpha_i(\boldsymbol{\kappa})\mu'(\mathrm{d}\boldsymbol{\kappa}).\end{aligned}$$

For Gaussian random fields N_{sto} is typically chosen such that the root mean square error fulfills a certain threshold. For example,

$$N_{\text{sto}} := \min\left\{N' \in \mathbb{N} : \sum_{i=1}^{N'} \alpha_i \geq A \cdot \sum_{i=1}^{\infty} \alpha_i\right\},$$

where A is a fixed factor. Looking at the error bound in (3.1) we see that A determines which amount of the total variance of the exact (Gaussian) random field is captured by the truncated KL expansion. The same strategy can be applied for parameterised Gaussian random fields. Let

$$N_{\text{sto}}^{\text{all}} := \min\left\{N' \in \mathbb{N} : \sum_{i=1}^{N'} \alpha_i(\kappa) \geq A \cdot \sum_{i=1}^{\infty} \alpha_i(\kappa) \quad (\kappa \in R, \mu'\text{-a.s.})\right\}$$

be the number of terms that fulfils the threshold A for μ'-a.e. parameter $\kappa \in R$. Then, the mean square error is bounded by

$$\mathbb{E}\left[\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{KL}}^{N_{\text{sto}}^{\text{all}}}\right\|_X^2\right] \leq (1 - A) \cdot \mathbb{E}\left[\|\boldsymbol{\theta}\|_X^2\right]. \tag{4.3}$$

Alternatively, it is possible to choose N_{sto} individually for μ'-a.e. $\kappa \in R$,

$$N_{\text{sto}}^{\kappa} := \min\left\{N' \in \mathbb{N} : \sum_{i=1}^{N'} \alpha_i(\kappa) \geq A \cdot \sum_{i=1}^{\infty} \alpha_i(\kappa) \quad\right\}.$$

This gives the truncated representation

$$\boldsymbol{\theta}_{\mathrm{KL}}^{N_{\mathrm{sto}}^{\boldsymbol{\kappa}}} := \sum_{i=1}^{N_{\mathrm{sto}}^{\boldsymbol{\kappa}}} \sqrt{\alpha_i(\boldsymbol{\kappa})}\boldsymbol{\xi}_i\psi_i(\boldsymbol{\kappa}) \quad (\boldsymbol{\xi} \sim \mathrm{N}(0, \mathrm{Id}_{N_{\mathrm{sto}}}), \boldsymbol{\kappa} \sim \mu').$$

Clearly, the mean square error of this expansion fulfils the exact same error bound as in (4.3), i.e.

$$\mathbb{E}\left[\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{KL}}^{N_{\mathrm{sto}}^{\boldsymbol{\kappa}}}\right\|_X^2\right] \leq (1-A) \cdot \mathbb{E}\left[\|\boldsymbol{\theta}\|_X^2\right].$$

However, the total number of terms in the expansion for a fixed parameter value $\boldsymbol{\kappa} = \kappa$ might be smaller. Recall that the cost of the sampling depends (linearly) on the number of KL terms. Observe that $N_{\mathrm{sto}}^{\mathrm{all}}$ is an upper bound for $N_{\mathrm{sto}}^{\kappa}$, $\kappa \in R$. Hence, using $N_{\mathrm{sto}}^{\kappa}$ is overall not more expensive than using $N_{\mathrm{sto}}^{\mathrm{all}}$, and the truncated expansion satisfies the same error bound. Moreover, the numbers $(N_{\mathrm{sto}}^{\kappa})_{\kappa\in R}$ are a priori unknown and have to be computed. To avoid this additional cost and to simplify the following discussion, we use $N_{\mathrm{sto}} := N_{\mathrm{sto}}^{\mathrm{all}}$ independently of $\kappa \in R$.

Optimal Karhunen–Loève expansion for $\boldsymbol{\theta}$. Note that the parametric KL expansion discussed above is a Gaussian KL expansion for fixed parameter κ. However, it is strictly speaking not the KL expansion of the non-Gaussian random field $\boldsymbol{\theta} \sim \mu''$. We now briefly discuss the actual Karhunen–Loève expansion of $\boldsymbol{\theta}$. Let

$$\widehat{\mathcal{C}} := \mathrm{Cov}(\boldsymbol{\theta}) := \int_R \mathcal{C}(\kappa)\mu'(\mathrm{d}\kappa)$$

be the covariance operator of $\boldsymbol{\theta}$. It is well-defined, since by assumption $\mathbb{E}[\|\boldsymbol{\theta}\|_X^2] < \infty$. Now, let $(\widehat{\alpha}_i, \widehat{\psi}_i)_{i=1}^{\infty}$ be the family of eigenpairs of $\widehat{\mathcal{C}}$. They are again ordered descendingly in the eigenvalue, and the eigenfunctions form an orthonormal basis of X.
Then, we can expand

$$\boldsymbol{\theta} = \int_R m(\kappa)\mu'(\mathrm{d}\kappa) + \sum_{i=1}^{\infty} \sqrt{\widehat{\alpha}_i}\widehat{\boldsymbol{\xi}}_i\widehat{\psi}_i. \tag{4.4}$$

To discretise the random field $\boldsymbol{\theta}$, we would now truncate the series in (4.4). Then, we obtain indeed the optimal linear function approximating the random field with N_{sto} random variables, similarly to (3.2). Unfortunately, $(\widehat{\boldsymbol{\xi}}_i)_{i\in\mathbb{N}}$ are not necessarily standard Gaussian random variables in this case. According to [96, §2.3.2], $\widehat{\boldsymbol{\xi}}_i$ is distributed as the random variable

$$\frac{1}{\sqrt{\widehat{\alpha}_i}}\left\langle \widehat{\psi}_i, \boldsymbol{\theta} - \int_R m(\kappa)\mu'(\mathrm{d}\kappa)\right\rangle_X,$$

for $i \in \mathbb{N}$. Hence, without already being able to sample from $\boldsymbol{\theta}$, we may not be able to compute the distribution of the $(\widehat{\boldsymbol{\xi}}_i)_{i=1}^{\infty}$. Thus, even though being optimal, the Karhunen–Loève expansion of $\boldsymbol{\theta}$ may be not of practical use in this setting.

4.1.2 Tasks and computational cost

In §1.4 we have discussed forward and inverse problems with hierarchical measures. These will now be considered in the setting of hierarchical Gaussian random fields. Compared to the non-hierarchical random field case, we have to tackle significantly larger computational costs *per sample*. This is the case since we can neither precompute the truncated KL expansion, nor perform all tasks with the standard Gaussian vector $\underline{\boldsymbol{\xi}} : \Omega \to \mathbb{R}^{N_{\text{sto}}}$. We will now list the computational tasks we need to execute to use Monte Carlo or MCMC in a hierarchical forward or inverse problem, respectively.
From now on, we assume that we have discretised the parameter space with finite elements; see §3.2.2. Hence, we set $X := X_\ell := \mathbb{R}^{N_\ell}$. In the non-hierarchical case, the following tasks have a computational cost that is independent of N_ℓ since we can perform all the tasks in $\mathbb{R}^{N_{\text{sto}}}$. These and other tasks are also listed in Schäfer et al. [224].

Sampling. We aim to use the parameterised KL expansion in (4.2) for sampling. Moreover, we use the most general algorithm that allows us to do so: we assemble the covariance operator and then compute the first N_{sto} leading eigenpairs. The cost for the assembly of the dense covariance matrix is of order $O(N_\ell^2; N_\ell \to \infty)$. We assume that the cost of a single function call $\text{eigs}(\cdot, N_{\text{sto}})$ is of order $O(N_\ell^2 \cdot N_{\text{sto}}; N_\ell \to \infty)$. This corresponds to an *Implictly Restarted Lanczos Method*, where $p = O(N_{\text{sto}})$. Note that this method is implemented in ARPACK (and thus, for instance, in MATLAB) as `eigs`; see [33, 108] for details.
Thus, the total computational cost of the sampling method in Algorithm 1 is

$$O(N_{\text{smp}} \cdot (N_\ell^2 \cdot (N_{\text{sto}} + 1)); N_\ell \to \infty).$$

The largest contribution to the computational cost is the repeated computation of the leading eigenpairs of $\mathcal{C}(\boldsymbol{\kappa}_n)$.

Memory. When solving a Bayesian inverse problem, we sample N_{smp} times from the posterior

$$(\boldsymbol{\kappa}_1, \boldsymbol{\theta}_1), (\boldsymbol{\kappa}_2, \boldsymbol{\theta}_2), \dots, (\boldsymbol{\kappa}_{N_{\text{smp}}}, \boldsymbol{\theta}_{N_{\text{smp}}}) \sim \mu_{\text{post}}^\dagger.$$

Those samples shall be used to approximate $\mu_{\text{post}}^\dagger$. Hence, we need to keep them in the memory. There are now two possibilities: We keep the hyperparameter and the complete random field $(\kappa_n, \theta_n)_{n=1}^{N_\ell} \in (R \times \mathbb{R}^{N_\ell})^{N_\ell}$ in the memory at a cost of

$$O((N_\ell + N_R) N_{\text{smp}}; N_{\text{smp}} \to \infty).$$

Alternatively, we keep only $(\kappa_n, \xi^{(n)})_{n=1}^{N_\ell} \in (R \times \mathbb{R}^{N_{\text{sto}}})^{N_\ell}$, where $\theta_n = m(\kappa_n) + \Psi(\kappa_n)\xi^{(n)}$. This is analoguous to working only with the KL mode ξ in the non-hierarchical case. This approach requires only a cost of

$$O((N_{\text{sto}} + N_R) N_{\text{smp}}; N_{\text{smp}} \to \infty);$$

note that $N_{\text{sto}} \ll N_\ell$. However, $\boldsymbol{\xi}^{(n)}$ has no particular meaning without knowing $\Psi(\boldsymbol{\kappa}_n)$, $n = 1, \dots, N_{\text{smp}}$. Hence, either $(\Psi(\boldsymbol{\kappa}_n))_{n=1}^{N_{\text{smp}}}$ have to be kept in memory or recomputed. Keeping them in the memory is at least as expensive as keeping the full random field in the memory. Recomputation is memory efficient, but requires the same cost per random field access from the memory as sampling. Overall, this should be more expensive, than keeping the full random field in the memory.

Pseudo-inverse and pseudo-determinants. Since X is finite-dimensional, we can define $K(\cdot\,|\kappa) = \mathrm{N}(m(\kappa), \mathcal{C}(\kappa))$ via the following density:

$$\mathrm{n}(\theta; m(\kappa), \mathcal{C}(\kappa)) = \left(\det{}^{+}\left(2\pi\mathcal{C}(\kappa)\right)\right)^{-1/2} \exp\left(-\frac{1}{2}\left\|(\mathcal{C}(\kappa)^{+})^{1/2}(\theta - m(\kappa))\right\|_X^2\right),$$

where A^+ is the (Moore–Penrose) pseudo-inverse of A and $\det^+(A)$ is its pseudo-determinant, i.e. the product of the non-zero eigenvalues of A. Here, $\mathrm{n}(\cdot; m(\kappa), \mathcal{C}(\kappa))$ is a density with respect to the product of the Lebesgue measure on $\mathrm{img}(\mathcal{C}(\kappa)^{1/2})$ and the Dirac measure concentrated in 0 on $\ker(\mathcal{C}(\kappa)^{1/2})$.
There are particular cases, in which we need to evaluate this density. The example we discuss in the following is a Metropolis-within-Gibbs algorithm in a hierarchical Bayesian inverse problem. Consider a BIP, with $\mu_{\text{prior}} := \mu := \mu' \odot K$ and likelihood $L(y^\dagger|\cdot)$. We apply a Metropolis-within-Gibbs algorithm where we have Gibbs moves alternating in θ- and κ-direction. In θ-direction, we use a pCN proposal; recall Example 3.13. Note that we have defined the pCN proposal only for mean-zero random fields. Therefore, we move the prior mean into the likelihood: We redefine $K(\cdot\,|\kappa) = \mathrm{N}(0, \mathcal{C}(\kappa))$ and $L(y^\dagger|\theta, \kappa) := L(y^\dagger|\theta + m(\kappa))$. In κ-direction, we apply a general Metropolis–Hastings kernel. In this κ-Metropolis–Hastings step, we need to evaluate the density $\mathrm{n}(\theta; m(\kappa), \mathcal{C}(\kappa))$, for any proposed κ. We give the resulting MWG method as pseudo-code in Algorithm 2. In the algorithm, we need to evaluate

Algorithm 2: Metropolis-within-Gibbs to sample from the posterior measure $\mu^\dagger_{\text{post}}$ of $(\boldsymbol{\kappa}, \theta)$

Let $(\boldsymbol{\kappa}_1, \boldsymbol{\theta}_1) := (\kappa_1, \theta_1)$ be the initial sample of the Markov chain.
for $n \in \{2, \dots, N_{\text{smp}}\}$ **do**
  Sample $\boldsymbol{\kappa}_* \sim q_R(\cdot\,|\boldsymbol{\kappa}_{n-1})$
  $a_R(\boldsymbol{\kappa}_{n-1}; \boldsymbol{\kappa}_*) \leftarrow \min\left\{1, \frac{q_R(\boldsymbol{\kappa}_{n-1}|\boldsymbol{\kappa}_*)}{q_R(\boldsymbol{\kappa}_*|\boldsymbol{\kappa}_{n-1})} \frac{\mathrm{n}(\boldsymbol{\theta}_{n-1}; 0, \mathcal{C}(\boldsymbol{\kappa}_*))}{\mathrm{n}(\boldsymbol{\theta}_{n-1}; 0, \mathcal{C}(\boldsymbol{\kappa}_{n-1}))}\right\}$
  Sample $\mathbf{U}_R \sim \mathrm{Unif}[0, 1]$
  if $\mathbf{U}_R \leq a_R$ **then**
    $\boldsymbol{\kappa}_n \leftarrow \boldsymbol{\kappa}_*$
  else
    $\boldsymbol{\kappa}_n \leftarrow \boldsymbol{\kappa}_{n-1}$
  end
  Sample $\boldsymbol{\theta}_* \sim \mathrm{N}(\sqrt{1-\beta^2}\boldsymbol{\theta}_{n-1}, \beta\mathcal{C}(\boldsymbol{\kappa}_n))$
  $a_X(\boldsymbol{\theta}_{n-1}; \boldsymbol{\theta}_*) \leftarrow \min\{1, L(y^\dagger|\boldsymbol{\theta}_* + m(\boldsymbol{\kappa}_n))/L(y^\dagger|\boldsymbol{\theta}_{n-1} + m(\boldsymbol{\kappa}_n))\}$
  Sample $\mathbf{U}_X \sim \mathrm{Unif}[0, 1]$
  if $\mathbf{U}_X \leq a_X$ **then**
    $\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_*$
  else
    $\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1}$
  end
end

the densities N_{smp} times. Pseudo-determinant and pseudo-inverse can for instance be computed with an incomplete spectral decomposition of $\mathcal{C}(\kappa)$; truncated after N_{sto} eigenpairs. This leads to an overall computational cost of

$$O(N_{\text{smp}} \cdot N_\ell^2 \cdot N_{\text{sto}}; N_{\text{smp}} \to \infty).$$

In this particular algorithm, the same eigendecomposition can then be reused without further cost to sample $\boldsymbol{\theta}_*$. Hence, the problem of computing pseudo-determinant and pseudo-inverse reduces to a spectral decomposition.

4.2 Low-rank approximation of parameterised covariances

In §4.1.2, we have discussed various computational tasks we need to perform when solving a hierarchical forward and inverse problem. In the following, we focus on the eigenproblem since we can solve all the other problems using the spectral decomposition we obtain from the eigenproblem. To solve the eigenproblems, we employ a reduced basis surrogate. This is the cornerstone in the algorithm we present in the following. To begin we explain the basic idea behind reduced basis approaches for eigenproblems in §4.2.1. This approach is very similar to RB approaches for parameterised PDEs, which we have briefly discussed in §3.4.1. There, we omitted discussing the offline-online decomposition of the operators. As we explain in §4.2.2, this step is vital for a fast construction of the reduced operators. For Matérn-type covariance operators, an exact offline-online decomposition is not possible. In §4.2.3, we discuss an approximate approach.

4.2.1 Basic idea

Let $\mathcal{C} : R \to \mathrm{CO}(X)$ be a measurable map, where $(X, \langle\cdot,\cdot\rangle_X) := (\mathbb{R}^{N_\ell}, \langle\cdot,\cdot\rangle_{M_\ell})$ is a finite-dimensional space arising from the discretisation of an infinite-dimensional function space (see §3.2). Recall that in Algorithm 1 we need to solve the generalised eigenproblem associated with $\mathcal{C}(\kappa)$ for multiple parameter values $\kappa \in R$; see (3.3). That is, we want to find $(\alpha_i(\kappa), \psi_i(\kappa))_{i=1}^{N_{\mathrm{sto}}} \in (\mathbb{R} \times X)^{N_{\mathrm{sto}}}$, such that

$$\mathcal{C}(\kappa)\psi_i(\kappa) = \alpha_i(\kappa) M_\ell \psi_i(\kappa). \tag{4.5}$$

X is in general high-dimensional, which results in a large computational cost for solving the eigenproblems. However, it is often not necessary to consider the full space X. If we assume that the eigenpairs corresponding to different parameter values are closely related, then the space

$$\mathrm{span}\{\psi_i(\kappa) : i = 1, \dots, N_{\mathrm{sto}}, \kappa \in R\} \subseteq X$$

can be approximated by a low dimensional subspace X_{RB}, where $N_{\mathrm{RB}} := \dim X_{\mathrm{RB}} \ll N_\ell$. We point out that the truncated KL expansion requires N_{sto} eigenpairs by assumption. However, the reduced operators are $N_{\mathrm{RB}} \times N_{\mathrm{RB}}$ matrices with N_{RB} eigenpairs. Hence, $N_{\mathrm{RB}} \geq N_{\mathrm{sto}}$ is required.
Now, let $W \in X_{\mathrm{RB}}^{N_{\mathrm{RB}}}$ be an orthonormal basis of X_{RB} with respect to the inner product $\langle\cdot,\cdot\rangle_{X_{\mathrm{RB}}} := \langle\cdot,\cdot\rangle_X := \langle\cdot,\cdot\rangle_{M_\ell}$. W is called *reduced basis* and X_{RB} is called *reduced space*. We can represent any function $\psi(\kappa) \in X_{\mathrm{RB}}$ by a coefficient vector $w(\kappa) \in \mathbb{R}^{N_{\mathrm{RB}}}$, such that $\psi(\kappa) = Ww(\kappa)$. The *reduced eigenproblem* is obtained by a Galerkin projection of the full eigenproblem in (4.5), and is again a generalised eigenproblem. The task is to find $(\alpha_i^{\mathrm{RB}}(\kappa), w_i^{\mathrm{RB}}(\kappa))_{i=1}^{N_{\mathrm{sto}}} \in (\mathbb{R} \times \mathbb{R}^{N_{\mathrm{RB}}})^{N_{\mathrm{sto}}}$, such that

$$\mathcal{C}^{\mathrm{RB}}(\kappa) w_i(\kappa) = \alpha_i^{\mathrm{RB}}(\kappa) M^{\mathrm{RB}} w_i(\kappa), \quad i = 1, \dots, N_{\mathrm{sto}}. \tag{4.6}$$

In (4.6), we have the *reduced operator* $\mathcal{C}^{\mathrm{RB}}(\kappa) := W^*\mathcal{C}(\kappa)W$ and the *reduced Gramian matrix* $M^{\mathrm{RB}} := W^*M_\ell W$, that are both $N_{\mathrm{RB}} \times N_{\mathrm{RB}}$ matrices. The eigenvector approximation in X_{RB} can then be obtained by

$$\psi_i^{\mathrm{RB}}(\kappa) = Ww_i(\kappa), \qquad i = 1, \dots, N_{\mathrm{sto}}.$$

4.2.2 Offline-online decomposition

A reduced basis method typically has two phases. In the offline phase, the reduced basis W is constructed. In the online phase, the reduced operator $\mathcal{C}^{\mathrm{RB}}(\kappa)$ is assembled and the reduced eigenproblem (4.6) is solved for selected parameter values $\kappa \in R$. To be able to shift a large part of the computational cost from the online to the offline phase, we assume that the following offline-online decomposition is available for the family of parameterised covariance operators.

Assumption 4.4. Let $N_{\mathrm{lin}} \in \mathbb{N}$. We assume that there are functions $F_k : R \to \mathbb{R}$ and linear operators $\mathcal{C}_k$, $k = 1, \dots, N_{\mathrm{lin}}$, such that

$$\mathcal{C}(\kappa) = \sum_{k=1}^{N_{\mathrm{lin}}} F_k(\kappa)\mathcal{C}_k, \quad \kappa \in R.$$

In this case, $\mathcal{C}(\kappa)$ is called a *linearly separable* operator. ◇

Offline phase. We use snapshots of the full eigenvectors to construct the reduced basis. Meaning that we choose a vector $\underline{\kappa}_{\mathrm{snap}} \in R^{N_{\mathrm{snap}}}$ and solve the full eigenproblem (4.5) for all entrys of $\underline{\kappa}_{\mathrm{snap}}$. We then have

$$W_{\mathrm{snap}} = (\psi_i(\kappa_{\mathrm{snap}}^{(s)}) : s = 1, \dots, N_{\mathrm{snap}}, i = 1, \dots, N_{\mathrm{sto}}),$$

where all of the computed eigenfunctions are included and here represented as column vectors. Hence, we obtain a matrix $W_{\mathrm{snap}} \in \mathbb{R}^{N \times N_{\mathrm{sto}} N_{\mathrm{snap}}}$. Moreover, we define the reduced space $X_{\mathrm{RB}} := \mathrm{span}(W_{\mathrm{snap}})$. Next, we construct an orthonormal basis for this vector space, using the *proper orthogonal decomposition*; see §3.4.1. As result of the POD we obtain a singular value decomposition of W_{snap} of the form

$$W_{\mathrm{snap}} = U \cdot \Sigma \cdot V^*,$$

where $\Sigma := \mathrm{diag}(\sigma_1, \dots, \sigma_N)$ is a diagonal matrix containing the singular values of W_{snap} and each column of $U = (u_1, u_2, \dots, u_{(N_{\mathrm{sto}} \cdot N_{\mathrm{snap}})})$ contains the associated orthonormal basis vectors. We use the basis vectors with non-zero singular values as basis vectors of X_{RB}, that is,

$$W := (u_i : \sigma_i > 0, i = 1, \dots, N_{\mathrm{sto}} \cdot N_{\mathrm{snap}}).$$

The magnitude of the singular values of W_{snap} is an indicator for the error when the corresponding basis vectors are not included in W. See the corresponding discussion in §3.2, for eigenvectors and eigenvalues. Neglecting reduced basis vectors, however, is beneficial due to the smaller dimension of the reduced basis. Depending on the pay-off of the dimension reduction compared to the approximation accuracy of the reduced basis one can choose a threshold $\underline{\sigma} > 0$ and work with the basis

$$W := (u_i : \sigma_i > \underline{\sigma}, i = 1, \dots, N_{\mathrm{sto}} \cdot N_{\mathrm{snap}}).$$

In this case, we redefine $X_{\mathrm{RB}} := \mathrm{span}(W)$.

Remark 4.5. We have discussed the importance of significant dimension reduction in reduced basis methods for PDEs in Example 3.14. This is mostly due to the full operators being sparse and the reduced operators being dense.
Notably, in most cases the discretised KL eigenproblem results in a dense matrix since the covariance integral operator is non-local. Hence, we expect a significant reduction of the total computational cost even if the size of the reduced basis is only slightly smaller than the number of unknowns in the unreduced eigenspace. ◇

As mentioned in §3.4.1, there are many options to choose the snapshot parameter values $\boldsymbol{\kappa}_{\text{snap}}$. In our applications κ is a random variable $\boldsymbol{\kappa}$ with probability measure μ'. Hence, a straightforward method is to sample independently $\boldsymbol{\kappa}_{\text{snap}}^{(s)} \sim \mu'$ ($s = 1, \dots, N_{\text{snap}}$). Alternatively, one can select deterministic points in R, e.g. quadrature nodes. We will come back to this question in §4.4 where we discuss some numerical experiments. Furthermore, note that it is generally possible to use different reduced bases $W_1, W_2, \dots$ for different subsets $R_1, R_2, \dots \subseteq R$ of hyperparameters and/or index sets $I_1, I_2, \dots \subseteq \{1, \dots, N_{\text{sto}}\}$ of eigenpairs. However, we do not pursue these ideas in this thesis.

Online phase. In the online phase we iterate over various hyperparameter values $\kappa \in R$. In every step, we assemble the operator $\mathcal{C}^{\text{RB}}(\kappa)$, and then we solve the associated eigenproblem (4.6). By Assumption 4.4 it holds

$$\mathcal{C}(\kappa) = \sum_{k=1}^{N_{\text{lin}}} F_k(\kappa)\mathcal{C}_k.$$

Hence, the reduced operator can be assembled efficiently as follows,

$$\mathcal{C}^{\text{RB}}(\kappa) = W^* \sum_{k=1}^{N_{\text{lin}}} F_k(\kappa)\mathcal{C}_k W = \sum_{k=1}^{N_{\text{lin}}} F_k(\kappa) W^* \mathcal{C}_k W = \sum_{k=1}^{N_{\text{lin}}} F_k(\kappa)\mathcal{C}_k^{\text{RB}}.$$

The reduced operators $\mathcal{C}_k^{\text{RB}}$ ($k = 1, \dots, N_{\text{lin}}$) can be computed in the offline phase and stored in the memory. In the online phase, we then only need to compute a certain linear combination of $(\mathcal{C}_k^{\text{RB}})_{k=1}^{N_{\text{lin}}}$. This reduces the computational cost of the assembly of the reduced operator significantly; we come back to this in §4.3.2. After the assembly step, we solve the reduced eigenproblem (4.6) to obtain the eigenfunctions $\psi_i^{\text{RB}}(\kappa) := W w_i(\kappa) \in X$ and eigenvalues $\alpha_i^{\text{RB}}(\boldsymbol{\kappa}), i = 1, \dots, N_{\text{sto}}$.

4.2.3 Matérn-type covariance operators

Matérn-type covariance operators are widely used in spatial statistics and uncertainty quantification. They are particularly popular for modelling spatially variable uncertainties in porous media. We are interested in solving the KL eigenproblem with Matérn covariance operators with hyperparameters, e.g. the correlation length. Unfortunately, the Matérn-type covariance operators are not linearly separable with respect to the hyperparameters of interest; see Assumption 4.4. For this reason we introduce and analyse a class of linearly separable covariance operators which can approximate Matérn-type covariance operators with arbitrary accuracy.

Definition 4.6 (Matérn)**.** Let $D \subseteq \mathbb{R}^d, d = 1, 2, 3$ be an open, bounded and connected domain, and let $X := \mathbf{L}^2(D; \mathbb{R})$. Furthermore, let $\lambda \in (0, \mathrm{diam}(D)), \nu \in (0, \infty], \sigma \in (0, \infty)$. Define the covariance kernel $c(\nu, \lambda, \sigma) : [0, \infty) \to [0, \infty)$ as

$$z \mapsto c(\nu, \lambda, \sigma)(z) = \frac{\sigma^2 \cdot 2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{z}{\lambda}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{z}{\lambda}\right),$$

where K_ν is the modified Bessel function of the second kind. Then, the *Matérn-type covariance operator with smoothness ν, standard deviation σ and correlation length λ* is given by

$$\mathcal{C}(\nu, \lambda, \sigma) : X \to X, \varphi \mapsto \int_D \varphi(x) c(\nu, \lambda, \sigma)(\mathrm{dist}(x, \cdot))\mathrm{d}x,$$

where $\mathrm{dist} : D \times D \to [0, \infty)$ is the Euclidean distance in D. ◇

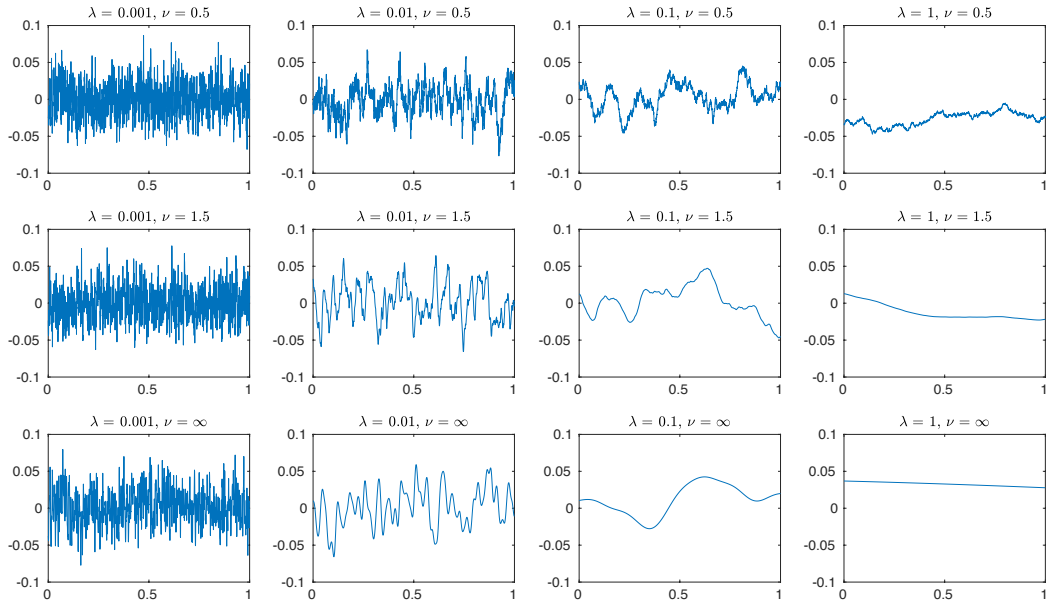


Figure 4.1. Samples of mean-zero Gaussian random fields in 1D with Matérn covariance, with $\nu \in \{0.5, 1.5, \infty\}$, $\lambda \in \{0.001, 0.01, 0.1, 1\}$, and $\sigma := 1$. The samples are discretised with 2000 piecewise constant finite elements, regularly distributed over $D = [0, 1]$.

We show samples of Gaussian random fields with Matérn-type covariance operators in Figure 4.1, with various correlation lengths and smoothness parameters. There, we see that the smoothness parameter indeed influences the smoothness of the random field samples. The correlation length has the same influence that we have already seen for the exponential covariance kernel in Figures 1.1–1.2. Actually, the exponential covariance kernel is contained in the Matérn class.

Remark 4.7. The exponential covariance operator in Examples 1.26 and 4.2 is a Matérn-type covariance operator. Indeed, $\mathcal{C}(1/2, \lambda, \sigma) \equiv \mathcal{C}_{\mathrm{exp}}^{(\lambda,\sigma)}$. ◇

In Example 4.2, we have discussed the possibility of using the standard deviation σ and the correlation length λ as hyperparameters in a Matérn-type covariance operator. *What are the computational implications for the KL expansion?* Changing

σ only rescales the KL eigenvalues and does not require a re-computation of the KL expansion. However, changing the correlation length clearly changes the KL eigenfunctions. We can see this in Figure 4.2. The good news is that the KL eigenfunctions for different correlation lengths are very similar, for example, the number and location of extrema is preserved. This suggests that we might be able to construct a useful reduced basis from selected snapshots of KL eigenpairs corresponding to different correlation lengths.

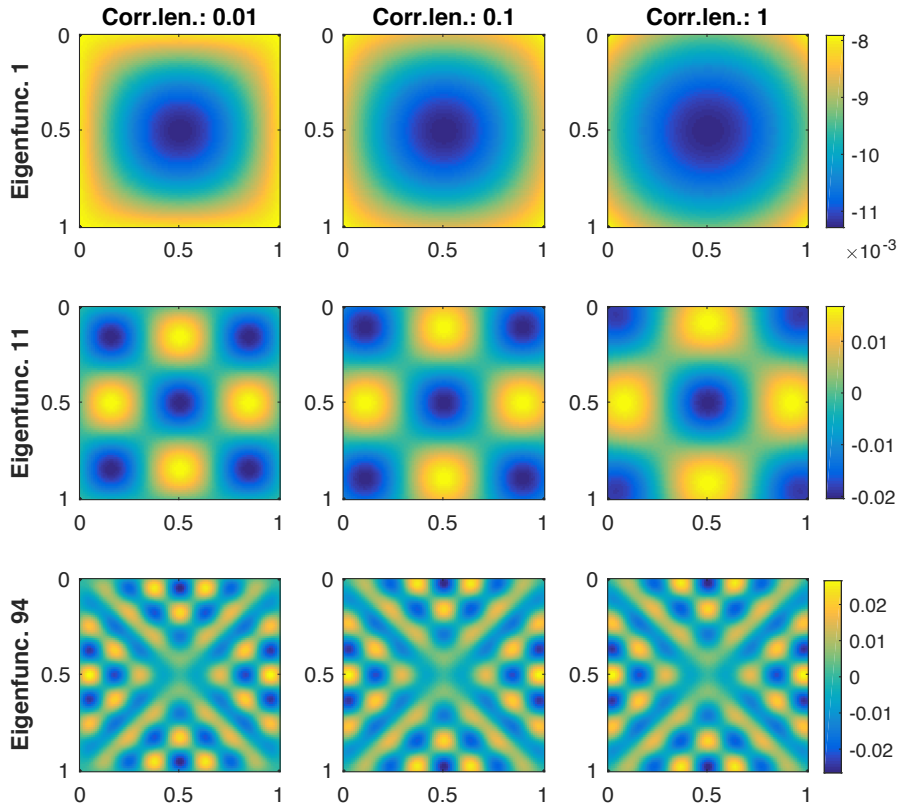


Figure 4.2. Eigenfunctions 1, 11 and 94 of the Matérn-type covariance operator with correlation lengths $\lambda = 0.01, 0.1, 1$ and smoothness $\nu = 1/2$.

Being able to construct and use the reduced basis efficiently requires the linear separability of the covariance operator, see Assumption 4.4. The Matérn operator is linearly separable with respect to σ. Unfortunately, it is not linearly separable with respect to λ since the covariance function $c(\nu, \lambda, \sigma)$ is not linearly separable. However, it is possible to approximate $\mathcal{C}(\nu, \lambda, \sigma)$ with any precision by a linearly separable operator. Using the approximate, linearly separable operator allows us to construct an offline-online decomposition for the exact Matérn covariance operator without the need to use advanced linearisation techniques, such as the discrete empirical interpolation method. We show this in the remainder of this section for $\nu \in (0, \infty) \backslash \mathbb{N}$. Similar approximations for $\nu \in \mathbb{N} \cup \{\infty\}$ follow from the analyticity of K_ν.

Assumption 4.8. The correlation length λ satisfies $0 < \underline{\lambda} \leq \lambda$ with fixed $\underline{\lambda}$. ◇

Definition 4.9. Let $\nu \in (0, \infty) \backslash \mathbb{N}$ and $N_{\text{lin}} \in 2\mathbb{N}$. Moreover, let Assumption 4.8

hold. We define the N_{lin}-*term approximation* of $c(\nu, \lambda, \sigma)$ by

$$\begin{aligned} &c(\nu, \lambda, \sigma, N_{\text{lin}})(z) \\ &:= \frac{\sigma^2 \pi \csc(\pi\nu)}{\Gamma(\nu)} \sum_{k=1}^{N_{\text{lin}}/2} \frac{(\sqrt{\nu} z/\lambda)^{2k-2}}{2^{k-1}\Gamma(k-\nu)(k-1)!} - \frac{(\sqrt{\nu} z/\lambda)^{2\nu+2k-2}}{2^{k+\nu-1}\Gamma(k+\nu)(k-1)!}. \end{aligned}$$

The associated operator is then defined as

$$\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}}) : X \to X, \varphi \mapsto \int_D \varphi(x) c(\nu, \lambda, \sigma, N_{\text{lin}})(\operatorname{dist}(x, \cdot)) \mathrm{d}x.$$

◇

Note that the operator $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}})$ is linearly separable w.r.t. λ. In particular, we have

$$\begin{aligned} \widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}}) &:= \sum_{k=1}^{N_{\text{lin}}} F_k(\nu, \lambda, \sigma) \mathcal{C}_k(\nu) \\ F_k(\nu, \lambda, \sigma) &:= \frac{\pi \csc(\pi\nu)}{\Gamma(\nu)} \cdot \begin{cases} \frac{\sigma^2}{\lambda^{k-2}}, & \text{if } k \in 2\mathbb{N}, \\ \frac{\sigma^2}{\lambda^{2\nu+k-1}}, & \text{if } k \in 2\mathbb{N}-1, \end{cases} \\ \mathcal{C}_k \varphi &:= \begin{cases} \int_D \varphi(x) \frac{(\sqrt{\nu} \cdot \operatorname{dist}(x,\cdot))^{k-2}}{2^{k/2-1}\Gamma(k/2-\nu)(k/2-1)!} \mathrm{d}x, & \text{if } k \in 2\mathbb{N}, \\ \int_D \varphi(x) \frac{(-1)\cdot(\sqrt{\nu} \cdot \operatorname{dist}(x,\cdot))^{2\nu+k-1}}{2^{k/2+\nu-1/2}\Gamma(k/2+1/2+\nu)(k/2-1/2)!} \mathrm{d}x, & \text{if } k \in 2\mathbb{N}-1, \end{cases} \end{aligned}$$

for any $\varphi \in X$.
The operator $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}})$ arises from a truncation of a series expansion of K_ν. This is detailed in the proof of the following lemma, where we derive an error bound between the exact Matérn covariance operator $\mathcal{C}(\nu, \lambda, \sigma)$ and the linearly separable approximation $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}})$.

Lemma 4.10. Let $\nu \in (0, \infty)\backslash\mathbb{N}$, let $N_{\text{lin}} \in 2\mathbb{N}$ and let Assumption 4.8 hold. Then,

$$\begin{aligned} &\|\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\text{lin}}) - \mathcal{C}(\nu, \lambda, \sigma)\|_X \\ &\qquad \leq \operatorname{diam}(D)^{2d} \frac{\pi |\csc(\pi\nu)|}{2^{1-\nu}} (1 + \zeta_{\max}^{2\nu}) \exp\left(\frac{\zeta_{\max}^2}{4}\right) \frac{\zeta_{\max}^{2N_{\text{lin}}}}{(N_{\text{lin}})!}, \end{aligned}$$

where $\zeta_{\max} := \operatorname{diam}(D)/\underline{\lambda}$.

Proof. Let $\nu \in (0, \infty)\backslash\mathbb{N}$. Consider the function

$$f_{(\nu)} : [0, \infty) \to [0, \infty), \quad \zeta \mapsto \zeta^\nu \cdot K_\nu(\zeta).$$

It holds $f_{(\nu)}(\sqrt{2\nu} z/\lambda) = \operatorname{const}(\nu, \sigma) c(\nu, \lambda, \sigma)(z)$, where $\operatorname{const}(\nu, \sigma) > 0$ is a constant that does not depend on the correlation length λ. Moreover, we assume that we work in a bounded computational domain D and that λ is bounded from below by a fixed positive constant $\underline{\lambda} > 0$ (see Assumption 4.8). Now, for $\nu \in (0, \infty)\backslash\mathbb{N}$ the function $f_{(\nu)}$ can be written in terms of a series

$$\begin{aligned} &f_{(\nu)}(\zeta) \\ &= \frac{\pi \csc(\pi\nu)}{2} \sum_{k=1}^{\infty} \left(\frac{1}{2^{2k-2-\nu}\Gamma(k-\nu)(k-1)!} - \frac{\zeta^{2\nu}}{2^{2k-2+\nu}\Gamma(k+\nu)(k-1)!} \right) \zeta^{2k-2}. \end{aligned}$$

This follows from the representations of K_ν and I_ν in [1, Equations 9.6.2, 9.6.10]. If we truncate the series after the first $N_{\text{lin}}/2$ terms, we obtain the following function:

$$f_{(\nu,N_{\text{lin}})}(\zeta) = \frac{\pi\csc(\pi\nu)}{2}\sum_{k=1}^{N_{\text{lin}}/2}\left(\frac{1}{2^{2k-2-\nu}\Gamma(k-\nu)(k-1)!} - \frac{\zeta^{2\nu}}{2^{2k-2+\nu}\Gamma(k+\nu)(k-1)!}\right)\zeta^{2k-2}.$$

This (truncated) series expansion is associated with the integral operator $\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\text{lin}})$ that is given by the kernel

$$\widetilde{c}^{(\nu,\lambda,\sigma,N_{\text{lin}})}(z) := \frac{f_{(\nu,N_{\text{lin}})}(\sqrt{2\nu}z/\lambda)}{\text{const}(\nu,\sigma)}. \tag{4.7}$$

Now, we bound the asymptotic truncation error. Assume w.l.o.g. that $N_{\text{lin}} > \nu$. Note that in this case $\Gamma(k-\nu) > 1$, and, moreover, $\zeta > 0$. Using the triangle inequality we arrive at

$$\begin{aligned}
|f_{(\nu,N_{\text{lin}})}(\zeta) - f_{(\nu)}(\zeta)| &\leq \frac{\pi|\csc(\pi\nu)|}{2}\left|\sum_{k=N_{\text{lin}}+1}^{\infty}\left(\frac{1-\zeta^{2\nu}}{2^{2k-2-\nu}\Gamma(k-\nu)(k-1)!}\right)\zeta^{2k-2}\right| \\
&\leq \frac{\pi|\csc(\pi\nu)|}{2^{1-\nu}}(1+\zeta^{2\nu})\sum_{k=N_{\text{lin}}+1}^{\infty}\left(\frac{1}{\Gamma(k-\nu)(k-1)!}\right)\left(\frac{\zeta}{2}\right)^{2k-2} \\
&\leq \frac{\pi|\csc(\pi\nu)|}{2^{1-\nu}}(1+\zeta^{2\nu})\sum_{k=N_{\text{lin}}+1}^{\infty}\left(\frac{1}{k-1!}\right)\left(\frac{\zeta}{2}\right)^{2k-2}.
\end{aligned}$$

The infinite sum on the right-hand side above can be bounded by the remainder term of a Taylor series expansion of the exponential function with N_{lin} terms and anchor point $\zeta = 0$. This gives

$$\begin{aligned}
|f_{(\nu,N_{\text{lin}})}(\zeta) - f_{(\nu)}(\zeta)| &\leq \frac{\pi|\csc(\pi\nu)|}{2^{1-\nu}}(1+\zeta^{2\nu})\exp\left(\frac{\zeta_{\max}^2}{4}\right)\frac{\zeta^{2N_{\text{lin}}}}{(N_{\text{lin}})!} \\
&\leq \frac{\pi|\csc(\pi\nu)|}{2^{1-\nu}}(1+\zeta_{\max}^{2\nu})\exp\left(\frac{\zeta_{\max}^2}{4}\right)\frac{\zeta_{\max}^{2N_{\text{lin}}}}{(N_{\text{lin}})!} =: \text{const}'(N_{\text{lin}}),
\end{aligned}$$

where $\zeta_{\max} = \text{diam}(D)/\underline{\lambda}$. Finally, let $\varphi \in X$. By the Cauchy–Schwarz inequality it holds

$$\begin{aligned}
&\|\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\text{lin}})\varphi - \mathcal{C}(\nu,\lambda,\sigma)\varphi\|_X^2 \\
&= \int_D\left(\int_D(\widetilde{c}^{(\nu,\lambda,\sigma,N_{\text{lin}})}(\text{dist}(x,x')) - c(\nu,\lambda,\sigma)(\text{dist}(x,x')))\varphi(x)\mathrm{d}x\right)^2\mathrm{d}x' \\
&\leq \int_D\left(\int_D(\widetilde{c}^{(\nu,\lambda,\sigma,N_{\text{lin}})}(\text{dist}(x,x')) - c(\nu,\lambda,\sigma)(\text{dist}(x,x')))^2\mathrm{d}x\right)\cdot\left(\int_D\varphi(x)^2\mathrm{d}x\right)\mathrm{d}x' \\
&\leq \text{Leb}(d)(D)^2\cdot\text{const}'(N_{\text{lin}})^2\cdot\|\varphi\|_X^2 \\
&\leq \text{diam}(D)^{2d}\cdot\text{const}'(N_{\text{lin}})^2\cdot\|\varphi\|_X^2.
\end{aligned}$$

Taking the square root on both sides and dividing by $\|\varphi\|_X$ gives the desired error bound. □

The covariance operator approximation brings new issues. The Matérn-type covariance operators $\mathcal{C}(\nu, \lambda, \sigma)$ are valid covariance operators in $\mathrm{CO}(X)$. However, this is not necessarily the case for $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$. One can easily verify the following.

Lemma 4.11. The operator $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ is self-adjoint, nuclear, and continuous.

Proof. The integral operator $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ is self-adjoint since the associated kernel function is symmetric. The operator is nuclear since D is a bounded domain, and

$$\int_D \widetilde{c}(\nu, \lambda, \sigma, N_{\mathrm{lin}})(\mathrm{dist}(x, x))\mathrm{d}x = \widetilde{c}(\nu, \lambda, \sigma, N_{\mathrm{lin}})(0) \cdot \mathrm{Leb}(d)(D) < \infty.$$

The boundedness of D also implies the continuity of the operator. □

However, $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ is not necessarily positive definite. Under weak assumptions we can cure this by replacing $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ with an operator $\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ which has the exact same eigenfunctions and positive eigenvalues as $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$, however, all negative eigenvalues are set to zero. Formally, we define $\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ by

$$\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}) = \sum_{i=1;\widetilde{\alpha}_i>0}^{\infty} \widetilde{\alpha}_i(\widetilde{\psi}_i \otimes \widetilde{\psi}_i), \tag{4.8}$$

where $(\widetilde{\alpha}_i, \widetilde{\psi}_i)_{i=1}^{\infty}$ are eigenpairs of $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ and the eigenfunctions are orthonormal. Note that the same technique has been applied in [42] to remove the degeneracy of multilevel sample covariance estimators. Fortunately, we can show that the approximation error of $\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ is of the same order as the error of $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$.

Lemma 4.12. The Matérn-type covariance operator $\mathcal{C}(\nu, \lambda, \sigma)$ and the approximate operator $\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ in (4.8) satisfy

$$\|\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}) - \mathcal{C}(\nu, \lambda, \sigma)\|_X \leq 2\|\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}}) - \mathcal{C}(\nu, \lambda, \sigma)\|_X.$$

Proof. Let $(\widetilde{\lambda}_i)_{i=1}^{\infty}$ denote the eigenvalues of the operator $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$. Without loss of generality, we assume that the spectrum of $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ contains a negative eigenvalue. Since $\widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})$ is trace-class, it holds $|\widetilde{\alpha}_i| \to 0$ for $i \to \infty$. Hence, there is an eigenpair $(\widetilde{\alpha}_{\max}, \widetilde{\psi}_{\max})$ which realises the maximum in the expression

$$\max_{i \in \mathbb{N}:\, \widetilde{\alpha}_i<0} |\widetilde{\alpha}_i|.$$

Thus,

$$\|\widetilde{\mathcal{C}}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}) - \widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}})\|_X = \left\| \sum_{i=1;\lambda_i<0}^{\infty} \widetilde{\alpha}_i \widetilde{\psi}_i \otimes \widetilde{\psi}_i \right\|_X = |\widetilde{\alpha}_{\max}|.$$

Moreover, since $\mathcal{C}(\nu, \lambda, \sigma)$ is positive definite, we have

$$\langle \widetilde{\psi}_{\max}, \widetilde{\mathcal{C}}(\nu, \lambda, \sigma, N_{\mathrm{lin}}) \widetilde{\psi}_{\max} \rangle_X \geq 0 > \langle \widetilde{\psi}_{\max}, \mathcal{C}(\nu, \lambda, \sigma) \widetilde{\psi}_{\max} \rangle.$$

Hence, we obtain

$$\begin{aligned}
&\|\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X\\
&\qquad\geq|\langle\widetilde{\psi}_{\max},(\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma))\widetilde{\psi}_{\max}\rangle_X|\\
&\qquad=\langle\widetilde{\psi}_{\max},\mathcal{C}(\nu,\lambda,\sigma)\widetilde{\psi}_{\max}\rangle_X-\langle\widetilde{\psi}_{\max},\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})\widetilde{\psi}_{\max}\rangle_X\\
&\qquad=\langle\widetilde{\psi}_{\max},\mathcal{C}(\nu,\lambda,\sigma)\widetilde{\psi}_{\max}\rangle_X-\widetilde{\alpha}_{\max}\langle\widetilde{\psi}_{\max},\widetilde{\psi}_{\max}\rangle_X\\
&\qquad\geq|\widetilde{\alpha}_{\max}|.
\end{aligned}$$

This gives the bound

$$\|\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})\|_X\leq\|\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X.$$

Finally, using the triangle inequality, we arrive at

$$\begin{aligned}
&\|\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X\\
&\quad\leq\|\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})\|_X+\|\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X\\
&\quad\leq 2\|\widetilde{\mathcal{C}}(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X.
\end{aligned}$$

□

We summarise the results in Lemma 4.10–4.12 as follows.

Proposition 4.13. Let $\nu\in(0,\infty)\setminus\mathbb{N}$. Under Assumption 4.8 there is a linearly separable, valid covariance operator $\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}})\in\mathrm{CO}(X)$ consisting of $N_{\mathrm{lin}}\in 2\mathbb{N}$ terms, such that

$$\|\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}})-\mathcal{C}(\nu,\lambda,\sigma)\|_X\leq\mathrm{const}''(\nu)(1+\zeta_{\max}^{2\nu})\exp\left(\frac{\zeta_{\max}^2}{4}\right)\frac{\zeta_{\max}^{2N_{\mathrm{lin}}}}{(N_{\mathrm{lin}})!},$$

where $\zeta_{\max}:=\mathrm{diam}(D)/\underline{\lambda}$ and $\mathrm{const}''(\nu)>0$ is a constant that depends only on ν. ◇

The expansion in (4.8) has infinitely many terms. We truncate this expansion and retain only the leading N_{sto} terms, denoting the resulting covariance operator by $\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}})$.

Finally, we discuss the sample path continuity; recall the discussion in §1.2.3. In the following proposition, we show that this also holds for the realisations of the Gaussian random fields with measure $\mathrm{N}(0,\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}}))$, when representing the measure on a space fine enough to consider continuity.

Let $Z:=\mathbb{R}^{\overline{D}}$ be the set of functions from $\overline{D}$ to $\mathbb{R}$. Moreover, let $\mathcal{F}$ be the cylindrical σ-algebra on Z. For the sake of simplicity, we assume that $D:=(0,1)$. We need to construct a random field on Z that corresponds to $\mathrm{N}(0,\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}}))\in\mathrm{Prob}(X)$. To this end, we construct a covariance function $\widetilde{c}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}})$ that corresponds to the operator $\widetilde{\mathcal{C}}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}})$. To define the covariance function, we proceed as in (4.8) and truncate after N_{sto} terms:

$$\widetilde{c}_0(\nu,\lambda,\sigma,N_{\mathrm{lin}},N_{\mathrm{sto}})(x,x'):=\sum_{i=1;\widetilde{\alpha}_i>0}^{N'_{\mathrm{sto}}}\widetilde{\alpha}_i(\widetilde{\psi}_i(x)\cdot\widetilde{\psi}_i(x'))\qquad(x,x'\in D),\tag{4.9}$$

where N'_{sto} is chosen such that the sum above has N_{sto} terms, and (α_i, ψ_i) satisfy

$$\widetilde{\alpha}_i \widetilde{\psi}_i(x') = \int_D \widetilde{\psi}_i(x) \widetilde{c}(\nu, \lambda, \sigma, N_{\mathrm{lin}})(\mathrm{dist}(x, x'))\mathrm{d}x \quad (x' \in D, i \in \mathbb{N}).$$

Moreover, $(\widetilde{\alpha}_i)_{i\in\mathbb{N}}$ are ordered descendingly with respect to their modulus. The existence of the eigenpairs can be shown by iteratively applying [259, Proposition 3.2.2]; the enumeration as a sequence is possible, since [259, Proposition 3.2.8]. Thus, $\widetilde{c}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}, N_{\mathrm{sto}})$ is well-defined. By construction it is indeed a covariance function; see Definition 1.27. Note that the continuity of $\widetilde{c}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}, N_{\mathrm{sto}})$ is implied by the $\widehat{\psi}_i$'s being continuous. This can be seen in (4.9), where $\widetilde{c}(\nu, \lambda, \sigma, N_{\mathrm{lin}})(\mathrm{dist}(\cdot, \cdot))$ is a continuous function. Let

$$\boldsymbol{\theta}(x) := \sum_{i=1;\widetilde{\alpha}_i>0}^{N'_{\mathrm{sto}}} \sqrt{\widetilde{\alpha}_i}\boldsymbol{\xi}_i \widetilde{\psi}_i(x) \quad (x \in D), \tag{4.10}$$

where $\boldsymbol{\xi}_i \sim \mathrm{N}(0, 1)$, $i \in \mathbb{N}$. According to [3, Theorem 3.3.2], $\boldsymbol{\theta}$ is then a Gaussian random field on $(Z, \mathcal{F})$ with mean zero and covariance kernel $\widetilde{c}_0(\nu, \lambda, \sigma, N_{\mathrm{lin}}, N_{\mathrm{sto}})$. Now, we can show that this Gaussian random field has continuous sample paths, i.e. every realisation has a continuous modification; see Theorem 1.29. Note that we will refer to the precisely defined random variable $\boldsymbol{\theta}$. Since this random variable is not implicitly defined by its probability measure, we can consider continuity of $\boldsymbol{\theta}$ immediately, not up to its modifications.

Proposition 4.14. Let $\boldsymbol{\theta}$ be given as in (4.10). Then, realisations of $\boldsymbol{\theta}$ are continuous.

Proof. We consider a realisation

$$\theta(x) := \sum_{i=1;\widetilde{\alpha}_i>0}^{N'_{\mathrm{sto}}} \sqrt{\widetilde{\alpha}_i}\xi_i \widetilde{\psi}_i(x) \quad (x \in D)$$

of $\boldsymbol{\theta}$. Hence, $\xi_i \in \mathbb{R}, i \in \mathbb{N}$ are deterministic values. Now, θ is a continuous function, if the $\widetilde{\psi}_i$'s are continuous functions. As mentioned before, this is true due to (4.9). □

We now comment on the error bound given in Lemma 4.10 and Proposition 4.13. As N_{lin} increases, the error bound goes to zero, asymptotically like $O(1/(N_{\mathrm{lin}}!); N_{\mathrm{lin}} \to \infty)$. However, when the lower bound of the correlation length $\underline{\lambda}$ is small, the prefactor of the error bound explodes like $O(\exp(\underline{\lambda}^{-2}); \underline{\lambda} \downarrow 0)$. Hence, for small $\underline{\lambda}$, a very large number of terms N_{lin} is required to obtain a small error. In addition, for large N_{lin}, numerical cancellations occur and reduce the accuracy of the approximation. We show this for the exponential covariance in Figure 4.3 where we plot the truncation error

$$\sup_{z\in[0,\sqrt{2}],\lambda\in[\underline{\lambda},\sqrt{2}]} |c(\nu, \lambda, \sigma, N_{\mathrm{lin}})(z) - c(\nu, \lambda, \sigma)(z)|, \tag{4.11}$$

where $\nu = 1/2$ refers to the exponential covariance and $\sigma = 1$, for different choices of N_{lin} and $\underline{\lambda}$. We clearly see that the linearisation technique in Definition 4.9 is not

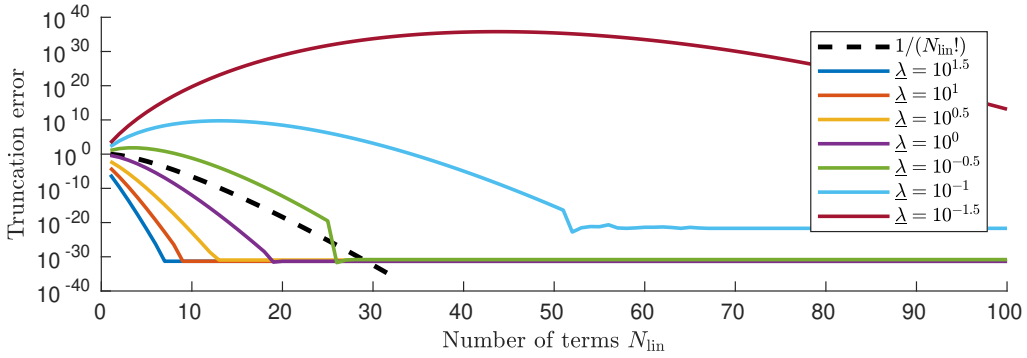


Figure 4.3. Error in the exponential covariance kernel when using the truncated linearisation in Definition 4.9, for different numbers of terms $N_{\text{lin}} = 1, \dots, 100$ and different minimal correlation lengths $\underline{\lambda} = 1\text{E-}1.5, \dots, 1\text{E}1.5$.

suitable for very small correlation lengths. In such a case, one could use alternative linearisation techniques, e.g. a polynomial chaos expansions in λ, a Taylor expansion of the Fourier representation of the Matérn kernel, or empirical interpolation.
Until now we considered the estimation of λ and σ, but not the estimation of ν. We comment on this in the following remark. Note further that the estimation of the smoothness ν of a Gaussian random field is studied, for instance, in the PhD thesis [116, §4], where a maximum-likelihood type estimation of the smoothness is performed.

Remark 4.15. For the Matérn kernel, the map $\nu \mapsto \zeta^\nu K_\nu(\zeta)$ is analytic for fixed $\zeta > 0$. Hence a linearisation as discussed in this section is generally possible. However, we expect that a reduced basis approach is not an efficient way to estimate the smoothness of a random field. Note that the smoothness parameter determines the smoothness of the random field realisations and shape of the eigenfunctions. To accurately represent functions with variable smoothnesses one would need separate reduced bases for each value of the smoothness parameter since otherwise we cannot guarantee mathematically that the random field sample is a.s. in the correct function space. We expect that over a small range of smoothness parameters a reduced basis could be constructed, however, the potential computational savings are limited. ◇

4.3 Reduced basis sampling

In §4.1.2, we have discussed sampling from a Markov kernel, or, equivalently, a parameterised Gaussian measure

$$K(\cdot\,|\boldsymbol{\kappa}) := \mathrm{N}(m(\boldsymbol{\kappa}), \mathcal{C}(\boldsymbol{\kappa})),$$

where $\boldsymbol{\kappa} \sim \mu'$. Now, in §4.3.1, to reduce the computational cost, we combine Algorithm 1 and reduced bases. We discuss the computational cost of the offline and the online phase of the suggested reduced basis sampling in §4.3.2. Finally, in §4.3.3 we explain how the reduced basis induces an alternative expansion for the parameterised Gaussian random field.

4.3.1 Algorithm

First, we describe the offline phase of the reduced basis sampling. We use a POD approach to compute a reduced basis W for $\mathcal{C}(\cdot)$. Here, it is important that the dimension of $X_{\mathrm{RB}} = \mathrm{span}(W)$ is larger than N_{sto}. Furthermore, we assume that $\mathcal{C}(\cdot)$ fulfils Assumption 4.4, i.e., it has the linearly separable form

$$\mathcal{C}(\cdot) := \sum_{k=1}^{N_{\mathrm{lin}}} F_k(\cdot)\mathcal{C}_k.$$

Having constructed the reduced basis W, we compute $\mathcal{C}_k^{\mathrm{RB}} = W^*\mathcal{C}_k W, k = 1, \ldots, N_{\mathrm{lin}}$. Then, we proceed with the online phase (see Algorithm 3). We iterate over $n =$

Algorithm 3: Reduced basis sampling from the parameterised Gaussian measure

for $n \in \{1, \ldots, N_{\mathrm{smp}}\}$ **do**
- Sample $\boldsymbol{\kappa}_n \sim \mu'$
- $\mathcal{C}^{\mathrm{RB}}(\boldsymbol{\kappa}_n) \leftarrow \sum_{k=1}^{N_{\mathrm{lin}}} F_k(\boldsymbol{\kappa}_n)\mathcal{C}_k^{\mathrm{RB}}$
- $\Psi_n^{\mathrm{RB}}(\boldsymbol{\kappa}_n) \leftarrow \mathrm{eigs}(\mathcal{C}^{\mathrm{RB}}(\boldsymbol{\kappa}_n), N_{\mathrm{sto}})$
- $\Psi_n(\boldsymbol{\kappa}_n) \leftarrow W\Psi_n^{\mathrm{RB}}(\boldsymbol{\kappa}_n)$
- Sample $\boldsymbol{\xi} \sim \mathrm{N}(0, \mathrm{Id}_{N_{\mathrm{sto}}})$
- $\theta_n \leftarrow m(\boldsymbol{\kappa}_n) + \Psi_n(\boldsymbol{\kappa}_n)\boldsymbol{\xi}$

end

$1, \ldots, N_{\mathrm{smp}}$. In each step, we first sample $\boldsymbol{\kappa}_n \sim \mu'$. Then, we evaluate the reduced covariance operator $\mathcal{C}^{\mathrm{RB}}(\boldsymbol{\kappa}_n)$, compute the eigenpairs $(\alpha_i^{\mathrm{RB}}(\boldsymbol{\kappa}_n), w_i^{\mathrm{RB}}(\boldsymbol{\kappa}_n))_{i=1}^{N_{\mathrm{sto}}}$ of $\mathcal{C}^{\mathrm{RB}}(\boldsymbol{\kappa}_n)$, and return $\Psi_n^{\mathrm{RB}}(\boldsymbol{\kappa}_n) := \left(\sqrt{\alpha_i^{\mathrm{RB}}(\boldsymbol{\kappa}_n)}w_i^{\mathrm{RB}}(\boldsymbol{\kappa}_n) : i = 1, \ldots, N_{\mathrm{sto}}\right)$. Next, we compute the representation of $\Psi_n^{\mathrm{RB}}(\boldsymbol{\kappa}_n)$ on the full space X, that is, $\Psi_n(\boldsymbol{\kappa}_n) := W\Psi_n^{\mathrm{RB}}(\boldsymbol{\kappa}_n)$. Finally, we proceed as in Algorithm 1: We sample a multivariate standard Gaussian random variable with N_{sto} components and return $m(\boldsymbol{\kappa}_n) + \Psi_n(\boldsymbol{\kappa}_n)\boldsymbol{\xi} \sim \mathrm{N}(m(\boldsymbol{\kappa}_n), \mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}_n))$. The covariance operator is given by

$$\mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}_n) = \Psi_n(\boldsymbol{\kappa}_n)\Psi_n(\boldsymbol{\kappa}_n)^*.$$

4.3.2 Computational cost of reduced basis sampling

We assume again that $X := X_\ell := \mathbb{R}^{N_\ell}$. The number of solves of the full eigenproblem in the offline phase is N_{snap}. We consider the following assumptions throughout the rest of the chapter. We will see that these are crucial to obtain a speed-up with reduced basis sampling.

Assumption 4.16. Let $N_{\mathrm{snap}} \ll N_{\mathrm{smp}}$, $N_{\mathrm{sto}} \leq N_{\mathrm{RB}} \ll N_\ell$. Moreover, assume that $\mathcal{C}(\kappa)$ and $\mathcal{C}^{\mathrm{RB}}(\kappa)$ are dense matrices for $\kappa \in R$. ◇

The computational cost of the tasks in the offline phase is given in Table 4.1. The total offline cost is

$$O(N_{\mathrm{snap}}N_\ell^2 + N_{\mathrm{snap}}N_\ell^2 N_{\mathrm{sto}} + N_{\mathrm{snap}}{N_\ell}^2 N_{\mathrm{sto}}; N_\ell \to \infty).$$

Task	Computational Cost
Construct the full operator	$O(N_{\text{snap}} N_\ell^2)$
Solve the full eigenproblem	$O(N_{\text{snap}} N_\ell^2 N_{\text{sto}})$
POD	$O(N_{\text{snap}} N_\ell^2 N_{\text{sto}})$

Table 4.1. Computational cost of the offline phase.

Since $N_{\text{snap}} \ll N_{\text{smp}}$ the offline cost is asymptotically much cheaper than the cost of Algorithm 1 where we solve the full eigenproblem for each sample.
The computational cost of the tasks in the online phase is given in Table 4.2. The total online cost is

$$O(N_{\text{smp}} N_{\text{RB}}^2 N_{\text{lin}} + N_{\text{smp}} N_{\text{RB}}^2 N_{\text{sto}} + N_{\text{smp}} N_{\text{RB}} N_\ell; N_\ell \to \infty).$$

In the online phase, we solve the covariance eigenproblems in the reduced space.

Task	Computational Cost
Construct the reduced operator	$O(N_{\text{smp}} N_{\text{RB}}^2 N_{\text{lin}})$
Solve the reduced eigenproblem	$O(N_{\text{smp}} N_{\text{RB}}^2 N_{\text{sto}})$
Map the reduced solution to the full space	$O(N_{\text{smp}} N_{\text{RB}} N_\ell)$

Table 4.2. Computational cost of the online phase.

The high-dimensional full space X is only required when we map the reduced sample to X. The cost of these steps is linear in the dimension N_ℓ of X and quadratic in the dimension of the reduced basis N_{RB} for every sample. In contrast, the cost of Algorithm 1 is at least quadratic in N_ℓ, for every sample. Hence, for every sample, we need to solve an $O(N_\ell; N_\ell \to \infty)$ problem using RB, but an $O(N_\ell^2; N_\ell \to \infty)$ problem in the full space. This clearly demonstrates the advantages of RB sampling.
In §4.1.2, we have also discussed computing pseudo-determinants and pseudo-inverses of $\mathcal{C}(\kappa)$, as further tasks. Recall that when we sample with the KL expansion, we compute the spectral expansion of $\mathcal{C}(\kappa)$. Using the spectral expansion, we can easily compute $\det^+(\mathcal{C}(\kappa)) = \det^+(\mathcal{C}^{\text{RB}}(\kappa))$ since W is orthonormal. We can also compute $\mathcal{C}^{\text{RB}}(\kappa)^+$. However, representing $\mathcal{C}(\kappa)^+ := W\mathcal{C}^{\text{RB}}(\kappa)^+ W^*$ is computationally inefficient. It costs $O(N_\ell^2; N_\ell \to \infty)$ in time and memory. In the next section, we will see that we can often work with $\mathcal{C}^{\text{RB}}(\kappa)^+$ and do not need to express the full size operator $\mathcal{C}(\kappa)^+$.

4.3.3 Reduced basis random field expansion and MCMC

In §4.1.2, we have discussed the challenging memory consumption in hierarchical UQ. In the following, we explain an efficient way to represent the random fields on the reduced basis.
Observe that the reduced basis enables a natural compression of the full random field $\boldsymbol{\theta}$. Let $(\boldsymbol{\kappa}, \boldsymbol{\theta}) \sim \mu = \mu' \odot K$. Moreover, we assume that the mean function $m(\cdot)$ can be represented on the reduced basis, i.e. $WW^* m(\kappa) = m(\kappa)$, $\kappa \in R$. The

reduced basis implies the representation

$$\boldsymbol{\theta} = m(\boldsymbol{\kappa}) + \Psi_0(\boldsymbol{\kappa})\boldsymbol{\xi} = W \cdot (W^* m(\boldsymbol{\kappa}) + \Psi_0^{\mathrm{RB}}(\boldsymbol{\kappa})\boldsymbol{\xi})$$

for some $\boldsymbol{\xi} \sim \mathrm{N}(0, \mathrm{Id}_{N_{\mathrm{sto}}})$. We can represent $\boldsymbol{\theta}$ in terms of

$$\boldsymbol{\theta}_{\mathrm{RB}} := (W^* m(\boldsymbol{\kappa}) + \Psi_0^{\mathrm{RB}}(\boldsymbol{\kappa})\boldsymbol{\xi}) \in \mathbf{L}^2(\Omega; \mathbb{R}^{N_{\mathrm{RB}}}), \tag{4.12}$$

which gives the *reduced basis expansion* $\boldsymbol{\theta} = W\boldsymbol{\theta}_{\mathrm{RB}}$, of the random field $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta}_{\mathrm{RB}} \sim \mathrm{N}(W^* m(\boldsymbol{\kappa}), \mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}))$ is a Gaussian random variable on the reduced space $X_{\mathrm{RB}} := \mathbb{R}^{N_{\mathrm{RB}}}$, for fixed $\boldsymbol{\kappa} = \kappa$. Hence, we can fully represent random field samples $\boldsymbol{\theta} : \Omega \to \mathbb{R}^{N_\ell}$ by $(\boldsymbol{\kappa}, \boldsymbol{\theta}_{\mathrm{RB}}) : \Omega \to \mathbb{R}^{N_R + N_{\mathrm{RB}}}$. Note that typically $N_R + N_{\mathrm{RB}} \ll N_\ell$. Moreover, since we keep W in the memory, we can reproduce $\theta = W\theta_{\mathrm{RB}}$ from θ_{RB} computationally efficiently in $O(N_{\mathrm{RB}}^2 N_\ell; N_\ell \to \infty)$.

We now consider again the hierarchical Bayesian inverse problem from §4.1.2. Recall the Metropolis-within-Gibbs method that we have described in Algorithm 2. To compute the acceptance probability of the $\boldsymbol{\kappa}$-proposal, we need to construct and evaluate the probability density function of $\boldsymbol{\theta}$. This involves computing the pseudo-determinant of $\mathcal{C}(\kappa)$, as well as its pseudo-inverse. At the end of §4.3.2, we have discussed that the pseudo-determinant can be approximated computationally fast, but that the pseudo-inverse on the full space X is inefficient. Instead, we can rephrase the Bayesian inverse problem in terms of the reduced basis expansion.

Let $L(y^\dagger|\cdot) : X \to \mathbb{R}$ be the likelihood, $K(\cdot\,|\kappa) := \mathrm{N}(m(\kappa), \mathcal{C}(\kappa))$, and $\mu_{\mathrm{prior}} := \mu' \odot K \in \mathrm{Prob}(R \times X)$ be the prior of the original problem. We again centre the prior measure, and set

$$L(y^\dagger|\kappa, \theta) := L(y^\dagger|m(\kappa) + \theta), \qquad K(\cdot\,|\kappa) := \mathrm{N}(0, \mathcal{C}(\kappa)) \qquad (\kappa \in R, \theta \in X).$$

We define $L^{\mathrm{RB}}(y^\dagger|\kappa, \theta^{\mathrm{RB}}) := L(y^\dagger|m(\kappa) + W\theta^{\mathrm{RB}})$, and $K^{\mathrm{RB}}(\cdot\,|\kappa) := \mathrm{N}(0, \mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\kappa))$. Note that here θ^{RB} given $\boldsymbol{\kappa} = \kappa$ is always centred Gaussian, and thus we do not need to be able to represent $m(\kappa)$ on W, $\kappa \in R$. We consider the resulting BIP with parameter space $R \times X_{\mathrm{RB}}$, prior $\mu_{\mathrm{prior}}^{\mathrm{RB}} = \mu' K^{\mathrm{RB}}$, and likelihood $L^{\mathrm{RB}}(y^\dagger|\cdot)$. Now, we approach this BIP with MWG. We call this method *Reduced Basis MCMC* and summarise it in Algorithm 4. Notably, the computational cost of the Gibbs step in κ-direction is indepedent of N_ℓ. In the θ-direction, we still need to evaluate the likelihood that requires the full random field $\boldsymbol{\theta} = W\boldsymbol{\theta}_{\mathrm{RB}}$.

4.4 Numerical experiments

In this section we illustrate and verify the reduced basis sampling for use with forward and Bayesian inverse problems. We start by measuring runtime and accuracy of the reduced basis approximation to the parametric KL eigenproblems. In Example 4.18, we consider a forward and a Bayesian inverse problem in a low-dimensional test setting. This allows us to compare the reduced basis sampling with the samples obtained by using the full, unreduced KL eigenproblems. We then move on to high-dimensional estimation problems in Examples 4.19–4.21. Note that we are not able to compute reference solutions in the high-dimensional test cases within a reasonable amount of time since these are computationally very expensive. Nevertheless, these examples are a proof-of-concept and showcase potential applications.

Algorithm 4: Reduced Basis Markov Chain Monte Carlo

Let $(\boldsymbol{\kappa}_0, \boldsymbol{\theta}_{\mathrm{RB},0}) \in \mathbb{R}^{N_{\mathrm{RB}}+1}$ be the initial state of the Markov chain.
for $n \in \{1, \dots, N_{\mathrm{smp}}\}$ **do**
  Sample $\boldsymbol{\kappa}_* \sim q_R(\cdot\,|\boldsymbol{\kappa}_{n-1})$
  $a_R(\boldsymbol{\kappa}_{n-1}; \boldsymbol{\kappa}_*) \leftarrow \min\left\{1, \frac{q_R(\boldsymbol{\kappa}_{n-1}|\boldsymbol{\kappa}_*)}{q_R(\boldsymbol{\kappa}_*|\boldsymbol{\kappa}_{n-1})} \frac{\mathrm{n}(\boldsymbol{\theta}_{\mathrm{RB},n-1}; W^* m(\boldsymbol{\kappa}_*), \mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}_*))}{\mathrm{n}(\boldsymbol{\theta}_{\mathrm{RB},n-1}; W^* m(\boldsymbol{\kappa}_{n-1}), \mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}_{n-1}))}\right\}$
  Sample $\boldsymbol{U}_R \sim \mathrm{Unif}[0,1]$
  if $\boldsymbol{U}_R \leq a_R$ **then**
    $\boldsymbol{\kappa}_n \leftarrow \boldsymbol{\kappa}_*$
  else
    $\boldsymbol{\kappa}_n \leftarrow \boldsymbol{\kappa}_{n-1}$
  end
  Sample $\boldsymbol{\theta}^*_{\mathrm{RB}} \sim \mathrm{N}(\sqrt{1-\beta^2}\boldsymbol{\theta}_{\mathrm{RB},n-1}, \beta\mathcal{C}^{\mathrm{RB},N_{\mathrm{sto}}}(\boldsymbol{\kappa}_n))$
  $a_X(\boldsymbol{\theta}_{\mathrm{RB},n-1}; \boldsymbol{\theta}^*_{\mathrm{RB}}) \leftarrow \min\{1, L^{\mathrm{RB}}(y^\dagger|\boldsymbol{\kappa}_n, \boldsymbol{\theta}^*_{\mathrm{RB}})/L^{\mathrm{RB}}(y^\dagger|\boldsymbol{\kappa}_n, \boldsymbol{\theta}_{\mathrm{RB},n-1})\}$
  Sample $\boldsymbol{U}_X \sim \mathrm{Unif}[0,1]$
  if $\boldsymbol{U}_X \leq a_X$ **then**
    $\boldsymbol{\theta}_{\mathrm{RB},n} \leftarrow \boldsymbol{\theta}^*_{\mathrm{RB}}$
  else
    $\boldsymbol{\theta}_{\mathrm{RB},n} \leftarrow \boldsymbol{\theta}_{\mathrm{RB},n-1}$
  end
end

In Examples 4.19–4.20, we consider the elliptic PDE

$$-\nabla \cdot (\exp(\theta(x))\nabla u(x)) = f(x) \qquad (x \in D) \tag{4.13}$$

on the unit square domain $D = (0,1)^2$ together with suitable boundary conditions. Subject to Dirichlet boundary conditions, we have discussed this PDE thoroughly in Examples 1.3, 1.4, and the stochastic version in Example 1.37. The PDE (4.13) is discretised with linear, continuous finite elements on a uniform, triangular mesh. The coefficient function θ is a parameterised Gaussian random field with exponential covariance operator, and random correlation length and standard deviation, respectively (see Example 4.2). The spatial discretisation of θ is done with piecewise constant finite elements on a uniform, rectangular mesh. The evaluation of the covariance operator on this finite element space requires to evaluate an integral. We approximate this integral using a composite midpoint rule, with one quadrature node in each finite element.

We further discretise θ by a truncated KL expansion where we retain the leading N_{sto} terms. The parameter N_{sto} is selected such that the truncated KL captures at least 90% of the total variance. We list the random field parameters for Examples 4.18–4.21 in Table 4.3. We introduce the estimation problems in more detail in the following subsections. Moreover, we refer to Remark 2.11 concerning the well-posedness of the elliptic Bayesian inverse problem and to Remark 2.9 for the well-posedness of hierarchical BIPs. Note that we solve the test problems in Examples 4.18–4.21 using the reduced basis samplers presented in §4.3.

	Example 4.18	Example 4.19	Example 4.20	Example 4.21
$\underline{\sigma}$	1	0.1	0.5	0.1
$\overline{\sigma}$	1	1	0.5	1
m_σ	1	0.5	0.5	0.5
σ_σ^2	0	0.1	0	0.1
$\underline{\lambda}$	0.3	0.3	0.3	0.1
N_{sto}	200	100	100	800

Table 4.3. Random field parameters in Examples 4.18–4.21.

4.4.1 Accuracy and speed up

First we assess the accuracy and time consumption of the reduced basis approximation. We measure the quality of the reduced basis surrogate by comparing reduced basis eigenvalues with full eigenvalues in a simplified setting. The *full matrix* is the finite element approximation of $\mathcal{C}_{\text{exp}}^{(\lambda,1)}$ with 100×100 piecewise constant finite elements. The goal is to compute the leading 100 eigenpairs of $\mathcal{C}_{\text{exp}}^{(\lambda,1)}$ for selected values $\lambda \in [0.1, \sqrt{2}]$. We compute reference solutions for $\lambda = 0.1, 0.5, 1.4$ using the full matrix. The reduced basis is constructed using 10 snapshots

$$\lambda_{\text{snap}} = \left(\frac{1}{2^{-1/2} + s} : s = 0, \dots, 9 \right).$$

We compute the leading 100 eigenpairs for all correlation lengths in λ_{snap}, and assemble the associated eigenvectors in a single matrix. Then we apply the POD and retain $N_{\text{RB}} = 2^1, \dots, 2^{13}$ orthonormal basis vectors. Recall that the offline-online decomposition requires a linearisation of the covariance operator (see §4.2.3). Throughout this section (§4.4) we retain $N_{\text{lin}} = 39$ linearisation terms. In this case, the truncation error defined in (4.11) is equal to 9.09E-5 for the linearisation.

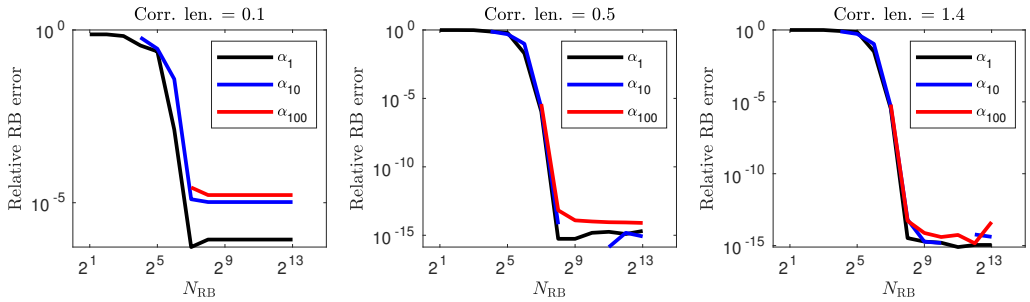


Figure 4.4. Relative reduced basis error of the eigenvalues $\alpha_1(\lambda), \alpha_{10}(\lambda), \alpha_{100}(\lambda)$ for correlation lengths $\lambda = 0.1, 0.5, 1.4$ and reduced basis dimensions $N_{\text{RB}} = 2^1, \dots, 2^{13}$.

We plot the relative error of the reduced eigenvalue compared to the exact eigenvalue in Figure 4.4 for various reduced basis dimensions N_{RB}. Note that it is not possible to compute eigenvalues with an index larger than N_{RB}. For $\lambda = 0.1$ the relative RB error stagnates at a level that is not smaller than 1E-6. In further experiments not reported here we observed that this stagnation is caused by the linearisation

error of the covariance kernel (recall that we use $N_{\text{lin}} = 39$ terms with an error of order 1E-5). We remark that the mean square error of the MC and MCMC estimation results in this section is of order $O(1\text{E-}4)$. Hence, an eigenvalue error of magnitude $O(1\text{E-}6)$ is acceptable. We point out, however, that a full error analysis of the reduced basis samplers (including the linearisation and RB error) is beyond the scope of this study; see only Remark 4.17. For $\lambda = 0.5$ and $\lambda = 1.4$, we achieve an accuracy of order 1E-6 for $N_{\text{RB}} \approx 128$. For $N_{\text{RB}} > 128$, the relative errors are of the size of the machine epsilon. This error is unnecessarily much smaller than the sampling error mentioned above and introduces a higher computational cost in the online phase. Hence, in our test problems $N_{\text{RB}} = 128$ would be a sufficient choice.

Remark 4.17. We note that the eigenvalues of the covariance operator correspond to the variance in the direction of the associated eigenvector. This is indeed implicitly contained in the statement (3.1). Hence, we anticipate to be able to balance between the eigenvalue error (here, $O(1\text{E-}6)$), the mean square sampling error (here, $O(1\text{E-}4)$), and the error in the covariance function (here, $O(1\text{E-}5)$); note that after all each of these values refers to variances. ◇

To explore the speed-ups that are possible with reduced basis sampling we repeat the experiment. This time we vary the dimension of the finite element space and use $N_\ell = 4^4, \dots, 4^7$. The dimension of the reduced basis is fixed with $N_{\text{RB}} = 256$. We plot the test results in Figure 4.5. The time measurements correspond to serial simulations in MATLAB with an Intel i7 (2.6 Ghz) CPU and 16 GB RAM memory. The dashed lines show the theoretical asymptotic behaviour, that is, $O(N_\ell; N_\ell \to \infty)$ for the reduced basis sampling and $O(N_\ell^2; N_\ell \to \infty)$ for the full sampling. We see

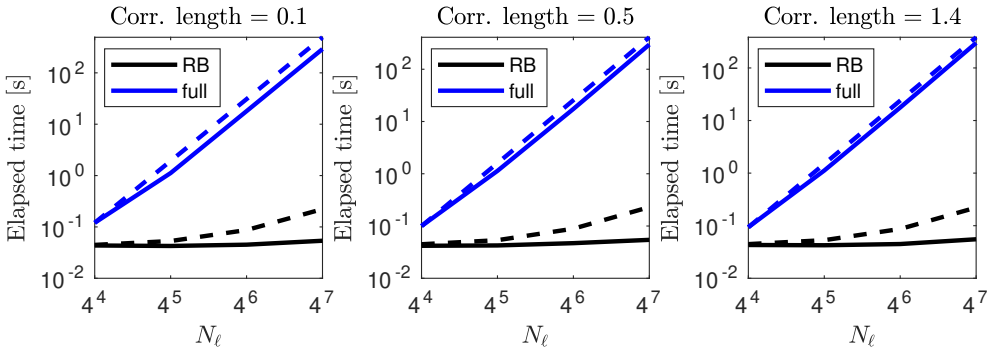


Figure 4.5. Timings for the full and reduced problem with different FE resolutions and correlation lengths. The elapsed time is shown as solid line, and the asymptotic behaviour is shown as dashed line.

that the theoretical and observed timings for the full sampling are almost identical. In contrast, the observed timings for RB sampling are smaller than predicted by the theory. This is caused by the fact that the dimension N_ℓ of the finite element space is quite small. As N_ℓ increases we observe a massive speed-up of the reduced basis sampling compared to the full sampling. For example, for $N_\ell = 4^7$ the reduced basis sampling requires less than 1E-1 seconds, while the full sampling requires several minutes. In this case the speed-up in the online phase is of order $O(1\text{E}3)$.
For the estimation problems in Examples 4.18–4.21, we use Monte Carlo and Markov Chain Monte Carlo with 1E4 up to 1.5E5 samples and a random field resolution with

$N_\ell = 256 \times 256$ finite elements in space. In these cases, MC and MCMC estimations based on the full KL eigenproblem would take from a couple of days up to a couple of years to terminate; considering the current computational capability. In contrast, the (serial) run-time of the reduced basis sampling is ~ 15 minutes in Example 4.19 and ~ 18 hours in Example 4.21. Of course, standard Monte Carlo simulations are trivially parallelisable. MCMC is a serial algorithm by design, and parallelisation is not trivial, see, e.g. [251] for suitable strategies. Our experiments show that RB sampling can reduce the computational cost without the need for parallelisation.

4.4.2 Verification of reduced basis sampling

Next, we test the accuracy of RB sampling using coarse spatial discretisations. This allows us to obtain reference solutions in a reasonable amount of time.

Example 4.18. Let μ be the joint probability measure from Example 4.2 together with the parameter values given in Table 4.3. We discretise the random field $\boldsymbol{\theta} \sim \mu''$ using 32^2 finite elements. The test problems are as follows.

(a) Forward uncertainty propagation: We consider a flow cell in 2D with log-permeability θ. See Example 4.19 for the definition of the flow cell. Given the random coefficient $\boldsymbol{\theta}$, we want to estimate the probability distribution of the outflow over the boundary $Q(\boldsymbol{\theta})$. We discretise the PDE with $2 \cdot 16^2$ finite elements.

(b) Bayesian inverse problem: We observe a random field realisation on $D = (0,1)^2$ at nine points in the spatial domain

$$D_{\text{obs}} := \{(n/4, m/4) : n, m = 1, 2, 3\}.$$

In each of the points we observe the value $y^\dagger := 0.1$, which we assume to be noisy. We want to reconstruct the random field. The prior measure is μ as specified above. The likelihood is given by

$$L(0.1|\theta) \propto \exp\left(-\frac{1}{2\text{E-}2} \sum_{x \in D_{\text{obs}}} (0.1 - \theta(x))^2\right).$$

We want to estimate the posterior mean and variance of the correlation length λ given the data y. In addition, we compute the model evidence. Note that this test problem is similar to Example 1.55 in Chapter 1, see also the corresponding Figure 1.3.

$\diamond$

We solve the test problems in Example 4.18(a)–(b) with Monte Carlo. In part (b), we use Importance Sampling with samples from the prior as proposal; recall §3.3.2 for an introduction and particularly Remark 3.4. We solve (a) and (b) with reduced basis sampling as well as standard sampling based on the full (discretised) eigenvectors of the parameterised covariance operator. The standard sampling serves as reference solution for the reduced basis sampling. The reduced basis is constructed using the snapshot correlation lengths $\lambda^{\text{snap}} = (0.322, 0.433, 0.664, 1.414)$; these are

simply the inverses of four equidistant points in the interval $[1/\sqrt{2}, 3.11]$ including the boundary points. This choice clusters snapshots near zero which is desirable due to the singularity of the exponential covariance at $\lambda = 0$. We apply the POD to construct three reduced bases with different accuracies $\underline{\sigma}^2 :=$ 1E-1, 1E-5, 1E-9. For each of the settings, we run 61 Monte Carlo simulations with 1E4 samples each to estimate the mean and the variance of the pushforward measure $\mu''(Q \in \cdot)$ in part (a), as well as the posterior mean, posterior variance and model evidence in the Bayesian inverse setting in part (b). We compute a reference solution for all those quantities by using 6.15E5 samples. With respect to the reference solutions, we compute the relative error of the 61 estimates in each setting. In Table 4.4, we give the means and the associated standard deviations (StD) of the relative errors. We observe that the (mean of the) relative error is of order O(1E-2) down to O(1E-3). Moreover, in Table 4.4 we list the sample mean of the error between the full covariance operators and their representations on the reduced basis, measured in the Frobenius norm. That is, we list the Monte Carlo estimate using 6.15E5 samples of the expression

$$\mathbb{E}\left[\|\mathcal{C}^{N_{\text{sto}}} - \mathcal{C}^{\text{RB},N_{\text{sto}}}\|_F\right] = \int_R \|\mathcal{C}^{N_{\text{sto}}}(\kappa) - \mathcal{C}^{\text{RB},N_{\text{sto}}}(\kappa)\|_F \, \mathrm{d}\mu'(\kappa).$$

We observe that the error decreases as we include more vectors in the reduced basis, as expected.

$\underline{\sigma}^2$	1E-1	(StD)	1E-5	(StD)	1E-9	(StD)
Pushforward mean	0.0055	(0.0041)	0.0057	(0.0051)	0.0054	(0.0040)
Pushforward variance	0.0451	(0.0262)	0.0442	(0.0352)	0.0321	(0.0282)
Evidence	0.0147	(0.0262)	0.0122	(0.0352)	0.0154	(0.0282)
Posterior mean	0.0087	(0.0070)	0.0075	(0.0062)	0.0082	(0.0063)
Posterior variance	0.0856	(0.0691)	0.0733	(0.0608)	0.0807	(0.0619)
Mean covariance error	2.609E-4		1.240E-7		3.606E-11	

Table 4.4. Verification. Relative errors in the Monte Carlo estimation of the mean and variance of the pushforward measure, and posterior mean, posterior variance, and model evidence in the Bayesian inverse problem (Example 4.18). Each error value is the mean taken over 61 simulations with 1E4 samples each. The simulations are performed with reduced basis sampling with POD accuracies $\underline{\sigma}^2 =$ 1E-1, 1E-5, 1E-9. The relative errors are computed with respect to a reference solution computed with 6.15E5 samples based on the full eigenproblem. In the last line of the table, we list the mean error between the full covariance operator and the operator represented in the reduced basis, measured in the Frobenius norm.

4.4.3 Forward uncertainty propagation

Next we study the forward uncertainty propagation of a hierarchical random field given as the diffusion coefficient in an elliptic PDE operator.

Example 4.19. Consider a flow cell problem on $D = (0,1)^2$ where the flow takes place in the x_1-direction. Hence, we consider the PDE from (4.13), subject to the

following boundary conditions,

$$\begin{aligned} u(x) &= 0 \quad (x \in \{1\} \times [0,1]), \\ u(x) &= 1 \quad (x \in \{0\} \times [0,1]), \\ \frac{\partial u}{\partial \vec{n}}(x) &= 0 \quad (x \in (0,1) \times \{0,1\}). \end{aligned}$$

There are no sources within the flow cell ($f \equiv 0$). The random field $\boldsymbol{\theta}$ is as in Example 4.2 with the parameters given in Table 4.3. The PDE is discretised with 2×128^2 finite elements, and the random field with 256^2 finite elements. The quantity of interest is the outflow over the (western) boundary $\Gamma_{\text{out}} := \{0\} \times [0,1]$. It can be approximated by

$$Q(\theta) := -\int_D a(\theta) \nabla u \cdot \nabla \psi \mathrm{d}x,$$

where $\psi|_{D \setminus \Gamma_{\text{out}}} \approx 0$ and $\psi|_{\Gamma_{\text{out}}} \approx 1$; see, e.g. [74]. We discretise the outflow using a piecewise linear, continuous finite element function ψ on the same mesh that we used for the PDE discretisation. ◇

The log-permeability is modelled as a parameterised Gaussian random field. We employ the reduced basis sampling with $N_{\text{RB}} = 191$. We construct the reduced basis analogously to the simple test setting in §4.4.2. However, now we let $N_{\text{sto}} = 100$, and remove vectors from the POD where the corresponding squared singular value is smaller than $\underline{\sigma}^2 = 1\text{E-}10$; see also §4.2.2.
We estimate the mean and variance of the output quantity of interest. We compare 24 estimations by computing the associated coefficient of variation (cv) for the mean and variance estimator, respectively. The cv is defined as the ratio of the standard deviation of the estimator and the absolute value of its mean. We present the estimation results in Table 4.5. The small cv tells us that $N_{\text{smp}} = 1\text{E}4$ samples were

	Mean	(cv)	Variance	(cv)
MC estimate	157.286	(0.0028)	3012.2	(0.0355)

Table 4.5. Hierarchical forward problem. Mean and variance estimates with 1E4 samples (Example 4.19). We compare these estimates to 23 further simulation results by computing the coefficient of variation within the 24 estimates.

sufficient to estimate the pushforward measure of the quantity of interest, as well as its mean and variance. Note that with the reduced basis sampling a single Monte Carlo simulation run took about 18 minutes.

4.4.4 Hierarchical Bayesian inverse problem

We consider two hierarchical Bayesian inverse problems based on random fields. Note that we use again 256^2 finite elements to discretise the random fields in both problems and 2×128^2 finite elements to discretise the elliptic PDE in Example 4.20.

Example 4.20. Consider the Bayesian estimation of a random field and its correlation length. The true underlying random field is propagated through the elliptic PDE (4.13) together with Dirichlet boundary conditions

$$u(x) = 0 \quad (x \in \partial D),$$

and 9 Gaussian-type source terms

$$f(x) = \sum_{n,m=1}^{3} \mathrm{n}\left(x_1; 0.25n, 0.001\right) \cdot \mathrm{n}\left(x_2; 0.25m, 0.001\right).$$

We observe the solution u at 49 locations. In particular, the observation operator is given by

$$\mathcal{O}(u) := (u(n/8, m/8) : n, m = 1, \ldots, 7).$$

The (synthetic) observations are generated with log-permeability fields that are

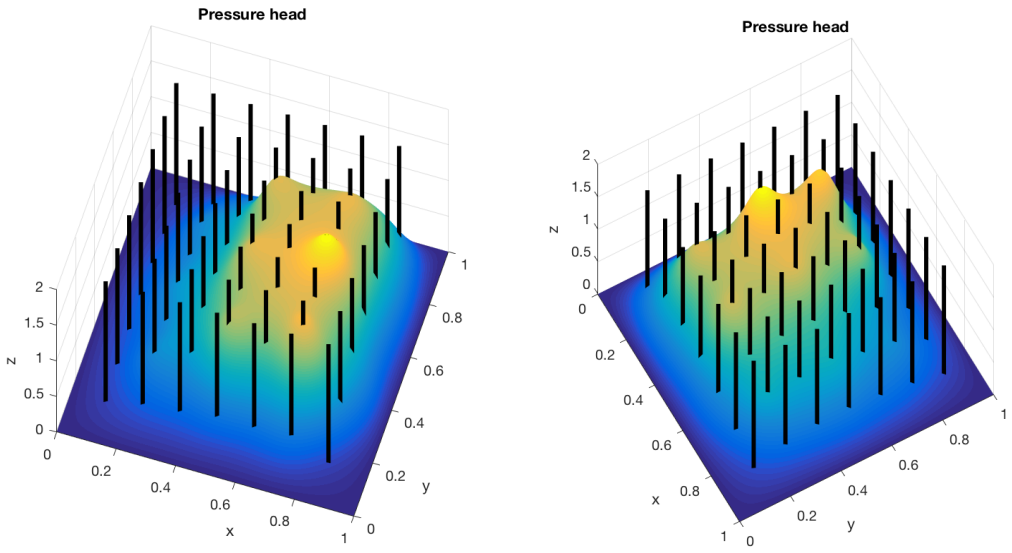


Figure 4.6. Synthetic solution p observed in the hierarchical Bayesian inverse problem (Example 4.20). The black lines indicate the measurement locations. The figures show the model output for a log-permeability with correlation length $\lambda = 0.5$ (left) and $\lambda = 1.1$ (right), respectively.

samples of a Gaussian random field with exponential covariance operator with $\lambda \in \{0.5, 1.1\}$ and $\sigma = \sqrt{1/2}$. We show the corresponding PDE outputs and the measurement locations in Figure 4.6. The Gaussian random fields have been sampled with the full (unreduced) $N_{\mathrm{sto}} = 100$ leading KL terms. Every observation is perturbed with i.i.d. Gaussian noise $\eta_1 \sim \mathrm{N}(0, 1\mathrm{E}\text{-}6)$. We use the measure μ in Example 4.2 with parameter values given in Table 4.3 as prior measure. ◇

This example matches the elliptic inverse problem discussed in Example 1.41.

Example 4.21. Consider the Bayesian estimation of a Gaussian random field together with its standard deviation and correlation length. We observe the field directly, however, the observations are again noisy. The estimations are performed with two data sets that have been generated with fixed hyperparameters $\lambda \in \{0.2, 1.1\}$ and $\sigma = 1/(\sqrt{2} \cdot 256)$. We set $\sigma = 1/\sqrt{2}$ and rescale the KL eigenfunctions by $1/256$. The random field discretisation uses an $N_{\mathrm{sto}} = 800$ dimensional full (unreduced) KL basis. We observe the random field at 2500 positions. Each observation is perturbed by i.i.d. Gaussian noise $\eta_1 \sim \mathrm{N}(0, 1\mathrm{E}\text{-}6)$. The prior measure is the measure μ in Example 4.2 with parameter values given Table 4.3. ◇

Note that in Examples 4.20–4.21 we use the same PDE and random field discretisation for the generation of the data and the estimation problem. The reason is that we are mainly interested in the reduced basis error and not in the reconstruction error of the inverse problem. Note further that in Examples 4.20–4.21 the standard deviation $\sigma = \sqrt{1/2}$ is fixed a priori, and is not estimated. The hierarchical Bayesian inverse problems in Examples 4.20–4.21 are well-posed since they satisfy the conditions of Corollary 2.6; see also Remark 2.9.

Observations from PDE output. We consider Example 4.20 and the settings in Table 4.3. The Reduced Basis MCMC method presented in Algorithm 4 is used to sample from the posterior measure. The correlation length $\lambda \in [0.3, \sqrt{2}]$. Since this is the same range as in Example 4.19 we reuse the reduced basis computed in Example 4.19. Recall that the standard deviation $\sigma = 1/\sqrt{2}$ of the random field θ is fixed and not estimated. Moreover, we assume that the observational noise is given by $\eta \sim \mathrm{N}(0, \text{1E-3} \cdot \mathrm{Id})$. This corresponds to a noise level of $\sqrt{\text{1E-3}}/\|y\|_Y \approx 0.6\%$. We perform experiments for two synthetic data sets with $\lambda = 0.5$ and $\lambda = 1.1$, respectively. For both data sets we compute a Markov chain of length $N_{\mathrm{smp}} = \text{1E5}$. To avoid burn-in effects the initial states are chosen close to the true parameter values for the Markov chains. In a setting with real world data it is often possible to obtain suitable initial states with Sequential Monte Carlo, see §5.

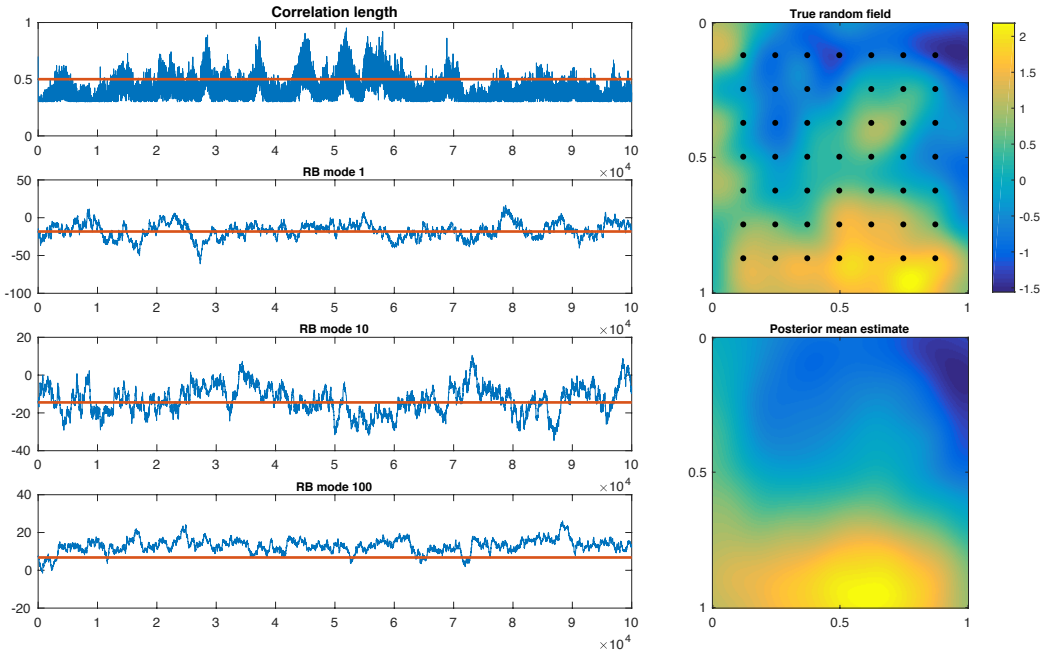


Figure 4.7. PDE-based hierarchical BIP with short true correlation length. Results of the MCMC estimation (Example 4.20, $\lambda = 0.5$). The top-right plot shows the synthetic truth together with the measurement locations (black dots). Below we plot the posterior mean estimate computed with MCMC. The four path plots on the left side of the figure show the Markov chains for the correlation length $\boldsymbol{\lambda}$, and the reduced basis modes $(\boldsymbol{\theta}_{\mathrm{RB}})_1$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{10}$, and $(\boldsymbol{\theta}_{\mathrm{RB}})_{100}$, respectively. The red lines mark the truth.

The estimation results are depicted in Figure 4.7 and Figure 4.8. We observe in both figures that the Markov chain for λ mixes very fast, however, it takes some

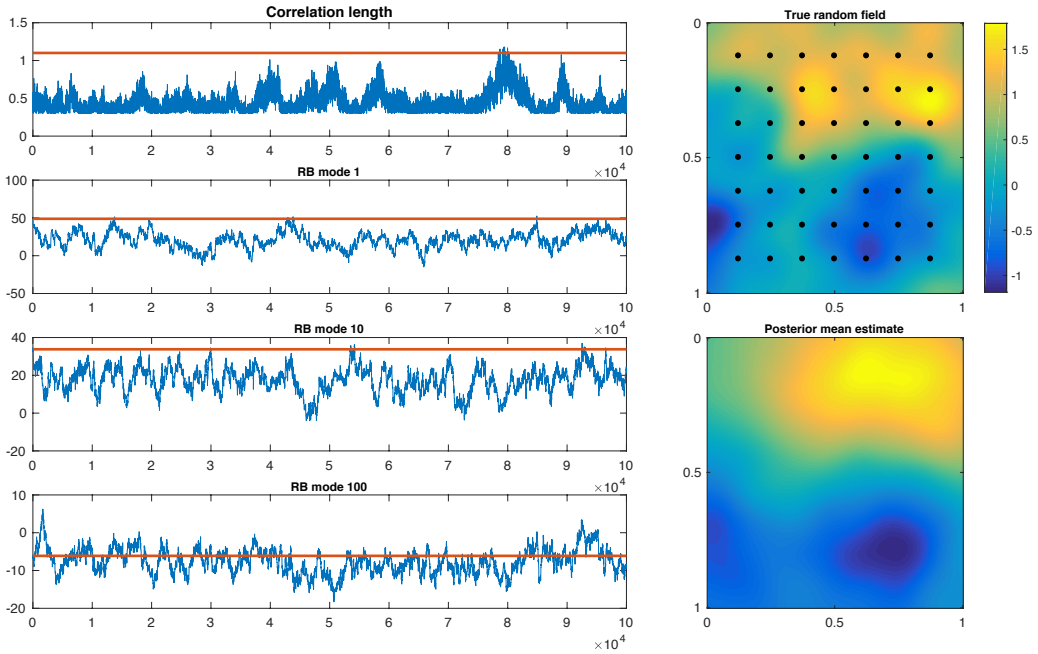


Figure 4.8. PDE-based hierarchical BIP with long true correlation length. Results of the MCMC estimation (Example 4.20, $\lambda = 1.1$). The top-right plot shows the synthetic truth together with the measurement locations (black dots). Below we plot the posterior mean estimate computed with MCMC. The four path plots on the left side of the figure show the Markov chains for the correlation length $\boldsymbol{\lambda}$, and the reduced basis modes $(\boldsymbol{\theta}_{\mathrm{RB}})_1$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{10}$, and $(\boldsymbol{\theta}_{\mathrm{RB}})_{100}$, respectively. The red lines mark the truth.

time for the Markov chains of the reduced basis modes to explore the whole space. To investigate this further we conduct a heuristic convergence analysis. To this end we consider multiple Markov chains; see in [213, §12.1.2] for a review of MCMC convergence analysis with multiple Markov chains. For each of the two test data sets, we compute 4 additional Markov chains starting at different initial states. In results not reported here, we observed a similar mixing and coverage of the parameter space of the additional chains. Given these mixing properties, it can reasonably be assumed that the Markov chains have reached the stationary regime.
Moreover, we have computed posterior mean and posterior variance of the correlation length parameter λ for each of the 5 Markov chains. The accuracy of these estimates is assessed by computing the coefficient of variation within these five estimates. We tabulate the posterior mean and variance estimates for λ of a single Markov chain as well as the associated cvs in Table 4.6. The single Markov chains in this table are the chains shown in Figures 4.7-4.8. The coefficients of variation of the posterior mean and variance estimates are considerably small. This tells us that the posterior mean and variance estimates are reasonably accurate.

Discussion of the estimation results. The correlation length is underestimated in both cases. In the first case, where the true parameter is given by $\lambda = 0.5$, the posterior mean is close to the true parameter. The relative distance between truth and posterior mean is about 18%. In the second setting, where in truth $\lambda = 1.1$, the posterior mean is far away from the true parameter. Here, the relative distance

	Mean	(cv)	Variance	(cv)
MCMC $\boldsymbol{\lambda}$ given y (Truth: $\lambda = 0.5$)	0.4105	(0.0040)	0.0081	(0.3235)
MCMC $\boldsymbol{\lambda}$ given y (Truth: $\lambda = 1.1$)	0.4403	(0.0524)	0.0157	(0.2346)

Table 4.6. Estimation results of the hierarchical Bayesian inverse problem with observations from PDE output (Example 4.20). We tabulate the posterior mean and variance estimates of the correlation length $\boldsymbol{\lambda}$ of one Markov chain each and the cvs within the estimates of 5 different Markov chains.

between truth and posterior mean is about 60%. In both cases, we conclude that the data likelihood was not sufficiently informative to estimate the correlation length more accurately. Then, the underestimation might be caused by the prior μ'. The prior measure μ' is concentrated close to the lower bound 0.3. This can be seen in Figure 4.9, where we show the probability density function $\pi' : R \to \mathbb{R}$ of μ'.

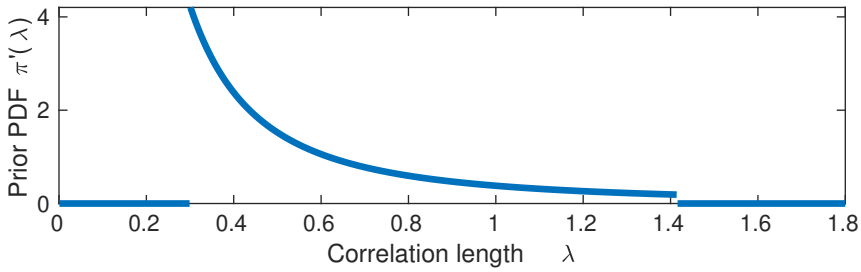


Figure 4.9. Probability density function π' of the prior measure μ' of the correlation length $\boldsymbol{\lambda}$. The measure is concentrated at the lower boundary 0.3.

Finally, we note that the succession of the estimates is correct: The posterior mean estimate in the problem with the larger true correlation is larger than the posterior mean estimate in the other case. Hence, we observe a certain consistency with the data in the estimation.

Observations from a random field. Finally, we consider Example 4.21. Here, we allow for much smaller correlation lengths $\lambda \in [0.1, \sqrt{2}]$. Moreover, we consider an uncertain standard deviation σ. This requires more KL terms for an accurate approximation, in particular, we use the leading 800 KL terms. This also means that we cannot reuse the reduced basis computed in Example 4.19. Instead, we construct a reduced basis as follows. We solve the KL eigenproblem for 5 snapshots

$$\lambda^{\text{snap}} = (0.1148, 0.1491, 0.2124, 0.3694, 1.4142)$$

of the correlation length. The rationale behind this choice is explained in §4.4.3. Given the collection of snapshot KL eigenvectors we apply a POD and retain only the basis vectors with $\sigma_i^2 \geq \underline{\sigma}^2 = 1\text{E-}10$. Note that we can compute the dependency of the standard deviation σ on the eigenpairs analytically and that we do not need to consider them when constructing the reduced basis.
Recall that in this example the observational noise is given by $\eta \sim \mathrm{N}(0, 1\text{E-}4 \cdot \mathrm{Id})$. This corresponds to a noise level of $\sqrt{1\text{E-}4}/\|y\|_Y \approx 6.6\%$. We employ the Reduced Basis MCMC sampler in Algorithm 4 to generate $N_{\text{smp}} = 1.5\text{E}5$ samples

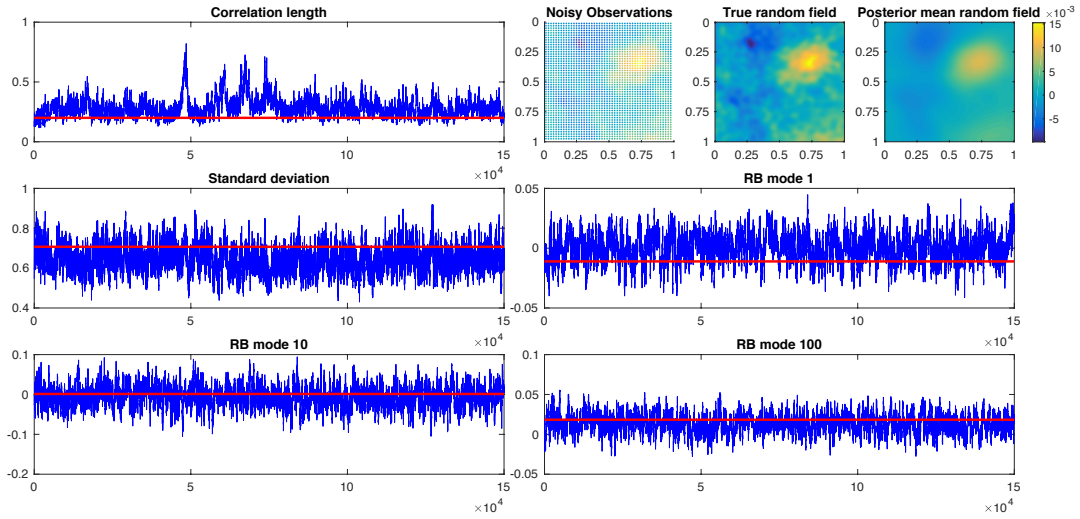


Figure 4.10. Results of the MCMC estimation given random field observations with short true correlation length (Example 4.21, $\lambda = 0.2$, $\sigma = 1/\sqrt{2}$). In the top-right corner we plot the positions and values of the noisy observations (left), the synthetic truth (middle), and the posterior mean (right). The five path plots show the Markov chains for $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$, and the reduced basis modes $(\boldsymbol{\theta}_{\mathrm{RB}})_1$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{10}$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{100}$, respectively. The red lines mark the truth.

of the posterior measure. We present the Markov chains and estimation results in Figure 4.10 and in Figure 4.11. We observe a fast mixing of the Markov chains. To conduct a heuristic convergence assessment, we again compute 4 additional Markov chains with $N_{\mathrm{smp}} = 1.5\mathrm{E}5$ samples each and different initial states. We found that the additional Markov chains mix similarly compared to the Markov chains shown in Figures 4.10–4.11. They also cover the same area of the parameter space. Hence, it can be reasonably concluded that the Markov chains reached a stationary regime.

	Mean	(cv)	Variance	(cv)
MCMC $\boldsymbol{\lambda}$ given y (Truth: $\lambda = 0.2$)	0.2847	(0.0465)	0.0064	(0.1520)
MCMC $\boldsymbol{\sigma}$ given y (Truth: $\sigma = 1/\sqrt{2}$)	0.6438	(0.0077)	0.0042	(0.0207)
MCMC $\boldsymbol{\lambda}$ given y (Truth: $\lambda = 1.1$)	0.7248	(0.0161)	0.0575	(0.1308)
MCMC $\boldsymbol{\sigma}$ given y (Truth: $\sigma = 1/\sqrt{2}$)	0.5484	(0.0096)	0.0052	(0.0453)

Table 4.7. Estimation results of the Bayesian inverse problem with observations from a random field (Example 4.21). We tabulate the posterior mean and variance of the correlation length $\boldsymbol{\lambda}$ and standard deviation $\boldsymbol{\sigma}$.

In addition, we present in Table 4.7 the posterior mean and posterior variance estimates of $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ associated with the Markov chains given in Figures 4.10–4.11. To assess the accuracy of these estimates we compare them with the posterior mean and variance estimates of the 4 other Markov chains by computing the coefficients of variations of the estimators. Again, the coefficients of variation are reasonably small.

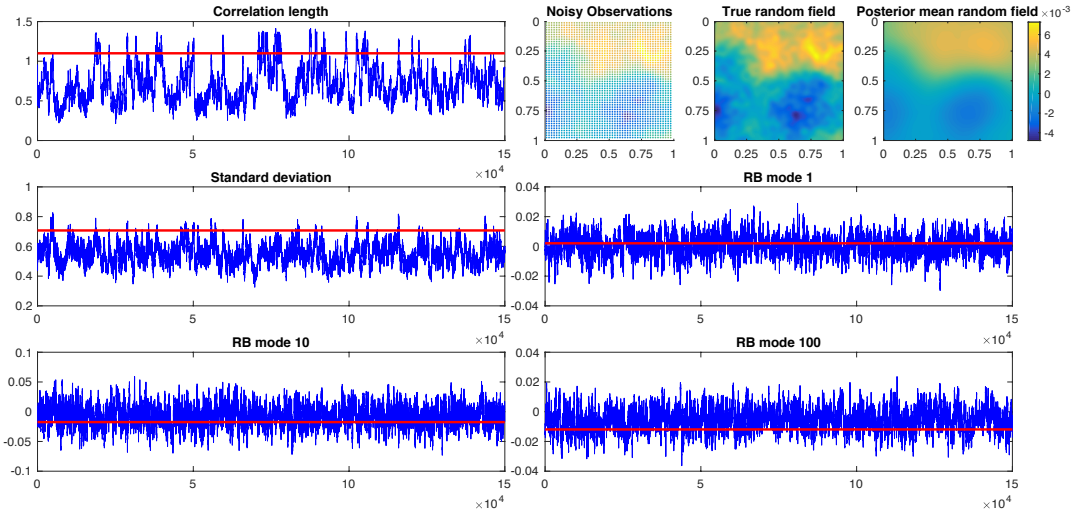


Figure 4.11. Results of the MCMC estimation given random field observations with long true correlation length (Example 4.21, $\lambda = 1.1$, $\sigma = 1/\sqrt{2}$). In the top-right corner we plot the positions and values of the noisy observations (left), the synthetic truth (middle), and the posterior mean (right). The five path plots show the Markov chains for $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$, and the reduced basis modes $(\boldsymbol{\theta}_{\mathrm{RB}})_1$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{10}$, $(\boldsymbol{\theta}_{\mathrm{RB}})_{100}$, respectively. The red lines mark the truth.

Discussion of the estimation results. While the likelihood was rather uninformative in the PDE-based Bayesian inverse problem, we see overall more consistent estimates in Example 4.21. For the short correlation length $\lambda = 0.2$, the relative distance between posterior mean and truth is 42%. The long correlation length $\lambda = 1.1$ is again underestimated. The relative distance between truth and posterior mean is 34% in this case. This result could be explained by the uncorrelated noise that has an influence on the observation of the correlation structure. In particular, we actually observe a generalised random field $\boldsymbol{\theta}' := \boldsymbol{\theta} + \boldsymbol{\eta}'$, where $\boldsymbol{\theta} \sim \mathrm{N}(0, \mathcal{C}_{\mathrm{exp}}^{(\lambda,\sigma)})$ and $\boldsymbol{\eta}' \sim \mathrm{N}(0, \sigma_*^2 \cdot \mathrm{Id}_X)$, for some $\sigma_*^2 > 0$. In this situation, the Gaussian white noise $\boldsymbol{\eta}'$ can be understood as a random field with correlation length 0, see §1.2.3 for a discussion of Gaussian white noise. This might explain the underestimation of the correlation lengths. The standard deviations are slightly underestimated and some of the reduced basis modes are overestimated – this is a consistent result. The posterior mean random fields appear to be smoother than the true random fields. This might be due to the high noise level.

Chapter 5

Sequential Monte Carlo samplers

> [...] there is no such thing as a random number – there are only methods to produce random numbers, and a strict arithmetic procedure of course is not such a method.
>
> John von Neumann [254, p. 768]

In this chapter, we discuss Sequential Monte Carlo (SMC) samplers as a way to approximate sequences of probability measures. We introduce the framework of measure-valued sequences in §5.1, where we consider Bayesian filtering as an example. In §5.2, we discuss Sequential Monte Carlo, and how it is derived starting from Importance Sampling and Sequential Importance Sampling (SIS). After this rather practical introduction, we discuss Sequential Monte Carlo in a different framework in §5.3, where it is represented as a sequence of random measures. Here, SIS and SMC construct measure-valued Markov chains, and can be understood and analysed as MCMC methods. Finally, in §5.4, we discuss the use of Sequential Monte Carlo in BIPs.

5.1 Framework

Let X be a measurable subset of a separable Banach space and $\underline{\mu} := (\mu_k)_{k=0}^{N_{\text{seq}}} \in \text{Prob}(X)^{N_{\text{seq}}}$ be a finite or infinite family of probability measures; i.e. $N_{\text{seq}} \in \mathbb{N} \cup \{\infty\}$. Moreover, let $\underline{\gamma} := (\gamma_k)_{k=1}^{N_{\text{seq}}}$ be a family of probability density functions, where $\gamma_k : X \to (0, \infty)$ and γ_k is bounded from above, for $k = 1, \dots, N_{\text{seq}}$. In the light of the wording *Sequential* Monte Carlo, we will refer to these families usually as *sequences*, even if they contain only a finite number of probability measures or densities.

We assume that the sequence of measures $\underline{\mu}$ starts with a given measure μ_0 and is otherwise defined recursively:

$$\frac{\mathrm{d}\mu_k}{\mathrm{d}\mu_{k-1}}(\theta) := \frac{\gamma_k(\theta)}{\int_X \gamma_k(\theta')\mu_{k-1}(\mathrm{d}\theta')} \propto \gamma_k(\theta) \qquad (\theta \in X, k = 1, \dots, N_{\text{seq}}). \tag{5.1}$$

Moreover, we define this update as a map $\mathrm{W}_k : \text{Prob}(X) \to \text{Prob}(X)$, where

$$\mu \longmapsto \mathrm{W}_k(\mu) := \left(\mathcal{B}X \ni A \mapsto \frac{\int_A \gamma_k(\theta)\mu(\mathrm{d}\theta)}{\int_X \gamma_k(\theta)\mu(\mathrm{d}\theta)} \in [0,1] \right), \tag{5.2}$$

and thus $\mu_k := \mathrm{W}_k(\mu_{k-1})$ for $k = 1, \dots, N_{\mathrm{seq}}$.

Bayesian filtering. Sequences like $\underline{\mu}$ appear, for instance, in Bayesian filtering. In Bayesian filtering, we solve a sequence of Bayesian inverse problems, where the posterior of the $(k-1)$-th BIP serves as the prior of the k-th BIP. This means, that we observe one dataset after another and use the current observation to refine our knowledge concerning the unknown parameter.
We now introduce this framework more precisely. Let $G : X \to H$ be a mathematical model. We aim to estimate the true model parameter $\theta^\dagger \in X$ given datasets $y^{(1)}, \dots, y^{(N_{\mathrm{seq}})}$. We observe independently the datasets at time points $t = 1, \dots, N_{\mathrm{seq}}$. The observation at each time t is modelled by a strictly positive and bounded likelihood function $L_t : X \times Y_t \to (0, \infty)$, where the data spaces Y_t may also vary over time.
Let now $T < N_{\mathrm{seq}} + 1$ be some finite point in time. In *Bayesian inversion* or *smoothing*, we consider all the data up to T at once and then estimate the posterior

$$\mu_{\mathrm{post}}^{(T)} := \mathbb{P}\left(\theta \in \cdot \,\middle|\, \boldsymbol{y}^{(1)} = y^{(1)}, \dots, \boldsymbol{y}^{(T)} = y^{(T)}\right).$$

We can express this measure as the posterior in a particular BIP. Let $\mu_{\mathrm{prior}} \in \mathrm{Prob}(X)$ be a prior measure. Since the observations are assumed to be independent, the likelihood is formed as the product of the likelihoods at times $t = 1, \dots, T$:

$$\overline{L}_T(y^{(1)}, \cdots, y^{(T)}|\theta) = \prod_{t=1}^{T} L_t(y^{(t)}|\theta) \qquad (\theta \in X, y^{(s)} \in Y_s, s = 1, \dots, T)$$

and $\overline{Y}_T := Y_1 \times \cdots Y_T$. We now obtain $\mu_{\mathrm{post}}^{(T)}$ by applying Theorem 1.45, with prior μ_{prior} and likelihood $\overline{L}_T : \overline{Y}_T \times X \to \mathbb{R}$.
In this time-dependent setting, a natural question to ask is: *Given $\mu_{\mathrm{post}}^{(T)}$ and a new data set $y^{(T+1)}$, how do we obtain $\mu_{\mathrm{post}}^{(T+1)}$?* Since the likelihoods are all strictly positive and the data are independent, we have

$$\frac{\mathrm{d}\mu_{\mathrm{post}}^{(T+1)}}{\mathrm{d}\mu_{\mathrm{post}}^{(T)}}(\theta) = \frac{\frac{\mathrm{d}\mu_{\mathrm{post}}^{(T+1)}}{\mathrm{d}\mu_{\mathrm{prior}}}(\theta)}{\frac{\mathrm{d}\mu_{\mathrm{post}}^{(T)}}{\mathrm{d}\mu_{\mathrm{prior}}}(\theta)} \propto \frac{\overline{L}_{T+1}(y^{(1)}, \cdots, y^{(T)}, y^{(T+1)}|\theta)}{\overline{L}_T(y^{(1)}, \cdots, y^{(T)}|\theta)} = L_{T+1}(y^{(T+1)}|\theta).$$

Hence, to incorporate the new data set into the posterior measure, we can apply Bayes' formula with $\mu_{\mathrm{prior}} := \mu_{\mathrm{post}}^{(T)}$ and likelihood $L := L_{T+1}$; see (1.24). The task of incorporating the new dataset is called *(Bayesian) filtering*. It maps

$$(y^{(T+1)}, \mu_{\mathrm{post}}^{(T)}) \longmapsto \mu_{\mathrm{post}}^{(T+1)}.$$

In filtering, the sequence of measures $\underline{\mu}$ is given by

$$\mu_0 := \mu_{\mathrm{prior}}, \quad \mu_k := \mu_{\mathrm{post}}^{(k)}, \text{ and } \gamma_k := L_k(y^{(k)}|\cdot),$$

where the index $k = 1, \dots, N_{\mathrm{seq}}$ refers to the time $t =: k$.
The filtering of a model parameter has for instance been discussed in [31, 236]. It is closely related to the filtering problem in *data assimilation*, see, e.g. [73, 128, 157,

189]. In data assimilation, the states of a dynamical system shall be estimated. This is in contrast to the *static* parameter estimation that is considered throughout this thesis, where the uncertain parameter is fixed.
In the next section, we introduce SIS and SMC to approximate sequences of measures. As these sequences often arise in (Bayesian) filtering problems, SIS and SMC are sometimes referred to as *particle filters*, even if actually no filtering is carried out.

5.2 Foundations of SMC

We consider the framework from §5.1. Hence, $\underline{\mu}$ is a sequence of measures given recursively in (5.1). We assume that we can sample independently from $\mu_0 \in \mathrm{Prob}(X)$ and that we can evaluate the densities in $\underline{\gamma}$. We now aim to approximate the sequence $\underline{\mu}$ with Monte Carlo. To this end, we introduce the Sequential Importance Sampling method, as well as the Sequential Monte Carlo method.

5.2.1 Sequential Importance Sampling

The most basic algorithm to approximate $\underline{\mu}$ is *Sequential Importance Sampling* (SIS). We apply Importance Sampling to approximate μ_1 with samples from μ_0. Then, we update the weights to approximate μ_k for $k = 2, 3, \ldots, N_{\mathrm{seq}}$; see §3.3.2 for an introduction to Importance Sampling. The algorithm proceeds as follows. Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_0$ be i.i.d. samples. Note that all approximate measures in the following are random measures and therefore printed bold. These random measures are indeed well-defined, as will be shown in §5.3. We approximate μ_0 by

$$\widehat{\boldsymbol{\mu}}_0 := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n) = \sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n^{(0)} \delta(\cdot - \boldsymbol{\theta}_n), \tag{5.3}$$

where $\boldsymbol{w}_n^{(0)} = 1/N_{\mathrm{smp}}$ for $n = 1, \ldots, N_{\mathrm{smp}}$ are uniform weights. We obtain the approximation of μ_k by

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_k &:= \sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n^{(k)} \delta(\cdot - \boldsymbol{\theta}_n), \\ \boldsymbol{w}_n^{(k)} &:= \frac{\gamma_k(\boldsymbol{\theta}_n) \boldsymbol{w}_n^{(k-1)}}{\sum_{m=1}^{N_{\mathrm{smp}}} \gamma_k(\boldsymbol{\theta}_m) \boldsymbol{w}_m^{(k-1)}}, \qquad (n = 1, \ldots, N_{\mathrm{smp}}), \end{aligned}$$

where $k = 1, \ldots, N_{\mathrm{seq}}$.
Interestingly, we can also represent this update using the operators $(\mathrm{W}_k)_{k=1}^{N_{\mathrm{seq}}}$ given in (5.2). $\widehat{\boldsymbol{\mu}}_0$ is given as above in (5.3). Then, we have

$$\widehat{\boldsymbol{\mu}}_k := \mathrm{W}_k\left(\widehat{\boldsymbol{\mu}}_{k-1}\right) \qquad (k = 1, \ldots, N_{\mathrm{seq}}),$$

since the sums in the definition of the weights are integrals with respect to a discrete measure. Indeed,

$$\int_A \gamma_k(\theta) \widehat{\boldsymbol{\mu}}_{k-1}(\mathrm{d}\theta) = \sum_{n=1}^{N_{\mathrm{smp}}} \mathbf{1}_A(\boldsymbol{\theta}_n) \gamma_k(\boldsymbol{\theta}_n) \boldsymbol{w}_n^{(k-1)} \qquad (A \in \mathcal{B}X).$$

Hence, in Sequential Importance Sampling, we simply apply the correct weighting operator W_k to an approximate measure.

Remark 5.1. Let $k = 1, \ldots, N_{\mathrm{seq}}$. The idea of SIS is to apply the exact operator $\mathrm{W}_k : \mathrm{Prob}(X) \to \mathrm{Prob}(X)$ to an approximation of μ_k rather than the actual measure. This idea is conceptually similar to, e.g. an FEM discretisation of an elliptic equation. In that case, we consider the exact weak operator $\mathcal{E} : H \times H' \to \mathbb{R}$ on finite dimensional subspaces $H_\ell \subseteq H$ and $H'_\ell \subseteq H'$, see §3.1. In Sequential Importance Sampling, we replace $\mathrm{Prob}(X)$ by the convex hull of the Dirac measures in the points $\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}$. Those points are the realisations of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$ in the initial step. We denote the convex hull by

$$C_{\underline{\theta}} := \mathrm{conv}\left\{\delta(\cdot - \theta_1), \ldots, \delta(\cdot - \theta_{N_{\mathrm{smp}}})\right\} = \left\{\sum_{n=1}^{N_{\mathrm{smp}}} w_n \delta(\cdot - \theta_n) : \underline{w} \geq 0, \|\underline{w}\|_1 = 1\right\}.$$

Note that $C_{\underline{\theta}} \subseteq \mathrm{Prob}(X)$ and that $\mathrm{img}(\mathrm{W}_k|_{C_{\underline{\theta}}}) \subseteq C_{\underline{\theta}}$ since the W_k operator only reweighs the particles, but does not move them. ◇

Long-time behaviour. Throughout the Sequential Importance Sampling process, the particles $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$ do not change their position. This has bad implications if μ_0 and μ_k are not similar for some $k = 1, \ldots, N$. Next, we specify and discuss this claim theoretically. Moreover, we give an example in which we see how the SIS approximation degenerates over time.
Note that the measure $\widehat{\boldsymbol{\mu}}_k$ is indeed equivalent to an Importance Sampling approximation of μ_k, where we use samples from μ_0 and the density $\overline{\gamma}_k := \prod_{i=1}^k \gamma_i$. Hence, we can use the Importance Sampling error bound that we have discussed in Proposition 3.5. We obtain

$$\mathrm{d}_{\mathrm{etv}}(\mu_k, \widehat{\boldsymbol{\mu}}_k) \leq 2 \cdot \sqrt{\frac{\rho_k}{N_{\mathrm{smp}}}}, \tag{5.4}$$

where $\rho_k := \int_X \overline{\gamma}_k^2(\theta)\mu_0(\mathrm{d}\theta) / \left(\int_X \overline{\gamma}_k(\theta)\mu_0(\mathrm{d}\theta)\right)^2$.
Consider the following example, where we consider a sequence of measures for which we can compute the sequence $(\rho_k)_{k=1}^\infty$. This example is a simplified version of [31, Example 4], where Bayesian filtering is considered, rather than a generic sequence of measures.

Example 5.2 (Gaussian example)**.** Let $\mu_k = \mathrm{N}(0, 1/(k+1))$, for $k = 0, \ldots, \infty$. We can construct this sequence starting from a standard Gaussian measure $\mu_0 = \mathrm{N}(0, 1^2)$ and updating the measure with another standard Gaussian density in every step. Hence, for $k \in \mathbb{N}$, we have

$$\gamma_k(\theta) \propto \mathrm{n}(\theta; 0, 1^2) \propto \exp\left(-\frac{1}{2}\theta^2\right) \qquad (\theta \in X).$$

In this case, we can compute ρ_k analytically and obtain

$$\rho_k = \frac{k+1}{\sqrt{2k+1}}.$$

It is easy to see that $\rho_k \to \infty$, as $k \to \infty$. Hence, over time the bound on the right-hand side in (5.4) goes to infinity. ◇

In Example 5.2, we expect that the measure approximation becomes arbitrarily bad as the algorithm proceeds over time. This result is independent of the number of samples that is used for the approximation. Intuitively, we can explain this with the behaviour of the sequence $\underline{\mu}$: We have $\mu_k \xrightarrow{\mathrm{w}} \delta(\cdot - 0)$, as $k \to \infty$. This assertion can be shown using the characteristic functions of the $(\mu_k)_{k\in\mathbb{N}}$; recall Theorem 1.21.
Hence, we are able to describe the limiting measure with our set of particles if and only if one of the particles $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$ equals 0. However, it is unlikely that this happens. Indeed, one can show that

$$\mathbb{P}\left(\bigcup_{n=1}^{N_{\mathrm{smp}}} \{\boldsymbol{\theta}_n = 0\}\right) \leq \sum_{n=1}^{N_{\mathrm{smp}}} \mathbb{P}\left(\{\boldsymbol{\theta}_n = 0\}\right) = 0.$$

Preasymptotically, the measures μ_k, with $k \gg 0$, are concentrated close to 0. Let $\varepsilon > 0$ be small. The probability that one of the particles has an ε-neighbourhood A, with $\mu_k(A) \gg 0$, is then not zero, but arbitrarily small.
Sequences of measures that become more concentrated over time and converge to a Dirac measure appear in Bayesian filtering. In mathematical terms, this has been shown in the *Bernstein–von Mises Theorem*, see [250, §10.2].
We have introduced SIS in this section as an iterated version of Importance Sampling, see §1.3.1. This method is obtained considering the algorithm from [157, §4.3.2] without underlying dynamical system and without resampling. In the static framework, the SIS algorithm can also be defined differently. [62, 73], for instance, introduce an *importance Markov kernel* to move the particles after each step. Hence, the importance distribution is not equal to μ_0 in every step. Unfortunately, these importance distributions have to be known a priori and the importance Markov kernels have to be constructed accordingly. The method is otherwise impractical [62, §2.4]. However, finding good importance distributions a priori may not be possible either in practical problems. Therefore, we have defined SIS as an iterated Importance Sampling, with only a single importance distribution, that is, μ_0.

5.2.2 From SIS to SMC

In SIS, we do not change the particle positions. Hence, they are distributed according to μ_0 throughout the entire algorithm. As shown in Example 5.2, this can lead to a bad asymptotic behaviour of the SIS estimates. Indeed, the effective sample size in this case converges to zero. In *Sequential Monte Carlo*, we aim to solve this problem and retain a large ESS, by moving and removing particles.
We now discuss how the algorithm proceeds. As in SIS, we sample N_{smp} times from μ_0 and construct the approximate measure $\widehat{\boldsymbol{\mu}}_0$ as in (5.3). We now iterate over $k = 1, \ldots, N_{\mathrm{seq}}$. First, we apply W_k and obtain $\widehat{\boldsymbol{\mu}}_k := \mathrm{W}_k(\widehat{\boldsymbol{\mu}}_{k-1})$. Then, we check the effective sample size in the measure $\widehat{\boldsymbol{\mu}}_k$. We approximate the effective sample size by

$$\widehat{\mathrm{ESS}}_k := J \cdot \frac{\left(\int_X \gamma_k(\theta)\widehat{\boldsymbol{\mu}}_{k-1}(\mathrm{d}\theta)\right)^2}{\int_X \gamma_k(\theta)^2 \widehat{\boldsymbol{\mu}}_{k-1}(\mathrm{d}\theta)} = J \cdot \frac{\left(\frac{1}{J}\sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n^{(k)}\right)^2}{\frac{1}{J}\sum_{i=1}^{N_{\mathrm{smp}}} \left(\boldsymbol{w}_n^{(k)}\right)^2} = \frac{\left(\sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n^{(k)}\right)^2}{\sum_{i=1}^{N_{\mathrm{smp}}} \left(\boldsymbol{w}_n^{(k)}\right)^2}, \quad (5.5)$$

where $\underline{\boldsymbol{w}^{(k)}}$ represents the weights in $\widehat{\boldsymbol{\mu}}_k$; see (3.8) for the definition of the effective sample size in Importance Sampling. If $\widehat{\mathrm{ESS}}_k$ is larger than a pre-defined threshold

$\tau_{\mathrm{ESS}} > 0$, we set $k := k + 1$ and continue in the loop. Otherwise, if $\widehat{\mathrm{ESS}}_k$ is smaller than the treshold τ_{ESS}, we proceed with the *selection step* and the *mutation step*.

Selection step. In the selection step, we aim to remove particles with small weights. Those particles barely contribute to the approximation of the probability measure and shall be deleted; recall that the weights sum to one. Deletion means, we move the particle from a particular position to set it at the position of a particle with a larger weight. As a selection strategy, we use *(multinomial) resampling*. In particular, we sample N_{smp} times from the discrete measure $\widehat{\boldsymbol{\mu}}_k$ and obtain a new set of samples $\widetilde{\boldsymbol{\theta}}_1, \dots, \widetilde{\boldsymbol{\theta}}_{N_{\mathrm{smp}}} \sim \widehat{\boldsymbol{\mu}}_k$. By resampling, we naturally obtain samples at positions with high weights. For other resampling (or selection) methods, we refer to [95].

Mutation step. In the mutation step, we now use a Markov kernel to change the position of the particles. Let K_k be a Markov kernel from $(X, \mathcal{B}X)$ into itself. Moreover, let K_k be stationary with respect to μ_k. To each of the particles $\widetilde{\boldsymbol{\theta}}_1, \dots, \widetilde{\boldsymbol{\theta}}_{N_{\mathrm{smp}}}$, we now apply K_k, i.e. we sample

$$\boldsymbol{\theta}_1 \sim K_k(\cdot\,|\widetilde{\boldsymbol{\theta}}_1), \dots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim K_k(\cdot\,|\widetilde{\boldsymbol{\theta}}_{N_{\mathrm{smp}}}).$$

If K_k satisfies ergodicity conditions, the particles are now approximately μ_k-distributed. Such ergodicity conditions are strong, geometric, and uniform ergodicity; see [213, §6.6.2-6.6.3]; see also §3.3.3. Thus, we redefine

$$\widehat{\boldsymbol{\mu}}_k := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n).$$

Next, we set $k := k + 1$ and continue in the loop. The SMC algorithm terminates if k exceeds N_{seq}.

Remark 5.3. We use the denominations *selection* and *mutation* to describe the SMC algorithm. These terms imply a genetic intuition behind the algorithm. In this case, the particles represent different individuals and the weights of the particles represent their fitness. Species with small weights, and thus a low fitness, are killed in the selection step. The mutation step represents a variety of possible random evolutions. This and other interpretations of Sequential Monte Carlo are discussed in [61, §4]. ◇

Similarly to the W_k in SIS, we define the selection and mutation step as a family of Markov kernels. Note that $\mathrm{W}_k : \mathrm{Prob}(X) \to \mathrm{Prob}(X)$ are operators since the reweighing step is deterministic. We can represent W_k as a Markov kernel in terms of a point-wise Dirac measure. Selection and mutation step form a Markov kernel S_k since both the selection and the mutation follow a randomised procedure. Let $\mathrm{S}_k : \mathrm{Prob}(X) \times \mathcal{P}(X) \to [0, 1]$ be the Markov kernel defined by

$$(\mu, A) \mapsto \iint_{X^{N_{\mathrm{smp}}} \times X^{N_{\mathrm{smp}}}} \mathbf{1}_A \left(\sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \theta_n) \right) K_k^{\otimes N_{\mathrm{smp}}}(\mathrm{d}\underline{\theta}|\underline{\widetilde{\theta}}) \mu^{\otimes N_{\mathrm{smp}}}(\mathrm{d}\underline{\widetilde{\theta}}). \quad (5.6)$$

We show the well-definedness of this Markov kernel in §5.3. There, we also discuss the definition of the σ-algebra $\mathcal{P}(X) \subseteq 2^{\mathrm{Prob}(X)}$. In the following remark, we briefly explain the construction of the function given in (5.6).

Remark 5.4. We consider the map given in (5.6) and describe successively how it is constructed. We consider

$$(\mu, A) \mapsto \iint_{X^{N_{\mathrm{smp}}} \times X^{N_{\mathrm{smp}}}} \mathbf{1}_A \left(\underbrace{\sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \theta_n)}_{\text{(iii)}} \right) \underbrace{K_k^{\otimes N_{\mathrm{smp}}}(\mathrm{d}\underline{\theta}|\widetilde{\underline{\theta}})}_{\text{(ii)}} \underbrace{\mu^{\otimes N_{\mathrm{smp}}}(\mathrm{d}\widetilde{\underline{\theta}})}_{\text{(i)}}.$$

(i) μ is the input of the kernel. We indicate with this term that $\widetilde{\underline{\theta}}$ is a realisation of a vector of N_{smp} independent random variables distributed according to μ. This represents the selection step.

(ii) We take the realisations $\widetilde{\underline{\theta}}$ and use them as inputs of the Markov kernels K_k, which we apply independently to each of them. The realisations we obtain from the K_k are in the vector $\underline{\theta}$. This represents the mutation step.

(iii) We construct an empirical measure from the realisations $\underline{\theta}$, with equal weights.

Finally, we integrate the indicator $\mathbf{1}_A(\dots)$ to obtain the probability of (iii) being in A. ◇

The SMC algorithm constructs the following random sequence:

$$\widehat{\boldsymbol{\mu}}_0 := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n), \qquad (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_0, \text{ i.i.d.}),$$

$$\widehat{\boldsymbol{\mu}}_k := \begin{cases} \mathrm{W}_k(\widehat{\boldsymbol{\mu}}_{k-1}), & \text{if } \widehat{\mathrm{ESS}}_k \geq \tau_{\mathrm{ESS}}, \\ \boldsymbol{\mu} \sim \mathrm{S}_k\left(\cdot \middle| \mathrm{W}_k(\widehat{\boldsymbol{\mu}}_{k-1})\right), & \text{otherwise;} \end{cases} \qquad (k = 1, \dots, N_{\mathrm{seq}}).$$

We call this method *SMC with adaptive resampling*. If we skip the ESS test and always apply $\mathrm{S_k}$, we obtain *SMC*. In this case, the sequence simplifies to

$$\widehat{\boldsymbol{\mu}}_0 := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n), \qquad (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_0, \text{ i.i.d.}),$$

$$\widehat{\boldsymbol{\mu}}_k \sim \mathrm{S}_k\left(\cdot \middle| \mathrm{W}_k(\widehat{\boldsymbol{\mu}}_{k-1})\right) \qquad (k = 1, \dots, N_{\mathrm{seq}}).$$

In Figure 5.1, we illustrate the steps in SMC to update $\widehat{\boldsymbol{\mu}}_{k-1}$ to $\widehat{\boldsymbol{\mu}}_k$. There, we use circles to represent the positions of the particles and antennas to represent their weight. We start with a measure having equal weights. In the weighting step, we change the weights without changing the particle positions. In the selection step, we apply resampling and remove the first and fifth particle, which are now placed at the positions of the third and fourth particle. All the weights are now identical again. Finally, we mutate the particle measure, meaning, we change the particles' positions.

The Sequential Monte Carlo methodology in the setting of static parameters has been first proposed by Del Moral et al. [62]. A more thorough introduction to interacting particle methods are given in the textbooks by Del Moral [60, 61]. As historical resources concerning particle filtering methods used for static parameter

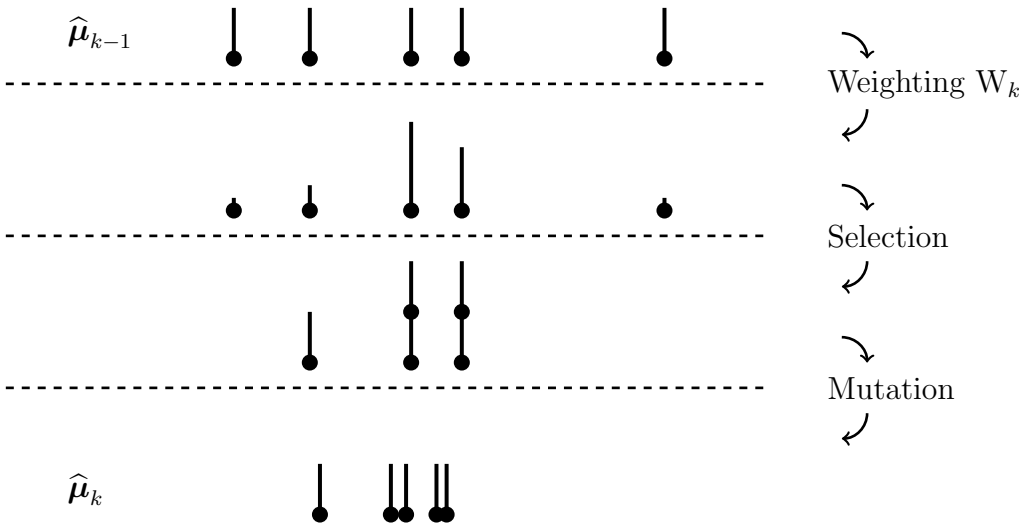


Figure 5.1. A cartoon of the SMC update from $\widehat{\boldsymbol{\mu}}_{k-1}$ to $\widehat{\boldsymbol{\mu}}_k$. Each of the compartments limited by the dashed lines marks one particle-based discrete probability measure. On the right-hand side, we note the functions applied to update the measures. The dots represent the positions of the particles in a one-dimensional space; the vertical lines (antennas) represent the weights associated to the particles. Dots lying vertically above each other have the exact same position.

estimation, we mention the articles by Chopin [43] and Neal [190, 191]. Various researchers have analysed Sequential Monte Carlo methods. To name a few: Chopin [44] considered the long-time behaviour of particle filters by showing a central limit theorem. Rebeschini and van Handel [208], as well as Beskos et al. [16, 20] discussed the use of particle filters in high-dimensional settings. Finally, we mention Whiteley [257], who gives stability results for SMC.

5.2.3 Simulating normalising constants with SMC

The *model evidence* $Z(y^\dagger)$ appears as a normalising constant in Bayes' rule; see Theorem 1.45. When introducing Importance Sampling in §3.3.2 and MCMC in §3.3.3, we were able to ignore its computation. In some applications, it is necessary to compute the model evidence; e.g. in Bayesian model selection which we have considered in §1.4.3.

Let $k = 1, \ldots, N_{\mathrm{seq}}$. Equivalently to model evidences, we can consider the estimation of the following normalising constants in the framework of §5.1:

$$\overline{Z}_k := \int_X \overline{\gamma}_k(\theta)\mu_0(\mathrm{d}\theta).$$

In SMC, we can easily obtain estimates of $\overline{Z}_k$. These estimates are based on the

following identity:

$$\begin{aligned}\overline{Z}_k = \int_X \prod_{i=1}^{k} \gamma_i(\theta)\mu_0(\mathrm{d}\theta) &= \int_X \left(\prod_{i=2}^{k} \gamma_i(\theta)\right) \frac{\gamma_1(\theta)}{\int_X \gamma_1(\theta')\mu_0(\mathrm{d}\theta')}\mu_0(\mathrm{d}\theta) \int_X \gamma_1(\theta)\mu_0(\mathrm{d}\theta) \\ &= \int_X \prod_{i=2}^{k} \gamma_i(\theta)\mu_1(\mathrm{d}\theta) \int_X \gamma_1(\theta)\mu_0(\mathrm{d}\theta) \\ &= \prod_{i=1}^{k} \int_X \gamma_i(\theta)\mu_{i-1}(\mathrm{d}\theta). \end{aligned} \tag{5.7}$$

In SMC, we obtain approximations for the factors in (5.7) when computing the normalisation of the weights for W_k. Indeed, we have the approximations

$$\overline{\boldsymbol{Z}}_k := \prod_{i=1}^{k} \int_X \gamma_i(\theta)\widehat{\boldsymbol{\mu}}_{i-1}(\mathrm{d}\theta) := \prod_{i=1}^{k} \left(\sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \gamma_i(\boldsymbol{\theta}_n)\right).$$

We refer to, e.g. [62, §3.2] for a discussion of this approach.

5.3 A random measure derivation of SMC

In several instances in the last chapter, we have discussed random probability measures. Those are random variables that take values in $\mathrm{Prob}(X)$. Indeed, we have defined Sequential Importance Sampling and Sequential Monte Carlo as algorithms that generate sequences of random measures. This point of view has been taken before by, e.g. Crisan [51] and Beskos et al. [20]. We now set it in a rigorous foundation. In §5.3.1, we discuss some foundations of random measures and show that the sequences generated in SIS and SMC are well-defined. Indeed, the sequences generated in SIS and SMC are Markov chains in $\mathrm{Prob}(X)$. In §5.3.2, we commence an MCMC-type discussion of the Markov chains generated by SIS and SMC. Here, we consider the homogeneity of the Markov chains, as well as their stationary points and stationary measures.

5.3.1 SMC and random measures

Let X be a separable Banach space. We discuss random variables taking values in $\mathrm{Prob}(X)$. We will represent random measures as Markov kernels, as it is usual in the literature; see, e.g. [137] for a detailed introduction.

Definition 5.5. A *random measure* $\boldsymbol{\mu}$ on $(X, \mathcal{B}X)$ is a Markov kernel from $(\Omega, \mathcal{A})$ to $(X, \mathcal{B}X)$. Hence, $\boldsymbol{\mu} : \Omega \times \mathcal{B}X \to [0, 1]$. ◇

Note that throughout this thesis, random measures are always *random probability measures*. We often denote Markov kernels M in a *conditional-measure*-sense, such as $M(\cdot\,|*) := M(*, \cdot)$; see §1.2.4 for a justification of this notation. As opposed to this, we treat random measures as random variables taking values in $\mathrm{Prob}(X)$. Hence, we usually drop the dependence on $\omega \in \Omega$ and use boldface letters.

Remark 5.6. Let $\boldsymbol{\mu}$ be a random measure on $(X, \mathcal{B}X)$. The probability distribution of $\boldsymbol{\mu}$ is defined on the σ-algebra generated by the evaluation maps $f_A : \mathrm{Prob}(X) \to [0,1]$, where $\mu \mapsto \mu(A)$. First, let

$$f_A^{-1}(B) := \{f_A \in B\} := \{\mu \in \mathrm{Prob}(X) : \mu(A) \in B\}$$

be the pre-image of $B \subseteq [0,1]$. We then consider the smallest σ-algebra generated by all these maps that is

$$\mathcal{P}(X) := \sigma_{\mathrm{Prob}(X)}\left(\{f_A^{-1}(B) : A \in \mathcal{B}X \text{ and } B \in \mathcal{B}[0,1]\}\right).$$

We can now define a random measure equivalently as a measurable map from $(\Omega, \mathcal{A})$ to $(\mathrm{Prob}(X), \mathcal{P}(X))$; see also Definition 1.10. ◇

We obtain a random measure for instance by sampling a random vector and constructing a discrete measure supported on the samples. Here, the weights can be deterministic or random.

Proposition 5.7. Let $\underline{\boldsymbol{\theta}} : \Omega \to X^{N_{\mathrm{smp}}}$ be a measurable function, and $\underline{w} \in [0,1]^{N_{\mathrm{smp}}}$ be a set of weights that sums to 1, i.e. $\|\underline{w}\|_1 = 1$. Then, the map

$$\boldsymbol{\mu} : \Omega \times \mathcal{B}X \to [0,1], \qquad (\omega, A) \mapsto \sum_{n=1}^{N_{\mathrm{smp}}} w_n \delta(A - \boldsymbol{\theta}_n(\omega)) \tag{5.8}$$

is a random measure. The same holds true if the weights are random variables as well, i.e., if $\underline{\boldsymbol{w}} : \Omega \to [0,1]^{N_{\mathrm{smp}}}$ is a measurable function, with $\|\underline{\boldsymbol{w}}(\omega)\|_1 = 1$ for $\omega \in \Omega$. In this case, the map

$$\boldsymbol{\mu}' : \Omega \times \mathcal{B}X \to [0,1], \qquad (\omega, A) \mapsto \sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n(\omega) \delta(A - \boldsymbol{\theta}_n(\omega)) \tag{5.9}$$

is a random measure.

Proof. In this proof, we only consider $\boldsymbol{\mu}'$, since $\boldsymbol{\mu}$ arises from $\boldsymbol{\mu}'$ if the weights are $\mathbb{P}$-a.s. constant random variables. We have to show that $\boldsymbol{\mu}'(\omega, \cdot) : \mathcal{B}X \to [0,1]$ is a probability measure ($\omega \in \Omega$), and that $\boldsymbol{\mu}'(\cdot, A) : \Omega \to [0,1]$ is measurable ($A \in \mathcal{B}X$). Let $\omega \in \Omega$. Then, $\sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n(\omega)\delta(\cdot - \boldsymbol{\theta}_n(\omega))$ is a probability measure by construction. To show the measurability, we first assume that $N_{\mathrm{smp}} = 1$ and $\boldsymbol{w}_1 \equiv 1$. Now, we need to show that $\omega \mapsto \delta(A - \boldsymbol{\theta}_1(\omega))$ for any $A \in \mathcal{B}X$.
Let $A \in \mathcal{B}X$. The image of $\omega \mapsto \delta(A - \boldsymbol{\theta}_1(\omega))$ is $\{0,1\}$. Hence, we only need to check that

$$\{\delta(A - \boldsymbol{\theta}_1(\cdot)) = 1\} \in \mathcal{A}, \qquad \{\delta(A - \boldsymbol{\theta}_1(\cdot)) = 0\} \in \mathcal{A}.$$

Indeed, since the two sets are complements of each other, one of the statements is sufficient. We have $\{\delta(A - \boldsymbol{\theta}_1(\cdot)) = 1\} = \{\boldsymbol{\theta}_1 \in A\}$. The latter set, however, is contained in $\mathcal{A}$ since $\boldsymbol{\theta}_1$ is measurable. Therefore, $\delta(A - \boldsymbol{\theta}_1(\cdot))$ is measurable.
This also implies the measurability of $\omega \mapsto \sum_{n=1}^{N_{\mathrm{smp}}} \boldsymbol{w}_n(\omega)\delta(A - \boldsymbol{\theta}_n(\omega))$ since this is a sum of products of measurable functions. As $A \in \mathcal{B}X$ was chosen arbitrarily, we have that $\boldsymbol{\mu}'$ is a random measure. □

Random measures as in (5.8) appear in standard Monte Carlo and Markov chain Monte Carlo. Random measures as in (5.9) appear in Importance Sampling. Thus, the random measures that appear in these Monte Carlo approximations are well-defined. This includes the initial measure $\widehat{\boldsymbol{\mu}}_0$ in SIS and SMC.
Next, we show the measurability of the operators W_k. In particular, we show that when applying W_k to a random measure, we obtain again a random measure. The measurability implies then that the random measures given in Sequential Importance Sampling are well-defined.

Proposition 5.8. Let W_k be the operator in (5.2) for some bounded $\gamma_k : X \to (0,\infty)$. Then, $\mathrm{W}_k : \mathrm{Prob}(X) \to \mathrm{Prob}(X)$ is measurable. In particular, for any random measure $\boldsymbol{\mu}$, the map

$$\boldsymbol{\mu}'(\omega, A) := \mathrm{W}_k(\boldsymbol{\mu})(\omega, A) := \frac{\int_A \gamma_k(\theta)\boldsymbol{\mu}(\omega, \mathrm{d}\theta)}{\int_X \gamma_k(\theta)\boldsymbol{\mu}(\omega, \mathrm{d}\theta)} \qquad (A \in \mathcal{B}X, \omega \in \Omega)$$

is a random measure.

Proof. We show the last assertion, i.e. we show that $\boldsymbol{\mu}'$ is a Markov kernel from $(\Omega, \mathcal{A})$ to $(X, \mathcal{B}X)$.
1. Let $\omega \in \Omega$ be fixed and $\boldsymbol{\mu}$ be a random measure. Then, $\mu' := \boldsymbol{\mu}'(\omega, \cdot) = \mathrm{W}_k(\boldsymbol{\mu})(\omega, \cdot)$ is by construction a measure. Since γ_k and $\boldsymbol{\mu}(\omega, \cdot)$ are bounded, the measure μ' is well-defined and finite. Due to the normalisation, we have $\mu'(X) = 1$. Thus, μ' is a probability measure.
2. Let $A \in \mathcal{B}X$ and $f_A : \Omega \to (0,\infty)$ be given by

$$f(\omega) := \int_A \gamma_k(\theta)\boldsymbol{\mu}(\omega, \mathrm{d}\theta) \qquad (\omega \in \Omega).$$

f_A is measurable according to [136, Lemma 1.38(i)] and strictly positive. Therefore, $f_A/f_X = \boldsymbol{\mu}'(A, \cdot)$ is also measurable. □

Hence, SIS constructs a well-defined sequence of random measures. We now move on to the two versions of SMC. In particular, we show that S_k is indeed a Markov kernel.

Proposition 5.9. Let S_k be the function defined in (5.6). Then S_k is a Markov kernel from $(\mathrm{Prob}(X), \mathcal{P}(X))$ into itself.

Proof. 1. We define three Markov kernels M_1, M_2, M_3, such that $\mathrm{S}_k = M_1 M_2 M_3$. In particular, M_1 is a Markov kernel from $(\mathrm{Prob}(X), \mathcal{P}(X))$ to $(X^{N_{\mathrm{smp}}}, \mathcal{B}X^{N_{\mathrm{smp}}})$, where

$$(\mathrm{Prob}(X) \times \mathcal{B}X^{N_{\mathrm{smp}}}) \ni (\mu, A) \mapsto \mu^{\otimes N_{\mathrm{smp}}}(A) \in [0,1].$$

This Markov kernel represents the *selection step*. M_2 is a Markov kernel from $(X^{N_{\mathrm{smp}}}, \mathcal{B}X^{N_{\mathrm{smp}}})$ into itself, where

$$(\underline{\theta}, A) \mapsto \int_X \cdots \int_X \mathbf{1}_A(\widetilde{\theta}_1, \ldots, \widetilde{\theta}_{N_{\mathrm{smp}}}) K(\mathrm{d}\widetilde{\theta}_1 | \theta_1) \cdots K(\mathrm{d}\widetilde{\theta}_{N_{\mathrm{smp}}} | \theta_{N_{\mathrm{smp}}}).$$

This Markov kernel represents the *mutation step*. Finally, M_3 is a Markov kernel, from $(X^{N_{\mathrm{smp}}}, \mathcal{B}X^{N_{\mathrm{smp}}})$ to $(\mathrm{Prob}(X), \mathcal{P}(X))$, where

$$(\underline{\theta}, A) \mapsto \mathbf{1}_A\left(\sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \theta_n)\right) = \delta\left(A - \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \theta_n)\right).$$

M_3 models the representation of $\underline{\theta}$ as a random measure concentrated in $\theta_1, \dots \theta_{N_{\text{smp}}}$.
2. We now show that M_1, M_2, M_3 are indeed Markov kernels. M_2 is a Markov kernel as it is a product of Markov kernels. M_3 is measurable for fixed $A \in \mathcal{P}(X)$ since indicator functions with respect to measurable sets are measurable and since empirical measures are measurable, see Proposition 5.7. By using the representation with the Dirac measure, we see that M_3 is a probability measure for fixed $\underline{\theta}$. M_1 is by construction a probability measure for fixed $\mu \in \text{Prob}(X)$.
For the measurability proof, with fixed $A \in \mathcal{B}X^{N_{\text{smp}}}$, we assume that $N_{\text{smp}} = 2$. The proof for $N_{\text{smp}} = 1$ is trivial since the M_1 is then the identity map. Moreover, $N_{\text{smp}} > 2$ follows inductively analoguously to the case $N_{\text{smp}} = 2$. The map $\mu \mapsto \mu^{\otimes 2}(A)$ is indeed measurable: Note that we can write

$$\begin{aligned}\mu^{\otimes 2}(A) &= \iint_{X\times X} \mathbf{1}_A(\theta_1, \theta_2)\mu(\mathrm{d}\theta_1)\mu(\mathrm{d}\theta_2) \\ &= \int_X \int_X \mathbf{1}_{A_{\theta_2}}(\theta_1, \theta_2)\mu(\mathrm{d}\theta_1)\mu(\mathrm{d}\theta_2) = \int_X \mu(A_{\theta_2})\mu(\mathrm{d}\theta_2),\end{aligned}$$

where $A_{\theta_2} = \{(\widetilde{\theta}_1, \widetilde{\theta}_2) \in A : \widetilde{\theta}_2 = \theta_2\} \in \mathcal{B}X^2$ is the θ_2-*section* of A. We can represent the integral $\int_X \mu(A_{\theta_2})\mu(\mathrm{d}\theta_2)$ as the limit of a weighted sum of products of $\mu(B)$, for various $B \in \mathcal{B}X$. This follows from the definition of the Bochner integral; see [26]. Due to the definition of the σ-algebra $\mathcal{P}(X)$ in Remark 5.6, mappings of the form $\text{Prob}(X) \ni \mu \mapsto \mu(B) \in [0, 1]$ are measurable. Therefore, this is also true for continuous combinations of these functions and their limits. □

Hence, the version of SMC (without adaptive resampling) that applies S_k in every step, constructs a well-defined sequence of random measures. We discuss the SMC with adaptive resampling in the following remark.

Remark 5.10. In SMC with adaptive resampling, we apply

$$\mathrm{R}_k(\cdot\,|\mu) = \mathbf{1}_{\{\widehat{\mathrm{ESS}}_k(\cdot)\geq\tau_{\text{ESS}}\}}(\mu)\mathrm{W}_k(\mu) + \mathbf{1}_{\{\widehat{\mathrm{ESS}}_k(\cdot)<\tau_{\text{ESS}}\}}(\mu)\mathrm{S}_k\left(\cdot\middle|\mathrm{W}_k(\mu)\right), \tag{5.10}$$

instead of S_k. We now explain why this function is a Markov kernel from $(\text{Prob}(X), \mathcal{P}(X))$ into itself. First, note that

$$\mu \mapsto \widehat{\mathrm{ESS}}_k(\mu) := \frac{J \cdot \left(\int_X \gamma_k(\theta)\mu(\mathrm{d}\theta)\right)^2}{\int_X \gamma_k(\theta)^2\mu(\mathrm{d}\theta)}$$

is a measurable map from $\text{Prob}(X)$ to $\mathbb{R}$. One can prove the measurability of $\widehat{\mathrm{ESS}}_k$ similarly to showing the measurability of M_3 in the proof of Proposition 5.9. Therefore,

$$\{\widehat{\mathrm{ESS}}_k(\cdot) \geq \tau_{\text{ESS}}\}, \{\widehat{\mathrm{ESS}}_k(\cdot) < \tau_{\text{ESS}}\} \in \mathcal{P}(X).$$

Thus, the indicator functions in (5.10) are measurable functions from $\text{Prob}(X)$ to $[0, 1]$. Finally, R_k is a Markov kernel, since it is a continuous combination of measurable functions. ◇

We have rigorously established that SIS and SMC construct sequences of random measures. These sequences are generated with Markov kernels. Hence, the sequences form Markov chains. Throughout the rest of this thesis, we will only consider the version of SMC without adaptive resampling.

5.3.2 MCMC analysis of SIS and SMC with few particles

We have referred to some analytical results concerning SIS and SMC in §5.2.1-5.2.2. Those results usually show consistency and Central Limit Theorems, which hold as the number of particles increases, i.e. $N_{\text{smp}} \to \infty$. SIS and SMC are particle filters, i.e. Monte Carlo methods. In general, we know that using a small number of particles is not advisable; see, e.g. the error bound in (3.5) or Proposition 3.3. In the former, we can hope for good estimation results when the standard deviation of the quantitiy of interest is very small. In the latter, where we consider measure approximations, a small number of samples always gives us a bad estimate. Considering a Monte Carlo approximation with few particles, we also mention von Bortkewitsch's *Law of Small Numbers* [253].
However, in practice, particle filters are used with very few particles. *Deutscher Wetterdienst*, that is the German meteorological service, uses an Ensemble Kalman Filter (EnKF) type method with only $N_{\text{smp}} = 40$ samples; see e.g. [228, Table 3] or the webpage of Deutscher Wetterdienst [66]. The small number of samples is likely caused by the immense computational cost of their models *COSMO* and *ICON*. The long-time behaviour of the EnKF with a finite number of particles has recently been studied by Blömker et al. [25] and Schillings and Stuart [226, 227]. We aim to extend such a discussion to SIS and SMC.
The finite sample behaviour of SIS and SMC has been studied before. We mention the very recent contributions by Marion and Schmiddler [174, 175], who, in particular, study the complexity of SMC methods with finitely many samples. For stability results for SMC with finite samples, we refer to [257].
In the following, we consider the framework introduced in §5.1. Now $\underline{\mu} \in \text{Prob}(X)^{\mathbb{N}}$ is always an infinite sequence of probability measures; hence $N_{\text{seq}} = \infty$. It is again defined by a sequence of functions $\underline{\gamma}$, as in the recursion formula (5.1). Considering only infinite sequences is not a restrictive assumption. If $N_{\text{seq}} < \infty$, we set $\gamma_k \equiv 1$ for $k > N_{\text{seq}}$. In the last section, we have shown that SIS and SMC indeed construct Markov chains in the space $\text{Prob}(X)$. Here, these Markov chains are now infinite and, thus conceptionally similar to the Markov chains that appear in MCMC methods; see §3.3.3. Hence, we can now use the theory of MCMC to investigate the long-time behaviour of SMC and SIS with a finite number of particles.
It has recently become popular to use MCMC theory to investigate objects that do not adhoc come from MCMC samplers. Dunlop et al. [76] show that certain deep Gaussian processes convergence quickly to a stationary measure; indicating that those are not particularly deep. The works [25, 226, 227] discuss the Ensemble Kalman Inversion using Langevin-type dynamics, which appear in, e.g. continuous-time MCMC. See e.g. [167, §9] or [49, §4] for continuous-time MCMC algorithms.
The following paragraphs are an introduction of a new approach to investigate the long-time behaviour of SMC and SIS. We collect various results and leave a more thorough investigation open for future work.

Inhomogeneity. First, we note that the Markov chains generated in SIS and SMC are in general inhomogeneous in time. Observe for instance, that the Markov kernel K_k used in the mutation step is always chosen such that it is stationary with respect to μ_k, $k \in \mathbb{N}$. If $\mu_j \neq \mu_k$ for some $j \neq k$, and $K_j \neq K_k$, the SMC Markov chain is inhomogeneous.

The SIS Markov chain is inhomogeneous, if $\gamma_k \neq \gamma_j$ for some $j \neq k$. This occurs, for instance, in Bayesian filtering, where a new dataset is available in every step. In contrast, in the setting of Example 5.2, the SIS algorithm constructs a homogeneous Markov chain.

Stationary points of W_k. First, we consider the operator W_k, recall (5.2), and determine stationary points in $\mathrm{Prob}(X)$.

Proposition 5.11. Let $\mu \in \mathrm{Prob}(X)$ be given such that a set $C \in \mathcal{B}X$ exists, with $\mu(C) = 1$ and $\gamma_k|_C$ is constant. Then, $\mathrm{W}_k(\mu) = \mu$.

Proof. Let C be given as above and $A \in \mathcal{B}X$. Then,

$$\mathrm{W}_k(\mu)(A) = \frac{\int_A \gamma_k(\theta)\mu(\mathrm{d}\theta)}{\int_X \gamma_k(\theta)\mu(\mathrm{d}\theta)} = \frac{\int_{A\cap C} \gamma_k(\theta)\mu(\mathrm{d}\theta)}{\int_C \gamma_k(\theta)\mu(\mathrm{d}\theta)} \overset{(*)}{=} \frac{\int_{A\cap C} c\mu(\mathrm{d}\theta)}{\int_C c\mu(\mathrm{d}\theta)} = \mu(A\cap C) = \mu(A),$$

where $c := \gamma_k(\theta)$, for some $\theta \in C$. The equality $(*)$ holds since $\gamma_k > 0$ by assumption and, thus, also $c > 0$. □

Note that the stationary points in Proposition 5.11 are measure-valued, but not stationary measures in the sense of §3.3.3. A stationary measure is a measure on $(\mathrm{Prob}(X), \mathcal{P}(X))$ that is preserved when applying W_k. In this case, such a measure is the Dirac in the stationary point: $\delta(\cdot - \mu) \in \mathrm{Prob}(\mathrm{Prob}(X))$.
In SIS and SMC, we consider discrete measures of the form $\mu := \sum_{n=1}^{N_{\mathrm{smp}}} w_n \delta(\cdot - \theta_n)$. Such a measure is a stationary point of W_k if for some index set $I \subseteq \{1, \ldots, N_{\mathrm{smp}}\}$, we have $\gamma_k(\theta_j) = \gamma_k(\theta_i)$ for $i, j \in I$ and $w_i = 0$ if $i \notin I$.

Reducibility of SIS. In §5.2.1, we have shown that in SIS, the particles are not moved. The SIS Markov chain moves only in the set

$$C_{\underline{\theta}} = \mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\}) = \left\{ \sum_{n=1}^{N_{\mathrm{smp}}} w_n \delta(\cdot - \theta_n) : \underline{w} \geq 0, \|\underline{w}\|_1 = 1 \right\},$$

where $\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}$ are the realisations of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_0$ drawn i.i.d. in the first step. We have mentioned this property of SIS already in Remark 5.1. In particular, for all $\mu \in \mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\})$, we have $(W_k \circ \cdots \circ W_j)(\mu) \in \mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\})$ for any $j, k \in \mathbb{N}$, $j \leq k$.
What does this mean for the reducibility of the SIS Markov chain? Let ν be a σ-finite measure on $(\mathrm{Prob}(X), \mathcal{P}(X))$. By the argument above, the SIS Markov chain is not ν-irreducible if

$$\nu(\mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\})) < \nu(\mathrm{Prob}(X)).$$

That means, the SIS Markov chain is reducible if measures not in $\mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\})$ have mass with respect to ν, i.e. if

$$\nu(\mathrm{Prob}(X) \backslash \mathrm{Prob}(\{\theta_1, \ldots, \theta_{N_{\mathrm{smp}}}\})) > 0.$$

Stationary measures in SMC. In the last paragraphs, we have discussed measures that are stationary with respect to W_k by finding stationary points of W_k. This gives us information about stationary measures of SIS.
Now we consider stationary measures of SMC. First note that an investigation of the Markov kernel S_k alone may be unrewarding. This is the case since SMC proceeds by applying S_k to weighted measures of type $\sum_{n=1}^{N_{\mathrm{smp}}} w_n \delta(\cdot - \theta_n)$ with arbitrary weights $\underline{w}$, but returns only measures with $w_n = 1/N_{\mathrm{smp}}$, $n = 1, ..., N_{\mathrm{smp}}$. Hence, the set of stationary measures would be strongly restricted. Therefore, we consider the composed Markov kernel

$$\mathrm{U}_k := \mathrm{S}_k(\cdot|\mathrm{W}_k(\cdot)) \qquad (k \in \mathbb{N}).$$

Here we naturally have only measures with equal weights $w_n = 1/N_{\mathrm{smp}}$, $n = 1, ..., N_{\mathrm{smp}}$ as inputs and outputs.
When approximating a sequence $\underline{\mu}$ with SMC, we obtain a Markov chain $\underline{\widehat{\boldsymbol{\mu}}}$. Ideally, this chain would contain measures of the following type:

$$\widehat{\boldsymbol{\mu}}_k := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n), \qquad \boldsymbol{\theta}_n \sim \mu_k, \qquad (n = 1, ..., N_{\mathrm{smp}}), \tag{5.11}$$

where $k \in \mathbb{N}$. Let $N_{\mathrm{smp}} \geq 2$. Note that the description in (5.11) is not sufficient to determine the probability distribution of the random object $\widehat{\boldsymbol{\mu}}_k$ since it is lacking a description of the interdependence of the particles $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N_{\mathrm{smp}}}$. We give two examples for interdependences and determine probability distributions of the associated random measure.

Example 5.12. (i) Consider the setting in (5.11) and let

$$\mathbb{P}(\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_{N_{\mathrm{smp}}}) = 1.$$

Then $\widehat{\boldsymbol{\mu}}_k = \delta(\cdot - \boldsymbol{\theta}_1)$, and

$$\widehat{\boldsymbol{\mu}}_k(A) = \mathbf{1}[\boldsymbol{\theta}_1 \in A] \sim \mathrm{Bi}(1, \mu_k(A)),$$

where $A \in \mathcal{B}X$. The last assertion follows from $\mathbb{P}(\boldsymbol{\theta}_1 \in \cdot) = \mu_k$.

(ii) Consider again the setting in (5.11), and let now $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$ be independent. Then,

$$(N_{\mathrm{smp}} \cdot \widehat{\boldsymbol{\mu}}_k(A)) = \sum_{n=1}^{N_{\mathrm{smp}}} \mathbf{1}[\boldsymbol{\theta}_n \in A] \sim \mathrm{Bi}(N_{\mathrm{smp}}, \mu_k(A))$$

for all $A \in \mathcal{B}X$. The statement about the distribution of $\widehat{\boldsymbol{\mu}}_k(A)$ follows again from $\mathbb{P}(\boldsymbol{\theta}_1 \in \cdot) = \mu_k$ and the independence of the $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$.

The probability distributions of the random measures in the settings (i) and (ii) are indeed different. Note that they are even concentrated on different spaces. In the first setting $\mathbb{P}(\widehat{\boldsymbol{\mu}}_k(A) \in \{0, 1\}) = 1$; in the second setting $\mathbb{P}(\widehat{\boldsymbol{\mu}}_k(A) \in \{0, 1/N_{\mathrm{smp}}, 2/N_{\mathrm{smp}}, \ldots, 1\}) = 1$ for any $A \in \mathcal{B}X$. ◇

In the following, we consider all probability measures on the space $(\mathrm{Prob}(X), \mathcal{P}(X))$ that describe random measures as described in (5.11). We define the set

$$\begin{aligned}\mathrm{P}_k := &\left\{ \mathbb{P}(\widehat{\boldsymbol{\mu}}_k \in \cdot\,) \,\middle|\, \widehat{\boldsymbol{\mu}}_k := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n), \quad \boldsymbol{\theta}_n \sim \mu_k, \quad n = 1, ..., N_{\mathrm{smp}} \right\} \\ &\subseteq \mathrm{Prob}(\mathrm{Prob}(X))\end{aligned}$$

of all of these probability distributions. We end this section by showing that when applying the Markov kernel U_k to a random measure $\widehat{\boldsymbol{\mu}}_{k-1}$ distributed according to $\mathcal{M} \in \mathrm{P}_k$, we obtain a random measure $\widehat{\boldsymbol{\mu}}_k$ distributed according to $\mathcal{M}' \in \mathrm{P}_k$. Hence, if we apply U_k to a random measure $\widehat{\boldsymbol{\mu}}_{k-1}$ of the form (5.11), we obtain a random measure $\widehat{\boldsymbol{\mu}}_k$ of the form (5.11).

Theorem 5.13. Let $k \in \mathbb{N}$ and $\mathcal{M} \in \mathrm{P}_k$. Then, some probability measure $\mathcal{M}' \in \mathrm{P}_k$ exists, such that

$$\mathcal{M}\mathrm{U}_k = \mathcal{M}'.$$

Proof. Let $\mathcal{M} \in \mathrm{P}_k$ and $\widehat{\boldsymbol{\mu}}_{k-1} \sim \mathcal{M}$. When evaluating U_k, we first apply W_k, and then successively the Markov kernels M_1, M_2, M_3, which we have defined in the proof of Proposition 5.9. When applying W_k, we do not change the distribution of the particles $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_k$. In the resampling step M_1, we obtain new particles $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}}$ that are only a permutation of the original particles. Hence, each of them is still distributed according to μ_k. In the mutation step M_2, we apply K_k to all of the particles. By definition, we have $\mu_k = \mu_k K_k$. Hence, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\mathrm{smp}}} \sim \mu_k$. In M_3, we put the resulting particles in a measure

$$\widehat{\boldsymbol{\mu}}_k := \sum_{n=1}^{N_{\mathrm{smp}}} \frac{1}{N_{\mathrm{smp}}} \delta(\cdot - \boldsymbol{\theta}_n).$$

By the discussion above, this is a random measure still of the form (5.11). Thus, we can choose $\mathcal{M}' =: \mathbb{P}(\widehat{\boldsymbol{\mu}}_k \in \cdot)$, proving our assertion. □

Hence, we have determined a stationary regime for SMC. In MCMC algorithms, we usually have a single stationary measure; recall Theorem 3.8. In SMC, we rather have a *stable set* of measures. If we can also show ergodic behaviour of these Markov chains, we can probably comment on whether this stable set is reached. This will give us a deeper understanding of the long-time behaviour of SMC with small sets of particles. However, this is not part of this thesis.

5.4 Sequences of measures in Bayesian inversion

We next discuss Sequential Monte Carlo methods that are used to discretise Bayesian inverse problems. Note that the sequences of measures discussed here do not appear naturally as in a Bayesian filtering setting. Instead, the sequences are constructed artificially for an efficient approximation of the Bayesian inversion. We introduce the sequences arising in tempering in §5.4.1, and bridging in §5.4.2, as well as their adaptive construction in §5.4.3. In the following, we consider a BIP with prior μ_{prior}, likelihood $L(y^\dagger|\theta) =: \exp(-\Phi(\theta))$, and posterior μ_{post}. We discuss Sequential Monte Carlo techniques that can be used in this setting.

5.4.1 Tempering

In BIPs the posterior measure is often concentrated in a small area of the high-dimensional parameter space X. *Tempering* (T) is a widely-used method to approximate such measures. The fundamental idea – borrowed from Statistical Physics – is to adjust the temperature $\mathcal{T}$ in the Boltzmann distribution.[a] In a Monte Carlo setting, tempering is the systematic raising of a density to some power $\beta := 1/\mathcal{T} \in (0, 1]$. Looking at the Boltzmann distribution, this means that $\mathcal{T} \in [1, \infty)$.
In Proposition 3.5, we have discussed the error of Importance Sampling. Let $\mu_0 \in \mathrm{Prob}(X)$ and $\mu_1 \ll \mu_0$ with $\gamma \propto \mathrm{d}\mu_1/\mathrm{d}\mu_0$. We bound the $\mathrm{d}_{\mathrm{etv}}$-distance between μ_1 and its Importance Sampling approximation with N_{smp} samples from μ_0 by $\sqrt{\rho/N_{\mathrm{smp}}}$, where $\rho := \int_X \gamma(\theta)^2 \mu_0(\mathrm{d}\theta) / \left(\int_X \gamma(\theta)\mu_0(\mathrm{d}\theta)\right)^2$. Tempering can decrease ρ.

Lemma 5.14. Let $\mu_0 \in \mathrm{Prob}(X)$ and $\gamma : X \to (0, \infty)$ bounded. Then, there is a constant $\beta \in (0, 1]$, such that

$$\rho_\beta := \frac{\int_X \gamma^{2\beta}(\theta)\mu_0(\mathrm{d}\theta)}{\left(\int_X \gamma^\beta(\theta)\mu_0(\mathrm{d}\theta)\right)^2} \leq \frac{\int_X \gamma^2(\theta)\mu_0(\mathrm{d}\theta)}{\left(\int_X \gamma(\theta)\mu_0(\mathrm{d}\theta)\right)^2} = \rho.$$

If $\rho > 1$, the inverse temperature $\beta \in (0, 1]$ can be chosen such that the inequality above is strict.

Proof. We have $\rho_\beta = 1$ for $\beta = 0$ and $\rho_\beta \geq 1$ for $\beta \in (0, 1]$ by Jensen's inequality. If $\rho = 1$, the statement holds with $\beta = 1$ since $\rho_1 = \rho$. Let now $\rho > 1$. Since γ is bounded, Lebesgue's Dominated Convergence Theorem implies that $[0, 1] \ni \beta \mapsto \rho_\beta$ is a continuous map. Hence, $\beta \mapsto \rho_\beta$ is a continuous function connecting $\rho_0 = 1$ and $\rho_1 > 1$. By the Intermediate Value Theorem, there is some $\beta \in (0, 1)$, such that $\rho_\beta < \rho_1 = \rho$. □

We note that a similar result was shown with the same technique in [17, Lemma 3.1]. In the BIP setting, we have $\mu_0 := \mu_{\mathrm{prior}}$ and $\gamma := \exp(-\Phi)$. Lemma 5.14 tells us that we can use tempering for a more efficient Importance Sampling approximation. However, when just applying tempering, we change the update density and approximate a different measure. To solve this problem, we apply tempering in combination with an SMC sampler with $N_{\mathrm{T}} \in \mathbb{N}$ intermediate steps. We start with the prior $\mu_0 := \mu_{\mathrm{prior}}$; this is equivalent to an infinite temperature $\mathcal{T} = \infty$ or an inverse temperature $\beta_0 = \mathcal{T}^{-1} = 0$. In the subsequent steps we scale down the temperature $\mathcal{T}$ successively until $\beta_{N_{\mathrm{T}}} = \mathcal{T}^{-1} = 1$, and we arrive at the posterior $\mu_{N_{\mathrm{T}}} = \mu_{\mathrm{post}}$. Formally, we define a finite, strictly increasing sequence of inverse

[a] The Boltzmann distribution is a discrete probability measure on the set of energy states S of some system of particles. Its #-density is proportional to

$$S \ni s \mapsto \exp\left(-\frac{E_s}{\mathcal{T} \cdot k_{\mathrm{Boltz}}}\right),$$

where E_s is the energy of state s and $k_{\mathrm{Boltz}} = 1.380649\text{E-}23$ J/K is the Boltzmann constant. A large temperature $\mathcal{T}$ allows the particles to move faster. See [100, Chapter VIII], [167, §1.1] and [183] for details.

temperatures $(\beta_k : k = 0, \dots, N_{\mathrm{T}})$, where $\beta_0 = 0$ and $\beta_{N_{\mathrm{T}}} = 1$. The SMC sequence of probability measures $(\mu_k : k = 0, \dots, N_{\mathrm{T}})$ is then given by

$$\frac{\mathrm{d}\mu_k}{\mathrm{d}\mu_0} \propto \overline{\gamma}_k := \exp(-\beta_k \Phi), \qquad (k = 1, \dots, N_T). \tag{5.12}$$

Remark 5.15. In §2.2, we consider inverse problems with additive Gaussian noise on a finite-dimensional data space. In this setting, the tempering can be understood as an upscaling of the observational noise. Note that in this case

$$\overline{\gamma}_k = \exp\left(-\tfrac{1}{2}\|(\beta_k^{-1}\Gamma)^{-\frac{1}{2}}(y^\dagger - \mathcal{G})\|_Y^2\right) \qquad (k = 1, \dots, N_T).$$

The last term on the right-hand side tells us that the measure

$$\mu_k = \mathbb{P}(\theta \in \cdot \,|\mathcal{G}(\theta) + \beta_k^{-1/2}\eta = y) \qquad (k = 1, \dots, N_T).$$

Hence, an upscaling of the temperature $\mathcal{T} = \beta_1^{-1}, \dots, \beta_{N_{\mathrm{T}}}^{-1}$ is equivalent to an upscaling of the noise level in BIPs. Moreover, $\beta_0 = 0$ corresponds to an infinitely large noise level, where the likelihood does not contain any information. Hence, $\mu_0 = \mu_{\mathrm{prior}}$ is consistent. ◇

The densities $(\overline{\gamma}_k : k = 1, \dots, N_{\mathrm{T}})$ are strictly positive. Hence, the intermediate densities $(\gamma_k : k = 1, \dots, N_{\mathrm{T}})$ in the SMC sampler (see 5.1) are given by

$$\frac{\mathrm{d}\mu_k}{\mathrm{d}\mu_{k-1}} \propto \gamma_k = \frac{\overline{\gamma}_k}{\overline{\gamma}_{k-1}} = \exp(-(\beta_k - \beta_{k-1})\Phi).$$

We refer to this method as either *SMC with Tempering* or simply *single-level SMC*. SMC with tempering has been used to solve Bayesian inverse problems in [20, 134, 138]. The idea of using different temperatures combined with particle filters has also been discussed by [62, 190]. Moreover, it is the basic idea of the Ensemble Kalman Inversion [25, 147, 226, 227]. Here, the Ensemble Kalman Filter is used to approximate the sequence defined in (5.12). Tempering is also used in Markov chain Monte Carlo, e.g., to overcome multimodalities in probability measures. See, e.g., [80] for a review on the *parallel tempering* MCMC technique.

5.4.2 Standard Bridging

Bridging (B) is an SMC type method, where the sequence of probability measures represents a smooth transition from one probability measure μ to another probability measure μ^*. We assume that both these probability measures are defined on a common measurable space $(X, \mathcal{B}X)$ and that they are equivalent, i.e. $\mu \ll \mu^*$ and $\mu^* \ll \mu$. Moreover, we assume that μ and μ^* are absolutely continuous with respect to a σ-finite measure ν_X on $(X, \mathcal{B}X)$. Then, the Radon–Nikodym Theorem tells us that $\mathrm{d}\mu/\mathrm{d}\nu_X$ and $\mathrm{d}\mu^*/\mathrm{d}\nu_X$ exist and are unique ν_X-a.e. Moreover, these densities are strictly positive a.e. on the support of μ and μ^*.
Now, let μ and μ^* be based on functions $f, f^* : X \to \mathbb{R}$ which are proportional to the Radon–Nikodym derivatives given above. That is,

$$f \propto \frac{\mathrm{d}\mu}{\mathrm{d}\nu_X} \quad \text{and} \quad f^* \propto \frac{\mathrm{d}\mu^*}{\mathrm{d}\nu_X}.$$

Let $N_{\mathrm{B}} \in \mathbb{N}$ and $(\zeta_k : k = 0, \ldots, N_{\mathrm{B}}) \in [0,1]^{(N_{\mathrm{B}}+1)}$ be a strictly increasing finite sequence, where $\zeta_0 = 0$ and $\zeta_{N_{\mathrm{B}}} = 1$. Then, the bridging sequence of measures $(\mu_k : k = 0, \ldots, N_{\mathrm{B}})$ is defined as

$$\frac{\mathrm{d}\mu_k}{\mathrm{d}\nu_X} \propto \overline{\gamma}_k := f^{(1-\zeta_k)} \cdot (f^*)^{\zeta_k},$$

or, equivalently,

$$\frac{\mathrm{d}\mu_k}{\mathrm{d}\mu_{k-1}} \propto \gamma_k := \frac{\overline{\gamma}_k}{\overline{\gamma}_{k-1}} = f^{(\zeta_{k-1}-\zeta_k)} \cdot (f^*)^{(\zeta_k-\zeta_{k-1})}. \tag{5.13}$$

Note that $\mu_0 = \mu$ and $\mu_{N_{\mathrm{B}}} = \mu^*$.
We will later use bridging to exchange one likelihood in a given posterior for another. Then, f, f^* are two likelihoods, $\nu_X := \mu_{\mathrm{prior}}$, and μ, μ^* are posterior measures. This will be necessary in multilevel settings, when having likelihoods that are based on different model discretisations; see §3.4.2. We discuss this in detail in §6.1.1. Aside from this multilevel setting, bridging has been discussed e.g. by Del Moral et al. [62] and Gelman and Meng [93]. The method that approximates the sequence (5.13) with SMC will be called *SMC with bridging*.

5.4.3 Adaptive Sequential Monte Carlo

The accuracy and computational cost of SMC with tempering and bridging depends crucially on the number of intermediate probability measures $N_{\mathrm{seq}} \in \{N_{\mathrm{B}}, N_{\mathrm{T}}\}$, respectively, and the choice of the inverse temperatures $(\beta_k : k = 0, \ldots, N_{\mathrm{seq}})$. Up to now we assumed that N_{seq} and $(\beta_k : k = 0, \ldots, N_{\mathrm{seq}})$ are given a priori. However, we can also determine the inverse temperatures and associated intermediate probability measures adaptively "on the fly". In the literature, several strategies for adapting the inverse temperatures are known. We review methods based on the coefficient of variation of the update weights.
To simplify the notation we drop the subscript k and consider the SMC update $\mu \mapsto \mu^*$ in the remainder of this section. Let γ^* denote the density of μ^* with respect to μ. The probability measures μ and μ^* are approximated by $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\mu}}^*$, respectively, and are based on N_{smp} particles each. Note that we can define the ESS for the SMC update step by

$$\widehat{\mathrm{ESS}} := \frac{N_{\mathrm{smp}}}{1 + \mathrm{cv}^2_{\widehat{\boldsymbol{\mu}}}(\gamma^*)}, \tag{5.14}$$

where

$$\mathrm{cv}_{\widehat{\boldsymbol{\mu}}}(\gamma^*) := \frac{\mathrm{StD}_{\widehat{\mu}}(\gamma^*)}{\mathbb{E}_{\widehat{\boldsymbol{\mu}}}[\gamma^*]} := \frac{\sqrt{\mathrm{Var}_{\widehat{\boldsymbol{\mu}}}(\gamma^*)}}{\mathbb{E}_{\widehat{\boldsymbol{\mu}}}[\gamma^*]}$$

is the *coefficient of variation* of γ^*. This definition indeed coincides with the definitions of the ESS in (3.8) and (5.5).
Now, the inverse temperature β associated with μ is known. Our task is to define the inverse temperature β^* associated with μ^*. Clearly, the density of μ^* with respect

to μ depends on β^*. For this reason we write $\gamma^* = \gamma^*(\beta^*)$. Then, the ESS of the SMC update also depends on β^* and we write

$$\widehat{\mathrm{ESS}}(\beta^*) := \frac{N_{\mathrm{smp}}}{1 + \mathrm{cv}^2_{\widehat{\boldsymbol{\mu}}}\left(\gamma^*(\beta^*)\right)}.$$

We use this ESS again as an estimate of the Importance Sampling error; see Proposition 3.5. Similarly to the SMC method with adaptive resampling, we now aim to retain a certain effective sample size $\tau_{\mathrm{ESS}} \in (0, N_{\mathrm{smp}})$.
Note that $\widehat{\mathrm{ESS}}(\beta^*)$ can be computed without further evaluations of the (potentially expensive) likelihood for any $\beta^* \in (\beta, 1]$. In our implementation, we choose β^* such that $\widehat{\mathrm{ESS}}(\beta^*)$ is equal to a target value τ_{ESS}. In practice, we would like to avoid inverse temperature choices that meet the target ESS, that is, $\widehat{\mathrm{ESS}}(\beta^*) = \tau_{\mathrm{ESS}}$, but do not increase the inverse temperature by at least some $\varepsilon = \beta^* - \beta > 0$. Thus, we define

$$\beta^* := \operatorname*{argmin}_{\beta' \in [\min\{\beta+\varepsilon,1\},1]} \left(\widehat{\mathrm{ESS}}(\beta') - \tau_{\mathrm{ESS}}\right)^2. \tag{5.15}$$

Note that the optimisation problem in (5.15) is equivalent to the following problem, which is also common in the literature:

$$\beta^* = \operatorname*{argmin}_{\beta' \in [\min\{\beta+\varepsilon,1\},1]} \left(\mathrm{cv}_{\widehat{\mu}}\left(\gamma^*(\beta')\right) - \tau^*\right)^2, \tag{5.16}$$

where $\tau^* := \sqrt{(N_{\mathrm{smp}} - \tau_{\mathrm{ESS}})/\tau_{\mathrm{ESS}}}$; see e.g. [196, §2.3]. Hence the fitting of the effective sample size is equivalent to a fitting of the coefficient of variation of the weights.
Note that both the optimisation problems (5.15) and (5.16) have a continuous target functional, see Lemma 5.14, and a compact domain. Hence, both optimisation problems have a solution. However, the solution may be not unique. We refer to [20, 132, 138] for discussions and applications of adaptive Sequential Monte Carlo. A recent analysis of adaptive SMC has been given by Beskos et al. [17].

Chapter 6

Exploiting hierarchies with SMC samplers

> In manchen Fällen werden Größen miteinander kombiniert, deren Fehler den Charakter zufälliger Variabler haben.
>
> Johann Pfanzagl [202, p. 215]

Already in §5.4.1, we have used SMC samplers to exploit hierarchies, i.e. introduced a hierarchy of measures to potentially speed-up the sampling from a posterior measure. There, the hierarchy was given by different temperatures in the likelihood. In this chapter, we are going to exploit a hierarchy of model discretisations, as in Multilevel Monte Carlo; see §3.4.2. We discuss SMC samplers that use multiple discretisation levels in §6.1. In practice, these methods can be actually inefficient since inaccurate model evaluations can lead to large errors in the posteriors. To solve this problem, we introduce the *Multilevel Sequential*^{2} *Monte Carlo* (MLS^{2}MC) algorithm in §6.2. This is a method that combines SMC with tempering and multilevel Bridging in an adaptive way. Finally, we present numerical result in §6.3, comparing SMC, MLS^{2}MC and Multilevel Bridging.

6.1 Sequential multilevel methods

In the following, we review two SMC-based multilevel methods that have been used in the past: Multilevel Bridging and Multilevel SMC.

6.1.1 Multilevel Bridging

We have discussed standard bridging in §5.4.2. It is possible to generalize the idea of standard bridging to a setting where the probability measures μ and μ^* depend on discretisation parameters ℓ, ℓ^*. In BIPs this is, e.g. the case if the forward response operator $\mathcal{G}$, respectively the likelihood, is discretised using two different mesh sizes that correspond to ℓ and ℓ^*. Here, $\ell^* > \ell$, hence ℓ^* refers to a more accurate model discretisation than ℓ. In particular, we have a BIP, with prior μ_{prior} and likelihood $L = \exp(-\Phi)$. The potential Φ contains a mathematical model that has to be discretised. We define the potentials that are based on discretised models by Φ_ℓ and Φ_{ℓ^*}. We assume that the likelihoods $L^{(\ell)} = \exp(-\Phi_\ell)$ and $L^{(\ell^*)} = \exp(-\Phi_{\ell^*})$ also imply well-posed BIPs.

Suppose now that the BIP has been solved on a coarse mesh ℓ, i.e. with respect to $L^{(\ell)}$, and that we wish to obtain a more accurate solution with $\ell^* > \ell$. This means that the posterior $\mu_{\text{post}}^{(\ell)}$ shall be refined to $\mu_{\text{post}}^{(\ell^*)}$. Koutsourelakis [146] proposes to bridge between the two probability measures, that is, apply standard bridging for $\mu = \mu_{\text{post}}^{(\ell)}$ and $\mu^* = \mu_{\text{post}}^{(\ell^*)}$. In fact, this idea is carried out in a multilevel way by bridging between a hierarchy of probability measures associated with a sequence of decreasing mesh sizes. For this reason, we refer to the method as *Multilevel Bridging* (MLB). Indeed, we define MLB as a combination of [146] and SMC with tempering from §5.4.1 on the coarsest discretisation level. See the following paragraph for an introduction of MLB.
Let $\ell = 1, \dots, N_{\text{L}}$ denote the hierarchy of discretisation levels, where N_{L} is the desired final level and $1, \dots, N_{\text{L}} - 1$ are the intermediate levels, increasing in accuracy. Starting with the prior, we first use tempering to compute the posterior measure $\mu_{\text{post}}^{(1)}$ associated with the potential Φ_1. This step is based on the following densities:

$$\frac{\mathrm{d}\mu_k^{\text{T}}}{\mathrm{d}\mu_{\text{prior}}} \propto \gamma_k^{\text{T}} := \exp(-\beta_k \Phi_1),$$

where $\mu_0^{\text{T}} := \mu_{\text{prior}}$ is the prior measure and $(\beta_k : k = 0, \dots, N_{\text{T}})$ is the vector of inverse temperatures. Then, we proceed iteratively by bridging $\mu_{\text{post}}^{(\ell-1)} \mapsto \mu_{\text{post}}^{(\ell)}$ for each $\ell = 2, \dots, N_{\text{L}}$. In every bridging update, we use $N_{\text{B}}^{(\ell)}$ intermediate steps based on the (bridging) inverse temperatures $(\zeta_k^{(\ell)} : k = 0, \dots, N_{\text{B}}^{(\ell)})$. In particular,

$$\frac{\mathrm{d}\mu_{\ell,k}^{\text{B}}}{\mathrm{d}\mu_{\text{prior}}} \propto \gamma_{\ell,k}^{\text{B}} := \exp\left(-\zeta_k^{(\ell)} \Phi_\ell - (1 - \zeta_k^{(\ell)}) \Phi_{\ell-1}\right) \qquad (\ell = 2, \dots, N_{\text{L}}),$$

where $\mu_{\ell+1,0}^{\text{B}} := \mu_{\ell,N_{\text{B}}^{(\ell)}}^{\text{B}}$ and $\mu_{2,0}^{\text{B}} := \mu_{N_{\text{T}}}^{\text{T}}$.
As opposed to the multilevel methods discussed in §3.4.2, MLB is not a variance reduction method per se. Indeed, since the method uses Sequential Monte Carlo with tempering, the error of the estimator (i.e. ρ as given in Proposition 3.5) is indeed reduced; see Lemma 5.14. However, the multiple levels are actually used to speed-up the tempering process. Other multilevel methods that are not used for variance reduction are, for instance, the following: Farcas et al. [87] consider multiple discretisation levels to place sparse quadrature nodes in a domain; [197, 247] propose strategies to compute the probabilities of rare events; [198] use a multifidelity approach to find efficient MCMC proposals. An SMC method with classical Multilevel-Monte-Carlo type variance reduction is given in Multilevel Sequential Monte Carlo (MLSMC), which we discuss next.

6.1.2 Multilevel Sequential Monte Carlo

Beskos et al. [19, 18] and Del Moral et al. [64, 63] proposed and improved the Multilevel Sequential Monte Carlo method for Bayesian inverse problems. Here, the different discretisation levels are used for a variance reduction in the classical Multilevel Monte Carlo sense. The intermediate measures are given by the posteriors on different discretisation levels $\mu_1 := \mu_{\text{post}}^{(1)}, \mu_2 := \mu_{\text{post}}^{(2)}, \dots, \mu_{N_{\text{L}}} := \mu_{\text{post}}^{(N_{\text{L}})}$. Those measures are now approximated with the SMC framework. In this particular case, we keep all the measures in the memory, including the measures obtained after a

weighting step (W_k), but before the selection and mutation step (S_k). We obtain the initial measure $\widehat{\boldsymbol{\mu}}_0$ by sampling independently from the prior. Then, we proceed

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_k &= \mathrm{W}_k(\widehat{\boldsymbol{\mu}}_{k-1}),\\ \widehat{\boldsymbol{\mu}}_k &\sim \mathrm{S}_k(\cdot\,|\widetilde{\boldsymbol{\mu}}_k) \qquad (k = 1, \dots, N_{\mathrm{L}}).\end{aligned}$$

Let $Q : X \to \mathbb{R}$ be a bounded quantity of interest. We obtain an approximation of the posterior integral using the telescopic sum:

$$\int_X Q(\theta)\mu_{\mathrm{post}}^{(N_{\mathrm{L}})}(\mathrm{d}\theta) \approx \int_X Q(\theta)\widetilde{\boldsymbol{\mu}}_1(\mathrm{d}\theta) + \sum_{k=2}^{N_{\mathrm{L}}} \left(\int_X Q(\theta)\widetilde{\boldsymbol{\mu}}_k(\mathrm{d}\theta) - \int_X Q(\theta)\widehat{\boldsymbol{\mu}}_{k-1}(\mathrm{d}\theta) \right). \tag{6.1}$$

In particular cases, one can show that the estimator on the right-hand side of (6.1) has a smaller variance than the standard SMC estimator

$$\int_X Q(\theta)\widehat{\boldsymbol{\mu}}_{N_{\mathrm{L}}}(\mathrm{d}\theta).$$

Note that this integral as well as the integrals on the right-hand side of (6.1) are indeed estimators since they integrate with respect to empirical measures. Due to the variance reduction, and similarly to MLMC, one can successively reduce the number of samples. A natural point to do that in SMC is the resampling step, where one can just successively draw fewer samples. By using few samples with the computationally expensive fine model discretisation, the overall computational cost can be reduced significantly.
Note that this method is taylored for the approximation of integrals, but does not generally give a valid measure approximation. Hence, the MLSMC method follows a different paradigm than MLB. In the following, we will rather focus on the approximation of the measures, not on the approximation of integrals.

6.1.3 Issues of concentration and bias

In MLB and also MLSMC, we aim to *bridge* from one posterior measure $\mu_{\mathrm{post}}^{(\ell-1)}$ to another posterior measure $\mu_{\mathrm{post}}^{(\ell)}$. Hence, we apply Importance Sampling or SMC with tempering to obtain approximate samples from $\mu_{\mathrm{post}}^{(\ell)}$ using samples from $\mu_{\mathrm{post}}^{(\ell-1)}$. We now consider a setting in which the measures $(\mu_{\mathrm{post}}^{(\ell)})_{\ell=1}^{N_{\mathrm{L}}}$ are based on forward models with differently discretised PDEs. It has been observed that the posterior measures in $(\mu_{\mathrm{post}}^{(\ell)})_{\ell=1}^{N_{\mathrm{L}}}$ are biased with respect to each other, see, e.g. Cockayne et al. [46, Figure 3]. On the other hand, the concentration of the posterior depends for a large part on the noise level in the likelihood, not so much on the model discretisation. Hence, two consecutive posterior measures can have a very small variance each and a bias with respect to each other. This however may lead to a large ρ, in the Importance Sampling error; see Proposition 3.5. We consider this in the following example.

Example 6.1. Let $\mu := \mathrm{N}(0, \sigma^2)$ and $\mu^* := \mathrm{N}(m, \sigma^2)$ for some $m \in X$ and $\sigma^2 > 0$. Moreover, from Theorem 1.23, we obtain

$$\gamma(\theta) := \exp\left(\frac{\theta m}{\sigma^2}\right) \propto \frac{\mathrm{d}\mu^*}{\mathrm{d}\mu}(\theta) \qquad (\theta \in X).$$

We are now applying Importance Sampling to approximate μ^*, using μ and γ. Using $N_{\text{smp}} \in \mathbb{N}$ samples from μ, we get the error

$$\mathrm{d}_{\text{etv}}(\mu^*, \widehat{\boldsymbol{\mu}}^*) \leq 2 \cdot \sqrt{\frac{\rho}{N_{\text{smp}}}}.$$

In this particular setting, we obtain

$$\rho = \frac{\int_X \gamma(\theta)^2 \mu(\mathrm{d}\theta)}{\left(\int_X \gamma(\theta)\mu(\mathrm{d}\theta)\right)^2} = \exp\left(\frac{3m^2}{2\sigma^2}\right).$$

Let now μ, μ^* be posterior measures on different discretisation levels. As discussed above, we then anticipate a bias (i.e. $m \gg 0$) and a concentration on a small area (i.e. $\sigma \ll 1$). This implies that we obtain a large $m : \sigma$ ratio and, hence, $\rho \gg 1$. $\diamond$

A large ρ implies that we either need a huge number of samples, or, in SMC, a huge number of intermediate tempering steps. Both eventually leads to an inefficient algorithm. Hence, MLSMC and MLB may be inefficient in practice.

6.2 Multilevel Sequential^2 Monte Carlo

In this section we generalize Multilevel Bridging and propose the *Multilevel Sequential^2 Monte Carlo* (MLS^{2}MC) sampler. We explain the advantages of this generalisation in §6.2.3, but before we do this, we introduce the sampler formally in §6.2.1 and discuss its accuracy and computational cost in §6.2.2.
MLS^{2}MC is a Sequential Monte Carlo method which combines Tempering and Multilevel Bridging. Sequential^2 ("sequential squared") refers to two individual sequences in a Sequential Monte Carlo sampler, namely a sequence of inverse temperatures $(\beta_k : k = 0, \ldots, N_{\text{T}})$ and a sequence of discretisation levels $\ell = 1, \ldots, N_{\text{L}}$. Starting with the prior measure μ_{prior} and discretisation level $\ell = 1$, the MLS^{2}MC update either increases the discretisation resolution $\ell \mapsto \ell + 1$ or the inverse temperature $\beta_k \mapsto \beta_{k+1}$ $(k = 1, \ldots, N_{\text{L}} - 1)$. This process is repeated until we arrive at the inverse temperature $\beta_{N_{\text{T}}} = 1$ and maximal discretisation level N_{L}. See Figure 6.1 for an illustration.

6.2.1 Formal introduction

We introduce a general framework to describe MLS^{2}MC update strategies. Let $N_{\text{S}^2} = N_{\text{T}} + N_{\text{L}}$ denote the total number of bridging steps (i.e. the number of discretisation levels) and inverse temperature updates. Let $\mathcal{U} : \{0, \ldots, N_{\text{S}^2}\} \to \{0, \ldots, N_{\text{T}}\} \times \{1, \ldots, N_{\text{L}}\}$ denote a function, where

$$\mathcal{U}_i(s) = \mathcal{U}_i(s-1) \Leftrightarrow \mathcal{U}_j(s) = \mathcal{U}_j(s-1) + 1, \quad (i, j = 1, 2, i \neq j) \tag{6.2}$$

$$\mathcal{U}(0) = (0, 1), \tag{6.3}$$

$$\mathcal{U}(N_{\text{S}^2}) = (N_{\text{T}}, N_{\text{L}}). \tag{6.4}$$

We refer to $\mathcal{U}$ as *update scheme*. In each step $s = 0, \ldots, N_{\text{S}^2}$ of the algorithm $\mathcal{U}_1(s) = k$ refers to the inverse temperature and $\mathcal{U}_2(s) = \ell$ refers to the discretisation level. The update function $\mathcal{U}$ is convenient for the discussion and analysis of various update

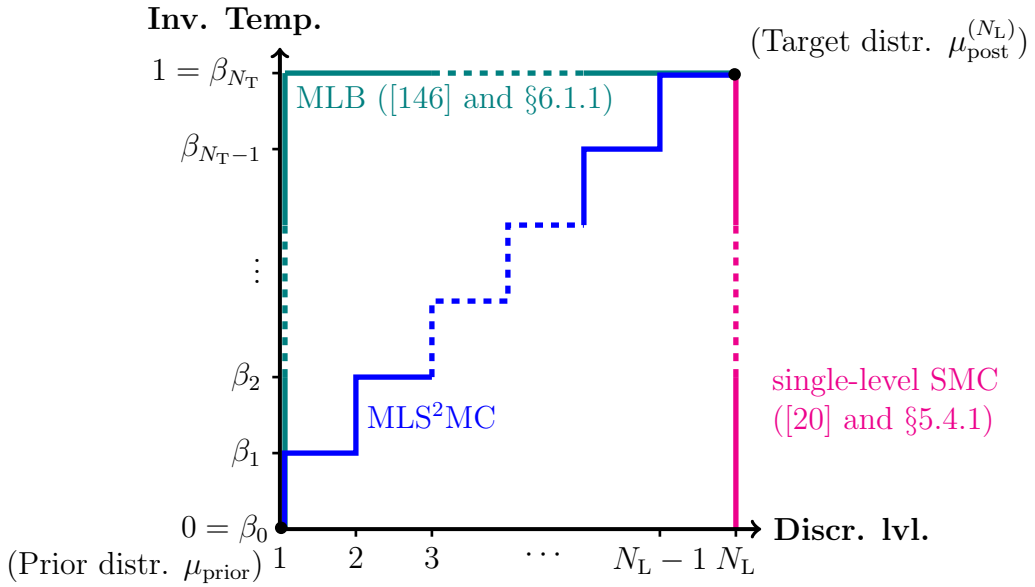


Figure 6.1. The update schemes associated with Multilevel Bridging, single-level SMC, and MLS^{2}MC.

schemes. If we consider only a single update scheme $\mathcal{U}$, we define $\mathcal{U}_1(s) =: \mathrm{T}(s)$ and $\mathcal{U}_2(s) =: \mathrm{B}(s)$. Furthermore, if it is clear whether s refers to $\mathrm{T}(s)$ or $\mathrm{B}(s)$ or to both, then we use the notation

$$\Phi^{(s)} := \Phi_{\mathrm{B}(s)}, \qquad \beta_s := \beta_{\mathrm{T}(s)}, \quad (s = 0, \ldots, N_{\mathrm{S}^2}).$$

Before we present the formal definition of MLS^{2}MC, we give two examples for alternative update schemes. See Figure 6.1 for an illustration.

Example 6.2. Let $N_{\mathrm{L}} = 1$ and define the update scheme $\mathcal{U}$, where $s \mapsto \mathcal{U}(s) := (s, 1)$. Then, the associated sampler is equivalent to single-level SMC. ◇

Example 6.3. Let $\mathcal{U} : \{0, \ldots, N_{\mathrm{S}^2}\} \to \{0, \ldots, N_{\mathrm{T}}\} \times \{1, \ldots, N_{\mathrm{L}}\}$, where

$$s \mapsto \begin{cases} (s, 1), & \text{if } s \leq N_{\mathrm{T}}, \\ (N_{\mathrm{T}}, s - N_{\mathrm{T}} + 1), & \text{otherwise.} \end{cases}$$

The corresponding sampler is equivalent to Multilevel Bridging. ◇

Now, we define MLS^{2}MC as a Sequential Monte Carlo sampler; see §5. Hence, we construct a sequence of probability measures $(\mu_{\mathcal{U}(s)} : s = 0, \ldots, N_{\mathrm{S}^2})$, where $\mu_{\mathcal{U}(0)} = \mu_{\text{prior}}$ and $\mu_{\mathcal{U}(N_{\mathrm{S}^2})} = \mu_{\text{post}}^{(N_{\mathrm{L}})}$. The intermediate probability measures are based on the update scheme u and are given once again by the Radon–Nikodym Theorem:

$$\frac{\mathrm{d}\mu_{\mathcal{U}(s)}}{\mathrm{d}\mu_{\text{prior}}}(\theta) \propto \exp\left(-\beta_s \Phi^{(s)}(\theta)\right) \quad (s = 1, \ldots, N_{\mathrm{S}^2},\ \theta \in X).$$

In the MLS^{2}MC sampler we distinguish two update types. Let $s = 1, \ldots, N_{\mathrm{S}^2}$. If $\mathrm{T}(s) = \mathrm{T}(s-1) + 1$, then

$$\frac{\mathrm{d}\mu_{\mathcal{U}(s)}}{\mathrm{d}\mu_{\mathcal{U}(s-1)}}(\theta) \propto \exp\left(-(\beta_s - \beta_{s-1})\Phi^{(s)}(\theta)\right) \quad (s = 1, \ldots, N_{\mathrm{S}^2}, \theta \in X).$$

We refer to this update as *inverse temperature update (ITU)*. Otherwise, if $\mathrm{B}(s) = \mathrm{B}(s-1)+1$, then the update is given by

$$\frac{\mathrm{d}\mu_{\mathcal{U}(s)}}{\mathrm{d}\mu_{\mathcal{U}(s-1)}}(\theta) \propto \exp\left(-\beta_s(\Phi^{(s)}(\theta) - \Phi^{(s-1)}(\theta))\right) \quad (s = 1, \ldots, N_{\mathrm{S}^2}, \theta \in X).$$

We refer to this update as *level update (LU)*. However, we usually perform more than one Bridging step from one discretisation level to the next (see §5.4.2). We can redefine the update by the following (telescoping) product of $N_{\mathrm{B}}^{(\mathrm{B}(s))} =: N_{\mathrm{B}}^{(s)} \in \mathbb{N}$ densities, each of which reflects a particular intermediate bridging measure that is based on bridging inverse temperatures $(\zeta_m^{(s)} : m = 1, \ldots, N_{\mathrm{B}}^{(s)})$:

$$\frac{\mathrm{d}\mu_{\mathcal{U}(s)}}{\mathrm{d}\mu_{\mathcal{U}(s-1)}}(\theta) \propto \prod_{m=1}^{N_{\mathrm{B}}^{(s)}} \exp\left(-\beta_s(\zeta_m^{(s)} - \zeta_{m-1}^{(s)})(\Phi^{(s)}(\theta) - \Phi^{(s-1)}(\theta))\right) \quad (s = 1, \ldots, N_{\mathrm{S}^2}, \theta \in X).$$

For clarity of presentation we do not include the intermediate bridging measures in the update scheme $\mathcal{U}$. Furthermore, if it is clear which update scheme is used, we write $\mu_s := \mu_{\mathcal{U}(s)}$.

6.2.2 Computational cost and accuracy

Before we propose an efficient update scheme for the MLS^{2}MC sampler, we briefly discuss its computational cost and accuracy. Let $\mathrm{C}_\ell \in (0, \infty)$ denote the computational cost of one evaluation of Φ_ℓ (for $\ell = 1, \ldots, N_{\mathrm{L}}$). Moreover, we denote the total cost of the MLS^{2}MC sampler with associated update scheme $\mathcal{U}$ by $\mathrm{Cost}(\mathcal{U})$. We typically measure C_ℓ in terms of the number of floating point operations that are required to evaluate Φ_ℓ. One could also think of estimating the elapsed time of model evaluations or, e.g. the memory requirement.

Example 6.4. Let Φ denote a potential that is based on the solution operator G of an elliptic boundary value problem in d-dimensional space ($d = 1, 2, 3$). Furthermore, let $h_\ell = 2^{-\ell} h_0, \ell \in \mathbb{N}, h_0 > 0$, denote the mesh size of the discretised potential Φ_ℓ. The mesh is regular and we use linear finite elements; see also Example 3.14, for a discussion of the computational cost. Then the ratio of the computational cost associated with two consecutive levels in terms of floating point operations is

$$\frac{\mathrm{C}_{\ell+1}}{\mathrm{C}_\ell} = 2^d, \quad \ell \in \mathbb{N}.$$

Given a maximal level $N_{\mathrm{L}} \in \mathbb{N}$, we normalize the values such that $\mathrm{C}_{N_{\mathrm{L}}} = 1$. We arrive at

$$\mathrm{C}_\ell := 2^{d(\ell - N_{\mathrm{L}})}, \quad \ell = 1, \ldots, N_{\mathrm{L}}.$$

$\diamond$

In the following we discuss the computational cost of MLS^{2}MC in terms of the update scheme $\mathcal{U}$ and the costs $(\mathrm{C}_\ell : \ell = 1, \ldots, N_{\mathrm{L}})$. To begin, we consider inverse temperature updates. If the Markov kernel update is performed by a Metropolis–Hastings scheme, then one PDE solve for each of the N_{smp} particles is required, to

evaluate the acceptance probability. The acceptance step also requires the model evaluations of the current particles. This however should remain in the memory, until the particles are updated. Hence, the computational cost of the inverse temperature updates is given by

$$\sum_{\substack{s=1 \\ s \text{ is an ITU}}}^{N_{\mathrm{S}^2}} N_{\mathrm{smp}} \mathrm{C}_{\mathrm{B}(s)}.$$

In Bridging, we also perform a Markov kernel step for each of the $N_{\mathrm{B}}^{(s)}$ intermediate Bridging steps and each of the N_{smp} particles. Here, the evaluation of the Markov update density requires two model evaluations in total, namely one on each discretisation level $\mathrm{B}(s-1)$ and $\mathrm{B}(s)$, respectively. Thus, we perform $N_{\mathrm{B}}^{(s)} \cdot N_{\mathrm{smp}}$ evaluations of the model on the two levels. In addition, we have to consider the first intermediate Bridging step. In contrast to the inverse temperature update, we do not yet have the model evaluation of the current particles on level $\mathrm{B}(s)$. Thus, we need to add the cost of $N_{\mathrm{smp}} \cdot \mathrm{C}_{\mathrm{B}(s)}$ to each of the level updates. In summary, the computational cost for a level update is given by

$$\sum_{\substack{s=1 \\ s \text{ is an LU}}}^{N_{\mathrm{S}^2}} N_{\mathrm{smp}} \left(\mathrm{C}_{\mathrm{B}(s)} + (N_{\mathrm{B}}^{(s)})(\mathrm{C}_{\mathrm{B}(s)} + \mathrm{C}_{\mathrm{B}(s-1)}) \right).$$

Adding the costs for bridging and inverse temperature updates, respectively, we arrive at the following total cost.

Proposition 6.5. Let the Markov kernels in the MLS^{2}MC sampler be given in terms of a single Metropolis–Hastings MCMC update. Then, the total computational cost of the Multilevel Sequential^2 Monte Carlo sampler is given by

$$\mathrm{Cost}(\mathcal{U}) = \sum_{\substack{s=1 \\ s \text{ is an ITU}}}^{N_{\mathrm{S}^2}} N_{\mathrm{smp}} \mathrm{C}_{\mathrm{B}(s)} + \sum_{\substack{s=1 \\ s \text{ is an LU}}}^{N_{\mathrm{S}^2}} N_{\mathrm{smp}} \left(\mathrm{C}_{\mathrm{B}(s)} + (N_{\mathrm{B}}^{(s)})(\mathrm{C}_{\mathrm{B}(s)} + \mathrm{C}_{\mathrm{B}(s-1)}) \right).$$

◇

Next we discuss the accuracy of the MLS^{2}MC sampler in terms of $\mathrm{d}_{\mathrm{etv}}(\mu_{\mathrm{post}}, \widehat{\boldsymbol{\mu}})$, where $\widehat{\boldsymbol{\mu}}$ is the measure based on the MLS^{2}MC approximation of μ_{post}. Note that $\widehat{\boldsymbol{\mu}}$ is a random measure, as discussed in §5.3.1. An accurate discussion of the SMC error is beyond the scope of this thesis. We refer to the references noted in §5.2.2. We make use of the following observation: In every MLS^{2}MC update we perform a Monte Carlo estimation with weighted samples. Hence, in each update the approximation accuracy measured in terms of the root mean square error is of order

$$O(\mathrm{ESS}^{-1/2}; \mathrm{ESS} \to \infty).$$

Here, ESS is the effective sample size defined in (3.8). Recall that we choose the update steps adaptively (see §5.4.3). Thus, the ESS is approximately constant in every step. Hence, every Bridging and Tempering step has the same influence on the accuracy. Thus, the total accuracy is bounded by the sum of the individual accuracies associated with the update steps. For this reason, we can maximise

the accuracy of the MLS^{2}MC approximation by minimising the total number of MLS^{2}MC update steps. The latter is given by

$$\#\mathrm{Upd}(\mathcal{U}) = N_{\mathrm{T}} + \sum_{\substack{s=1 \\ s \text{ is an LU}}}^{N_{\mathrm{S}^2}} N_{\mathrm{B}}^{(s)}.$$

In summary, we wish to design an update scheme which minimises both $\#\mathrm{Upd}(\cdot)$ and $\mathrm{Cost}(\cdot)$.

6.2.3 Is Multilevel Bridging optimal?

We next discuss the computational cost of Multilevel Bridging; see §6.1.1 for details. We do this to motivate our generalisation, the MLS^{2}MC sampler. First, we state two assumptions on the inverse temperatures and number of intermediate bridging steps.

Assumption 6.6. In the MLS^{2}MC sampler,

(i) the inverse temperature $\beta_{\mathrm{T}(s)}$ is independent of the discretisation level $\mathrm{B}(s-1)$ for any $s = 1, \dots, N_{\mathrm{S}^2}$, where s refers to an ITU, and

(ii) the number of intermediate bridging steps $N_{\mathrm{B}}^{(s)}$ is independent of the inverse temperature $\beta_{\mathrm{T}(s-1)}$ for any $s = 1, \dots, N_{\mathrm{S}^2}$, where s refers to an LU.

◇

Given these assumptions, $\#\mathrm{Upd}(\mathcal{U})$ is constant for every possible update scheme $\mathcal{U}$. Hence, we expect the same accuracy for any MLS^{2}MC sampler independently of $\mathcal{U}$. One can argue analogously for the cost of the bridging steps: Due to the Assumption 6.6(ii) the number of Bridging steps is fixed throughout all feasible update schemes. Thus, the crucial factor contributing to the total cost is the tempering. In MLB, the tempering is performed completely on level $\ell = 1$ which requires the least computational effort. We summarize this paragraph in the following proposition.

Proposition 6.7. Let $\mathcal{U}$ be the Multilevel Bridging update scheme defined in Example 6.3. If Assumption 6.6 are satisfied, then $\mathcal{U}$ minimizes both $\#\mathrm{Upd}(\cdot)$ and $\mathrm{Cost}(\cdot)$. ◇

We now comment on Assumption 6.6, starting with (i). The major reason for performing the tempering is the concentrated support of the posterior in the small noise limit. The width of this concentrated support is associated with the posterior variance which, in turn, reflects the certainty in the considered parameter. This certainty in the parameter is based on the precision Γ^{-1} of the data which we define a priori in the likelihood. Since Γ^{-1} is chosen independently of the discretisation resolution h, Assumption 6.6(i) is likely satisfied.
In contrast, Assumption 6.6(ii) is not always justified. If $\Phi_{\ell-1}$ is a good approximation to Φ_ℓ, then also $\exp(-\beta\Phi_{\ell-1}) \approx \exp(-\beta\Phi_\ell)$, independently of the inverse temperature β. Hence, the support of the associated posterior measures differs only in a small area of the parameter space, and a small number of intermediate bridging steps is required from $\ell - 1 \to \ell$. However, on two consecutive coarse discretisation

levels the discrepancy of $\Phi_{\ell-1}$ and Φ_ℓ might be large. This is not a big problem for small inverse temperatures associated with a larger noise level in the likelihood; it is still possible that there is a substantial overlap of the support of the associated probability measures, and hence a moderate number of intermediate bridging steps is required. However, for large inverse temperatures and a small noise level the associated probability measures are likely highly concentrated, and their supports might have a small intersection. Thus, either a large number of intermediate bridging steps is required, or bridging might not be possible at all. We have discussed this issue also in §6.1.3. In any case, Assumption 6.6(ii) is hardly justified. Note that a large number of intermediate bridging steps also reduces the overall accuracy of the SMC sampler.

When is bridging practically impossible? Let s refer to an LU. Given a fixed number of particles N_{smp}, it is possible that the update density $\mathrm{d}\mu_s/\mathrm{d}\mu_{s-1}$ is numerically zero for all particles. Then, we refer to μ_{s-1} and μ_s as *numerically singular*. Importantly, we are not able to carry out an MLS^{2}MC update from μ_{s-1} to μ_s in this case.

Now, in MLB all level updates are performed with the untempered likelihood, i.e. $\beta = 1$, even for coarse discretisations. As explained above, this might result in an inaccurate or expensive estimate, or the estimation might not be possible at all. In the numerical experiments in §6.3, we will illustrate these issues.

Of course, these problems can be cured by starting the MLB on a fine discretisation level, where Assumption 6.6(ii) is satisfied. If the operator $\Phi_{(\cdot)}$ is well understood, it might even be possible to define a suitable minimal starting level. In this case, the cost of MLB might often be cheaper than the cost of the adaptive update scheme that we propose in the next section. However, in most cases the operator $\Phi_{(\cdot)}$ is not well understood or even only given in a black box sense. In this case, determining a sufficiently fine starting level for MLB is not possible. This motivates us to introduce an efficient, parameter-free, adaptive update scheme which does not require a priori information on the model resolution.

6.2.4 An efficient update scheme

Now we discuss the major component in the proposed MLS^{2}MC sampler, namely, the choice of the inverse temperature and level updates, respectively. Balancing these updates with the computational cost is a non-trivial task. If we increase the discretisation level too early in the update scheme, then many inverse temperature updates on an expensive level are required. Increasing the discretisation level too late could result in the undesirable situation that many intermediate bridging steps might be required later on (see the discussions in §6.1.3 and §6.2.3). To simplify the derivation of the computational cost we work under Assumption 6.6(i) which is likely satisfied for a large class of relevant BIPs. In this case, to obtain a good accuracy of the MLS^{2}MC approximation, we aim at minimizing the number of bridging steps. However, we also need to consider the computational cost associated with the proposed path since MLS^{2}MC should operate with minimal cost.

Suppose we are in the update step from μ_{s-1} to μ_s, where $s \in \{2, \dots, N_{\text{S}^2}\}$ and $\mathcal{U}(s-1) =: (k-1, \ell-1)$. Under Assumption 6.6(i) we study the following *decision problem:*

Do we update the discretisation level $\ell - 1 \mapsto \ell$ or the temperature $k - 1 \mapsto k$?

To account for the full impact of this decision, we consider the cost and the loss

in accuracy of *all* future update steps. We split the future path into two parts, namely, from μ_{s-1} to μ_s and from μ_s to $\mu_{N_{S^2}}$. For simplification, we suppose that for the second part both Assumption 6.6(i) and (ii) are satisfied. By Proposition 6.7, the optimal strategy for the second part starting in s is to first increase the inverse temperature to $\beta = 1$ (in multiple steps) and then to bridge to the maximal level N_L. This is equivalent to carrying out the Multilevel Bridging with initial probability measure μ_s. Hence the second part of the path is determined, and we only need to decide on the path from μ_{s-1} to μ_s. We now investigate this.
Let $s_T := \min\{s \in \{1, \dots, N_{S^2}\} : \beta_s = 1\}$. In Figure 6.2, we show the two options:

- Path $\mathcal{W}$: Update the level $\ell - 1 \mapsto \ell$, in step $s - 1 \mapsto s$, then proceed as in MLB,
- Path $\mathcal{V}$: Update the inverse temperature $k - 1 \mapsto k$ in step $s - 1 \mapsto s$, then proceed as in MLB,

where $\mathcal{W}(s-1) = \mathcal{V}(s-1) = (k-1, \ell-1)$, $\mathcal{W}(s) = (k-1, \ell)$, and $\mathcal{V}(s) = (k, \ell-1)$. Note that s_T differs for the paths $\mathcal{V}$ and $\mathcal{W}$.

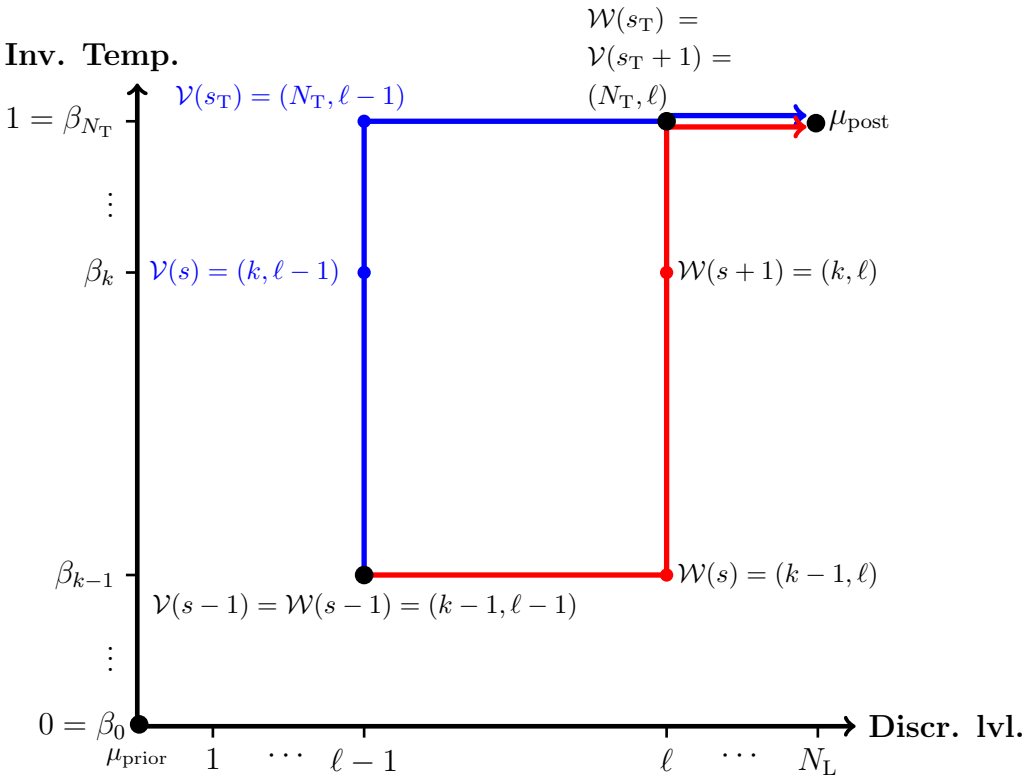


Figure 6.2. Decision problem in MLS^{2}MC: *Which path is cost-optimal?* First level, then inverse temperature update ($\mathcal{W}$, red) or first inverse temperature, then level update ($\mathcal{V}$, blue).

We assume that $N_B^{(s)} \leq N_B^{(s+1)}$. This is reasonable since the probability measures that are bridged in path $\mathcal{V}$ contain a smaller noise level and are thus more concentrated. The computational costs associated with paths $\mathcal{V}$ and $\mathcal{W}$ between μ_{s-1} and

$\mu_{(N_\mathrm{T},\ell)}$ are given by

$$\mathrm{Cost}\left(\mathcal{W}|_{\{s-1,\dots,s_\mathrm{T}\}}\right) = \underbrace{N_\mathrm{smp}N_\mathrm{B}^{(s)}(\mathrm{C}_{\ell-1}+\mathrm{C}_\ell)+N_\mathrm{smp}\mathrm{C}_{\ell-1}}_{\text{LU } \ell-1\mapsto\ell} + \underbrace{N_\mathrm{smp}(N_\mathrm{T}-k+1)\mathrm{C}_\ell}_{\text{ITU } k-1\mapsto N_\mathrm{T}}, \tag{6.5}$$

$$\mathrm{Cost}\left(\mathcal{V}|_{\{s-1,\dots,s_\mathrm{T}+1\}}\right) = \underbrace{N_\mathrm{smp}(N_\mathrm{T}-k+1)\mathrm{C}_{\ell-1}}_{\text{ITU } k-1\mapsto N_\mathrm{T}} + \underbrace{N_\mathrm{smp}N_\mathrm{B}^{(s+1)}(\mathrm{C}_{\ell-1}+\mathrm{C}_\ell)+N_\mathrm{smp}\mathrm{C}_{\ell-1}}_{\text{LU } \ell-1\mapsto\ell}. \tag{6.6}$$

Note that we do not consider the cost of Bridging from $\mu_{(N_\mathrm{T},\ell)}$ to $\mu_{(N_\mathrm{T},N_\mathrm{L})} = \mu_{N_{\mathrm{S}^2}}$ since this cost is identical for both paths. Given our assumptions we can reformulate the decision problem in terms of computational cost as follows:

Is the number of additional bridging steps $N_\mathrm{B}^{(s+1)}$ needed in comparison to $N_\mathrm{B}^{(s)}$ more expensive than the increased computational cost of the inverse temperature update on level ℓ compared to level $\ell-1$?

This question corresponds directly to the expressions in (6.5) and (6.6). However, we need to minimise both the computational cost and the number of updates #Upd. If bridging and tempering are performed non-adaptively, then all quantities in (6.5) and (6.6) are known, as are $\#\mathrm{Upd}(\mathcal{V})$ and $\#\mathrm{Upd}(\mathcal{W})$. Hence, we can simply compare the costs and the number of update steps and choose the path that is more appropriate. However, this is not the focus of this thesis. From now on we consider adaptive tempering and bridging only.
Without loss of generality we assume that $\tau^* > 0$ is the target value for the coefficient of variation of the weights in *every* tempering and bridging update. Unfortunately, there are in general no simple analytic expressions for the interdependency of τ^*, β_s and $N_\mathrm{B}^{(s)}$. Furthermore, given the probability measure μ_{s-1}, it is difficult to estimate how many intermediate bridging steps $N_\mathrm{B}^{(s)}$ are required for the bridging $\ell-1\mapsto\ell$. To make progress we continue as follows. We select a small proportion $N_{\widetilde{\mathrm{smp}}}$ of the N_smp samples and estimate the coefficient of variation associated with a bridging update using $N_\mathrm{B}^{(s)}=1$ steps based on these $N_{\widetilde{\mathrm{smp}}}$ samples. We obtain

$$\mathrm{cv}_{\mu_{s-1}}\left[\exp\left(-\beta_s(\Phi^{(s)}(\theta)-\Phi^{(s-1)}(\theta))\right)\right] =: \mathrm{cv}_s^{\mathrm{LU}}.$$

This estimation requires $N_{\widetilde{\mathrm{smp}}}$ additional evaluations of Φ_ℓ. If we update the discretisation level immediately afterwards, then these evaluations can be re-used for the bridging update. If this is not the case, then the additional samples are discarded. In §6.3, we consider various proportions $N_{\widetilde{\mathrm{smp}}}/N_\mathrm{smp}$.
To continue we make the following observation. If $\mathrm{cv}_s^{\mathrm{LU}} < \tau_{\mathrm{LU}}$, where $\tau_{\mathrm{LU}} \in (0,\tau^*]$, then the bridging can be performed with only one intermediate step. We use this observation as a measure of the accuracy of the approximation $\Phi_{\ell-1}\approx\Phi_\ell$.

- If the accuracy is small (i.e. $\mathrm{cv}_s^{\mathrm{LU}} > \tau_{\mathrm{LU}}$), then we bridge immediately, since we would otherwise propagate the inaccurate model to an inverse temperature that is unreasonably small.
- If the accuracy is high (i.e. $\mathrm{cv}_s^{\mathrm{LU}} < \tau_{\mathrm{LU}}$), we know that $N_\mathrm{B}^{(s)}=1$. Moreover, we define $N_\mathrm{B}^* := N_\mathrm{B}^{(s+1)}$ and $N_\mathrm{T}^* := N_\mathrm{T}-k+1$. Based on comparing the costs

in (6.5) and (6.6), we perform an inverse temperature update from $s-1 \mapsto s$ if the condition

$$N_{\text{smp}}(\mathrm{C}_{\ell-1}+\mathrm{C}_\ell)+N_{\text{smp}}N_{\mathrm{T}}^*\mathrm{C}_\ell \geq N_{\text{smp}}N_{\mathrm{T}}^*\mathrm{C}_{\ell-1}+N_{\text{smp}}N_{\mathrm{B}}^*(\mathrm{C}_{\ell-1}+\mathrm{C}_\ell) \quad (6.7)$$

is satisfied (since then the ITU cost is cheaper than the LU cost). If (6.7) is not satisfied, then we perform a level update.

Note that the condition in (6.7) is equivalent to

$$\frac{\mathrm{C}_\ell}{\mathrm{C}_{\ell-1}} \geq \frac{N_{\mathrm{T}}^*+N_{\mathrm{B}}^*-1}{N_{\mathrm{T}}^*-N_{\mathrm{B}}^*+1}, \quad (6.8)$$

where we define $\frac{1}{0}:=\infty$. We visualise the condition in (6.8) in Figure 6.3 where we show which combinations of N_{B}^* and N_{T}^* satisfy (6.8) for $\mathrm{C}_\ell/\mathrm{C}_{\ell-1} \in \{2,4,8\}$. These three cases refer to solves of elliptic PDEs in 1D, 2D, and 3D; see Example 6.4. We see that condition (6.7) in the 3D case implies $N_{\mathrm{B}}^*+1 \approx N_{\mathrm{T}}^*$.

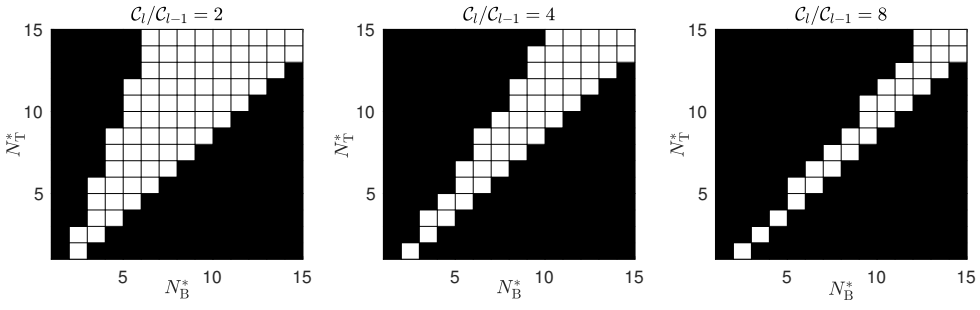


Figure 6.3. Visualisation of the combinations of N_{T}^* and N_{B}^* satisfying condition (6.8) (black squares).

Of course, the evaluation of (6.8) requires N_{T}^* and N_{B}^*. However, these quantities are (still) not known a priori, since bridging and tempering are performed adaptively. In practice, we do not necessarily need to know N_{T}^* and N_{B}^*. If the increase of the computational cost $\mathrm{C}_\ell/\mathrm{C}_{\ell-1}$ is large (as is the case in realistic applications), then most combinations of $(N_{\mathrm{B}}^*, N_{\mathrm{T}}^*)$ satisfy (6.8); see Figure 6.3. Thus, if $N_{\mathrm{B}}^{(s)}=1$, one might as well skip checking condition (6.8) and always perform an inverse temperature update. We follow this strategy from now on.

The noise in the BIP can be understood as a combination of observational noise and model (discretisation) error. This point of view fits very well with our MLS^{2}MC framework. Indeed, we reduce the noise level while increasing the accuracy of our model evaluation. See also the method presented in [32] for a further discussion of this interpretation.

Suppose now that in the update scheme the inverse temperature has not yet reached its maximum $\beta_{(\cdot)}=1$. Given the argument above, it is a good idea to increase the inverse temperature after every level update (non-adaptively). This reduces the total computational cost since we save $N_{\widetilde{\text{smp}}}$ model evaluations in situations where a level update is very unlikely. We implement both these ideas in our algorithm. That is, we do not check the condition (6.8), and we update the inverse temperature

automatically after every level update. This update scheme is given by the formula

$$\mathcal{U}(s) = \begin{cases} (\mathrm{T}(s-1)+1, \mathrm{B}(s-1)), & \text{if } B(s-1) = B(s-2)+1, \\ & \underline{\text{or}}\ \mathrm{cv}_s^{\mathrm{LU}} < \tau_{\mathrm{LU}}, \\ (\mathrm{T}(s-1), \mathrm{B}(s-1)+1), & \text{otherwise}, \end{cases} \tag{6.9}$$

for any $s = 1, \ldots, N_{\mathrm{S}^2}$, where $\mathcal{U}(0) := (0,1)$ and $\mathcal{U}(-1) := (0,0)$. Note that the update scheme in (6.9) is independent of $\mathrm{C}_\ell/\mathrm{C}_{\ell-1}$. Clearly, its effectiveness depends on the computational costs at each level, but the cost does not determine the adaptive choice between level update and inverse temperature update. Recalling Figure 6.3, we expect that the efficiency of the algorithm increases as the multiplicative increment $\mathrm{C}_\ell/\mathrm{C}_{\ell-1}$ increases. The algorithm reduces the total number of intermediate bridging steps N_{B}^*, and thus also increases the accuracy. The numerical results in §6.3 show that the update scheme in (6.9) implements a compromise between computational cost and accuracy.

In summary, our proposed update scheme uses a heuristic backtracking type method to find a suitable inverse temperature for the bridging from discretisation level $\ell - 1$ to ℓ. This means that we bridge on the smallest of the adaptively determined inverse temperatures where bridging is necessary.

Let $s - 1 \mapsto s$ refer to a level update. The approximation $\mu_{s-1} \approx \mu_s$ is accurate if ℓ is sufficiently large, or if β_k is sufficiently small. Hence, we deduce that our method leads to a small number of required intermediate bridging updates $N_{\mathrm{B}}^{(\ell)}$. This in turn implies a small total computational cost and a good accuracy of the measure approximation.

6.2.5 Maximum level $N_{\mathrm{L}}^{\max}$

A natural question in the context of the adaptive update scheme (6.9) is: *Do we need to go to the level $\ell = N_{\mathrm{L}}$ or can we stop earlier?* To address this question, we proceed as follows.

Let $F := \{s = 1, \ldots, N_{\mathrm{S}^2} : \mathrm{T}(s) = N_{\mathrm{T}}, \mathrm{B}(s) \leq N_{\mathrm{L}}\}$ denote the subset of the domain of the update scheme $\mathcal{U}$ where the maximal inverse temperature $\beta_{N_{\mathrm{T}}} = 1$ is reached and some Bridging steps remain. Note that F can be the empty set. If this is not the case, we refer to F as the set of *final level updates*. We reformulate the question above as follows. *Is there an $s \in F$, such that the intermediate probability measure μ_s is a sufficiently accurate approximation to the target posterior measure $\mu_{\mathrm{post}}^{(N_{\mathrm{L}})}$?*

We assess the necessity of updating the discretisation level in terms of the information gain associated with the update. If the information gain of the level update is smaller than a certain threshold, then the algorithm terminates. To be consistent with the update scheme (6.9), we measure the information gain in terms of $\mathrm{cv}_s^{\mathrm{LU}}$. This coefficient of variation gives an upper bound for the Kullback–Leibler divergence from μ_{s-1} to μ_s; see [5] for details. Let $\tau_{\min} > 0$, where $\tau_{\min} < \tau_{\mathrm{LU}}$, denote a threshold parameter. The modified update scheme $\mathcal{U}'$ reads as follows:

$$\mathcal{U}'(s) = \begin{cases} \mathcal{U}'(s-1) \text{ and terminate}, & \text{if } s-1 \in F \text{ and } \mathrm{cv}_s^{\mathrm{LU}} < \tau_{\min}, \\ (\mathrm{T}(s-1)+1, \mathrm{B}(s-1)), & \text{if } B(s-1) = B(s-2)+1, \\ & \text{or } \mathrm{cv}_s^{\mathrm{LU}} < \tau_{\mathrm{LU}}, \\ (\mathrm{T}(s-1), \mathrm{B}(s-1)+1), & \text{otherwise}, \end{cases} \tag{6.10}$$

for any $s = 1, \ldots, N_{\mathrm{S}^2}$, where $\mathcal{U}'(0) := (0,1)$ and $\mathcal{U}'(-1) := (0,0)$. If the algorithm terminates for $s < N_{\mathrm{S}^2}$, we define $N_{\mathrm{S}^2} := N_{\mathrm{T}} + N_{\mathrm{L}}^{\max}$ and $N_{\mathrm{L}}^{\max} := \mathrm{B}(s-1)$. Otherwise, we let $N_{\mathrm{L}}^{\max} := N_{\mathrm{L}}$. We test the performance of the modified update scheme $\mathcal{U}'$ in §6.3.

6.3 Numerical experiments

We consider the elliptic inverse problem discussed in Example 1.41 on the unit square domain $D = (0,1)^2$. Hence, the permeability $\exp(\theta)$ and the hydrostatic pressure p are coupled via the elliptic PDE

$$-\nabla \cdot (\exp(\theta(x)) \nabla u(x)) = f(x) \qquad (x \in D).$$

The source term f and the boundary conditions are specified below. We observe the pressure at N_{obs} points $(o_n : n = 1, \ldots, N_{\mathrm{obs}})$ in the domain D. Thus the observation operator $\mathcal{O}$ maps $u \mapsto (u(o_n) : n = 1, \ldots, N_{\mathrm{obs}})$.
The parameter $\exp(\boldsymbol{\theta})$ is a log-normal random field. In particular, we assume that the prior distribution of $\boldsymbol{\theta}$ is a Gaussian random field with mean and covariance operator specified below. This Gaussian random field is discretised by a truncated KL expansion, which takes the form

$$\boldsymbol{\theta} \approx \boldsymbol{\theta}_{\mathrm{KL}}^{N_{\mathrm{sto}}} := m_0(x) + \sum_{n=1}^{N_{\mathrm{sto}}} m_n(x) \boldsymbol{\theta}_n^{\mathrm{KL}}, \tag{6.11}$$

where $\boldsymbol{\theta}_1^{\mathrm{KL}}, \ldots, \boldsymbol{\theta}_{N_{\mathrm{sto}}}^{\mathrm{KL}}$ denote standard Gaussian random variables; see §3.2.1. We generate the true parameter by sampling from the discretised prior random field. The observations y are given by the model evaluation of the true parameter plus (additive) Gaussian measurement noise $\boldsymbol{\eta} \sim \mathrm{N}(0, 0.01 \cdot \mathrm{Id})$.
We consider three estimation problems.

Example 6.8. Here the pressure on the boundary of D is zero,

$$u(x) = 0 \qquad (x \in \partial D).$$

The source term f models nine smoothed point sources that are distributed uniformly over the domain:

$$f(x) = \sum_{n,m=1}^{3} \mathrm{n}\left(x_1; 0.25n, 0.001\right) \mathrm{n}\left(x_2; 0.25m, 0.001\right),$$

where $\mathrm{n}(\cdot; E, V)$ is the probability density function of the one-dimensional Gaussian measure with mean E and variance V; see also Proposition 1.22. The prior

random field $\boldsymbol{\theta} \sim \mu_{\text{prior}} = \mathrm{N}(m, \mathcal{C})$, where $m \equiv 0$ and $\mathcal{C}$ is the Matérn covariance operator with correlation length $\lambda = 0.65$, smoothness parameter $\nu = 1.5$, and variance $\sigma^2 = 1$; see Definition 4.6 for details. The random field $\boldsymbol{\theta}$ is discretised by a truncated KL expansion using the $N_{\text{sto}} := 10$ leading terms which capture 94.5% of the variance. Note that for the prior field these random variables are uncorrelated. For the posterior random field, this is not necessarily the case. However, we only consider the marginals of the posterior distribution. The action of the operator G is approximated by piecewise linear, triangular, continuous finite elements on uniform meshes with $2 \cdot 8^2, 2 \cdot 16^2, 2 \cdot 32^2, 2 \cdot 64^2$ and $2 \cdot 128^2$ elements. The observation operator $\mathcal{O}$ returns the pressure at 25 points in the spatial domain. The 25 points are shown in Figure 6.4 along with the actual pressure given the true underlying permeability. Finally, the covariance operator of the noise is given by the matrix $\Gamma = 0.07^2 \cdot \mathrm{Id}$. ◇

Example 6.9. The inverse problem and its discretisation is the same as in Example 6.8, but with noise covariance matrix $\Gamma = 0.035^2 \cdot \mathrm{Id}$. ◇

Example 6.10. We consider a flow cell problem on $D = (0,1)^2$; recall Example 4.19. We have flow in the x_1-direction and no-flow boundaries along the x_2-direction,

$$\begin{aligned} u(x) &= 0 && (x \in \{0\} \times [0,1]), \\ u(x) &= 1 && (x \in \{1\} \times [0,1]), \\ \frac{\partial u}{\partial \vec{n}}(x) &= 0 && (x \in (0,1) \times \{0,1\}). \end{aligned}$$

Furthermore, the source term $f \equiv 0$. The prior random field is $\boldsymbol{\theta} \sim \mu_{\text{prior}} = \mathrm{N}(m', \mathcal{C}')$, where $m' \equiv 2$ and $\mathcal{C}'$ is the Matérn covariance operator with correlation length $\lambda = 0.1$, smoothness parameter $\nu = 1.5$, and variance $\sigma^2 = 1$. The random field $\boldsymbol{\theta}$ is discretised by a truncated KL expansion of the form (6.11) using the leading $N_{\text{sto}} := 320$ terms which capture 95% of the variance. The action of the operator G is approximated by piecewise linear, triangular, continuous finite elements on uniform meshes with $2 \cdot 16^2, 2 \cdot 32^2, 2 \cdot 64^2, 2 \cdot 128^2$ and $2 \cdot 256^2$ elements. The measurement locations are uniformly distributed as in Example 6.8, however, we use 49 measurements (see Figure 6.4). ◇

In all examples the KL basis functions $(m_n)_{n=0}^{N_{\text{sto}}}$ can be chosen to be continuous. Moreover, the observational noise is finite dimensional and non-degenerate Gaussian. See Remark 2.11 for a discussion of the well-posedness of the BIPs in Examples 6.8–6.10.

In all examples we test the performance of single-level SMC on the finest mesh (from now on simply 'SMC') as well as MLB and MLS^{2}MC on the given mesh hierarchy. We observe that in Example 6.10 the adaptive update scheme of MLS^{2}MC is identical to the MLB update scheme. Interestingly, in Example 6.9 it is impossible to perform the update $\ell = 1$ to $\ell = 2$ with MLB since the probability measures $\mu_{\text{post}}^{(1)}$ and $\mu_{\text{post}}^{(2)}$ are numerically singular. We anticipated this situation in §6.2.3.

For each of the tests above, we consider different numbers of particles and different target values τ^* for the coefficient of variation in the adaptive bridging and tempering updates. Furthermore, we choose the maximal discretisation level $N_{\text{L}}^{\max}$ adaptively in Example 6.10, using the modified update scheme $\mathcal{U}'$ in (6.10). The simulation setups are summarised in Table 6.1.

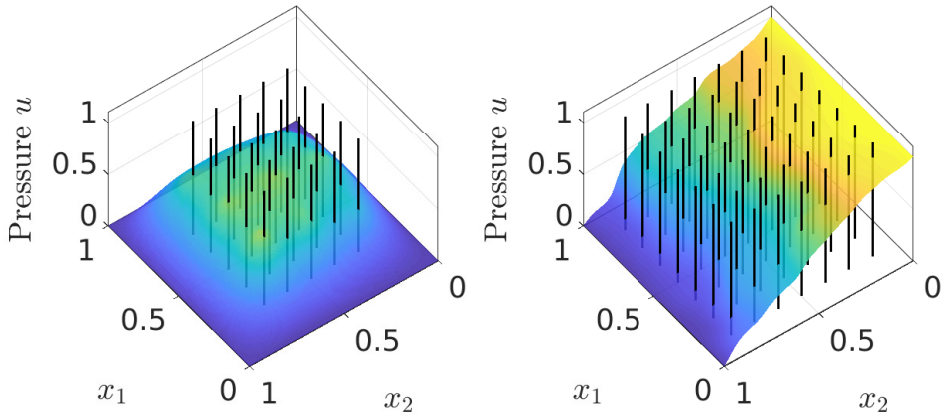


Figure 6.4. Measurement locations and pressure. The surface plots show the hydrostatic pressure given the true permeability. The vertical lines indicate the measurement points. On the left: Examples 6.8 and 6.9. On the right: Example 6.10.

Example	6.8	6.9	6.10
# runs	50 per setup		
N_{smp}	156, 312, 625, 1250, 2500		250, 500, 1000, 2000
τ^*	0.5, 1		1
τ_{LU}	τ^*		
N_{sto}	10		320
h^{-1}	(8, 16, 32, 64, 128)		(16, 32, 64, 128, 256)
Update scheme	$\mathcal{U}$ in (6.9)		$\mathcal{U}'$ in (6.10); $\tau_{\min} = 0.001$
Γ	$0.07^2 \cdot \text{Id}$	$0.035^2 \cdot \text{Id}$	$0.045^2 \cdot \text{Id}$

Table 6.1. Parameter settings in Examples 6.8-6.10

All SMC samplers use a Markov kernel in the mutation step. We choose a single step of a Random Walk Metropolis MCMC sampler with Gaussian proposal density; see Example 3.12. The covariance operator of this proposal density is given by $C^{\text{prop}} = \frac{2.38^2}{N_{\text{sto}}}\text{Id}$. It remains unchanged for all intermediate measures. In high dimensions it would be a good idea to employ the preconditioned Crank-Nicolson MCMC sampler, see Example 3.13, however, we do not implement this here.

6.3.1 Zero boundary pressure

First, we consider the Examples 6.8 and 6.9. Recall that the solution of a BIP is the posterior measure. The mean of the posterior measure is the best unbiased point estimator of the true underlying parameter in the $\mathbf{L}^2$-sense; see §1.3.3 and [179] for details on conditional expectations and their properties. It is important to note that unbiasedness refers only to the stochastic approximation. The discretised PDE solution introduces a bias compared to the exact PDE solution. For this reason we measure the approximation accuracy of the posterior measure and also the accuracy of the posterior mean when used as point estimator. In addition, for each Sequential

Monte Carlo sampler we compare the estimated model evidences and the associated computational costs.

Posterior mean. We consider synthetic data and thus the true (spatially varying) parameter θ_{true} is known. θ_{true} is identical in Examples 6.8 and 6.9 (their setup differs only in the noise covariances). Note that θ_{true} is generated using the truncated KL expansion in (6.11). Hence the KL truncation error is not included in our experiments. In the top row of Figure 6.5, we plot θ_{true} together with typical posterior means estimated with SMC, MLB, and MLS^{2}MC, respectively. In the bottom row of Figure 6.5, we plot the corresponding hydrostatic pressure. We observe that SMC and MLS^{2}MC give similar results. In contrast, the estimate delivered by MLB differs (visually) from the SMC estimate. We discuss this below.

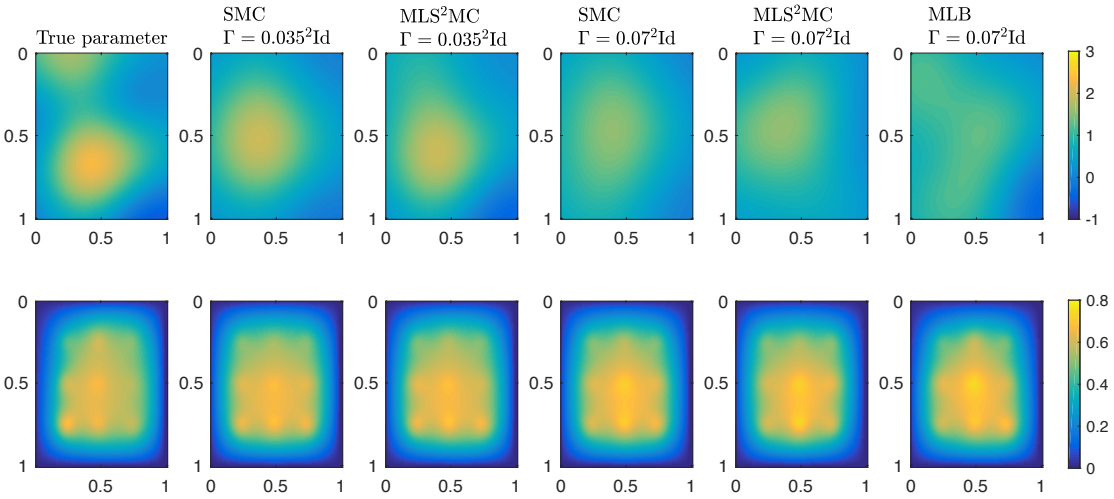


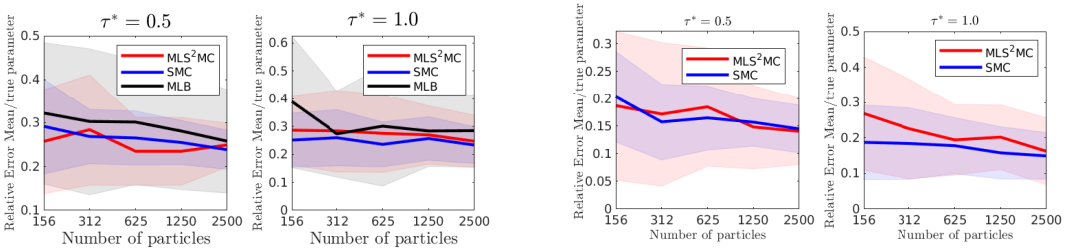
Figure 6.5. True parameter and posterior mean random fields. Top row: The true underlying log-permeability and various posterior mean estimates based on $N_{\text{smp}} = 1250$ particles and $\tau = 0.5$ for Examples 6.8 and 6.9. Bottom row: The hydrostatic pressure corresponding to the log-permeability in the top row.

Now we evaluate the posterior mean estimates more systematically, and quantitatively. We use the following error metric:

$$\text{RelErr}(\widehat{\boldsymbol{\theta}}^{\text{KL}}, \theta^{\text{KL}}_{\text{true}}) := \|\Lambda^{1/2}(\widehat{\boldsymbol{\theta}}^{\text{KL}} - \theta^{\text{KL}}_{\text{true}})\|_1 / \|\Lambda^{1/2}\theta^{\text{KL}}_{\text{true}}\|_1, \tag{6.12}$$

where $\widehat{\boldsymbol{\theta}}^{\text{KL}} \in \mathbb{R}^{10}$ is the estimate of the posterior mean (column) vector and $\theta^{\text{KL}}_{\text{true}} \in \mathbb{R}^{10}$ is the (column) vector of the true parameter values. The (row) vector $\Lambda^{1/2} := (\lambda_1^{1/2}, \ldots, \lambda_{10}^{1/2})$ contains the square roots of the 10 leading KL eigenvalues. Hence, the error measure is a weighted ℓ^1-distance, where we weigh the particles according to their contribution in the KL expansion. We plot the results in Figure 6.6.
As expected the estimation quality is better for a smaller noise level, consistently for all methods. We see that SMC is the most accurate method, while MLS^{2}MC performs slightly worse than SMC, and MLB performs slightly worse than MLS^{2}MC. This is more pronounced for small numbers of particles N_{smp} and a relatively large coefficient of variation $\tau^* = 1$. The results are consistent with the fact that in every Importance Sampling update we introduce a sampling error. A large number of updates gives a large sampling error. The number of updates is minimal in SMC and

maximal in MLB. Hence we expect SMC to give a better estimation result compared to MLB. The estimates obtained with MLS^{2}MC are similar to the estimation results of SMC. Overall, these experiments confirm our motivation for MLS^{2}MC given in §6.2.



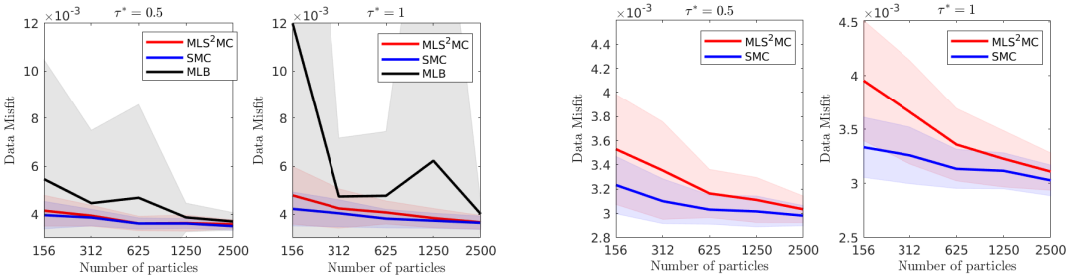
(a) Example 6.8: $\Gamma = 0.07^2\mathrm{Id}$ **(b)** Example 6.9: $\Gamma = 0.035^2\mathrm{Id}$

Figure 6.6. RelErr of the posterior mean estimate compared to the true parameter. The bold lines show the sample mean of the error taken over 50 runs. The shaded areas show the associated standard deviation, again taken over 50 runs.

Next we consider the misfit of the (discretised) model output $\mathcal{G}(\widehat{\boldsymbol{\theta}}^{\mathrm{KL}})$ and the observed data:

$$\mathrm{RelMisfit}(\widehat{\boldsymbol{\theta}}^{\mathrm{KL}}) := \|\Gamma^{-1/2}(y - \mathcal{G}^{(N_{\mathrm{L}})}(\widehat{\boldsymbol{\theta}}^{\mathrm{KL}}))\|_2^2 / \|\Gamma^{-1/2} y\|_2^2. \tag{6.13}$$

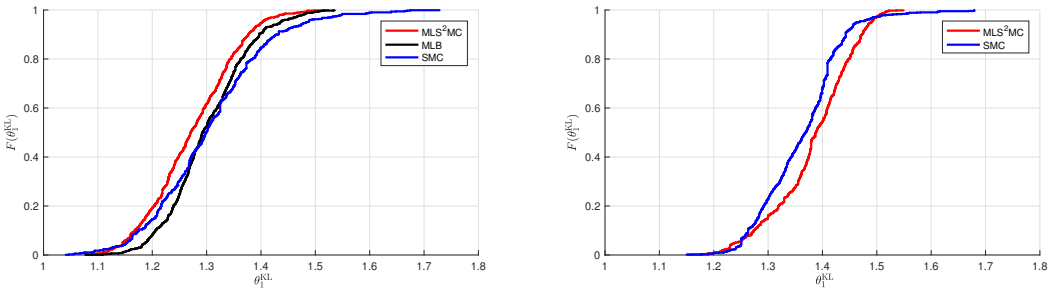
We plot the relative misfit in Figure 6.7. As expected we do not observe significant differences for the two noise levels since the noise precision Γ^{-1} cancels in the relative expression. For all methods we see that the misfit is reasonably small. Therefore, the posterior mean estimate is a good approximation to the maximum a posteriori estimator.



(a) Example 6.8: $\Gamma = 0.07^2\mathrm{Id}$ **(b)** Example 6.9: $\Gamma = 0.035^2\mathrm{Id}$

Figure 6.7. RelMisfit of the posterior mean estimate compared to the observations y. The bold lines show the sample mean of the error taken over 50 runs. The shaded areas show the associated standard deviation, again taken over 50 runs.

Posterior measure. Now we only consider the leading three KL random variables $\boldsymbol{\theta}_1^{\mathrm{KL}}, \boldsymbol{\theta}_2^{\mathrm{KL}}$, and $\boldsymbol{\theta}_3^{\mathrm{KL}}$. These parameters capture 76% of the variance of the prior random field. In Figure 6.8, we plot the empirical cumulative distribution functions (ecdfs) of θ_1^{KL} for representative simulations in Example 6.8 and 6.9.
We assess the accuracy of the posterior measure approximations produced by MLS^{2}MC and MLB by comparing it with the associated (single-level) SMC method,



(a) Example 6.8 **(b)** Example 6.9

Figure 6.8. Empirical cumulative distribution function of the posterior measure of the leading KL random variable estimated with $N_{\text{smp}} = 1250$ particles and $\tau^* = 0.5$.

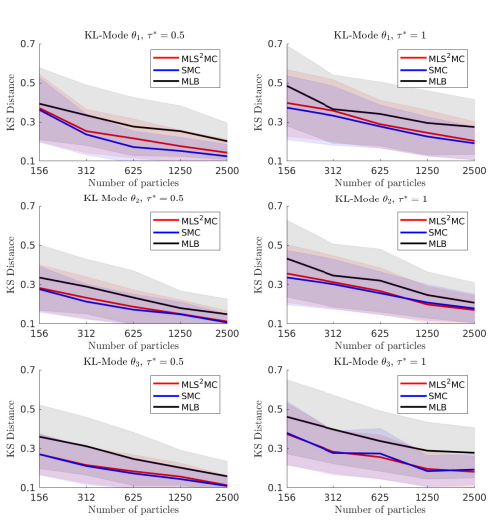
using the same values for N_{smp} and τ^*. We compute the *Kolmogorov-Smirnoff (KS) distance*[a] of all $50 \cdot 50 = 2500$ pairs of simulations of (MLS^{2}MC, SMC) and (MLB, SMC), respectively.
We plot the sample means and standard deviations of the 2500 KS distances of the leading three KL random variables in Figure 6.9. Since we expect some scattering within the reference SMC approximation itself, we also show the 2500 KS distances within the SMC simulations. This line can be used as baseline to account for the intrinsic scattering within the stochastic methods. The results are similar to the observations we made for the posterior mean approximation in the previous subsection. In Example 6.8, there is no significant difference between SMC and MLS^{2}MC. MLB performs slightly worse; we suspect that this is again caused by the larger number of intermediate Importance Sampling updates. In Example 6.9, we observe a larger discrepancy of the approximate posterior measures compared to Example 6.8.

Model evidence. Every SMC-type method delivers automatically an estimate $\widehat{\boldsymbol{Z}}$ of the model evidence $Z(y)$ in Theorem 1.45. We have discussed this in §5.2.3.
$\widehat{\boldsymbol{Z}}$ is a random variable, and in each simulation run of SMC, MLB or MLS^{2}MC we obtain a realisation of it. We plot the ecdfs for 50 runs of SMC, MLB and MLS^{2}MC each in Figure 6.10. Note that the random variable $\widehat{\boldsymbol{Z}}$ is a biased estimator for the model evidence due to the adaptivity of the algorithm; see [17].
In addition, we compute the distance of $\widehat{\boldsymbol{Z}}$ to a reference solution Z^{ref} given by the geometric mean of 50 estimates produced by single-level SMC. We consider the geometric mean since the model evidence is a prefactor. For the same reason we consider the log of the model evidence rather than the model evidence itself from now on. We use the error metric

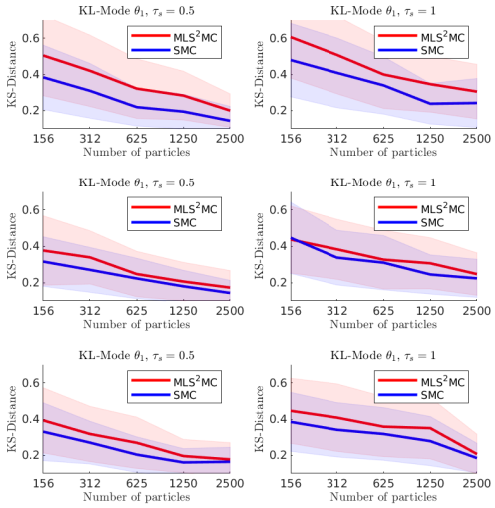
$$\text{RelErrEvid}(\widehat{\boldsymbol{Z}}, Z^{\text{ref}}) := \|\log(\widehat{\boldsymbol{Z}}) - \log(Z^{\text{ref}})\|_1 / \|\log(Z^{\text{ref}})\|_1.$$

Again we compare the SMC estimates with the reference solution to obtain a base value for the dispersion within the stochastic algorithms. The results are given in Figure 6.11. We see that MLB gives poor estimates of the model evidence compared

[a] The KS distance has several applications in statistics. It is often used to compare two discrete probability measures or a continuous and a discrete probability measure. For example, the KS distance is the test statistic used in the Kolmogorov-Smirnoff test. See [55] for details.



(a) Example 6.8: $\Gamma = 0.07^2\mathrm{Id}$



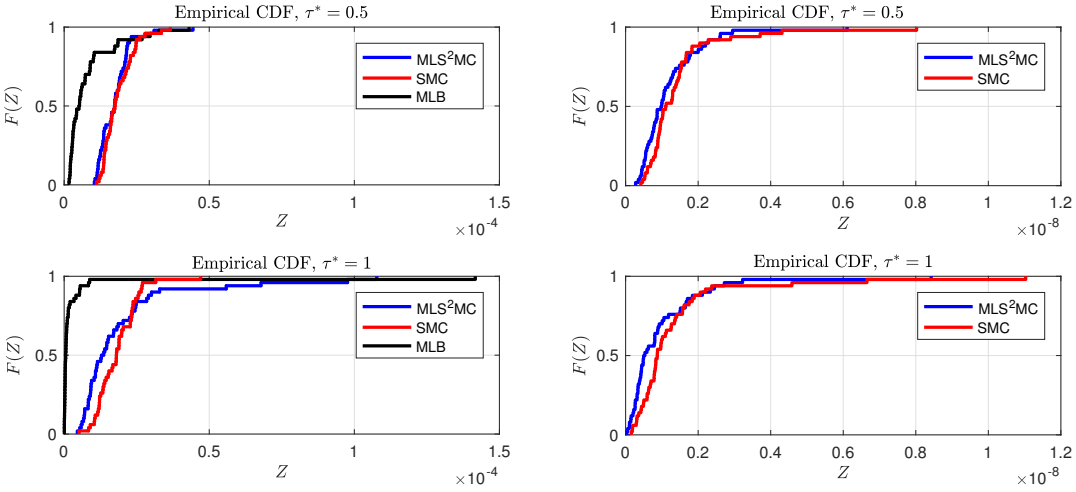
(b) Example 6.9: $\Gamma = 0.035^2\mathrm{Id}$

Figure 6.9. KS distances of posterior measure approximations. The bold lines show the sample means of KS distances of $50 \cdot 50$ combinations of SMC and either MLS^{2}MC, MLB or SMC. The shaded areas show the associated standard deviations.

to SMC and MLS^{2}MC. This is consistent with the results for the KS distances of the posterior measures of $\boldsymbol{\theta}_1^{\mathrm{KL}}$, $\boldsymbol{\theta}_2^{\mathrm{KL}}$, and $\boldsymbol{\theta}_3^{\mathrm{KL}}$ where MLB produced significantly different approximations compared to SMC and MLS^{2}MC.

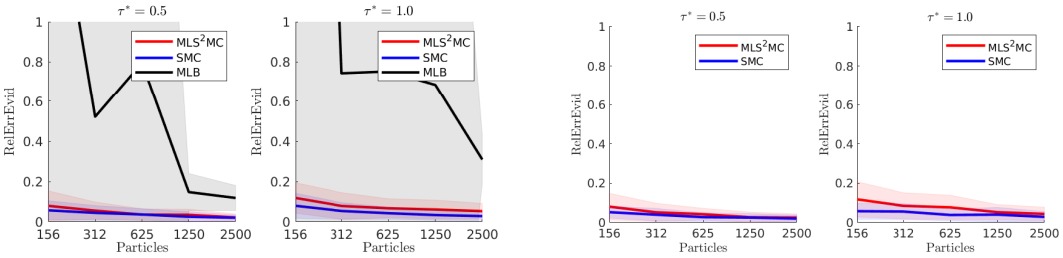
Adaptive Update Scheme. In MLS^{2}MC we apply the adaptive update scheme $\mathcal{U}$ introduced in (6.9). We always use $N_{\widetilde{\mathrm{smp}}} := 100$ particles to predict the number of intermediate bridging steps. In Figure 6.12, we present realisations of the adaptive update scheme. Note that these are realisations of the schematic sketch in Figure 6.1. We observe that the first discretisation level $\ell = 1$ is very inaccurate. In all realisations, the update scheme leaves this level with a very small inverse temperature. This might be the reason why Multilevel Bridging performs poorly here. Indeed, given the inverse temperature $\beta = 1$, the bridging from $\ell = 1$ to $\ell = 2$ requires many intermediate bridging steps. This in turn induces a large sample error in MLB as observed throughout this section. Since the evaluation of Φ_1 and Φ_2 is cheap, the influence on the computational cost of MLB is negligible.
Observe that for $\tau^* = 0.5$ the algorithm might choose to go to $\ell = 3$ before arriving at the maximal inverse temperature $\beta = 1$. For $\tau^* = 1$, the algorithm goes to $\beta = 1$ first, before moving to the discretisation level $\ell = 3$. We anticipated this situation. In the first case, for a small value of τ^*, the algorithm is more conservative, meaning that the level updates are performed early. This strategy increases the accuracy but also the computational cost of the method. The path selected for the larger value $\tau^* = 1$ is computationally cheaper, however, it might give a larger sample error. Note that we do, in fact, observe a larger error in the examples where $\tau^* = 1$.

Computational Cost. Our implementation of the SMC-type samplers and the finite element approximation is not optimized. For these reasons, we compare the



(a) Example 6.8 **(b)** Example 6.9

Figure 6.10. Empirical cumulative distribution functions of the model evidences of 50 posterior measures, each computed with $N_{\text{smp}} = 2500$ particles.



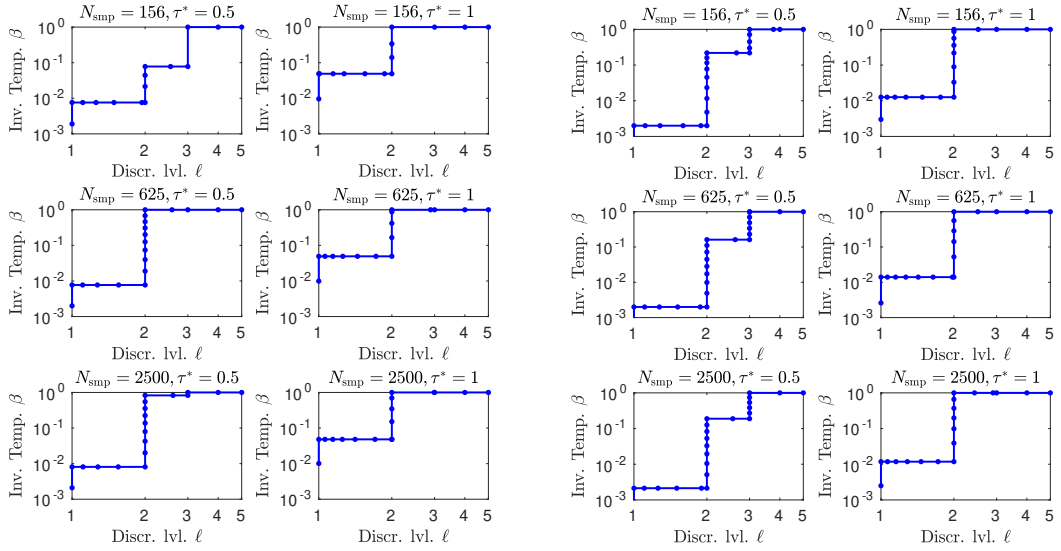
(a) Example 6.8 **(b)** Example 6.9

Figure 6.11. Relative error of the estimated model evidences. The bold lines show the sample mean of the error taken over 50 experiments. The shaded areas show the associated standard deviation, again taken over 50 runs.

computational cost in terms of floating point operations, and not in terms of the elapsed time. The cost of a single evaluation of Φ_ℓ is

$$\mathrm{C}_\ell := 2^{2(\ell-5)}, \quad \ell = 1, \ldots, 5.$$

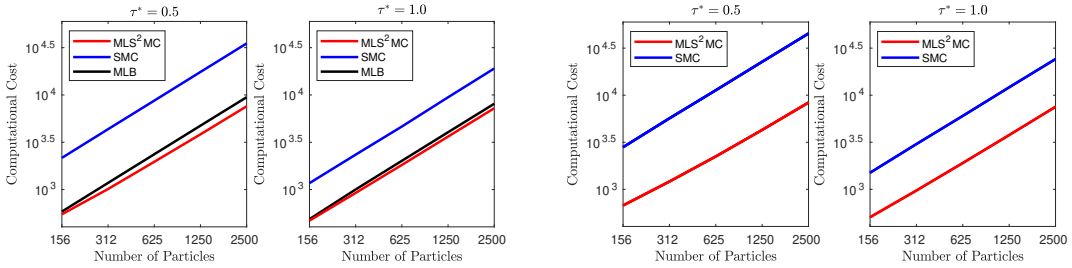
This is motivated in Example 6.4, where we take $d = 2$. In Figure 6.13, we plot C_ℓ against the number of particles N_{smp}. As expected, the cost scales linearly in N_{smp}. If N_{smp} is fixed, then we observe a speed-up of factor 4 for both MLB and MLS^{2}MC compared to single-level SMC. Increasing the discretisation level by one unit increases the cost by a factor of 4 in single-level SMC. Hence, using either of the multilevel methods gives us one discretisation level more for the same computational cost as single-level SMC. However, in the preceding sections, we observed that the MLS^{2}MC samplers are more accurate compared to MLB. In Figure 6.14, we compare computational cost and accuracy directly. We measure the accuracy in terms of the relative error of the model evidence. Given the relatively large $\tau^* = 1$, the additional stochastic error that is introduced in MLS^{2}MC outweighs the advantages in terms of computational cost. For $\tau^* = 1$ we see that MLS^{2}MC is not as accurate as SMC



(a) Example 6.8 **(b)** Example 6.9

Figure 6.12. Realisations of the adaptive update scheme $\mathcal{U}$ (6.9) within the MLS^{2}MC algorithm. Each dot corresponds to one intermediate probability measure.

because in MLS^{2}MC we perform a much larger number of intermediate update steps. On the contrary, for the smaller value $\tau^* = 0.5$ and a fixed accuracy of the estimator, MLS^{2}MC is strictly cheaper than SMC. Overall, this demonstrates the advantages of MLS^{2}MC in terms of both cost and accuracy.

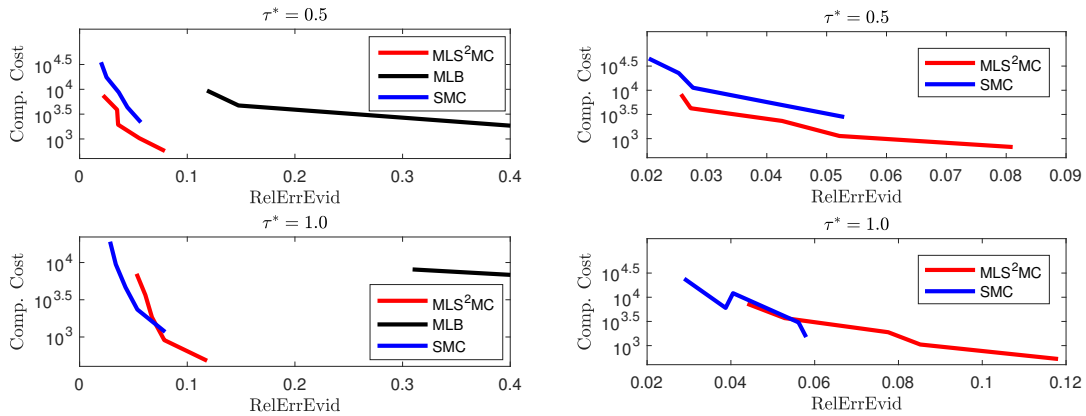


(a) Example 6.8 **(b)** Example 6.9

Figure 6.13. Computational cost of the SMC-type samplers. Each of the bold lines represents the mean computational cost throughout 50 simulations. The costs are measured in terms of the theoretical number of floating point operations per PDE solve on the given discretisation level. These costs are normalised such that $\mathrm{C}_{N_\mathrm{L}} = 1$.

6.3.2 Flow cell

Now we consider Example 6.10. We are particularly interested in the performance of MLS^{2}MC in high dimensions. We compare only MLS^{2}MC and single-level SMC since the adaptive update scheme in MLS^{2}MC delivers the same sequence of intermediate



(a) Example 6.8 **(b)** Example 6.9

Figure 6.14. Comparison of the computational cost and accuracy of MLS^{2}MC, SMC and MLB for $\tau^* \in \{0.5, 1\}$. The different levels of accuracy are associated with different numbers of samples N_{smp}. This combines Figures 6.11 and 6.13.

probability measures as MLB. In addition, we choose the maximal discretisation level adaptively within MLS^{2}MC; see §6.2.5 for a discussion. Hence, we apply the update scheme $\mathcal{U}'$ defined in (6.10). Note that we use 16 rather than 8 finite elements in each spatial direction on the coarsest level.

Posterior approximation in high dimensions. We present the posterior mean estimates and the true underlying parameter in Figure 6.15. We see that the estimation results are visually not as informative as the previous examples. Indeed, one can only recognize the coarse-scale structure of the true parameter. Recall that in §6.3.1, we considered the three leading KL terms. In this example, however, the three leading KL terms capture only about 8% of the prior variance. Informative results would require the consideration of a large number of marginal distributions. However, since this is not illustrative for the reader, we consider the random field at two fixed points in the spatial domain; these points are $x^{(1)} = (0.5, 0.5)$ and $x^{(2)} = (0.75, 0.25)$.
Before looking at the KS distances of the distributions of $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(1)})$ and $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(2)})$, we assess their posterior mean estimates. The relative error of the posterior means in these points compared to the true values $\theta_{\text{true}}(x^{(1)})$ and $\theta_{\text{true}}(x^{(2)})$ is given in Figure 6.16. While the estimate of $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(1)})$ is quite accurate, the estimate of $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(2)})$ is very inaccurate – consistently in both methods. This is consistent with the plots of the posterior means in Figure 6.15.
Next we consider the relative misfit defined in (6.13). We plot this error metric in Figure 6.17. Even though the parameters are approximated quite poorly, the relative misfit is fairly small. Hence, the data might be not sufficient to identify the underlying parameter more precisely.
We now move on to assessing the approximation accuracy of the posterior measures. To this end, we consider again the random variables $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(1)})$ and $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(2)})$. We compute the KS distances of their posterior measures as discussed in §6.3.1. That is, we compare 50 MLS^{2}MC approximations with 50 SMC approximations, using the identical numbers of particles. To obtain a base value for the KS distance, we again

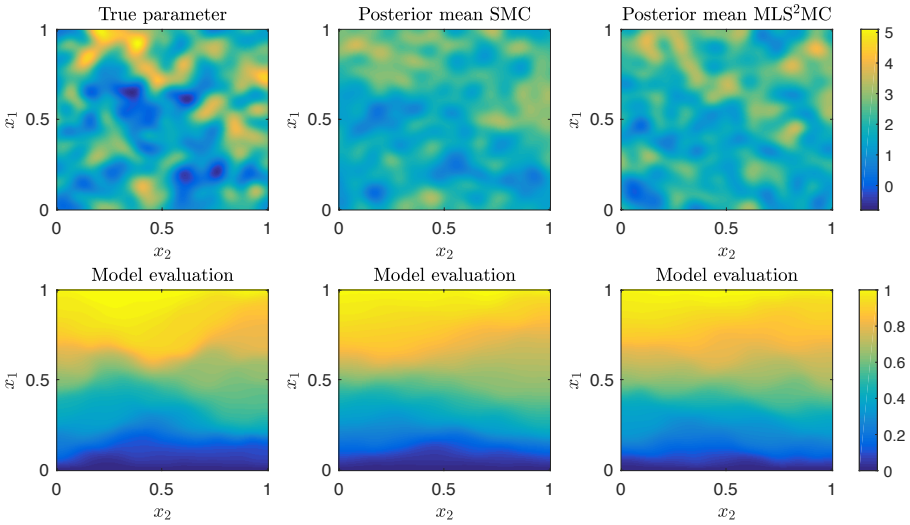


Figure 6.15. Parameter estimation results with estimations from flow cell. Top row: True underlying parameter (left) and posterior mean estimates of SMC (center) and MLS^{2}MC (right) in Example 6.10. The estimations are based on $N_{\text{smp}} = 1000$ particles. Bottom row: Hydrostatic pressure corresponding to the log-permeability in the top row.

compare also the SMC approximations to one another. The results are presented in Figure 6.18. As in Examples 6.8 and 6.9, we see that MLS^{2}MC approximates the SMC reference solution very well.

Adaptive update scheme. We present again some representative update schemes in Figure 6.19. We see that MLS^{2}MC chooses the same updates as MLB. This can be justified as follows: First of all, we started with a finer PDE discretisation on the initial level. Hence, the Bridging with large inverse temperatures should be genuinely easier. Moreover, the noise level in this Example 6.10 is not as small as in Example 6.9. In such a setting, MLB is optimal.
Recall that the maximal discretisation level is chosen adaptively. The samplers using $N_{\text{smp}} \in \{250, 500, 1000\}$ particles stop on level 4, whereas the samplers using $N_{\text{smp}} = 2000$ particles continue to level 5. Hence it might not be possible to capture the difference between the discretisations $\Phi^{(4)}$ and $\Phi^{(5)}$ using a small number of particles. It is possible that the *necessary sample size* is too small in this setting; see Sanz-Alonso [223]. In Figure 6.18, we do not see a significant difference between the MLS^{2}MC approximations using $N_{\text{smp}} \in \{250, 500, 1000\}$ particles and the respective SMC approximations. This might be surprising since the posterior approximations are based on different PDE discretisations. However, SMC also uses $N_{\text{smp}} \in \{250, 500, 1000\}$ particles for its approximation. If the N_{smp} particles were not able to capture the difference between the potentials $\Phi^{(4)}$ and $\Phi^{(5)}$ in MLS^{2}MC, this should also be the case in SMC. Hence, by using the adaptive update scheme, we can reduce the final discretisation level without losing accuracy.

Computational cost. We give the computational cost again in terms of number of PDE evaluations with their respective theoretical number of floating point operations. Furthermore, we normalize $\mathrm{C}_4 = 1$ to be consistent with Examples 6.8

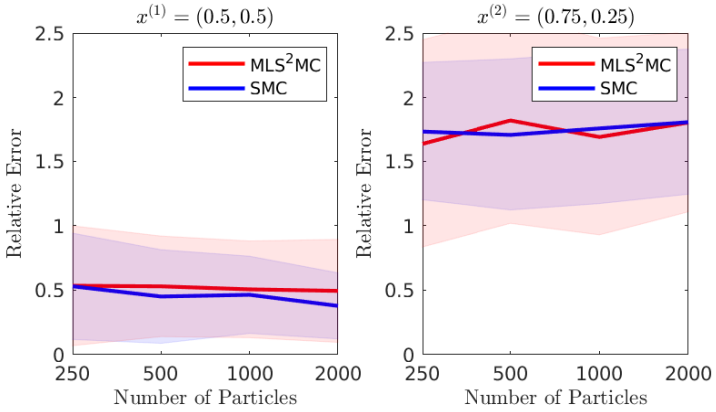


Figure 6.16. Relative error of posterior mean estimates compared to the true parameter in $x^{(1)}$ (left) and $x^{(2)}$(right) in Example 6.10. The bold lines show the sample mean of the error taken over 50 experiments. The shaded areas show the associated standard deviation, again taken over 50 runs.

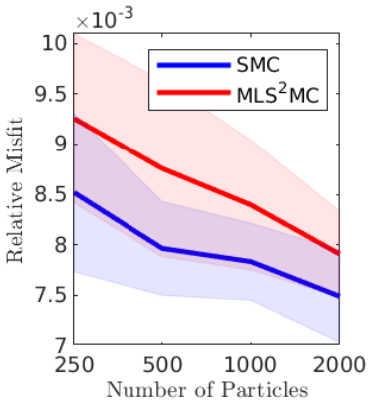


Figure 6.17. Relative misfit of the posterior mean estimates compared to the observations in Example 6.10. The bold lines show the sample mean of the error taken over 50 experiments. The shaded areas show the associated standard deviation, again taken over 50 runs.

and 6.9. Hence,

$$C_\ell = 2^{2(4-\ell)} \qquad (\ell = 1, ..., 5).$$

We present the cost of the simulations in Figure 6.20. We observe a speed-up of a factor 4 compared to single-level SMC, considering the number of particles. This is similar to the results in Example 6.8 and 6.9.
Furthermore, in this figure we see a kink at $N_{\text{smp}} = 1000$ in the graph representing the MLS^{2}MC method. This corresponds to a disproportional increment in logarithmic computational cost we observe when using $N_{\text{smp}} = 2000$ particles. It is caused by the larger maximal discretisation level our algorithm chooses adaptively. In Figure 6.13, we also compare computational cost and accuracy of the posterior mean estimates in terms of the relative misfit. We see that MLS^{2}MC is less accurate than SMC. This is consistent with the numerical results in Examples 6.8 and 6.9; see §6.3.1. There we have noticed that the large $\tau^* = 1$ leads to a large stochastic error in MLS^{2}MC, but not in SMC. We expect that this problem can be solved by choosing a small τ^*.

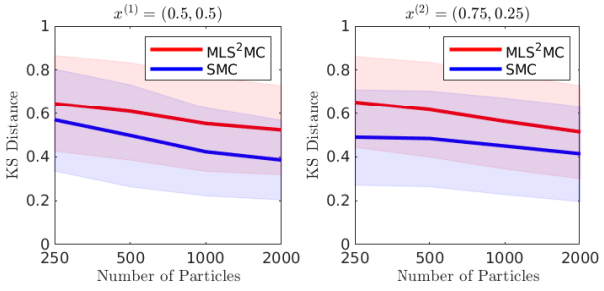


Figure 6.18. KS distances of the marginal posterior distributions of $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(1)})$ (left) and $\boldsymbol{\theta}_{N_{\text{sto}}}(x^{(2)})$ (right) in Example 6.10. We compare the MLS^{2}MC approximation with the SMC approximations and also the SMC approximations to one another. The bold lines show the sample means of KS distances of $50 \cdot 50$ combinations of SMC and either MLS^{2}MC, or SMC. The shaded areas show the associated standard deviations.

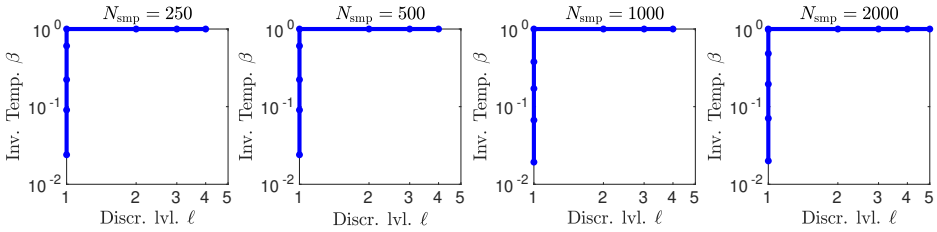


Figure 6.19. Realisations of the adaptive update scheme $\mathcal{U}'$ (6.10) in MLS^{2}MC applied to Example 6.10. Each dot represents one intermediate probability measure.

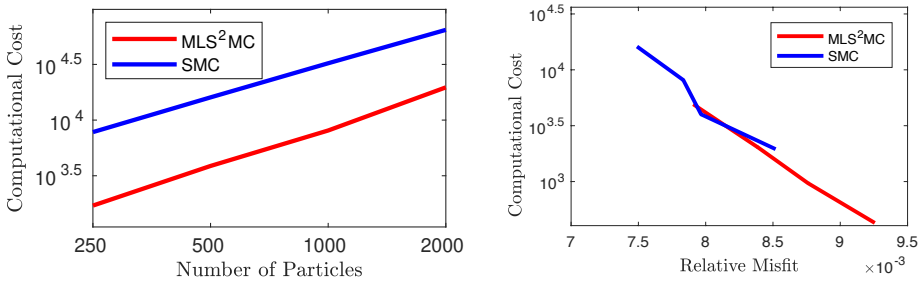


Figure 6.20. Cost and accuracy of SMC and MLS^{2}MC evaluations in Example 6.10. Each of the bold lines represents the mean computational cost throughout 50 simulations. The y-axes represents the costs in terms of the theoretical number of floating point operations per PDE solve on the given discretisation level. These costs are normalised such that $\text{C}_{N_{\text{L}}} = 1$. The x-axes represents either the number of particles N_{smp} (left) or the relative misfits that are also given in Figure 6.17 (right).

Chapter 7

Conclusions and outlook

> A problem never exists in isolation;
> it is surrounded by other problems
> in space and time.
>
> Russell L. Ackoff [2, p. 429]

In the following, we briefly summarise the main contributions of this thesis. We also point the reader to possible directions for future research.

Well-posedness. We have introduced a new concept of well-posedness for Bayesian inverse problems. Our concept is slightly weaker than the state of the art, but allows for well-posedness statements without fully analysing the possibly complicated or unknown forward model. Here, 'unknown' refers to models that are hidden in software. Hence, we obtain very general statements using our concept.
In §2, we have mainly considered the Hellinger well-posedness. Weak well-posedness is a strictly weaker statement than Hellinger well-posedness. Hence, we would assume that we can relax Assumption 2.4 to, e.g. problems in which the Bayesian inverse problem is solved via disintegration.
In this thesis, we only consider the stability of the posterior measure with respect to the data. Other authors have also considered stability, see, e.g. Sprungk [233], and instability, see, e.g. Owhadi et al. [195], with respect to the prior and the model. Of particular interest for our future work is the effect of discretisations of prior and the likelihood; see e.g. Cockayne et al. [47]. Such a setting occurs, if, e.g. the prior is a random field and the likelihood is based on a PDE. The results of [233, 236] are only of asymptotic nature, i.e. no accurate bounds are known. For practical use however, we need accurate a priori and a posteriori error estimators. Such an error estimator would bound the error between the true posterior measure and the semidiscrete posterior that is based on, e.g. a discretised PDE. It could be used to balance the sampling error and the discretisation error in, e.g. an Importance Sampling approximation of a PDE-based BIP. Moreover, it would be a great addition for adaptive algorithms, such as MLS^{2}MC: In §6.2.5, we discuss an adaptive stopping criterion for this method, which is based on an approximation of the Kullback–Leibler divergence. An accurate error estimator would potentially be beneficial here.

Hierarchical measures. Having motivated the importance of hierarchical models in UQ, we discuss the computational challenges that occur with hierarchical random

fields. The main challenge is the continued reconstruction of the KL expansion, whenever the hyperparameters of the covariance operators change. We approach this challenges with a low-rank strategy. More precisely, we employ a reduced basis approach for the fast recomputation of KL eigenpairs. In the online phase, this gives us a linear cost strategy, whereas the standard strategy has quadratic cost. We construct the reduced basis with a proper orthogonal decomposition. In numerical experiments, we verify our strategy and illustrate its applicability. Hence, we show how to explore such hierarchies in random fields.

Our proposed algorithms has two drawbacks. First, we need to solve full eigenproblems in the offline phase. This may be not possible for large scale problems. Second, we have no error estimator to determine the set of snapshots. Hence, we can only choose the snapshots heuristically and have no guarantee that the resulting reduced basis leads to a good approximation. Error estimators from reduced basis methods for PDE eigenproblems may not be useable from a computational point of view. The error estimator introduced by Horger et al. [123], for instance, require to store a number of full operators in the memory during the basis construction. While this may be easily possible for local PDE operators, it is impossible for the non-local operators we typically consider as covariance operators. Note that the local operators can be represented by sparse matrices; non-local operators, however, typically lead to representations by dense matrices. Ideas from [115, 224] may help to solve these drawbacks in our reduced basis sampling algorithm.

For an efficient use of the reduced basis algorithm, we need an offline-online decomposition of the covariance operator. The approximate method we propose in §4.2.3 does not allow for very small correlation lengths when computing in the double precision format. This problem may be solved by expanding the covariance kernel in a different way, e.g. in Chebyshev polynomials, Chebyshev rational functions, or Legendre polynomials. Moreover, it would be interesting to consider other classes of covariance kernels; not only those of Matérn-type.

When motivating hierarchical measures in the introduction of this thesis, we have shown a complicated network structure in Figure 0.2. As opposed to this relatively deep hierarchy, the hierarchical measures, we consider in §4 are rather shallow – they have only one hidden layer. More complicated structures could be studied. Moreover, the hidden layers we consider are low-dimensional. In contrast, deep Gaussian processes have hidden layers that are also Gaussian random fields; see, e.g. Dunlop et al. [76]. Our approach will likely not scale well when the dimension of the hyperparameter space increases. A high-dimensional hyperparameter space may lead to a large number of snapshots and to a large reduced basis. Both mitigate the effectiveness of the reduced basis. Hence, other strategies need to be discussed in this setting.

Another open problem is the use of reduced basis sampling within multilevel algorithms, such as MLS^{2}MC. In multilevel algorithms for forward and inverse problems, the random field often has to be discretised on a hierarchy of different meshes. Moreover, random fields have to be projected onto different meshes. A trivial approach to use reduced basis sampling in such an algorithm is to always work with the most accurate random field discretisation. In this case however, the cost may not be balanced well between random field discretisation and model discretisation.

SMC and random measures. Sequential Monte Carlo methods and other particle filters have been applied in Bayesian filtering and also in static Bayesian inverse problems. We have rigorously introduced a new framework to study and analyse Sequential Monte Carlo methods. This random measure framework allows for investigations and interpretations of SMC with MCMC methods. This is particularly useful for the study of the long-time behaviour of particle filters given a fixed number of particles. Thus, it will enable us to study the use of hierarchical procedures in Bayesian estimation problems.
So far, we have discussed stationary measures with respect to the SIS and SMC Markov kernels. An obvious next step would be the investigation of ergodic behaviour of the SIS and SMC Markov chains; and hence the question, whether we converge to one of the stationary regimes that we have determined. Of particular interest would be the influence of the weighting and resampling steps. For simplicity, it might make sense to first consider a simple setting, in which analytical solutions are possible. We would start with a linear Gaussian inverse problem setting with noise-free data; see Schillings and Stuart [226]. Here, SMC with tempering could be applied as an annealing method, with fixed β increment and with the tempering going on until $\beta = \infty$. In this example, one could also compare the convergence behaviour of the Ensemble Kalman Inversion with that of SMC.
Having started from the example above, we should look at more complicated sequences of measures, i.e. appearing in non-linear Bayesian filtering. Moreover, we should consider sequences of measures that do not immediately fit in our framework; see §5.1. Those are, for instance, SMC with adapatively chosen temperatures; see §5.4.3. Another level of adaptivity can be introduced by choosing the MCMC sampler in the mutation step adaptively as well. Also, we have not discussed SMC with adaptive resampling any further, even though we have shown that it constructs a Markov chain. Additionally, the long-time behaviour of Multilevel Bridging and MLS^{2}MC should be studied. Here, one could assume that the number of discretisation levels is infinite. In the limit, the discretisation would converge to the correct model.
There are various algorithms that can be understood as generalisations of the presented SMC algorithms. Such are particle filters in data assimilation, see, e.g. [157], or Subset Simulation in reliability analysis, see, e.g. [10]. Natural generalisations to our discussion would include random measure representation of these algorithms; and a long-time, finite particle study of those.

Multilevel Sequential^2 Monte Carlo We have proposed the MLS^{2}MC method to efficiently solve Bayesian inverse problems. It adaptively combines the Multilevel Bridging algorithm and SMC with tempering. In numerical experiments, we have shown that it can outperform both basic algorithms. Hence, we can indeed exploit hierarchies of discretisations for an efficient approximation of a posterior measure.
So far, we have discussed MLS^{2}MC as a method to produce posterior samples. Assume that we aim to integrate a quantity of interest with respect to the posterior. During the process of MLS^{2}MC, we produce samples of posterior measures on various discretisation and temperature levels. Moreover, these samples have a certain correlation structure. In a classical Multilevel Monte Carlo manner, it may be possible to use these samples to reduce the variance of the estimator for the integral. This would combine MLS^{2}MC with Multilevel Sequential Monte Carlo, which we

have reviewed in §6.1.2.
MLS^{2}MC may be a good tool for Bayesian model selection. Note that as in all SMC algorithms, we can approximate the evidences for different likelihoods and obtain the posterior on the space of models by the formula in (1.26). This approach may be expensive in practice: we need to solve the full BIP for every likelihood. Another possibility are across-model simulations, i.e. sampling immediately from the joint posterior of models and model parameters; see, e.g. Green [106] and Uribe et al. [248]. Here, we automatically focus on the posteriors of models that have a high posterior probability. To the best of our knowledge, multilevel approaches have not yet been applied to across-model simulations. Indeed, most multilevel methods focus on computing integrals with respect to a quantity of interest. However, since the set of models is typically discrete, one may not be interested in computing integrals with respect to the model posterior. Since MLS^{2}MC focuses on the fast generation of posterior samples, it is easily applicable in model selection problems with across-model simulations.
In across-model simulations, a typical issue are different parameter spaces for different models; see §1.4.3. However, also in multilevel settings, we can have different parameter spaces for different model discretisations. In the elliptic PDE example, the discretisation of the unknown diffusion coefficient would typically depend on the PDE discretisation. Hence, for every discretisation level, the diffusion coefficient lives in a different parameter space. We solve this issue by using a KL expansion, which we consider to be more or less discretisation invariant. In some settings however, this may be not the possible. Then, we need to be able to change parameter spaces when changing the discretisation. This problem has for instance been discussed by Dodwell et al. [71] for Multilevel Markov chain Monte Carlo.
Problems closely connected to the Bayesian inverse problem appear in reliability analysis. Here, the probability of a failure event should be estimated. In practice, these events have very small probabilities. Since standard Monte Carlo would require too many samples, other methods have to be used for such estimation tasks. Such methods are Subset Simulation [10] and the Cross-Entropy method [216]. While Subset Simulation is an adaptive Sequential Monte Carlo method, the sequence of densities usually consists of indicator functions. Hence, it does not satisfy our assumptions on $\underline{\gamma}$; see §5.1. However, a Sequential Monte Carlo algorithm for reliability analysis that fits into our framework has been proposed by Papaioannou et al. [196]. Speeding-up reliability analyses with a multilevel or multifidelity approach has already been discussed by [197, 247]. By applying MLS^{2}MC in the setting of Papaioannou et al. [196], one can construct a further multilevel method for reliability analysis. Such a method would easily overcome the nestedness issues of Ullmann and Papaiannou's [247] multilevel estimator. Hence, it could be applied in more general settings.
Finally, we mention that we have not analysed the MLS^{2}MC algorithm yet. Such an analysis would include a Central Limit Theorem, as the number of particles goes to infinity, and a derivation of error bounds. The method can be seen as a generalisation of the adaptive SMC algorithm analysed by Beskos et al. [17]. Hence, we expect that an analysis of MLS^{2}MC can be performed in a similar way. A further aspect that should be analysed for MLS^{2}MC is its computational cost, probably by taking the work of Marion and Schmidler [174] into account.

Bibliography

[1] M Abramowitz and I A Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Dept. of Commerce, National Bureau of Standards, Washington, D.C., tenth edition, 1972.

[2] R L Ackoff. *Scientific method: optimizing applied research decisions*. John Wiley & Sons, New York, 1962.

[3] R J Adler. *The geometry of random fields*. John Wiley & Sons, Chichester, 1981.

[4] R J Adler. *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Institute of Mathematical Statistics, Hayward, CA, 1990.

[5] S Agapiou, O Papaspiliopoulos, D Sanz-Alonso, and A M Stuart. Importance sampling: intrinsic dimension and computational cost. *Statist. Sci.*, 32(3):405–431, 2017.

[6] S Agapiou, A M Stuart, and Y-X Zhang. Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.*, 22(3):297–321, 2014.

[7] A Ali, E Ullmann, and M Hinze. Multilevel Monte Carlo analysis for optimal control of elliptic PDEs with random coefficients. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):466–492, 2017.

[8] R B Ash. *Probability and measure theory*. Harcourt/Academic Press, Burlington, MA, 2000.

[9] G Astarita. *Thermodynamics*. Springer, New York, 1989.

[10] S-K Au and J L James L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probab. Eng. Mech.*, 16(4):263–277, 2001.

[11] M Bachmayr, I G Graham, V Kien Nguyen, and R Scheichl. Unified Analysis of Periodization-Based Sampling Methods for Matérn Covariances. *arXiv e-prints*, 1905.13522, 2019.

[12] S Bartels. *Numerical methods for nonlinear partial differential equations*. Springer, Cham, 2015.

[13] T Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc.*, 53:370–418, 1763.

[14] R V Belavkin. Asymmetric topologies on statistical manifolds. In F Nielsen and F Barbaresco, editors, *Geometric science of information*, pages 203–210, Cham, 2015. Springer.

[15] P Benner, Y Qiu, and M Stoll. Low-rank eigenvector compression of posterior covariance matrices for linear Gaussian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):965–989, 2018.

[16] A Beskos, D Crisan, and A Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, 24(4):1396–1445, 2014.

[17] A Beskos, A Jasra, N Kantas, and A Thiery. On the convergence of adaptive sequential Monte Carlo methods. *Ann. Appl. Probab.*, 26(2):1111–1146, 2016.

[18] A Beskos, A Jasra, K Law, Y Marzouk, and Y Zhou. Multilevel sequential monte carlo with dimension-independent likelihood-informed proposals. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):762–786, 2018.

[19] A Beskos, A Jasra, K J H Law, R Tempone, and Y Zhou. Multilevel sequential Monte Carlo samplers. *Stochastic Process. Appl.*, 127(5):1417–1440, 2017.

[20] A Beskos, A Jasra, E A Muzaffer, and A M Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Stat. Comput.*, 25(4):727–737, 2015.

[21] A Beskos, G O Roberts, A M Stuart, and J Voss. MCMC methods for diffusion bridges. *Stoch. Dyn.*, 8(3):319–350, 2008.

[22] W Betz, I Papaioannou, and D Straub. Numerical methods for the discretization of random fields by means of the Karhunen-Loève expansion. *Comput. Methods Appl. Mech. Eng.*, 271:109–129, 2014.

[23] P Billingsley. *Probability and measure*. John Wiley & Sons, Hoboken, NJ, anniversary edition, 2012.

[24] C M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[25] D Blömker, C Schillings, P Wacker, and S Weissmann. Well Posedness and Convergence Analysis of the Ensemble Kalman Inversion. *Inverse Problems (accepted)*, 2019.

[26] S Bochner. Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind. *Fundamenta Mathematicae*, 20:262–276, 1933.

[27] V I Bogachev. *Gaussian measures*. American Mathematical Society, Providence, RI, 1998.

[28] D Bolin and K Kirchner. The rational SPDE approach for Gaussian random fields with general smoothness. *arXiv e-prints*, 1711.04333v2, 2017.

[29] D Bolin, K Kirchner, and M Kovács. Weak convergence of Galerkin approximations for fractional elliptic stochastic PDEs with spatial white noise. *BIT*, 58(4):881–906, 2018.

[30] D Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, third edition, 2007.

[31] M Bulté, J Latz, and E Ullmann. A practical example for the non-linear Bayesian filtering of model parameters. *arXiv e-prints*, 1807.08713, 2018.

[32] D Calvetti, M M Dunlop, E Somersalo, and A M Stuart. Iterative updating of model error for Bayesian inversion. *Inverse Problems*, 34(2):025008, 2018.

[33] D Calvetti, L Reichel, and D C Sorensen. An implicitly restarted Lanczos method for large symmetric eigenvalue problems. *Electron. Trans. Numer. Anal*, 2:1–21, 1994.

[34] G Chan and A T A Wood. An Algorithm for Simulating Stationary Gaussian Random Fields. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 46(1):171–181, 1997.

[35] J Charrier. Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(1):216–246, 2012.

[36] G Chavent. *Nonlinear least squares for inverse problems*. Springer, New York, 2009.

[37] J Chen and M L Stein. Linear-cost covariance functions for Gaussian random fields. *arXiv e-prints*, 1711.05895, 2017.

[38] P Chen, A Quarteroni, and G Rozza. Reduced basis methods for uncertainty quantification. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):813–869, 2017.

[39] P Chen and C Schwab. Sparse-grid, reduced-basis Bayesian inversion. *Comput. Methods Appl. Mech. Engrg.*, 297:84–115, 2015.

[40] P Chen and C Schwab. Model order reduction methods in computational uncertainty quantification. In R Ghanem, D Higdon, and H Owhadi, editors, *Handbook of uncertainty quantification*, pages 937–990. Springer, Cham, 2017.

[41] V Chen, M M Dunlop, O Papaspiliopoulos, and A M Stuart. Dimension-Robust MCMC in Bayesian Inverse Problems. *arXiv e-prints*, 1803.03344, 2018.

[42] A Chernov, H Hoel, K J H Law, F Nobile, and R Tempone. Multilevel Ensemble Kalman Filtering for spatially extended models. *arXiv e-prints*, 1608.08558, 2016.

[43] N Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–551, 2002.

[44] N Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, 32(6):2385–2411, 2004.

[45] K A Cliffe, M B Giles, R Scheichl, and A L Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011.

[46] J Cockayne, C J Oates, T J Sullivan, and M Girolami. Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In G Verdoolaege, editor, *Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1853 of *AIP Conference Proceedings*, page 060001, 2017.

[47] J Cockayne, C J Oates, T J Sullivan, and M Girolami. Bayesian probabilistic numerical methods. *SIAM Rev. (accepted)*, 2019.

[48] A Contreras, P Mycek, O P Le Maître, F Rizzi, B Debusschere, and O M Knio. Parallel domain decomposition strategies for stochastic elliptic equations. Part A: Local Karhunen-Love representations. *SIAM J. Sci. Comput.*, 40(4):C520–C546, 2018.

[49] S L Cotter, G O Roberts, A M Stuart, and D White. MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 2013.

[50] R T Cox. Probability, frequency and reasonable expectation. *Amer. J. Phys.*, 14:1–13, 1946.

[51] D Crisan. Particle filters – a theoretical perspective. In A Doucet, N de Freitas, and N Gordon, editors, *Sequential Monte Carlo methods in practice*, pages 17–41. Springer, 2001.

[52] J A Cumming and M Goldstein. Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments. In *The Oxford handbook of applied Bayesian analysis*, pages 242–270. Oxford University Press, Oxford, 2010.

[53] G Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4):303–314, 1989.

[54] A I Dale. Bayes or Laplace? An examination of the origin and early applications of Bayes' theorem. *Arch. Hist. Exact Sci.*, 27(1):23–47, 1982.

[55] W W Daniel. *Applied nonparametric statistics*. PWS-Kent Publ., Boston, MA, second edition, 1990.

[56] M Dashti, K J H Law, A M Stuart, and J Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017, 2013.

[57] M Dashti and A M Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM J. Numer. Anal.*, 49(6):2524–2542, 2011.

[58] M Dashti and A M Stuart. The Bayesian approach to inverse problems. In R Ghanem, D Higdon, and H Owhadi, editors, *Handbook of uncertainty quantification*, pages 311–428. Springer, Cham, 2017.

[59] R Dechter. Learning while searching in constraint-satisfaction-problems. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, AAAI'86, pages 178–183, 1986.

[60] P Del Moral. *Feynman-Kac formulae. Genealogical and interacting particle systems with applications*. Springer, New York, 2004.

[61] P Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, Boca Raton, FL, 2013.

[62] P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):411–436, 2006.

[63] P Del Moral, A Jasra, and K J H Law. Multilevel sequential Monte Carlo: Mean square error bounds under verifiable conditions. *Stoch. Anal. Appl.*, 35(3):478–498, 2017.

[64] P Del Moral, A Jasra, K J H Law, and Y Zhou. Multilevel sequential Monte Carlo samplers for normalizing constants. *ACM Trans. Model. Comput. Simul.*, 27(3):20:1–20:22, 2017.

[65] M D'Elia and M Gunzburger. Coarse-grid sampling interpolatory methods for approximating Gaussian random fields. *SIAM/ASA J. Uncertain. Quantif.*, 1(1):270–296, 2013.

[66] Deutscher Wetterdienst. Ensemble-Data Assimilation. *https://www.dwd.de/EN/research/weatherforecasting/num_modelling/04_ensemble_methods/ensemble_data_assimilation/ensemble_data_assimilation_node.html [accessed July 23rd 2019]*, 2019.

[67] J Dick, R N Gantner, Q T Le Gia, and C Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017.

[68] J Dick, R N Gantner, Q T Le Gia, and C Schwab. Higher order quasi-Monte Carlo integration for Bayesian PDE inversion. *Comput. Math. Appl.*, 77(1):144–172, 2019.

[69] C R Dietrich and G N Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4):1088–1107, 1997.

[70] R L Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.*, 15(3):458–486, 1970.

[71] T J Dodwell, C Ketelsen, R Scheichl, and A L Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):1075–1108, 2015.

[72] T J Dodwell, C Ketelsen, R Scheichl, and A L Teckentrup. Multilevel Markov Chain Monte Carlo. *SIAM Rev.*, 61(3):509–545, 2019.

[73] A Doucet and A M Johansen. A tutorial on particle filtering and smoothing: fifteen years later. In D Crisan and B Rozovskiĭ, editors, *The Oxford handbook of nonlinear filtering*, pages 656–704. Oxford University Press, Oxford, 2011.

[74] J Douglas, Jr., T Dupont, and M F Wheeler. A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):47–59, 1974.

[75] M. Drohmann and K. Carlberg. The ROMES method for statistical modeling of reduced-order-model error. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):116–145, 2015.

[76] M M Dunlop, M A Girolami, A M Stuart, and A L Teckentrup. How deep are deep Gaussian processes? *J. Mach. Learn. Res.*, 19(54):1–46, 2018.

[77] M M Dunlop, T Helin, and A M Stuart. Hyperparameter Estimation in Bayesian MAP Estimation: Parameterizations and Consistency. *arXiv e-prints*, 1905.04365, 2019.

[78] M M Dunlop, M A Iglesias, and A M Stuart. Hierarchical Bayesian level set inversion. *Stat. Comput.*, 27(6):1555–1584, 2017.

[79] D K Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

[80] D J Earl and M W Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005.

[81] H Elman and Q Liao. Reduced basis collocation methods for partial differential equations with random coefficients. *SIAM/ASA J. Uncertain. Quantif.*, 1(1):192–217, 2013.

[82] S Engel, D Hafemeyer, C Münch, and D Schaden. An application of sparse measure valued Bayesian inversion to acoustic sound source identification. *Inverse Problems*, 35(7):075005, 2019.

[83] O G Ernst and B Sprungk. Stochastic collocation for elliptic PDEs with random data: the lognormal case. In J Garcke and D Pflüger, editors, *Sparse Grids and Applications – Munich 2012*, pages 29–53. Springer, Cham, 2014.

[84] O G Ernst, B Sprungk, and H-J Starkloff. Analysis of the Ensemble and Polynomial Chaos Kalman Filters in Bayesian Inverse Problems. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):823–851, 2015.

[85] L Euler. *Institutionum calculi integralis, Volumen primum, Editio tertia*. Impensis Academiae Imperialis Scientiarum, Petropolis, 1824.

[86] B S Everitt and D J Hand. *Finite mixture distributions*. Chapman & Hall, London-New York, 1981.

[87] I-G Farcaş, J Latz, E Ullmann, T Neckel, and H-J Bungartz. Multilevel adaptive sparse Leja approximations for Bayesian inverse problems. *arXiv e-prints*, 1904.12204, 2019.

[88] I-G Farcaş, P C Sârbu, H-J Bungartz, T Neckel, and B Uekermann. Multilevel adaptive stochastic collocation with dimensionality reduction. In J Garcke, D Pflüger, C G Webster, and G Zhang, editors, *Sparse Grids and applications – Miami 2016*, pages 43–68. Springer, Cham, 2018.

[89] I-G Farcaş, B Uekermann, T Neckel, and H-J Bungartz. Nonintrusive uncertainty analysis of fluid-structure interaction with spatially adaptive sparse grids and polynomial chaos expansion. *SIAM J. Sci. Comput.*, 40(2):B457–B482, 2018.

[90] M Feischl, F Y Kuo, and I H Sloan. Fast random field generation with H-matrices. *Numer. Math.*, 140(3):639–676, 2018.

[91] A Fick. Ueber Diffusion. *Ann. Phys.*, 170(1):59–86, 1855.

[92] Fumagalli, I, Manzoni, A, Parolini, N, and Verani, M. Reduced basis approximation and a posteriori error estimates for parametrized elliptic eigenvalue problems. *ESAIM Math. Model. Numer. Anal.*, 50(6):1857–1885, 2016.

[93] A Gelman and X-L Meng. imulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185, 1998.

[94] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6(6):721–741, 1984.

[95] M Gerber, N Chopin, and N Whiteley. Negative association, ordering and convergence of resampling methods. *Ann. Statist.*, 47(4):2236–2260, 2019.

[96] R G Ghanem and P D Spanos. *Stochastic finite elements: a spectral approach*. Springer, New York, 1991.

[97] S Ghosal and A van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, Cambridge, 2017.

[98] A L Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stoch. Models*, 20(4):473–492, 2004.

[99] A L Gibbs and F E Su. On Choosing and Bounding Probability Metrics. *Int. Stat. Rev.*, 70(3):419–435, 2002.

[100] J W Gibbs. *Elementary principles in statistical mechanics*. Scribner's Sons., New York, 1902.

[101] D Gilbarg and N S Trudinger. *Elliptic partial differential equations of second order*. Springer, Berlin, 2001.

[102] M B Giles. Multilevel Monte Carlo Path Simulation. *Oper. Res.*, 56(3):607–617, 2008.

[103] M B Giles. Multilevel Monte Carlo methods. In *Acta Numer.*, volume 24, pages 259–328. Cambridge University Press, Cambridge, 2015.

[104] G H Golub and C F van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[105] I G Graham, F Y Kuo, D Nuyens, R Scheichl, and I H Sloan. Analysis of circulant embedding methods for sampling stationary random fields. *SIAM J. Numer. Anal.*, 56(3):1871–1895, 2018.

[106] P J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[107] C Groetsch. Linear inverse problems. In O Scherzer, editor, *Handbook of mathematical methods in imaging*, pages 3–46. Springer, New York, 2015.

[108] M Gu, A Ruhe, R Lehoucq, D Sorensen, R Freund, G Sleijpen, H van der Vorst, Z Bai, and R Li. Hermitian Eigenvalue Problems. In Z Bai, J Demmel, J Dongarra, A Ruhe, and H van der Vorst, editors, *Templates for the solution of algebraic eigenvalue problems*, chapter 4, pages 45–107. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[109] W Hackbusch. *Multigrid methods and applications*. Springer, Berlin, 1985.

[110] J Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[111] A-L Haji-Ali, F Nobile, L Tamellini, and R Tempone. Multi-index stochastic collocation for random pdes. *Comput. Methods Appl. Mech. Engrg.*, 306:95 – 122, 2016.

[112] A-L Haji-Ali, F Nobile, and R Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numer. Math.*, 132(4):767–806, 2016.

[113] A-L Haji-Ali and R Tempone. Multilevel and multi-index monte carlo methods for the mckean-vlasov equation. *Stat. Comput.*, 28(4):923–935, 2018.

[114] S O Hansson. Decision Theory: An Overview. In M Lovric, editor, *International Encyclopedia of Statistical Science*, pages 349–355. Springer, Berlin, Heidelberg, 2011.

[115] H Harbrecht, M Peters, and M Siebenmorgen. Efficient approximation of random fields for numerical applications. *Numer. Linear Algebra Appl.*, 22(4):596–617, 2015.

[116] K A Haskard. *An anisotropic Matérn spatial covariance model: REML estimation and properties*. PhD thesis, University of Adelaide, 2007.

[117] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[118] E Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.*, 136:210–271, 1909.

[119] J S Hesthaven, G Rozza, and B Stamm. *Certified reduced basis methods for parametrized partial differential equations*. Springer, Cham, 2016.

[120] C F Higham and D J Higham. Deep learning: An introduction for applied mathematicians. *SIAM Rev. (accepted)*, 2019.

[121] J Hoffmann-Jørgensen and G Pisier. The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probability*, 4(4):587–599, 1976.

[122] B Hölting and W G Coldewey. *Hydrogeology*. Springer, Berlin, Heidelberg, 2019.

[123] T Horger, B Wohlmuth, and T Dickopf. Simultaneous reduced basis approximation of parameterized elliptic eigenvalue problems. *ESAIM Math. Model. Numer. Anal.*, 51(2):443–465, 2017.

[124] T Horger, B Wohlmuth, and L Wunderlich. Reduced basis isogeometric mortar approximations for eigenvalue problems in vibroacoustics. In P Benner, M Ohlberger, A Patera, G Rozza, and K Urban, editors, *Model reduction of parametrized systems*, pages 91–106. Springer, Cham, 2017.

[125] A S Hornby, E V Gatenby, and H Wakefield. uncertain. In *The Advanced Learner's Dictionary of Current English*. Oxford University Press, London, 2nd edition, 1963.

[126] B Hosseini. Well-posed Bayesian inverse problems with infinitely divisible and heavy-tailed prior measures. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):1024–1060, 2017.

[127] B Hosseini and N Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):436–465, 2017.

[128] J Humpherys, P Redd, and J West. A fresh look at the Kalman filter. *SIAM Rev.*, 54(4):801–823, 2012.

[129] M Iglesias, Y Lu, and A M Stuart. A Bayesian level set method for geometric inverse problems. *Interfaces Free Bound.*, 18:181–217, 2016.

[130] M A Iglesias, K Lin, and A M Stuart. Well-posed Bayesian geometric inverse problems arising in subsurface flow. *Inverse Probl.*, 30(11):114001, oct 2014.

[131] E Ising. Beitrag zur Theorie des Ferromagnetismus. *Z. Physik*, 31(1):253–258, 1925.

[132] A Jasra, D A Stephens, A Doucet, and T Tsagaris. Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Stat.*, 38(1):1–22, 2011.

[133] L Jiang and N Ou. Multiscale model reduction method for Bayesian inverse problems of subsurface flow. *J. Comput. Appl. Math.*, 319:188–209, 2017.

[134] C Kahle, K F Lam, J Latz, and E Ullmann. Bayesian parameter identification in Cahn–Hilliard models for biological growth. *SIAM/ASA J. Uncertain. Quantif.*, 7(2):526–552, 2019.

[135] J Kaipio and E Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.

[136] O Kallenberg. *Foundations of modern probability*. Springer, New York, first edition, 1997.

[137] O Kallenberg. *Random measures, theory and applications*. Springer, Cham, 2017.

[138] N Kantas, A Beskos, and A Jasra. Sequential Monte Carlo methods for high-dimensional inverse problems: a case study for the Navier-Stokes equations. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):464–489, 2014.

[139] K Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37:1–79, 1947.

[140] B N Khoromskij, A Litvinenko, and H G Matthies. Application of hierarchical matrices for computing the Karhunen-Loève expansion. *Computing*, 84(1-2):49–67, 2009.

[141] U Khristenko, L Scarabosio, P Swierczynski, E Ullmann, and B Wohlmuth. Analysis of Boundary Effects on PDE-Based Sampling of Whittle-Matérn Random Fields. *SIAM/ASA J. Uncertain. Quantif.*, 7(3):948–974, 2019.

[142] S C Kleene. Representation of events in nerve nets and finite automata. In C E Shannon and J McCarthy, editors, *Automata Studies (AM-34)*, pages 3–42. Princeton University Press, Princeton, NJ, 1956.

[143] A Klenke. *Probability theory: A comprehensive course*. Springer, London, 2014.

[144] A Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitstheorie*. Springer, Berlin, Heidelberg, 1933.

[145] G Köthe. *Topological Vector Spaces I*. Springer, Berlin, Heidelberg, 1983.

[146] P S Koutsourelakis. A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters. *J. Comput. Phys.*, 228(17):6184–6211, 2009.

[147] N B Kovachki and A M Stuart. Ensemble Kalman Inversion: A Derivative-Free Technique For Machine Learning Tasks. *Inverse Probl. (accepted)*, 2019.

[148] S. Krumscheid and F. Nobile. Multilevel Monte Carlo approximation of functions. *SIAM/ASA J. Uncertain. Quantif.*, 6(3):1256–1293, 2018.

[149] F Y Kuo, R Scheichl, C Schwab, I H Sloan, and E Ullmann. Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. *Math. Comput.*, 86(308):2827–2860, 2017.

[150] F Y Kuo, C Schwab, and I H Sloan. Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. *Found. Comput. Math.*, 15(2):411–449, 2015.

[151] H-H Kuo. *White Noise Distribution Theory*. CRC Press, Boca Raton, FL, 1996.

[152] P S Laplace. *Théorie analytique des probabilités*. Courcier, Paris, 1812.

[153] J Latz. *Bayes Linear Methods for Inverse Problems*. Master's Thesis, University of Warwick, 2016.

[154] J Latz. On the well-posedness of Bayesian inverse problems. *arXiv e-prints*, 1902.10257, 2019.

[155] J Latz, M Eisenberger, and E Ullmann. Fast sampling of parameterised Gaussian random fields. *Comput. Methods Appl. Mech. Engrg.*, 348:978–1012, 2019.

[156] J Latz, I Papaioannou, and E Ullmann. Multilevel Sequential^2 Monte Carlo for Bayesian inverse problems. *J. Comput. Phys.*, 368:154–178, 2018.

[157] K J H Law, A M Stuart, and K Zygalakis. *Data assimilation: A mathematical introduction*. Springer, Cham, 2015.

[158] Jr Leao, D, M Fragoso, and P Ruffino. Regular conditional probability, disintegration of probability and Radon spaces. *Proyecciones*, 23(1):15–29, 2004.

[159] Y LeCun, Y Bengio, and G Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[160] Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE*, 86(11):2278–2324, November 1998.

[161] Y LeCun, L D Jackel, L Bottou, C Cortes, J S Denker, H Drucker, I Guyon, U A Muller, E Sackinger, P Simard, and V Vapnik. Learning Algorithms For Classification: A Comparison On Handwritten Digit Recognition. In J H Oh, C Kwon, and S Cho, editors, *Neural Networks: The Statistical Mechanics Perspective*, pages 261–276. World Scientific, 1995.

[162] H C Lie and T J Sullivan. Equivalence of weak and strong modes of measures on topological vector spaces. *Inverse Problems*, 34(11):115013, 2018.

[163] H C Lie, T J Sullivan, and A L Teckentrup. Random forward models and log-likelihoods in bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 6(4):1600–1629, 2018.

[164] C Lieberman, K Willcox, and O Ghattas. Parameter and state model reduction for large-scale statistical inverse problems. *SIAM J. Sci. Comput.*, 32(5):2523–2542, 2010.

[165] E A B F Lima, J T Oden, D A Hormuth II, T E Yankeelov, and R C Almeida. Selection, calibration, and validation of models of tumor growth. *Math. Models Methods Appl. Sci.*, 26(12):2341–2368, 2016.

[166] E A B F Lima, J T Oden, B Wohlmuth, A Shahmoradi, D A Hormuth II, T E Yankeelov, L Scarabosio, and T Horger. Selection and validation of predictive models of radiation effects on tumor growth based on noninvasive imaging data. *Comput. Methods Appl. Mech. Eng.*, 327:277–305, 2017.

[167] J S Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2004.

[168] W K Liu, T Belytschko, and A Mani. Random field finite elements. *Internat. J. Numer. Methods Engrg.*, 23(10):1831–1845, 1986.

[169] M Loève. *Probability theory II*. Springer, New York, Heidelberg, 1978.

[170] G J Lord, C E Powell, and T Shardlow. *An introduction to computational stochastic PDEs*. Cambridge University Press, New York, 2014.

[171] P Loskot and N C Beaulieu. On monotonicity of the hypersphere volume and area. *Journal of Geometry*, 87(1):96–98, 2007.

[172] L Machiels, Y Maday, I B Oliveira, A T Patera, and D V Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 331(2):153–158, 2000.

[173] A Manzoni, S Pagani, and T Lassila. Accurate solution of Bayesian inverse uncertainty quantification problems combining reduced basis methods and reduction error models. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):380–412, 2016.

[174] J Marion and S C Schmidler. Finite Sample Complexity of Sequential Monte Carlo Estimators. *arXiv e-prints*, 1803.09365, 2018.

[175] J Marion and S C Schmidler. Finite Sample L_2 Bounds for Sequential Monte Carlo and Adaptive Path Selection. *arXiv e-prints*, 1807.01346, 2018.

[176] C Mark, C Metzner, L Lautscham, P L Strissel, R Strick, and B Fabry. Bayesian model selection for complex dynamic systems. *Nat. Commun.*, 9:1803, 2018.

[177] Y M Marzouk and H N Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *J. Comput. Phys.*, 228(6):1862–1902, 2009.

[178] B Matérn. *Spatial Variation*. Springer, New York, second edition, 1986.

[179] H G Matthies, E Zander, B V Rosić, A Litvinenko, and O Pajonk. Inverse Problems in a Bayesian Setting. In A Ibrahimbegovic, editor, *Computational Methods for Solids and Fluids: Multiscale Analysis, Probability Aspects and Model Reduction*, pages 245–286. Springer International Publishing, Cham, 2016.

[180] S A Mattis and B Wohlmuth. Goal-oriented adaptive surrogate construction for stochastic inversion. *Comput. Methods Appl. Mech. Engrg.*, 339:36–60, 2018.

[181] P McCullagh. What is a statistical model? With comments and a rejoinder by the author. *Ann. Statist.*, 30(5):1225–1310, 2002.

[182] W S McCulloch and W Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.

[183] N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, and E Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[184] N Metropolis and S Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44(247):335–341, 1949.

[185] S P Meyn and R L Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, first edition, 1993.

[186] V Minden, A Damle, K Ho, and L Ying. Fast spatial Gaussian process maximum likelihood estimation via skeletonization factorizations. *Multiscale Model. Simul.*, 15(4):1584–1611, 2017.

[187] A Mondal, B Mallick, Y Efendiev, and A Datta-Gupta. Bayesian uncertainty quantification for subsurface inversion using a multiscale hierarchical model. *Technometrics*, 56(3):381–392, 2014.

[188] K P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, Cambridge, MA, London, 2012.

[189] G Nakamura and R Potthast. *Inverse modeling. An introduction to the theory and methods of inverse problems and data assimilation*. IOP Publishing, Bristol, 2015.

[190] R M Neal. Annealed importance sampling. *Stat. Comput.*, 11(2):125–139, 2001.

[191] R M Neal. Estimating Ratios of Normalizing Constants Using Linked Importance Sampling. *arXiv e-prints*, math/05112, 2005.

[192] A K Noor and J M Peters. Reduced Basis Technique for Nonlinear Analysis of Structures. *AIAA Journal*, 18(4):455–462, 1980.

[193] S Osborn, P S Vassilevski, and U Villa. A multilevel, hierarchical sampling technique for spatially correlated random fields. *SIAM J. Sci. Comput.*, 39(5):S543–S562, 2017.

[194] S Osborn, P Zulian, T Benson, U Villa, R Krause, and P S Vassilevski. Scalable hierarchical PDE sampler for generating spatially correlated random fields using nonmatching meshes. *Numer. Linear Algebra Appl.*, 25(3):e2146, 2018.

[195] H Owhadi, C Scovel, and T Sullivan. On the Brittleness of Bayesian Inference. *SIAM Rev.*, 57(4):566 582, 2015.

[196] I Papaioannou, C Papadimitriou, and D Straub. Sequential importance sampling for structural reliability analysis. *Struct. Saf.*, 62:66–75, 2016.

[197] B Peherstorfer, B Kramer, and K Willcox. Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):737–761, 2018.

[198] B Peherstorfer and Y Marzouk. A transport-based multifidelity preconditioner for markov chain monte carlo. *Adv. Comput. Math. (accepted)*, 2019.

[199] B Peherstorfer, K Willcox, and M Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM J. Sci. Comput.*, 38(5):A3163–A3194, 2016.

[200] B Peherstorfer, K Willcox, and M Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.*, 60(3):550–591, 2018.

[201] T Petrila and D Trif. *Basics of Fluid Mechanics and Introduction to Computational Fluid Dynamics*. Springer, Boston, MA, 2005.

[202] J Pfanzagl. *Allgemeine Methodenlehre der Statisik I*. de Gruyter, Berlin, New York, sixth edition, 1983.

[203] J Potthoff. Sample properties of random fields. II. Continuity. *Commun. Stoch. Anal.*, 3(3):331–348, 2009.

[204] S Pranesh and D Ghosh. Faster computation of the Karhunen-Loève expansion using its domain independence property. *Comput. Methods Appl. Mech. Engrg.*, 285:125–145, 2015.

[205] Y V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.*, 1(2):157–214, 1956.

[206] A Quarteroni. *Numerical Models for Differential Problems*. Springer, Cham, third edition, 2017.

[207] A Quarteroni, A Manzoni, and F Negri. *Reduced basis methods for partial differential equations. An introduction*. Springer, Cham, 2016.

[208] P Rebeschini and R van Handel. Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.*, 25(5):2809–2866, 2015.

[209] A C Rencher and G B Schaalje. *Linear models in statistics*. Wiley, Hoboken, NJ, second edition, 2008.

[210] G R Richter. An inverse problem for the steady state diffusion equation. *SIAM J. Appl. Math.*, 41(2):210–221, 1981.

[211] C P Robert. *The Bayesian Choice. From decision-theoretic foundations to computational implementation*. Springer, New York, corrected paperback reprint of the second edition, 2007.

[212] C P Robert. Reading Théorie Analytique des Probabilités. *arXiv e-prints*, 1203.6249, 2012.

[213] C P Robert and G Casella. *Monte Carlo statistical methods*. Springer, New York, second edition, 2004.

[214] L Roininen, M Girolami, S Lasanen, and M Markkanen. Hyperpriors for Matérn fields with applications in Bayesian inversion. *Inverse Probl. Imaging*, 13(1):1–29, 2019.

[215] L Roininen, S Lasanen, M Orispää, and S Särkkä. Sparse approximations of fractional Matérn fields. *Scand. J. Stat.*, 45(1):194–216, 2018.

[216] R Y Rubinstein. Optimization of computer simulation models with rare events. *Eur. J. Oper. Res.*, 99(1):89–112, 1997.

[217] R Y Rubinstein and Kroese D P. *Simulation and the Monte Carlo method*. John Wiley & Sons, Hoboken, NJ, third edition, 2017.

[218] P-B Rubio, F Louf, and L Chamoin. Fast model updating coupling Bayesian inference and PGD model reduction. *Comput. Mech.*, 62:1485–1509, 2018.

[219] W Rudin. *Real and complex analysis*. McGraw-Hill, New York, third edition, 1987.

[220] D Rudolf and N Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 11 2018.

[221] D Rudolf and B Sprungk. On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm. *Found. Comput. Math.*, 18(2):309–343, 2018.

[222] A K Saibaba, J Lee, and P K Kitanidis. Randomized algorithms for generalized Hermitian eigenvalue problems with application to computing Karhunen-Loève expansion. *Numer. Linear Algebra Appl.*, 23(2):314–339, 2016.

[223] D Sanz-Alonso. Importance sampling and necessary sample size: an information theory approach. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):867–879, 2018.

[224] F Schäfer, T J Sullivan, and H Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *arXiv e-prints*, 1706.02205, 2017.

[225] C Schillings and C Schwab. Sparsity in Bayesian inversion of parametric operator equations. *Inverse Problems*, 30(6):065007, 2014.

[226] C Schillings and A M Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, 55(3):1264–1290, 2017.

[227] C Schillings and A M Stuart. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl. Anal.*, 97(1):107–123, 2018.

[228] C Schraff, H Reich, A Rhodin, A Schomburg, K Stephan, A Periez, and R Potthast. Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Q. J. Royal Meteorol. Soc.*, 142(696):1453–1472, 2016.

[229] C Schwab and A M Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Probl.*, 28(4):45003–32, 2012.

[230] C Schwab and R A Todor. Karhunen-Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122, 2006.

[231] P Sirković. *Low-rank methods for parameter-dependent eigenvalue problems and matrix equations*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2016.

[232] R C Smith. *Uncertainty quantification. Theory, implementation, and applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014.

[233] B Sprungk. On the Locally Lipschitz Robustness of Bayesian Inverse Problems. *arXiv e-prints*, 1906.07120, 2019.

[234] I Sraj, O P Le Maître, O M Knio, and I Hoteit. Coordinate transformation and polynomial chaos for the Bayesian inference of a Gaussian process with parametrized prior covariance function. *Comput. Methods Appl. Mech. Engrg.*, 298:205–228, 2016.

[235] J Stoer and R Bulirsch. *Introduction to numerical analysis*. Springer, New York, third edition, 2002.

[236] A M Stuart. Inverse problems: a Bayesian perspective. In *Acta Numer.*, volume 19, pages 451–559. Cambridge University Press, Cambridge, 2010.

[237] A M Stuart and A L Teckentrup. Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Math. Comp.*, 87:721–753, 2018.

[238] T J Sullivan. *Introduction to Uncertainty Quantification*. Springer, 2015.

[239] T J Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874, 2017.

[240] T J Sullivan. Well-posedness of Bayesian inverse problems in quasi-Banach spaces with stable priors. In C Könke and C Trunk, editors, *88th Annual Meeting of the International Association of Applied Mathematics and Mechanics (GAMM), Weimar 2017*, volume 17(1) of *Proc. Appl. Math. and Mech.*, pages 871–874. Wiley, Weinheim, 2017.

[241] P M Tagade and H-L Choi. A generalized polynomial chaos-based method for efficient bayesian calibration of uncertain computational models. *Inverse Probl. Sci. Eng.*, 22(4):602–624, 2014.

[242] A Tarantola. *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.

[243] A L Teckentrup, P Jantsch, C G Webster, and M Gunzburger. A Multilevel Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):1046–1074, 2015.

[244] A L Teckentrup, R Scheichl, M B Giles, and E Ullmann. Further analysis of multilevel monte carlo methods for elliptic pdes with random coefficients. *Numer. Math.*, 125(3):569–600, 2013.

[245] D Teets and K Whitehead. The discovery of Ceres: how Gauss became famous. *Math. Mag.*, 72(2):83–93, 1999.

[246] U Trottenberg, C W Oosterlee, and A Schüller. *Multigrid*. Academic Press, San Diego, CA, 2001.

[247] E Ullmann and I Papaioannou. Multilevel estimation of rare events. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):922–953, 2015.

[248] F Uribe, I Papaioannou, J Latz, W Betz, E Ullmann, and D Straub. Bayesian inference with subset simulation in spaces of varying dimension. *Under review*, 2019.

[249] S Vallaghé, P Huynh, D J Knezevic, L Nguyen, and A T Patera. Component-based reduced basis for parametrized symmetric eigenproblems. *Adv. Model. Simul. Eng. Sci.*, 2:7, 2015.

[250] A W van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998.

[251] D N VanDerwerken and S C. Schmidler. Parallel Markov Chain Monte Carlo. *arXiv e-prints*, 1312.7479, 2013.

[252] C Villani. *Optimal transport. Old and new*. Springer, Berlin, 2009.

[253] L von Bortkewitsch. *Das Gesetz der kleinen Zahlen*. B. G. Teubner, Leipzig, 1898.

[254] J von Neumann. Various techniques used in connection with random digits. In A S Householder, G E Forsythe, and H H Germond, editors, *Monte Carlo Method*, pages 36–38. U.S. Government Printing Office, Washington, D.C., 1951.

[255] L Wasserman. Bayesian model selection and model averaging. *J. Math. Psych.*, 44(1):92–107, 2000.

[256] S Whitaker. Flow in porous media I: A theoretical derivation of Darcy's law. *Transport Porous Med.*, 1(1):3–25, 1986.

[257] N Whiteley. Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions. *Stoch. Anal. Appl.*, 30(5):774–798, 2012.

[258] C K Wikle. Hierarchical models for uncertainty quantification. In R Ghanem, D Higdon, and H Owhadi, editors, *Handbook of uncertainty quantification*, pages 193–218. Springer, Cham, 2017.

[259] S M Zemyan. *The classical theory of integral equations. A concise treatment.* Birkhäuser/Springer, New York, 2012.

[260] Z Zheng and H Dai. Simulation of multi-dimensional random fields by Karhunen-Loève expansion. *Comput. Methods Appl. Mech. Engrg.*, 324:221–247, 2017.

List of Figures

List of Tables

List of Abbreviations

a.e. almost every(where)

ANN Artificial Neural Network

a.s. almost surely

B bridging

BIP Bayesian inverse problem

CM Cameron–Martin

CNN Convolutional Neural Network

cv coefficient of variation

DCT Lebesgue's Dominated Convergence Theorem

DNN Deep Neural Network

ecdf empirical cumulative distribution function

EnKF Ensemble Kalman Filter

ESS effective sample size

flops floating point operation per second

i.i.d. independent and identically distributed

ITU inverse temperature update

KL Karhunen–Loève

KLD Kullback–Leibler divergence

KS Kolmogorov–Smirnoff

LU level update

MCMC Markov chain Monte Carlo

MH Metropolis–Hastings

MLB Multilevel Bridging

MLMC Multilevel Monte Carlo

MLSMC Multilevel Sequential Monte Carlo

MLS^{2}MC Multilevel Sequential^2 Monte Carlo

MWG Metropolis-within-Gibbs

ODE ordinary differential equation

pCN preconditioned Crank–Nicolson

PDE partial differential equation

pdf probability density function

POD proper orthogonal decompositiion

RB reduced basis

RWM random walk Metropolis

SDE stochastic differential equation

SIS Sequential Importance Sampling

SMC Sequential Monte Carlo

SPDE stochastic partial differential equation

SVD singular value decomposition

T Tempering

tv total variation

UQ uncertainty quantification