EPFL

Thèse n° 8711

Probabilistic and Bayesian methods for uncertainty quantification of deterministic and stochastic differential equations

Présentée le 3 septembre 2021

Faculté des sciences de base
Chaire d'analyse numérique et mathématiques computationnelles
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Giacomo GAREGNANI

Acceptée sur proposition du jury

Prof. F. Eisenbrand, président du jury
Prof. A. Abdulle, directeur de thèse
Prof. T. Lelièvre, rapporteur
Prof. E. Vanden-Eijnden, rapporteur
Prof. M. Picasso, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

To my brother

Acknowledgements

When asked, I describe the path towards obtaining a PhD as being hilly, full of ups and downs in motivation, success, and self-satisfaction. Even when the downs hit me, though, I never forgot the immense privilege that is to freely explore whichever research question popped into my head, and to scratch the surface of the infinite world of mathematics. Therefore, all in all, the ups greatly outnumbered the downs. Despite my enjoyment of these last four years, I never took the time to really thank the people who guided me, who walked alongside me, and who put me in a good position to start with: *I guess that this must be the place.*

First and foremost, I thank my advisor *Monsieur le Professeur* Assyr Abdulle, and for numerous reasons. First, thank you for proposing an interesting research project, which kept me captivated throughout my doctoral studies. Second, thank you for sharing your unbounded knowledge on the integration of differential equations, homogenization theory, and uncertainty quantification among others. Moreover, thank you for pushing me towards the pursuit of mathematical rigour, clarity, and *elegance* in the exposition of complex concepts. Furthermore, thank you for teaching me how to be a good teacher, and for being an example of how much an instructor should care about his students. Finally, thank you for being an exceptional boss, always careful that everyone in your group can thrive and enjoy their time to the fullest (*On fait mieux de la science quand on rigole*, cit).

I'm thankful to professors Tony Lelièvre, Marco Picasso, and Eric Vanden-Eijnden for accepting to be on the jury of my oral examination, and for sharing their thoughts on this thesis. I moreover thank professor Friedrich Eisenbrand for presiding the jury.

My deepest gratitude goes to professors Greg Pavliotis and Andrew Stuart, with whom I had the privilege to collaborate scientifically and who welcomed me during scientific visits. Thanks are also due to professor Philipp Hennig and my friend professor Sebastian Krumscheid, for inviting me to give seminars in their institutes.

I'd like to thank my colleagues of ANMC: Adrian, Andrea DB, Andrea Z, Doghonay, Edoardo P, Giacomino, Orane, and Simon, for always participating in the good atmosphere of the lab. I need to spend some extra words on Timothée: thanks for helping me through my first year as a PhD student, for scaling down my vastly insignificant issues with your good humour, and ultimately for our friendship. Virginie deserves all the praise in the world for her wizardly administrative skills, and for rescuing me from typhoon Maria in 2018.

A big thank-you goes to professor Simone Deparis, for the two semesters spent teaching together, and for the skiing classes in Verbier.

It feels natural to mention here my peers Simon Bartels and Hans Kersting, with whom I shared interesting discussions about mathematics (and not mathematics) on multiple occasions during our almost-synchronized doctoral journeys. Moreover, I thank professor Andrew Holbrook for the good times in Dobbiaco and LA.

My life aside the PhD would not have been as light-hearted, fun, and interesting without the group

of extraordinary people whom I am proud to call my friends. Some have already been mentioned above. The others: Elisabetta, Eva, Fabian, Filippo, Gonzalo, Luca, Philippe, Riccardo, and Stéphane. I especially thank Edoardo R, whose friendship has been a solid and reassuring certainty for the last ten years.

I may not have pursued doctoral studies without the support and encouragement I have always received from my family. Therefore, I thank my parents, for teaching me the importance of culture and curiosity, and my brother, for leading the way.

To conclude, thank you, Victoria, for your patience, your affection, and the laughs. Some steps are already behind our backs, I look forward to walking many more by your side.

Lausanne, June 23, 2021 GG

Abstract

In this thesis we explore uncertainty quantification of forward and inverse problems involving differential equations. Differential equations are widely employed for modeling natural and social phenomena, with applications in engineering, chemistry, meteorology, and economics. Mathematical models of complex systems in these fields require numerical methods, which introduce uncertainties in the outcome. Moreover, there has recently been a steep rise in the availability of data, which also come with an uncertainty. Therefore, blending mathematical models, data, and their respective uncertainties is nowadays of the utmost importance.

The first part of this thesis is dedicated to two novel methods for multiscale inverse problems. We first consider an elliptic partial differential equation (PDE), with a diffusion tensor oscillating at a small scale. Given noisy observations of the solution, we consider the problem of inferring a slow-scale parametrization of the multiscale tensor. For this purpose, we combine numerical homogenization, which yields a single-scale surrogate of the full model, and the ensemble Kalman filter. The scheme we propose is accurate in the homogenized limit, and outperforms existing methods in terms of computational cost. We then study the error due to the mismatch between the full and the homogenized models, and show how to combine statistical techniques for model misspecification and our scheme. We then move to multiscale diffusion processes, and consider the problem of inferring effective dynamics from multiscale observations. A homogenized single-scale equation reproducing the full model exists also in this case. The effective model, though, is the subject of the inference procedure, and not only a computational tool. The resulting issue of model misspecification is usually bypassed by subsampling at an appropriate rate, which is non-trivial to choose, and which may give misleading results. We avoid subsampling by designing a novel technique based on filtered data, and show how to modify classical estimators and obtain an effective equation consistently with homogenization. Our technique is robust and can be employed as a black-box tool for inferring effective surrogates of complex stochastic models.

In the second part we present two novel schemes belonging to the field of probabilistic numerics, whose purpose is to provide a statistical description of the uncertainty due to numerical discretization. We first consider ordinary differential equations (ODEs), and introduce a probabilistic integrator based on random time steps and Runge–Kutta methods (RTS-RK). Tuning the distribution of the time steps, we generate a probability measure on the solution which allows for a consistent uncertainty quantification of numerical errors. Unlike previous probabilistic methods in literature, our scheme inherits the geometric properties of the underlying deterministic integrators. In particular, we show long-time energy conservation when the RTS-RK is applied to Hamiltonian ODEs. We employ the idea of randomizing the discretization to propose a random mesh finite element method (RM-FEM) for elliptic PDEs. We prove that the measure induced by the RM-FEM on the solution can be employed to derive a posteriori error estimators. Hence, the RM-FEM provides a consistent statistical characterization of numerical errors. For both our novel schemes, we demonstrate the usefulness of the probabilistic approach in Bayesian inverse problems.

Abstract

Key Words: Uncertainty quantification, Inverse problems, Multiscale differential equations, Probabilistic numerical methods, Model misspecification.

Sommario

L'argomento di questa tesi è la quantificazione dell'incertezza in problemi diretti e inversi per equazioni differenziali. Le equazioni differenziali sono ampiamente utilizzate per la modellazione di fenomeni naturali e sociali, con applicazioni in ingegneria, chimica, meteorologia ed economia. In questi campi, si è registrato recentemente un forte aumento nella disponibilità di dati. Pertanto, combinare i modelli matematici, i dati e le loro rispettive incertezze è oggi della massima importanza. Inoltre, è spesso necessario impiegare metodi numerici per risolvere sistemi complessi, il che introduce ulteriori incertezze nel risultato.

La prima parte è dedicata a due nuovi metodi per problemi inversi multiscala. Il primo problema riguarda un'equazione ellittica alle derivate parziali (EDP), con un tensore di diffusione rapidamente oscillante a una scala ε, e in particolare l'inferenza di una parametrizzazione lenta del tensore multiscala date osservazioni corrotte della soluzione. L'omogeneizzazione numerica, che produce un surrogato del modello completo, è combinata con il filtro di Kalman d'insieme. Lo schema risultante è accurato nel limite $\varepsilon \to 0$, ed è computazionalmente vantaggioso rispetto a metodi esistenti. L'errore dovuto al disallineamento tra il modello completo e quello omogeneizzato è trattato con una tecnica statistica combinata in maniera ottimale con il nostro schema. Il secondo problema concerne processi di diffusione multiscala e l'inferenza di dinamiche effettive date osservazioni multiscala. Anche in questo scenario, esiste un'equazione omogeneizzata che sintetizza il modello completo. L'equazione efficace, però, è l'oggetto dell'inferenza, e non uno strumento di calcolo. Il problema di discordanza nel modello risultante è solitamente aggirato campionando i dati a un tasso appropriato, che non è scontato da scegliere e che può dare risultati fuorvianti. Il campionamento è aggirato tramite una nuova tecnica basata sul filtraggio dei dati, che permette di ottenere un'equazione efficace coerente con l'omogeneizzazione. La tecnica proposta è robusta e può essere impiegata in modo diretto per dedurre surrogati semplici di modelli stocastici complessi.

La seconda parte è dedicata a due nuovi metodi numerici probabilistici, il cui scopo è caratterizzare l'incertezza data dalla discretizzazione numerica in modo statistico. Il primo problema riguarda equazioni differenziali ordinarie (EDO), per le quali è introdotto un integratore probabilistico basato su passi temporali casuali e metodi Runge–Kutta (RTS-RK). Controllando la distribuzione dei passi temporali, è possibile generare una misura di probabilità sulla soluzione che quantifica l'incertezza dovuta agli errori numerici. A differenza di precedenti metodi probabilistici in letteratura, il metodo RTS-RK eredita le proprietà geometriche degli integratori deterministici sottostanti. In particolare, l'energia è conservata per intervalli di tempo lunghi in EDO hamiltoniane. L'idea di randomizzare la discretizzazione può essere utilizzata anche per EDP ellittiche, per le quali è introdotto un metodo degli elementi finiti a maglia casuale (RM-FEM). La misura indotta dal metodo RM-FEM sulla soluzione è impiegata per ricavare stimatori a posteriori dell'errore. Il metodo RM-FEM fornisce quindi una caratterizzazione statistica consistente degli errori numerici. Per entrambi i nuovi metodi, è dimostrata l'utilità dell'approccio probabilistico nei problemi inversi bayesiani.

Sommario

Parole Chiave: Quantificazione dell'incertezza, Problemi inversi, Equazioni differenziali multi-scala, Metodi numerici probabilistici, Discordanza di modello.

Notation

Standard sets of numbers

$\mathbb{N}$	set of positive integers
$\mathbb{R}$	set of real numbers

Differentials

∇	gradient operator
$\nabla\cdot$	divergence operator
Δ	Laplacian operator
∇^2	Hessian operator

Functional spaces

Let D be an open domain of $\mathbb{R}^d$, d a positive integers, and consider functions $D \to \mathbb{R}$.

$C^k(D)$	space of k-times continuously differentiable functions
$L^p(D)$	usual Lebesgue space with $p \in [1, \infty]$
$W^{k,p}(D)$	usual Sobolev space with $k \in \mathbb{N}$ and $p \in [1, \infty]$
$H^k(D)$	Sobolev space $W^{k,2}(D)$
$H^1_0(D)$	closure in $H^1(D)$ of $C^\infty(D)$ functions with compact support in D

Probability theory

Ω	event space
$(\Omega, \mathcal{A})$	measurable space, also $(\Omega, \mathcal{F})$
$(\Omega, \mathcal{A}, P)$	probability space, also $(\Omega, \mathcal{F}, P)$
$\mathcal{B}(\mathcal{H})$	Borel σ-algebra of a metric space $\mathcal{H}$
$X \colon \Omega \to \mathcal{H}$	with $\mathcal{H}$ a metric space and X measurable, $\mathcal{H}$-valued random variable
$\mu_X \colon \mathcal{B}(\mathcal{H}) \to [0, 1]$	probability measure induced by random variable $X \colon \Omega \to \mathcal{H}$
a.s.	almost surely. If $A \in \mathcal{A}$ and $P(A) = 1$ then A a.s.
$L^p(\Omega)$	Lebesgue space of p-integrable random variables defined on Ω

Acronyms

ODE	ordinary differential equation
PDE	partial differential equation
SDE	stochastic differential equation
FEM	finite element method
FE-HMM	finite element heterogeneous multiscale method
MCMC	Markov chain Monte Carlo
KL	Karhunen–Loève as in KL expansion
BKE	Backward Kolmogorov equation
FPE	Fokker–Planck equation

Contents

Introduction

The keyword in the title of this thesis – *Probabilistic and Bayesian methods for uncertainty quantification of deterministic and stochastic differential equations* – is certainly "uncertainty quantification" (UQ). The field of UQ is broad, and finds applications ranging in various and diverse fields of natural and social sciences, such as geology, meteorology, oceanography, economics, and medicine. In this thesis, we focus on forward and inverse UQ for differential problems and their numerical solution.

Let us consider an input/output relation, that we will often call in this thesis a forward map. We are especially interested here in forward maps which involve the solution of a deterministic or stochastic differential equation. We then refer to forward UQ when we model the uncertainty due to the approximations which are necessary to reproduce the forward map. Conversely, we refer to inverse UQ in contexts of data assimilation, that is when we give a reconstruction of the input of the forward map modeling a phenomenon given observations of its outcome.

It is customary to adopt a UQ approach for inverse problems. Indeed, inverse problems are often ill-posed and corrupted by noise, which makes a UQ approach not only recommendable, but almost indispensable. Moreover, with data-driven applications dominating nowadays most fields of applied sciences, a statistical approach guarantees fast and reliable assimilation of data into models for analytic and predictive purposes. We mainly consider in this thesis, and in particular in Part I, inverse UQ for forward maps involving multiscale differential equations, i.e., equations whose parameters vary on different and separated scales. In this field, it is often possible to obtain a single-scale effective representation of the full multiscale problem, which allows for fast evaluations of the otherwise prohibitively-expensive forward map. Additionally to the corruption of the data and the ill-posedness, we therefore employ a UQ approach to deal with the further difficulty stemming from the mismatch between the multiscale model and its effective representation.

Forward UQ for deterministic differential problems is far less prominent in applied sciences. When one is confronted with the problem of solving a differential equation, the common approach is to apply the computationally cheapest numerical scheme to obtain a solution up to a tolerance in its uncertainty. In this thesis, and in particular in Part II, we present two methods for forward UQ belonging to the rapidly-emerging field of probabilistic numerics (PN), whose purpose is to quantify the uncertainty due to approximate computations in a statistical manner, rather than with traditional point estimates. The PN approach for deterministic forward problems can be helpful in a wide range of applications, and especially when the output of the forward map serves as the input for a subsequent problem in a complex system.

In the remainder of the introduction we present separately the two parts that compose this thesis. For both parts, we follow the same strategy: first we present the general framework of the topic of interest, then we give an overview of the state of the art and of the relevant literature, and finally we conclude by outlining what are our main contributions.

Part I: Multiscale Inverse Problems and Parameter Inference

In the first part of the thesis we introduce methods to solve inverse problems featuring multiple scales. Let X and Y be an input and an output space, respectively, let $\varepsilon > 0$ be a small real number, which we call the scale-separation parameter, and let us call the forward map a function $\mathcal{G}^\varepsilon \colon X \to Y$ depending on the scale-separation parameter ε. We are then interested in the solution of inverse problems of the form

$$\text{find } u \in X \text{ given observations } y = \mathcal{G}^\varepsilon(u) + \beta \in Y, \tag{1}$$

where β is a possibly degenerate Y-valued random variable modeling observational noise. Solving (1) engenders two main difficulties. First, the inverse problem (1) is ill-posed. Indeed, a possible mismatch between the dimension of the input and output spaces and the randomness of the noise prevent to define a unique solution $u \in X$ to (1). Ill-posedness is traditionally circumvented by employing some form of regularization or by recasting the problem in a Bayesian framework [131]. The second difficulty is due to the multiscale essence of the problem. In particular, we are interested in forward models $\mathcal{G}^\varepsilon$ which are highly-oscillatory at a small scale ε and which require a numerical routine to be evaluated. The level of refinement in the numerical discretization is constrained by the smallest scale ε, which in many applications leads to a severe increment in computational cost. In the settings we consider in the following, the model summarized by the forward map $\mathcal{G}^\varepsilon$ can be replaced by a surrogate $\mathcal{G}^0$, which approximates $\mathcal{G}^\varepsilon$ when ε is small, and which is cheaper to evaluate numerically. Such a surrogate model can have two roles: either one employs $\mathcal{G}^0$ as a computational tool to solve the inverse problem and retrieve the full multiscale model, or one can desire to directly fit the cheaper surrogate to the given multiscale data, thus obtaining a well-calibrated simple model for prediction purposes. In both scenarios, though, we are confronted with a problem of model misspecification. Indeed, in the first case we have to deal with the difference between $\mathcal{G}^\varepsilon$ and $\mathcal{G}^0$ in the computational process that leads to the solution of (1), while in the second case we have to make sure that the inferred simple model is a valid surrogate for the full observed phenomenon.

An Inverse Problem Involving Multiscale PDEs. Let us be more specific and introduce the first inverse problems which we treat in the first part of the thesis. Given a d-dimensional domain $D \subset \mathbb{R}^d$, we consider the elliptic partial differential equation (PDE)

$$\begin{aligned} -\nabla \cdot (A_u^\varepsilon \nabla p^\varepsilon) &= f, \quad \text{in } D, \\ p^\varepsilon &= 0, \quad \text{on } \partial D, \end{aligned} \tag{2}$$

for a given force term $f \colon D \to \mathbb{R}$ and for an elliptic diffusion tensor $A_u^\varepsilon \colon D \to \mathbb{R}^{d\times d}$, highly oscillatory at a micro scale of characteristic size ε. We suppose that the macro and micro scales in the tensor A_u^ε are clearly separated, and in particular that it holds

$$A_u^\varepsilon(x) = A\left(u(x), \frac{x}{\varepsilon}\right),$$

for a function $A \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{d\times d}$ which is periodic in its second argument. In Fig. 1 we represent graphically an example of such a tensor, and its associated solution p^ε of (2). In particular, we consider $D = (0,1)^2$, the force $f = 1$ in D, and

$$\begin{aligned} u(x) &= 1 - \frac{9}{10}\sin(2\pi x_1)\sin(2\pi x_2), \\ A_u^\varepsilon(x) &= \left(1 + \cos\left(\frac{2x_1}{\varepsilon}\right)^2 + \cos\left(\frac{2x_2}{\varepsilon}\right)^2\right) u(x)\, I, \end{aligned}$$

where I is the identity matrix on $\mathbb{R}^2$ and where we fix $\varepsilon = 1/16$. The inverse problem we are interested in is then to retrieve the function u, controlling the macro-scale variations of A_u^ε,

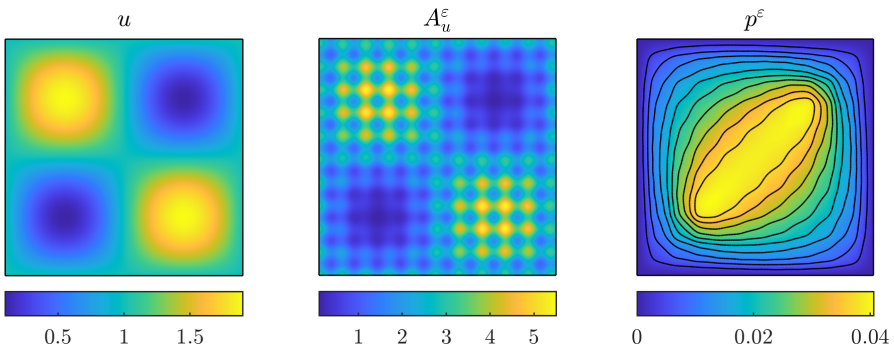


Figure 1 – Example of a multiscale elliptic PDE of the form (2). On the left, the function u controlling the slow-scale variations of the diffusion field A_u^ε, given in the center. On the right, the solution p^ε of (2).

given a discrete set of observations derived from the solution p^ε. Summarizing, we can write the problem in the form (1) by considering $\mathcal{G}^\varepsilon(u) = \mathcal{O}(p^\varepsilon)$, where p^ε depends on u through the PDE (2) and where $\mathcal{O}\colon p^\varepsilon \mapsto y$ is a given observation operator. Assuming the scale-separation parameter ε and the map $(t, x) \mapsto A(t, x/\varepsilon)$ to be known, solving this inverse problem thus allows to retrieve the full multiscale diffusion tensor.

As we will see in the remainder of this thesis, solving inverse problems of the form (1) often involves evaluating the forward map $\mathcal{G}^\varepsilon$ on several values $u \in X$, and therefore in this specific instance solving the PDE (2) multiple times. Unfortunately, in order to obtain convergent approximations of the solution p^ε it is necessary to discretize the domain D at a level of refinement which allows to resolve A_u^ε at its micro scale ε. In practice, if we have a computational mesh of D with characteristic size h, then we need to impose $h \ll \varepsilon$. This numerical constraint clearly leads to an unbearable computational cost in case $\varepsilon \ll |D|$, which is a realistic scenario in a wide range of applications. Fortunately, the theory of homogenization [22, 34] comes to our aid, and guarantees that there exists an elliptic single-scale tensor A_u^0 such that the solution p^0 of the PDE

$$\begin{aligned} -\nabla \cdot \left(A_u^0 \nabla p^0\right) &= f, \quad \text{in } D, \\ p^0 &= 0, \quad \text{on } \partial D, \end{aligned} \tag{3}$$

is the weak limit of the solution p^ε of (2) for $\varepsilon \to 0$. Therefore, in case $\varepsilon \ll |D|$ we can replace the full model (2) by the homogenized equation (3), which can be solved numerically with reasonable computational resources. In particular, given a multiscale tensor A_u^ε, there exist numerical schemes such as the finite element heterogeneous multiscale method (FE-HMM) [2, 5] which allow to approximate the homogenized tensor A_u^0 and thus the solution p^0 of (3). Hence, the forward map $\mathcal{G}^\varepsilon$ can be replaced in (1), too, by the homogenized forward map $\mathcal{G}^0\colon u \mapsto \mathcal{O}(p^0)$, where $\mathcal{O}$ is the same observation operator as above, and where the dependence of p^0 on u is defined through (3). It is then interesting to study what are the effects of employing the cheaper surrogate $\mathcal{G}^0$ instead of the full model $\mathcal{G}^\varepsilon$ in the solution of the inverse problem.

On top of the mathematical relevance per se of the topic introduced above, inverse problems involving multiscale PDEs are central in several applied fields, such as thermal engineering, geology, and medical sciences. Electrical impedance tomography [30], a technique for non-invasive medical imaging and underground inspection among others, is one of the uttermost applications. In this case, the operator $\mathcal{O}\colon p^\varepsilon \mapsto y$ yields discrete observations of the outwards flow of p^ε on portions of the boundary ∂D of the domain, and the inverse problem follows Calderón's mathematical formulation [27].

Parameter Inference for Multiscale SDEs. The second problem we consider involves the d-dimensional multiscale Itô stochastic differential equation (SDE)

$$\mathrm{d}X_t^\varepsilon = -\nabla V(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}\nabla p\left(\frac{X_t^\varepsilon}{\varepsilon}\right)\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \qquad X_0^\varepsilon = x \in \mathbb{R}^d, \tag{4}$$

where we call $V\colon \mathbb{R}^d \to \mathbb{R}$ the slow-scale confining potential, the function $p\colon \mathbb{R}^d \to \mathbb{R}$ the fast-scale potential, which we assume to be periodic, where $\sigma > 0$ is the diffusion coefficient and where W_t is a d-dimensional standard Brownian motion. Equation (4) describes the motion of a particle, whose position is given by the stochastic process X_t^ε, subject to the multiscale potential

$$V_\varepsilon(x) = V(x) + p\left(\frac{x}{\varepsilon}\right),$$

and perturbed by a source of Brownian noise. Let us remark that the SDE (4) is often referred to in literature as an overdamped Langevin equation. The multiscale structure of this problem is unveiled by introducing the rescaled process $Y_t^\varepsilon \coloneqq X_t^\varepsilon/\varepsilon$, so that (4) can be rewritten as the system

$$\begin{aligned}
\mathrm{d}X_t^\varepsilon &= -\nabla V(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}\nabla p\left(Y_t^\varepsilon\right)\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, & X_0^\varepsilon &= x \\
\mathrm{d}Y_t^\varepsilon &= -\frac{1}{\varepsilon}\nabla V(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon^2}\nabla p\left(Y_t^\varepsilon\right)\,\mathrm{d}t + \sqrt{\frac{2\sigma}{\varepsilon^2}}\,\mathrm{d}W_t, & Y_0^\varepsilon &= \frac{x}{\varepsilon}.
\end{aligned}$$

The variations of the fast component Y_t^ε of the system above are then one order of magnitude faster – with respect to the scale-separation parameter ε – than those of the slow component X_t^ε. In Fig. 2 we show the multiscale potential $V_\varepsilon(x) = x^2/2 + \sin(x/\varepsilon)$, where $\varepsilon = 1/10$, and a realization of the corresponding stochastic process X_t^ε, generated on the time interval $0 \le t \le 100$, given an initial condition $X_0^\varepsilon = 4$ and for a diffusion coefficient $\sigma = 0.4$. Loosely speaking, the process X_t^ε drifts towards the global minimum of the potential V_ε, jumping out of its local minima due to the Brownian term.

Similarly to the multiscale PDE (2), simulating one or multiple trajectories of the process X_t^ε up to some final time $T > 0$ can lead to unbearable computational cost in case $\varepsilon \ll T$. Again as in the PDE case, the theory of homogenization applies to the SDE (4) and it is possible to obtain coarse-grained models approximating the solution when ε is small. Let us narrow our interest to multiscale SDEs of the form (4) with the potential V given by

$$V(x) = \sum_{i=1}^{N} \alpha_i V_i(x), \tag{5}$$

where N is a positive integer and where for $i = 1, \dots, N$ the coefficients α_i are scalars and the functions $V_i\colon \mathbb{R}^d \to \mathbb{R}$ are independent of the scale-separation parameter ε. In this setting, there exists a matrix $K \in \mathbb{R}^{d\times d}$ such that the solution X_t^0 of the SDE

$$\mathrm{d}X_t^0 = -\sum_{i=1}^{N} \alpha_i K\,\nabla V_i(X_t^0)\,\mathrm{d}t + \sqrt{2\sigma K}\,\mathrm{d}W_t, \qquad X_0^0 = x, \tag{6}$$

where W_t is the same d-dimensional Brownian motion as in (4), is the weak limit of X_t^ε for $\varepsilon \to 0$ [22, Chapter 3]. Let us remark that the concept of weak convergence is different in the SDE and the PDE cases. Indeed, in PDEs the weak limit is meant in a functional sense, whereas here the homogenization result holds in a probabilistic manner.

In this setting, we are interested in the inverse problem of determining the effective drift coefficients $A_i \coloneqq \alpha_i K$ of the homogenized SDE (6) given observations from the full multiscale model (4). In

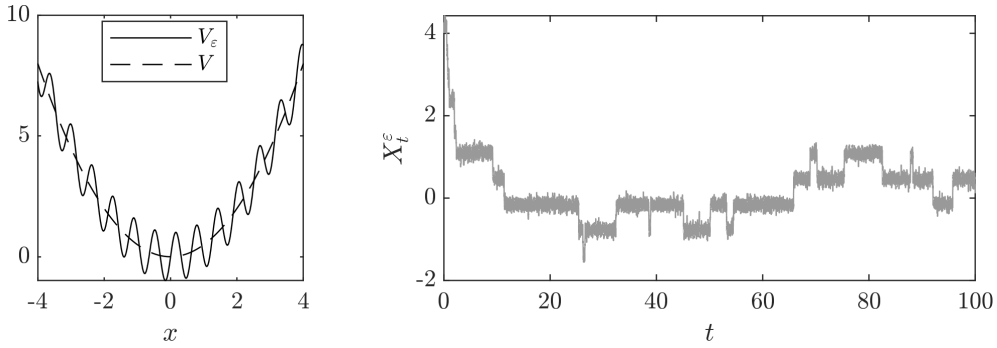


Figure 2 – Example of a multiscale SDE of the form (4). On the left, the multiscale potential V_ε, in this case given by $V_\varepsilon(x) = x^2/2 + \sin(x/\varepsilon)$. On the right, a realization of the solution X_t^ε of (4) for $0 \leq t \leq 100$.

particular, we consider data in the form of a continuous trajectory $X^\varepsilon := (X_t^\varepsilon, 0 \leq t \leq T)$ of the solution of (4) up to a final time $T > 0$. Denoting by $\alpha \in \mathbb{R}^N$ the vector of the drift coefficients appearing in (5), the forward model could be written as $\mathcal{G}^\varepsilon \colon \alpha \mapsto X^\varepsilon$, where X^ε depends on α through (4). Let us remark that due to the presence of Brownian noise, the forward model itself is random. Denoting by $A = \{A_i\}_{i=1}^N$ the coefficients $A_i = \alpha_i K$ of (6), the inverse problem then reads

$$\text{find } A \text{ given observations } X^\varepsilon = \mathcal{G}^\varepsilon(\alpha), \tag{7}$$

where with a slight abuse of notation we confound the random variable X^ε and one of its realizations. This inverse problem only partly fits into the definition (1). Indeed, the parameter we infer does not coincide with the one that we suppose to generate the observations, and we are confronted with a problem of model misspecification. Moreover, we assume to observe exactly the path X^ε, i.e., the noise β in (1) to be degenerate.

Efficient parameter estimation for stochastic models is essential in a wide range of applications in natural and social sciences. In several areas, the data originate from phenomena which vary continuously in time and which are endowed with a multiscale structure, and therefore fit the setting introduced above. This is the case, for example, in molecular dynamics, oceanography and atmosphere science or in econometrics. In all those areas, extracting from data a simple effective equation which describes the dominant slow dynamics is paramount, and leads to cheap models which can be employed for analysis and prediction.

Comparison of the two Problems. The two problems we introduced above are similar: both involve a multiscale differential equation and both hint at the theory of homogenization, at some stage. We could therefore say that the two settings belong beneath the wide umbrella of "multiscale inverse problems". There are though some differences, both in spirit and in mathematical formulation, which we highlight in the following.

The first fundamental difference is the role of the homogenized model. In the inverse problem involving PDEs, the effective equation (3) is a computational tool. Indeed, the full multiscale equation is in some situations impossible to solve numerically, and employing the single-scale model is eventually the only viable alternative. Conversely, in the SDE setting the object of interest of the inference procedure is the effective equation (6) itself. In fact, problem (7) can be interpreted as an instance of data-driven homogenization: the coarse-grained model is not deduced analytically or numerically, but via an inference procedure. In any case, in both scenarios we have to deal with an issue model misspecification and account for the difference between the multiscale and homogenized equations when solving the inverse problem.

Another difference, more related to the mathematical formulation of the two problems, is related to dimensionality. The unknown u of the PDE inverse problem has to be identified in an infinite-dimensional space from finite-dimensional noisy information, whereas in the SDE case we are interested in inferring a finite set of coefficients given infinite-dimensional observations. Hence, while the former is an instance of a non-parametric inverse problem, the latter could be referred to as a parameter inference problem in a statistical context, instead of an inverse problem. Nevertheless, taking the limit for $N \to \infty$ in the sum appearing at the right-hand side of (5) and choosing $\{V_i\}_{i\geq 1}$ to be the basis of a functional space X would bring us back to an infinite-dimensional setting. Still, the limit for $N \to \infty$ entails theoretical difficulties that we do not address in this thesis, and we thus only have considered a truncated sum in (5). Since N is arbitrary in our analysis, though, the setting we consider lays between the fully parametric – where N is fixed – and the non-parametric – where $N = \infty$ – cases, and is therefore referred to in literature as semi-parametric.

Literature Review. The issue of model misspecification in the context of inverse problems involving multiscale PDEs has already been considered in the literature. In particular, it has been shown that it is possible to infer a coarse-grained equation from data coming from the full model and to retrieve, in the large data limit, the correct result [100]. The papers [3, 4] focus instead on the same inverse problem which is presented above, i.e., on retrieving the slow-scale parametrization of the multiscale coefficient, and hence the full model. In particular, in the paper [3] the authors solve the issue of ill-posedness of the inverse problem by applying Tikhonov regularization, whereas in [4] they adopt a Bayesian approach for the same purpose. Let us moreover remark that the issue of model misspecification in PDE-driven inverse problems is an active area of research regardless of the multiscale setting. We cite [28, 29], where the authors propose statistical techniques to treat the modeling error as an additional source of random noise. Estimating the statistics of the modeling error and blending them with those of the observational noise, they are able to solve effectively inverse problems such as the one we consider here, with a special regard to applications in electrical impedance tomography. Let us furthermore notice that the techniques of [28, 29] have already been successfully applied as a black-box tool in [3] to multiscale inverse problems involving elliptic PDEs.

In the context of multiscale SDEs, the effect of model misspecification was studied in a series of papers [17, 18, 56, 57, 109, 111, 112] under the assumption of scale separation. We notice that all these contributions are particularly centered on simple applications in molecular dynamics (see e.g. [84] for a review). In particular, for Brownian particles moving in two-scale potentials it was shown that, when fitting data from the full dynamics to the homogenized equation, the maximum likelihood estimator (MLE) is asymptotically biased [112, Theorem 3.4]. To be more precise, in the large sample size limit, the data remains consistent with the multi-scale problem at small scale. Ostensibly this would seem related only to the estimation of the diffusion coefficient. However, because of detail balance, it also has the effect that the MLE for the drift in a parameter fit of a single-scale model, incorrectly identifies the coefficient of the homogenized equation. The bias of the MLE can be eliminated by subsampling at an appropriate rate, which lies between the two characteristic time scales of the problem [112, Theorems 3.5 and 3.6].

Similar techniques can be employed in econometrics, in particular for the estimation of the integrated stochastic volatility in the presence of market microstructure noise. In this case, too, the data have to be subsampled at an appropriate rate [14,104]. The correct subsampling rate can, in some instances, be rather extreme with respect to the frequency of the data itself, resulting in ignoring as much as 99% of the time-series. As the intuition suggests, this increases significantly the variance of the estimator, which is usually taken care of with additional bias corrections and variance reduction procedures. The need of such methodology is accentuated by data being obtained at high-frequency [13, 146]. Moreover, the problem of extracting large-scale variations from multiscale data is studied in atmosphere and ocean science. In this field, too, subsampling

the data is necessary to obtain an accurate coarse-grained model [40, 145].

The necessity to subsample the data can be alleviated by using appropriate martingale estimators, as was done in [72, 81]. This class of estimators can be applied to the case where the noise is multiplicative and also given by a deterministic chaotic system, as opposed to white noise. Estimators of this family have been applied to time series from paleoclimatic data and marine biology and augmented with appropriate model selection methodologies [82].

In case the data consists of discrete observations and not of continuous samples from the SDE solution, it is possible to employ estimators based on a spectral decomposition of the generator of the SDE. Methodologies of this kind have been applied successfully to inference problems for single-scale SDEs [44, 45, 78], for jump diffusions [46], as well as for multiscale SDEs [47]. A technique that combines filtered data and a spectral decomposition of the generator for multiscale SDEs has been developed recently in [11].

Inference of diffusion processes can be naturally performed under a Bayesian perspective. If one focuses on the drift coefficient, the form of the likelihood function guarantees, under a Gaussian prior hypothesis, that the posterior distribution is itself a Gaussian. The versatility of the Bayesian approach in the infinite-dimensional case [49, 131] gives the possibility to extend the study of inferring the drift of a diffusion process to the non-parametric case [115, 116].

Main Contributions. In the first part of this thesis, we introduce two novel methodologies for multiscale inverse problems. In particular, our main focus is alleviating the issue of model misspecification which affects both the PDE and the SDE models presented above, and which is due to the mismatch between the multiscale and the homogenized equations. We remark that the two methods are based respectively on our articles [9] and [8].

The first contribution of this thesis is a novel scheme based on numerical homogenization and on the ensemble Kalman filter (EnKF) to solve inverse problems for multiscale PDEs such as (2). The EnKF, first introduced in [53], is a Monte Carlo approximation of the standard Kalman filter [73], and is widely employed in the engineering community for the estimation of the state of partially-observed dynamical systems governed by a nonlinear agent. Let us consider discrete dynamics of the form

$$z_{n+1} = \Xi(z_n), \tag{8}$$

where $z_n \in Z$ for a vector space Z and where $\Xi \colon Z \to Z$ is a nonlinear map. Assume that incomplete and corrupted observations of z_n are available in the form $y_n = Hz_n + \beta_n$ in another vector space Y, for a linear function $H \colon Z \to Y$ and for a source of noise β_n. Then, the EnKF proceeds by updating recursively an ensemble of particles in Z, such that at time n the ensemble summarizes statistically the knowledge on the state z_n. In particular, the prediction of the ensemble blends at each update the dynamics (8) with observations in a Bayesian-like procedure. Kalman filters have long been used successfully in meteorology, oceanography and automation applications.

In [68], the authors propose the application of the EnKF method to obtain a point-wise solution to inverse problems involving PDEs. The technique they introduce, which they call ensemble Kalman inversion (EKI), is based on an appropriate definition of the space Z and of the dynamics (8) above, which are then employed to create a collection of replicas of approximate solutions of the inverse problem (1). The advantages of the EKI with respect to other techniques for inverse problems are twofold. First, it is possible to control the number of evaluations of the forward map by parallel computations. Second, the empirical measure induced by the ensemble of particles allows to fit the EKI into the Bayesian paradigm for the solution of (1) [125].

In this thesis, specifically in Chapter 2, we combine the well-established techniques of homog-

enization and the EKI to build a novel scheme for solving multiscale inverse problems in an efficient and reliable manner. Inspired by [3, 4], we first replace the expensive multiscale PDE (2) with its homogenized surrogate (3), which is simultaneously computed and solved numerically via the FE-HMM. We then apply the EKI employing such a cheap numerical and homogenized approximation of the PDE model, thus benefiting from both the computational savings of homogenization – in the discretization of the PDE – and of the EKI – in the solution of the inverse problem. We present a rigorous analysis of this novel numerical scheme, both from a point-wise and from a Bayesian perspectives, and prove results of convergence in the homogenization regime, i.e., when the scale-separation parameter ε vanishes, and with respect to the refinement of the FE-HMM.

A further original contribution presented in Chapter 2 consists of combining our new method and the statistical techniques of [28, 29] for assimilating modeling error into inverse problems. Let us remark that these techniques had already been applied in [4] to the multiscale setting. Here, we prove two novel results which allow to quantify the minimal number of solves of the full model so that the modeling error is approximated to a desired accuracy. These results are of the uttermost importance, since they apply to cases when $0 \ll \varepsilon \ll |D|$, i.e., for "mid-range" values of the scale-separation parameter, where the homogenization result does not hold in practice, and where the full model is still too expensive to be employed. In other words, we can tackle in this case the issue of model misspecification without renouncing to neither accuracy nor computational efficiency.

The second contribution of this thesis is a methodology based on filtered data for efficient estimation of the effective drift parameter of multiscale SDEs such as (4). As it is highlighted in the literature review above, most methods for avoiding the biasedness in the inference which is caused by model misspecification are based to some extent on subsampling, which has been proven to be successful on a wide range of applications. Still, subsampling the data presents some difficulties and issues which are arduous to circumvent. Indeed, the subsampling width should be chosen, as stated above, between the slow and the fast characteristic time scales of (4), but no clearer indication is given by the theory to this regard. In practice, the inference results are extremely sensitive to the subsampling width, which undermines the robustness of this method. Moreover, the scale-separation parameter ε may not be known in advance, which makes it impossible to fix a priori a subsampling rate which accounts correctly for the model misspecification. Finally, subsampling may lead to estimators featuring a high variance due to the great amount of data that is discarded in the process.

In this thesis, and in particular in Chapter 4, we bypass subsampling by designing a methodology based on filtered data. Let us consider for simplicity equation (4) in the one-dimensional case, so that X_t^ε is a scalar-valued stochastic process. We then consider a kernel $k\colon \mathbb{R} \to \mathbb{R}$ and the process

$$Z_t^\varepsilon \coloneqq \int_0^t k(t-s) X_s^\varepsilon \, \mathrm{d}s, \tag{9}$$

that is, the truncated convolution between the kernel k and the process X_t^ε. We consider in particular k to be a low-pass filter of the exponential kind, e.g., the function

$$k(t) = \frac{1}{\delta} \exp\left(-\frac{t}{\delta}\right), \tag{10}$$

where $\delta > 0$ is the filtering width, is a valid kernel function. Indeed, the analysis presented in Chapter 4 is mainly developed for the kernel (10), but we nevertheless demonstrated through numerical experiments the effectiveness of a wider class of filters. The convolution above acts as a smoother, eliminating the fast oscillations of the original trajectory, and can be computed in a computationally cheap fashion. In the language of signal processing, we pass the original

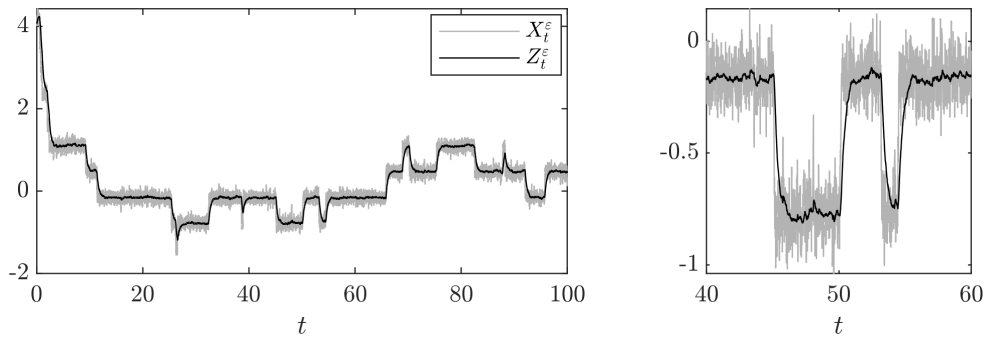


Figure 3 – Filtered trajectory Z_t^ε with an exponential kernel k in (9), from $0 \leq t \leq 100$ on the left and a zoom for the interval $40 \leq t \leq 60$ on the left. The original trajectory X_t^ε is the same as Fig. 2.

trajectory X_t^ε through a linear time invariant low-pass filter. We show in Fig. 3 the effect of this filtering procedure by choosing the kernel (10) and the same trajectory as the one shown in Fig. 2. We then combine the original trajectory X_t^ε and its filtered version Z_t^ε into a novel estimator for the drift of the homogenized equation.

The main theoretical result we present in Chapter 4 consists of showing that our novel estimator is asymptotically unbiased for the estimation of the drift of the homogenized equation (6). In particular, we show that the smoothing width δ of the filter (10) can be alternatively tuned to be proportional to the speed of the slow process or to smaller scales and provide in both cases unbiased results. We furthermore provide sharp estimates on the minimal width with respect to the multiscale parameter.

The problem of drift estimation can be naturally recast into a Bayesian framework. If no pre-processing of the data is applied the posterior distribution on the unknown is asymptotically biased, i.e., it shrinks towards the wrong value in the limit of infinite data and for vanishing ε. By carefully replacing the filtered trajectory Z_t^ε of (9) into a modified likelihood function, we are able to propose a corrected Bayesian computation which allows to get a correct estimation, asymptotically, and a homogenization-aware uncertainty quantification in the pre-asymptotic regime.

The advantages of our filtering approach with respect to subsampling are made evident by a series of numerical experiments on academic test cases. In particular, we show how the filtering width δ in (9), as well as the other parameters of the filter, play a much milder role in the inference result than does the subsampling width. This seems to be particularly accentuated when estimating a multi-dimensional parameter, i.e., when $N > 1$ in (5). Hence, the methodology we propose is not only theoretically justified, but seems to outperform existing techniques in terms of robustness and accuracy, and can therefore be employed as a black-box tool for parameter estimation of multiscale diffusion processes.

Part II: Probabilistic Methods for Differential Equations

In the second part of the thesis we present two novel methods for differential equations belonging to the field of probabilistic numerics (PN). Recycling the notation we employed above, let X and Y be an input and an output space, which we assume to be Banach spaces once equipped with the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively, and let $\mathcal{G}\colon X \to Y$ be a forward map. In particular, we focus in this part on scenarios where evaluating $\mathcal{G}$ involves the solution of a differential equation.

In this case, it is often necessary to resort to numerical approximations in order to evaluate the forward map. Let $h > 0$ denote the characteristic size of the discretization employed in the numerical approximation of the differential equation. Then, we denote by $\mathcal{G}_h \colon X \to Y$ the surrogate numerical forward map, which should satisfy the minimal requirement

$$\lim_{h \to 0} \|\mathcal{G}(x) - \mathcal{G}_h(x)\|_Y = 0,$$

for all $x \in X$. Moreover, it is possible in a wide range of situations to find a function $\psi \colon \mathbb{R} \to \mathbb{R}$ with $\psi(h) \to 0$ when $h \to 0$, often a polynomial, such that

$$\|\mathcal{G}(x) - \mathcal{G}_h(x)\|_Y \leq C(x, \mathcal{G}(x))\psi(h), \tag{11}$$

where C is a bounded positive function of input and output. Results of this form are fundamental and guarantee the good approximation of $\mathcal{G}$ by $\mathcal{G}_h$ in the asymptotic regime of $h \to 0$. Nevertheless, in several situations the combination of a point-wise approximation $\mathcal{G}_h$ and of a convergence result of the form (11) may not be sufficient. A notable examples of such a situation is given by chaotic differential equations, for which the constant $C(x, \mathcal{G}(x))$ in (11) may behave unpredictably over the space X. A further example is given by scenarios in which the map $\mathcal{G}$ is part of a pipeline of computations, and the output of the approximate map $\mathcal{G}_h$ has to be employed as the input of a subsequent computation. In this case, it may be complicated, impossible, or insufficient to transmit to the receiver the information contained in (11). The value $\mathcal{G}_h(x)$ could then be employed with absolute certainty, thus leading to an overconfident and possibly biased solution of the subsequent analysis. A typical example of computational pipelines for which applying blindly a numerical approximation may lead to overconfident solutions is given by Bayesian inverse problems, where the posterior distribution over the parameter of interest could be biased and "thin" regardless of the quality of the approximation.

Probabilistic numerical methods provide a solution to the issues illustrated above. In particular, all methods belonging to the field of PN share the idea of replacing the numerical forward map $\mathcal{G}_h$ by a forward map $\widetilde{\mathcal{G}}_h$, which more or less directly induces a probability distribution on Y which reflects the quality of the numerical approximation. In some cases, the random forward map $\widetilde{\mathcal{G}}_h$ is obtained as the combination of a classical numerical method and of random variables which are injected to account for uncertainty. In other cases, the function $\widetilde{\mathcal{G}}_h$ builds deterministically a probability measure – often belonging to a parameterized family – on the space Y. The output of the method is therefore not only a point endowed with an estimate such as (11), but a full probability measure, which can be pushed through pipelines of computations such as Bayesian inverse problems.

Motivation. Before introducing the probabilistic methods we treat in this thesis, let us give concrete examples motivating the need for PN.

We first consider the Lorenz system [91], which is defined by the following ordinary differential equation (ODE)

$$\begin{aligned} y_1' &= \eta(y_2 - y_1), & y_1(0) &= -10, \\ y_2' &= y_1(\rho - y_3) - y_2, & y_2(0) &= -1, \\ y_3' &= y_1 y_2 - \beta y_3, & y_3(0) &= 40. \end{aligned} \tag{12}$$

It is well-known that for $\rho = 28$, $\eta = 10$, $\beta = 8/3$, this equation has a chaotic behaviour, i.e., the solution is extremely sensitive to small perturbations. Integrating numerically (12) the error which is introduced at each time step is indeed a perturbation, thus any numerical solution cannot be considered reliable. We force a chaotic behavior by introducing random perturbation on the initial condition, implemented as a scalar Gaussian random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and artificially added to the first component $y_1(t)$ at time $t = 0$. In Fig. 4 we show $M = 20$ numerical trajectories given by a second-order Runge–Kutta method for three different scales of the noise.

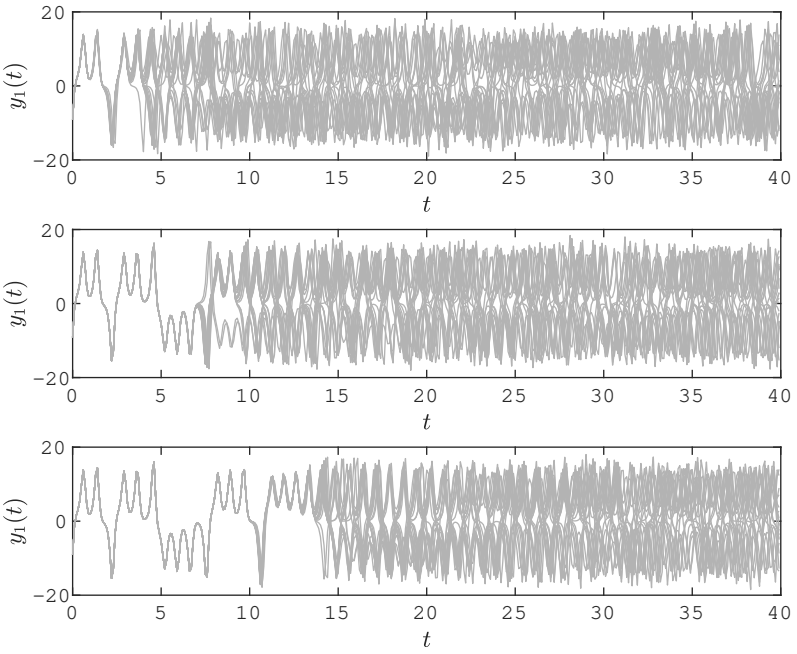


Figure 4 – First component $y_1(t)$ of the solution of (12) with decreasing zero-mean Gaussian perturbations on the initial condition from top to bottom (with standard deviation $\sigma = 10^{-1}, 10^{-3}, 10^{-5}$, respectively).

It is possible to remark that in each case, the numerical solutions almost coincide up to some time $\bar{t}$, thus diverging and unveiling the chaotic nature of the Lorenz system. It could be argued that up to time $\bar{t}$, the numerical solution offers a reliable approximation of the true solution as the dynamics have not yet switched to the chaotic regime. Still, by reducing the perturbation on the initial condition we delay the moment when numerical solutions diverge, and it is therefore unclear from a single trajectory what the time $\bar{t}$ is in practice. It is therefore not recommended to rely on a numerical approximation for a chaotic problem, and introducing a probability measure over the solution of the Lorenz system could be advantageous.

As we stated above, problems of Bayesian inference are most often employed to justify the usefulness of probabilistic methods for differential equations. The impact of a probabilistic component in the numerical approximation of inverse problems involving differential equations has been presented in several works [32, 36, 37, 39, 87, 101]. Let us give here a proof-of-concept example, which helps us to hint explicitly at the beneficial effect of the probabilistic approach. We consider the simple inverse problem

$$\text{find } x \in \mathbb{R}^n \text{ given observations } y = Ax + \beta \in \mathbb{R}^m, \tag{13}$$

where n and m are positive integers, where $A \in \mathbb{R}^{m\times n}$ is a non-singular matrix, and where $\beta \sim \mathcal{N}(0, \Gamma)$ is a Gaussian source of noise, with Γ a positive-definite covariance on $\mathbb{R}^m$. With the notation introduced above, we have $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and the exact forward map $\mathcal{G}(x) = Ax$. Given a scalar $h > 0$, we then consider the naif numerical method which is given by the forward map $\mathcal{G}_h(x) = (1 + h)Ax$. Clearly, we have for all $x \in \mathbb{R}^n$

$$\|\mathcal{G}(x) - \mathcal{G}_h(x)\|_2 \leq h \|Ax\|_2 ,$$

where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^m$, and the method fits the framework of (11). We furthermore consider, for the same $h > 0$, the probabilistic numerical method which is given by

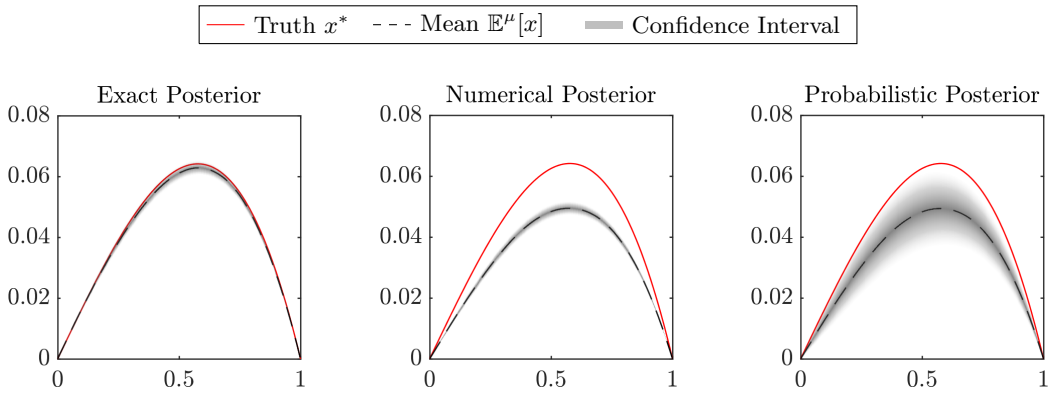


Figure 5 – Posterior distribution for (13) associated to the exact, numerical and probabilistic forward models. The true value of the unknown x^* is shown together with the posterior mean (dashed black line) and with a 95% confidence interval (shaded gray area).

the random forward map $\widetilde{\mathcal{G}}_h(x) = (1+h)Ax + h\xi$, where $\xi \sim \mathcal{N}(0, I)$ is a $\mathbb{R}^m$-valued standard Gaussian random variable. Let us remark that for all $x \in \mathbb{R}^n$ it holds

$$\begin{aligned} &\left\| \mathbb{E}\left[\widetilde{\mathcal{G}}_h(x)\right] - \mathcal{G}(x) \right\|_2 \leq h \left\| Ax \right\|_2, \\ &\mathbb{E}\left[\left\| \widetilde{\mathcal{G}}_h(x) - \mathcal{G}(x) \right\|_2^2 \right]^{1/2} \leq \sqrt{2}h \left(\mathbb{E}\left[\|\xi\|_2^2 \right]^{1/2} + \left\| Ax \right\|_2 \right) = \sqrt{2}h \left(\sqrt{d} + \left\| Ax \right\|_2 \right), \end{aligned} \tag{14}$$

where $\mathbb{E}$ denotes expectation with respect to the random variable ξ. In this case, as we will detail in Chapter 6, we say that the method is of first weak and mean-square order of convergence, respectively. Let us return to the inverse problem (13). In the Bayesian framework, we set a prior distribution μ_0 on the unknown x, that we assume to be a Gaussian $\mu_0 = \mathcal{N}(m_0, C_0)$, and then update it to obtain a posterior measure μ by conditioning on the observations $y \in \mathbb{R}^m$. In this linear and Gaussian case, the true posterior – i.e., associated to the true forward map $\mathcal{G}$ – is a Gaussian $\mu = \mathcal{N}(m, C)$, where

$$\begin{aligned} C^{-1} &= A^\top \Gamma^{-1} A + C_0^{-1}, \\ m &= C\left(A^\top \Gamma^{-1} y + C_0^{-1} m_0 \right), \end{aligned} \tag{15}$$

and where we recall that Γ is the covariance of the noise. Details on how the posterior is obtained in this linear case are given e.g. in [131, Example 6.23]. We now consider the numerical forward map $\mathcal{G}_h$, and notice that $\mathcal{G}_h(x) = A_h x$, where $A_h = (1+h)A$ is an approximation of the matrix A. The posterior μ_h associated to $\mathcal{G}_h$ is therefore a Gaussian $\mu_h = \mathcal{N}(m_h, C_h)$, where m_h and C_h are obtained by replacing A with A_h in the formulas (15). Finally, we consider the probabilistic forward map $\widetilde{\mathcal{G}}_h$, and remark that replacing in (13) we obtain

$$y = \widetilde{\mathcal{G}}_h(x) + \beta = \mathcal{G}_h(x) + h\xi + \beta.$$

Under the reasonable assumption that ξ and β are independent random variables, we therefore have the model $y = \mathcal{G}_h(x) + \beta_h$, where $\beta_h \sim \mathcal{N}(0, \Gamma_h)$ and where $\Gamma_h = \Gamma + h^2 I$ is the modified covariance of the noise. Hence, the posterior $\widetilde{\mu}_h = \mathcal{N}(\widetilde{m}_h, \widetilde{C}_h)$ associated to the probabilistic forward model is obtained by computing $\widetilde{m}_h$ and $\widetilde{C}_h$ with (15) replacing A with A_h and Γ with Γ_h. Let us consider an example. We fix $n = m = 99$ and choose the matrix $A \in \mathbb{R}^{n \times n}$ as

$$A = (n+1)^2 \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

i.e., the finite difference discretization of the negative second derivative on $(0,1)$ with an equal spacing of $1/(n+1)$ and with Dirichlet homogeneous boundary conditions. We then fix the true value $x^* \in \mathbb{R}^n$ of the unknown to be given component-wise by

$$(x^*)_i = -\frac{1}{6} t_i (t_i^2 - 1), \quad t_i = \frac{i}{n+1}, \quad i = 1, \dots, n,$$

and generate synthetic observations $y = Ax^* + \beta$, where $\beta \sim \mathcal{N}(0, \Gamma)$ and $\Gamma = 10^{-2} I$. We then consider the non-informative prior $\mu_0 = \mathcal{N}(0, I)$, fix $h = 0.3$ and compute the posterior distributions μ, μ_h and $\widetilde{\mu}_h$ on $\mathbb{R}^n$ associated to the true forward map $\mathcal{G}$, to its numerical approximation $\mathcal{G}_h$ and to the probabilistic map $\widetilde{\mathcal{G}}_h$, respectively. Results are given in Fig. 5, where we notice that the exact posterior μ is almost unbiased and the true value lays within a 95% confidence interval. Indeed, the problem is fairly simple, and it is possible to expect the true posterior to be precise and confident. Conversely, the posterior μ_h is clearly biased in the mean, and one can notice how despite this biasedness the posterior is overconfident about the inference result. This is the issue that the PN approach is designed to solve. Indeed, applying the naive probabilistic method associated to the forward map $\widetilde{\mathcal{G}}_h$ yields a posterior $\widetilde{\mu}_h$ which is still biased in the mean, but whose underconfidence on the result clearly helps quantify the uncertainty due to the numerical approximation.

A third motivation for the field of PN, which has been relatively overlooked in the literature, is the design of probabilistic a posteriori error estimators for differential equations. An a posteriori error estimator for the numerical approximation $\mathcal{G}_h(x)$ is a quantity $\mathcal{E}_h$ which is computable without knowledge of the true solution $\mathcal{G}(x)$, and which gives information about the numerical error. We say moreover that if it holds

$$C_{\text{low}} \mathcal{E}_h \leq \|\mathcal{G}(x) - \mathcal{G}_h(x)\|_Y \leq C_{\text{up}} \mathcal{E}_h, \tag{16}$$

then the estimator is reliable (upper bound) and efficient (lower bound), respectively. The constants C_{low} and C_{up} may depend on $x \in X$, that is, on the data, but not on the true output $\mathcal{G}(x)$ itself. If a quantity $\mathcal{E}_h$ satisfying (16) is available, then it is possible to control the quality of the numerical approximation and often to adapt the discretization in order to achieve an optimal scheme. Quantifying numerical errors is one of the overt goals of the PN community, and, in our opinion, it is therefore relevant to employ the statistical information given by probabilistic methods to build a posteriori error estimators. The design and analysis of such an error estimator for PDEs is presented in Chapter 8, and introduced with the main contributions below.

Literature Review. Despite the relative young age of the field, there have been several contributions to the literature of PN. We notice that the contributors to PN come from different backgrounds, such as numerical analysis, statistics, machine learning, and optimization, and that as a consequence the literature is diverse. In particular, several problems of linear algebra, optimization, numerical quadrature, and differential equations have been reinterpreted under a probabilistic perspective. We give here an overview of the literature on probabilistic methods for ODEs and PDEs, and we refer the reader to the articles [38, 65, 102] for a complete review of the other fields of application of PN.

For ordinary differential equations (ODEs), the methodologies can be roughly split in two different areas. In [32, 76, 77, 92, 126, 127, 129, 137] the authors present a series of schemes which rely in different measure on Bayesian filtering techniques. These methodologies proceed by updating Gaussian measures over the numerical solution with filtering formulae and evaluations of the right-hand side of the ODE, which are interpreted as observations. While being not involved computationally, analyzing the convergence properties of this class of methods is not always possible, and one can only marginally rely on standard techniques for this purpose. A valuable effort in this sense can be found in [77], where the authors show rates of convergence of the

mean of the Gaussian measure towards the exact solution. A different approach is presented in the series of works [39, 86, 134, 135], where the authors propose probabilistic schemes which are based on perturbing randomly the approximate solution and on letting evolve these perturbations through the dynamics of the ODE. In this manner, it is possible to obtain empirical probability measures over the otherwise deterministic numerical solution. In particular, an appropriate random perturbation can be directly added to the state at each step of the time integration, as it was presented and analyzed for one-step methods in [39, 85, 86], with a particular focus on implicit schemes in [134] and for multistep methods in [135].

There has been a keen interest from the PN community on developing probabilistic numerical solvers for partial differential equations (PDEs), too [32, 36, 37, 39, 58, 101, 106–108, 118, 119]. In [36], the authors present a meshless Bayesian method for PDEs, which they then apply to inverse problems in [37], and in particular to a challenging time-dependent instance drawn from an engineering application in [101]. Their methodology consists of imposing a Gaussian prior on the space of solutions, thus updating it with evaluations of the right-hand side, which are interpreted as noisy observations. A similar idea has been presented in [32], where the main focus are time-dependent problems, and in [118, 119], where the method is recast in the framework of machine learning algorithms. In [107, 108], a probabilistic approach involving gamblets is applied to the solution of PDEs with rough coefficients and by multigrid schemes, with a particular interest to reducing the complexity of implicit algorithms for time-dependent problems [108]. Moreover, in [106] the author presents a Bayesian reinterpretation of the theory of homogenization for PDEs, which can be seen as a contribution to the field of PN. To our knowledge, the only perturbation-based finite element (FE) probabilistic scheme for PDEs is presented in [39], where the authors randomize FE bases by adding random fields endowed with appropriate boundary conditions, thus obtaining an empirical measure over the space of solutions. By tuning the covariance of these random fields, they obtain a consistent characterization of the numerical error, which can then be employed to solve Bayesian inverse problems and to quantify the uncertainty over their numerical solution.

Main Contributions. In the second part of this thesis, we introduce two novel probabilistic methods, respectively for ODEs and PDEs, both based on a randomization of the discretization. In the ODE setting, our main focus is geometric integration. In particular, the probabilistic method we propose is tailored for ODE systems endowed with a certain geometric structure, such as the conservation of polynomial first integrals of motion or Hamiltonian systems. In the PDE setting, we consider an elliptic equation and present probabilistic a posteriori error estimators, demonstrating that statistical information on the numerical solution can be readily employed for mesh adaptation. We remark that the two methods are the subject of our articles [6] and [7], respectively.

In Chapter 7 we introduce a probabilistic method for the autonomous ODE on $\mathbb{R}^d$

$$y'(t) = f(y(t)), \qquad y(0) = y_0, \tag{17}$$

where d is a positive integer, the right-hand side $f\colon \mathbb{R}^d \to \mathbb{R}^d$ and $y_0 \in \mathbb{R}^d$ is a given initial condition. We consider moreover one-step Runge–Kutta (RK) approximations of the solution of (17), which, given a time step $h > 0$, read for $n = 1, 2, \dots$

$$y_n = \psi_h(y_{n-1}),$$

where the function $\psi_h\colon \mathbb{R}^d \to \mathbb{R}^d$ is the numerical flow of the RK method and where $y_n \approx y(nh)$. The numerical flow is built to mimic the exact flow of (17), i.e., the function $\varphi_t\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $y(t) = \varphi_t(y_0)$. The first PN method based on perturbations of traditional numerical scheme is the additive-noise Runge–Kutta method (AN-RK) proposed in [39] and further analyzed in [85, 86]. In particular, the AN-RK proceeds as

$$Y_n = \psi_h(Y_{n-1}) + \xi_n,$$

where ξ_n are appropriately scaled independent and identically distributed (i.i.d.) Gaussian random variables and where $Y_n \approx y(nh)$ in the weak and mean-square sense of (14). By tuning the random variables ξ_n, it is possible to prove that the random perturbations are on the one hand not spoiling the convergence properties of the underlying RK method and on the other hand capturing the numerical errors in a statistical sense. The method is moreover shown to be applicable with success to inference problems where the right-hand side f of (17) is parameterized by some unknown parameter.

Despite its favorable properties in forward and inverse UQ, the AN-RK lacks in some applications of robustness and accuracy. Indeed, an additive noise contribution could produce disruptive effects on favorable geometric features of the underlying RK method. A direct example of this non-robust behavior is given by ODEs for which the solution is supposed to stay positive and small. In this case, the addition of a random contribution could force the solution in the negative plane, hence the numerical solution could be physically meaningless and numerically unstable. Chemical reactions with small population size for one species at some time of the evolution are typical physical examples. Moreover, quantities which are invariant on the solution of the ODE (17) and on appropriately-chosen RK surrogates are not conserved by the AN-RK solution, always due to the random additive source of noise. This is a critical limitation for the AN-RK, since a large variety of physical phenomena are modeled by dynamical systems which are in some sense geometric.

Motivated by these issues, we introduce the random time step Runge–Kutta method (RTS-RK) a novel probabilistic method for ODEs based on a random selection of the time steps. The RTS-RK builds a sequence of random variables employing the numerical flow ψ_h of a RK method and the recursion

$$Y_n = \psi_{H_n}(Y_{n-1}),$$

where H_n are appropriate i.i.d. random variables all with mean $\mathbb{E}[H_n] = h$ for some fixed $h > 0$, and where, again, $Y_n \approx y(nh)$. Hence, the randomness artificially injected in the RK method is in the RTS-RK intrinsic to the scheme itself, in contrast to the AN-RK. The first property of the RTS-RK we analyze is its convergence. Indeed, we prove that the RTS-RK converges to the true solution y of (17) in the weak and the mean-square senses, similarly to the AN-RK. The RTS-RK method is moreover more robust than the AN-RK in case the solution of the ODE should be constrained to a set due to the physics of the problem. For these reasons, the RTS-RK has been successfully applied to ODEs arising in neuroscience, such as the Hodgkin–Huxley model, in [103].

The main advantage of the RTS-RK with respect to the AN-RK, as well as to other methods for ODEs in the PN literature, is its favorable geometric properties. Let us consider an ODE of the form (17) and such that there exists $Q\colon \mathbb{R}^d \to \mathbb{R}$, which we call an invariant or first integral of the ODE, such that $Q(y(t)) = Q(y(0))$ for all $t \geq 0$. The first geometric property we consider concerns the exact conservation of first integrals, i.e., numerical solutions such that $Q(y_n) = Q(y_0)$ for all $n = 1, 2, \ldots$. If Q is a linear function, then it is conserved by any RK method. This is notably important for ODEs modeling chemical reactions, where the total mass of the system is conserved. If Q is quadratic, then there exist RK methods which conserve exactly Q along the numerical trajectories. Most importantly, the class of Gauss collocation RK methods conserve quadratic first integrals. A typical example of systems with quadratic first integrals are found in astronomy, where planetary systems conserve the total angular momentum. In these situations, we are able to prove that if a RK method exactly conserves a first integral Q, so does the RTS-RK based on the same numerical flow map, in a path-wise sense, while the AN-RK does not.

We then consider Hamiltonian systems, which are employed for modeling a variety of physical

phenomena, and which can be written for an energy function $Q\colon \mathbb{R}^{2d} \to \mathbb{R}$ as the ODE on $\mathbb{R}^{2d}$

$$y' = J^{-1}\nabla Q(y), \quad y(0) = y_0, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{18}$$

where $y_0 \in \mathbb{R}^{2d}$ is a given initial condition. The energy Q is then a first integral, meaning that it is conserved on the solution y of (18). Moreover, the exact flow φ_t of (18) is symplectic, which means that for all $y \in \mathbb{R}^{2d}$ it holds

$$D_y\varphi_t(y)^\top J D_y\varphi_t(y) = J, \tag{19}$$

where $D_y\varphi_t$ is the Jacobian of φ_t and J is the same matrix as above. Geometrically, a symplectic flow map φ_t has the property of preserving volumes in the state space. If the numerical flow map ψ_h of a RK method satisfies (19) when it is applied to a Hamiltonian system, we say that the RK method is symplectic. Symplectic methods are widely employed for Hamiltonian systems because the numerical solution they provide approximately conserves the energy Q over exponentially long times, which in turn implies a good approximation of the solution over the same time span. A necessary condition for the approximation property, though, is that the time step is kept constant over the time integration. In the probabilistic setting, we consider the RTS-RK built on a symplectic integrator, and first show that the flow map associated to the RTS-RK is almost surely symplectic. We then consider the long-time approximation of the energy function, and show that Q is approximately conserved by the RTS-RK over polynomially long times. The importance of this result is twofold. First, it implies that the RTS-RK method inherits the favorable geometric properties of its deterministic counterpart, and therefore truly quantifies the uncertainty due to numerical discretization. Second, independently of the probabilistic paradigm, it implies that the condition of having a fixed time step to achieve long-time conservation of the energy can be relaxed to an appropriate sequence of random time steps. This, to our knowledge, was previously unknown in the literature.

To conclude, we apply the RTS-RK to inference problems involving Hamiltonian ODEs, and demonstrate the advantages of adopting a probabilistic approach combined with a geometry-aware integrator in this context. Indeed, employing e.g. the AN-RK for such a problem would introduce an unreasonable bias in the solution due to the lack of symplecticity of its flow.

In Chapter 8 we introduce a probabilistic finite element method (FEM) for the elliptic PDE

$$\begin{aligned} -\nabla \cdot (A\nabla u) &= f, \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{20}$$

where $A\colon D \to \mathbb{R}^{d\times d}$ is the diffusion coefficient, and $f\colon D \to \mathbb{R}^d$ the force term. Similarly to the ODE case and the RTS-RK, we build a probability measure on the numerical solution by perturbing randomly the discretization itself, which in the PDE setting is a mesh of the domain D. We then call our probabilistic method the random mesh finite element method (RM-FEM). In particular, denoting by V_h a linear finite element space based on a mesh $\mathcal{T}_h$ of D with characteristic size $h > 0$, we consider a perturbation $\widetilde{\mathcal{T}}_h$ of the mesh, obtained by moving the vertices of $\mathcal{T}_h$ randomly in an appropriate fashion, and the random finite element space $\widetilde{V}_h$ which is built on the mesh $\widetilde{\mathcal{T}}_h$. We then construct a probabilistic approximation of the solution u of (20) by computing alternatively either the RM-FEM solution $\widetilde{u}_h \in \widetilde{V}_h$ or the RM-FEM interpolant $\widetilde{\mathcal{I}}u_h \in \widetilde{V}_h$, where $\widetilde{\mathcal{I}}\colon V_h \to \widetilde{V}_h$ is an interpolation operator, and where u_h is the FEM solution on V_h.

Keeping in mind the fundamental goal of PN, i.e., giving a statistical characterization of numerical errors, we consider the problem of employing probabilistic methods to build a posteriori error estimators for the problem (20). Indeed, for the RTS-RK and other PN methods it is possible to show that the solution converges in the weak and mean-square senses of (14) with the same rate

as the deterministic method they are built on. This is a consistency result, and does not imply in practice that the error can be effectively controlled, or described in a statistical manner, by the probabilistic approach. Some forms of adaptivity for nonlinear ODEs based on probabilistic information can be found in [25, 31, 127], where the arguments are based on heuristics but are not rigorously analyzed. Employing the RM-FEM, we are able to construct a posteriori error estimators which can be readily employed for mesh adaptation in elliptic PDEs, and which are the main contribution of Chapter 8. Our estimators are entirely based on probabilistic information, are simple to compute and do not entail considerable computational cost. We present an analysis in the one-dimensional case that shows that our error estimators based on the RM-FEM are equivalent to a classical estimator by Babuška and Rheinboldt [19], which employs the jumps of the derivative of the solution at the nodes to quantify the numerical errors. Our one-dimensional theoretical analysis is complemented by a series of numerical experiments confirming the validity of our theory in higher dimensions.

Similarly to the RTS-RK in the ODE case, and to other methods in the field of PN, we show via numerical experiments the usefulness of the RM-FEM in the context of Bayesian inverse problems. In particular, we consider elliptic PDEs of the form (20), where the diffusion coefficient A is unknown and has to be retrieved through discrete observations of the solution u. We show that the solution of the inverse problem is consistent asymptotically with respect to the mesh spacing, and that its quality is enhanced if the latter is relatively large, i.e., if the forward model is approximated cheaply.

Outline

The thesis is divided in two parts. Part I is made of Chapters 1 to 5, and Part II is made of Chapters 6 to 9. We remark that Chapters 2, 4, 7 and 8 contain the original contributions of this thesis.

In Chapter 1 we introduce inverse problems and their Bayesian interpretation. We notice that there exist several introductions to inverse problems in the literature, most notably [49, 71, 131]. Since inverse problems are the main topic of Part I and they partly motivate Part II, we set nonetheless our notation and state basic results in this thesis.

In Chapter 2 we present the multiscale ensemble Kalman inversion for elliptic PDEs. After a general introduction on Kalman and ensemble Kalman filtering, we present the multiscale setting and results of convergence. We conclude with a novel quantitative characterization of classical approximations of the modeling error, and with numerical experiments.

In Chapter 3 we introduce multiscale diffusion processes and the problem of extracting effective diffusions from multiscale data. In particular, we report a proof of the homogenization result in this context, together with properties of maximum likelihood estimators for the drift coefficient.

In Chapter 4 we present a new framework based on filtered data for fitting effective diffusions to multiscale observations. In particular, we show that applying a low-pass filter to the data allows to circumvent the model misspecification and to infer models which are consistent with the theory of homogenization.

In Chapter 5 we draw our conclusions for Part I and present possible directions of future research.

In Chapter 6 we give a general introduction to the field of probabilistic numerics. We focus on defining the desired properties of a probabilistic method, and show how the PN approach can be combined with the Bayesian paradigm to enhance the solution of inverse problems. We

particularize the discussion by giving examples of ODE solvers taken from recent literature.

In Chapter 7 we present the RTS-RK, a probabilistic Runge–Kutta method for ODEs based on random time steps. An extensive theoretical analysis covers properties of convergence, geometric integration and robustness, which are further demonstrated by a series of numerical examples.

In Chapter 8 we present a probabilistic FEM based on random meshes, the RM-FEM. The main focus of the chapter is deriving a posteriori error estimators based on statistical information drawn from the probabilistic solution. In particular, we show that our estimators are equivalent to classical estimators in the literature, which may open the door for further developments in PN.

In Chapter 9 we draw our conclusions for Part II and present possible directions of future research.

Part I

Multiscale Inverse Problems and Parameter Inference

This first part of the thesis is devoted to inverse problems involving deterministic and stochastic multiscale problems.

Inverse problems are the main topic of this first part of the thesis and are as well partly covered in its second part. Therefore, we give in Chapter 1 a general introduction to inverse problems and their Bayesian interpretation. Inverse problems are generally ill-posed in the infinite-dimensional setting, and need therefore to be regularized. The Bayesian approach indeed allows to regularize the problem and to simultaneously obtain a full quantification of the uncertainty over its solution, which is achieved introducing probability measures over the unknown.

Chapter 2 is devoted to the application of the ensemble Kalman filters (EnKF) to elliptic multiscale inverse problems, which is one of the main original contributions of this thesis. In particular, we combine numerical homogenization techniques and the EnKF to provide effective solutions to inverse problems involving rapidly-oscillating tensors, both pointwise and in a Bayesian fashion. Neither the application of the EnKF nor the introduction of homogenization tools are novelties in the literature on inverse problems. Nevertheless, we successfully combine these two methodologies and demonstrate the validity and efficiency of our approach via a careful convergence analysis and numerical experiments, thus providing a reliable tool for the simulation and the uncertainty quantification of multiscale phenomena.

In Chapter 3 and Chapter 4 we present a novel methodology for the estimation of the drift coefficient of multiscale stochastic processes of the diffusion type. Unlike elliptic partial differential equations, where homogenization theory translates seamlessly to inverse problems, in the context of diffusion processes one has to recur to additional treatments of the multiscale data in order to fit an effective model. We present in detail in Chapter 3 the reasons why homogenization theory is not sufficient to obtain a correct solution of the multiscale inverse problem in this setting. In particular, we focus on both a maximum-likelihood and hence pointwise approach, and on the solution of the inference problem in the Bayesian sense. Our novel contribution is given in Chapter 4, where we present a methodology based on filtering the data which allows to obtain efficiently and robustly an effective solution to this multiscale inverse problem. In particular, comparing our approach with the widely-employed technique of subsampling shows numerically the advantages of our method.

Finally, in Chapter 5 we draw our conclusions and give suggestion for possible future developments.

Let us remark that a reader interested mainly in stochastic multiscale models could skip Chapter 2 without compromising the understanding of Chapters 3 and 4. If this reading strategy is adopted, we suggest nevertheless to at least skim through Chapter 1 in order to get acquainted with our notation.

1 An Introduction to Bayesian Inverse Problems

In this short chapter we introduce inverse problems and their Bayesian interpretation. Besides being of the uttermost relevance in several areas of applied sciences, inverse problems are one of the leitmotifs of this thesis, and in this chapter we provide a common framework for the remainder of this manuscript. Standard references for inverse problems and their Bayesian interpretation are [49, 71, 131, 132]. Throughout the chapter, we point the reader to relevant references which are specific to the treated subject. Let us remark that some phrasings employed here are borrowed from our original research articles [6, 7, 9].

The outline of this chapter is as follows. After introducing in Section 1.1 the general framework, we give in Section 1.2 details on finite-dimensional approximations of infinite-dimensional inverse problems. We conclude with a discussion on Markov chain Monte Carlo methods in Section 1.3.

1.1 The Bayesian Interpretation of Inverse Problems

Let X and Y be Banach spaces, with associated norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. Let us moreover consider a map $\mathcal{G}\colon X \to Y$, which we call the forward map. We suppose that quantities in Y are observable up to a source of noise, whereas those in X are not. To be precise, we consider the observational model

$$y = \mathcal{G}(u) + \eta, \tag{1.1}$$

where for simplicity we assume the source of noise η to be Gaussian, and in particular $\eta \sim \mathcal{N}(0, \Gamma)$, with Γ a valid covariance operator on Y. We then suppose to be given observations $y^* \in Y$ and we consider the inverse problem

$$\text{find } u^* \in X \text{ given observations } y^* = \mathcal{G}(u^*) + \eta, \tag{1.2}$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ is an unknown realization of the noise, and where $u^* \in X$ is the true value of the unknown, which we wish to retrieve. The inverse problem (1.2) above is ill-posed for two reasons. First, due the source of noise it is not possible to identify u^* exactly. Indeed, given any $u \in X$, one can obtain y^* by claiming the noise to be given by $\eta = y^* - \mathcal{G}(u)$. Second, the dimension of the spaces X and Y can present a mismatch, e.g., with X being infinite-dimensional and $Y \equiv \mathbb{R}^L$, so that the problem is naturally ill-posed. Indeed, in practice it is oftentimes the case that the observation space Y is finite-dimensional, whereas the space X of the unknown is an infinite-dimensional function space. For this reason, we choose for simplicity to identify $Y \equiv \mathbb{R}^L$ and consider therefore observations to be finite-dimensional. Furthermore, we have that the observational covariance $\Gamma \in \mathbb{R}^{L \times L}$, and we consider again for simplicity Γ to be positive definite.

Adopting a Bayesian approach allows to regularize and therefore to solve (1.2). In the Bayesian framework, the object of interest are not point values in the spaces X, but probability measures which express prior and posterior knowledge of the unknown u. In particular, let us consider a probability measure μ_0 on the measurable space $(X, \mathcal{B}(X))$, where $\mathcal{B}(X)$ denotes the Borel σ-algebra on X, which encapsulates all the knowledge on the unknown which is known a priori, and that we refer to as the prior in the following. For simplicity, we consider the prior to be a Gaussian measure $\mu_0 = \mathcal{N}(m_0, C_0)$, where $m_0 \in X$ and where C_0 is a covariance operator on X. A broader class of prior measures could be employed, such as Besov or heavy-tailed measures (see e.g. [49, 133]), but we restrict ourselves to the Gaussian case for simplicity. We then compute a probability measure μ^y on $(X, \mathcal{B}(X))$ which summarizes the knowledge on the unknown conditional on observations $y \in \mathbb{R}^L$. The measure μ^y is called the posterior and its Radon–Nykodim derivative with respect to the prior (see Section A.2) is formally given by

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z^y} \exp(-\Phi(u; y)), \tag{1.3}$$

where Z^y is the normalization constant

$$Z^y = \int_X \exp(-\Phi(u; y)) \, \mathrm{d}\mu_0(u), \tag{1.4}$$

and where for any $y \in \mathbb{R}^L$ the potential $\Phi(\cdot; y) \colon X \to \mathbb{R}$ is given due to the Gaussian assumption on the noise by

$$\Phi(u; y) = \frac{1}{2} \left\| \Gamma^{-1/2} \left(\mathcal{G}(u) - y \right) \right\|_2^2, \tag{1.5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^L$. The well-posedness of (1.3) is subject to conditions on the forward map $\mathcal{G}$ and on the prior measure μ_0. Minimal sets of assumptions in various scenarios are treated extensively in [49, 131]. We consider here the following assumption (see [131, Assumption 2.7]) which is sufficient if Y is finite-dimensional case.

Assumption 1.1. The forward map $\mathcal{G} \colon X \to \mathbb{R}^L$ satisfies

(i) For all $\beta > 0$ there exists $C_1 = C_1(\beta) \in \mathbb{R}$ such that for all $u \in X$

$$\left\| \Gamma^{-1/2} \mathcal{G}(u) \right\|_2 \leq \exp\left(\beta \|u\|_X^2 + C_1 \right);$$

(ii) For all $R > 0$ there exists $C_2 = C_2(R) > 0$ such that for all $u_1, u_2 \in X$ satisfying $\max\{\|u_1\|_X, \|u_2\|_X\} < R$ it holds

$$\left\| \Gamma^{-1/2} \left(\mathcal{G}(u_1) - \mathcal{G}(u_2) \right) \right\|_2 \leq C_2 \|u_1 - u_2\|_X .$$

In words, the first assumption above guarantees exponential boundedness of the forward operator, whereas the second requirement is that of $\mathcal{G}$ to be Lipschitz. Let us stress that Assumption 1.1 is an assumption on the forward map, and does not involve the inversion procedure. Under Assumption 1.1 and if μ_0 is a valid Gaussian probability measure on X, the posterior is indeed given by (1.3), and the inverse problem is well-posed and Lipschitz with respect to the data in the Hellinger metrics d_{Hell}, which for two probability measure ν_1 and ν_2 on $(X, \mathcal{B}(X))$ reads

$$d_{\mathrm{Hell}}(\nu_1, \nu_2) \coloneqq \sqrt{\frac{1}{2} \int_X \left(\sqrt{\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_0}} - \sqrt{\frac{\mathrm{d}\nu_2}{\mathrm{d}\nu_0}} \right)^2 \mathrm{d}\nu_0},$$

where ν_0 is a probability measure on X such that $\nu_1 \ll \nu_0$ and $\nu_2 \ll \nu_0$ (see Section A.2). This is formalized by the following result (see [131, Corollary 4.4] for a proof).

Proposition 1.2. *Let Assumption 1.1 hold, and let μ_0 be a Gaussian measure on X such that $\mu_0(X) = 1$. Then, the posterior μ^y given in (1.3) exists and is unique. Moreover, for all $R > 0$ there exists $C = C(R)$ such that for all $y_1, y_2 \in \mathbb{R}^L$ satisfying $\max\{\|y_1\|_2, \|y_2\|_2\} < R$ it holds*

$$d_{\text{Hell}}(\mu^{y_1}, \mu^{y_2}) \leq C \left\| \Gamma^{-1/2}(y_1 - y_2) \right\|_2 .$$

In the following, we drop for economy of notation the dependence on the data y from the posterior μ^y and the normalization constant Z^y of (1.4), and simply write μ and Z.

In practice, it is not always possible to evaluate exactly the forward map $\mathcal{G}$, but it is possible to approximate its action on X numerically. A notable example is the one of inverse problems involving differential equations. Let $h > 0$ be a discretization parameter, as, for example, the mesh size in a finite element method, or the time step in a ODE solver. We then denote by $\mathcal{G}_h \colon X \to \mathbb{R}^L$ an approximation of the forward map $\mathcal{G}$, whose quality is driven by the parameter h. Maintaining the observation model (1.1) and the same prior for the parameter as above, we consider the approximate posterior μ_h whose Radon–Nikodym derivative with respect to the prior is formally given by

$$\frac{\mathrm{d}\mu_h}{\mathrm{d}\mu_0}(u) = \frac{1}{Z_h} \exp(-\Phi_h(u; y)), \tag{1.6}$$

where the potential Φ_h is given by

$$\Phi_h(u; y) = \frac{1}{2} \left\| \Gamma^{-1/2} \left(\mathcal{G}_h(u) - y \right) \right\|_2^2 ,$$

and where the normalization constant Z_h is defined equivalently to (1.4). A natural question arising from this setting is whether the approximate posterior μ_h converges to the true posterior μ in the limit $h \to 0$. This is indeed guaranteed, and the following result holds (see [131, Corollary 4.9] for a proof).

Proposition 1.3. *Let Assumption 1.1 hold for $\mathcal{G}$ and for $\mathcal{G}_h$ uniformly in h, and let the prior μ_0 be a measure on $(X, \mathcal{B}(X))$ satisfying $\mu_0(X) = 1$. Then, the posterior measures μ and μ_h are well-defined in the sense of Proposition 1.2. Moreover, if for all $\beta > 0$ there exists $C_1 = C_1(\beta) > 0$ such that for all $u \in X$ it holds*

$$\|\mathcal{G}(u) - \mathcal{G}_h(u)\|_2 \leq C_1 \exp\left(\beta \|u\|_X^2 \right) \psi(h),$$

where $\psi(h) \to 0$ for $h \to 0$, then

$$d_{\text{Hell}}(\mu, \mu_h) \leq C\psi(h),$$

where $C > 0$ is a constant independent of h.

1.1.1 A Simple ODE Example

We conclude this section with an example in a simple setting, which shows how one should proceed in order to apply Propositions 1.2 and 1.3. Let us consider the ordinary differential equation (ODE) on $\mathbb{R}$

$$\begin{aligned} z'(t) &= -\exp(u) z(t), \quad t \geq 0, \\ z(0) &= z_0 \in \mathbb{R}, \end{aligned} \tag{1.7}$$

where $u \in \mathbb{R}$ is a real parameter. We choose this exponential model for the parameter multiplying the right-hand side so to make sure that the coefficient $\lambda = \exp(u)$ is positive given any sample from a Gaussian prior on u, and in turn the ODE (1.7) is Lyapunov stable. Imposing a prior on the logarithm of the parameter of interest is a known trick if the well-posedness of the forward

model depends on the sign of the coefficient of interest, and is referred to as the log-normal prior in literature. Given $T > 0$, we consider the forward map $\mathcal{G}\colon \mathbb{R} \to \mathbb{R}$ such that

$$\mathcal{G}(u) = z(T) = \exp(-\exp(u)T)z_0. \tag{1.8}$$

We then consider the inverse problem

$$\text{find } u \in \mathbb{R} \text{ given an observation } y^* = \mathcal{G}(u^*) + \eta,$$

where $\eta \sim \mathcal{N}(0, \gamma^2)$ and $\gamma > 0$ denotes the observational standard deviation. We start by verifying the assumptions for well-posedness on the forward map.

Lemma 1.4. *The forward map $\mathcal{G}$ given in* (1.8) *satisfies Assumption 1.1.*

Proof. For Assumption 1.1(i), it is sufficient to notice that $-\exp(u) < 0 \leq |u|^2$ so that

$$\gamma^{-1} |\mathcal{G}(u)| \leq \exp\left(T |u|^2 + \log |z_0| - \log \gamma\right).$$

Maximizing $|\mathcal{G}'(u)|$ it is easy to verify that the Lipschitz constant of $\mathcal{G}$ is $\exp(-1) |z_0|$, which shows that $\mathcal{G}$ satisfies Assumption 1.1(ii). □

Posing a Gaussian prior $\mu_0 = \mathcal{N}(m_0, C_0)$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for the parameter u, it is clear that $\mu_0(\mathbb{R}) = 1$, and therefore by Lemma 1.4 the posterior is well-defined and given by (1.3). Let us now consider a positive integer N, and a fixed time step $h = T/N$. We then consider the implicit Euler approximation of (1.7), which setting z_0 to the initial condition reads for all $n = 0, \dots, N-1$

$$z_{n+1} = (1 + h\exp(u))^{-1} z_n.$$

We then have the approximate forward map

$$\mathcal{G}_h(u) = z_N = (1 + h\exp(u))^{-N} z_0. \tag{1.9}$$

We verify the assumptions for well-posedness of the inverse problem for the numerical forward map.

Lemma 1.5. *The forward map $\mathcal{G}_h$ of* (1.9) *satisfies Assumption 1.1 uniformly in h.*

Proof. For Assumption 1.1(i), we remark that since $h > 0$, it holds

$$\frac{1}{1 + h\exp(u)} < 1 \leq \exp\left(h |u|^2\right),$$

so that replacing $T = Nh$

$$\gamma^{-1} |\mathcal{G}_h(u)| \leq \exp\left(T |u|^2 + \log |z_0| - \log \gamma\right).$$

For Assumption 1.1(ii), direct calculations show that the maximum of $|\mathcal{G}'(u)|$ is attained in $\bar{u} = -\log T$ and that

$$|\mathcal{G}'(\bar{u})| = \left(1 + \frac{1}{N}\right)^{-(N+1)} |z_0|.$$

Therefore the Lipschitz constant of $\mathcal{G}$ is a growing function of N, which for $N = 1$ equals $\mathcal{G}'(\bar{u}) = |z_0|/4$ and for $N \to \infty$ is bounded by $\exp(-1) |z_0|$, which proves the desired result. □

For the same reasons as above, we can therefore conclude that the posterior μ_h given by (1.6) is well-defined and Lipschitz in the data. We conclude this simple example with the following result, which is a corollary of Proposition 1.3 in this simple setting.

Corollary 1.6. *Let μ and μ_h be the posterior distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for the parameter u of (1.7) with the exact and approximate forward maps $\mathcal{G}$ and $\mathcal{G}_h$ defined in (1.8) and (1.9). Then, it holds*

$$d_{\mathrm{Hell}}(\mu, \mu_h) \leq Ch,$$

for a constant $C > 0$ independent of h.

Proof. Since both maps $\mathcal{G}$ and $\mathcal{G}_h$ are contractive, and by the order-one convergence of the implicit Euler method, it holds

$$|\mathcal{G}(u) - \mathcal{G}_h(u)| \leq Ch\,|z_0|\,,$$

for a constant $C > 0$ independent of u and h (see [62] or [136, Chapter 7]). Hence, the desired result follows from Proposition 1.3. □

1.2 Finite-Dimensional Approximations

In this section, we introduce a finite-dimensional approximation based on the Karhunen–Loève expansion (KL) of the otherwise infinite-dimensional inverse problem (1.2), which is employed in practice to compute its solution. We refer the reader to the research article [48], and more marginally to [63, 93] for further details and results concerning the content of this section.

Let for simplicity the prior $\mu_0 = \mathcal{N}(0, C_0)$ and let us denote by $\{(\lambda_i, \varphi_i)\}_{i\geq 1}$ the ordered eigenvalues/eigenfunctions of the prior covariance C_0, and remark that $\{\varphi_i\}_{i\geq 1}$ is an orthonormal basis for X. Then, the KL expansion guarantees that the random variable

$$u = \sum_{i\geq 1} \sqrt{\lambda_i}\varphi_i \xi_i,$$

where $\xi \coloneqq \{\xi_i\}_{i\geq 1}$ are independent and identically distributed (i.i.d.) random variables distributed as $\mathcal{N}(0,1)$, satisfies $u \sim \mu_0$. We then let M be a positive integer and truncate the sum above as

$$u = \sum_{i=1}^{M} \sqrt{\lambda_i}\varphi_i \xi_i, \tag{1.10}$$

thus obtaining a function $u \in X$ which is approximately distributed as μ_0. To be precise, let us introduce the space $X^M = \mathrm{span}\{\varphi_i\}_{i=1}^M$, and denote by $X^\perp$ the orthogonal complement of X^M in X. We then denote by $P^M \colon X \to X^M$ the projection operator over the first M elements of the basis $\{\varphi_i\}_{i\geq 1}$ of X, i.e., for any u in X

$$P^M u = \sum_{i=1}^{M} (u, \varphi_i)\varphi_i,$$

with $(\cdot,\cdot)$ the inner product defined on X. Let now $\mathcal{K} \colon \mathbb{R}^M \to X$, $\mathcal{K} \colon \xi \mapsto u$ be the map defined by (1.10), and let $\nu_0^M = \mathcal{N}(0, I)$, with I being the $M \times M$-dimensional identity matrix, be the measure induced by the random variable $\xi = \{\xi_i\}_{i=1}^M$ on the measurable space $(\mathbb{R}^M, \mathcal{B}(\mathbb{R}^M))$. Due to orthogonality, the prior measure μ_0 can then be decomposed as $\mu_0 = \mu_0^M \otimes \mu_0^\perp$, where μ_0^M and $\mu_0^\perp$ are independent probability measures on $(X^M, \mathcal{B}(X^M))$ and $(X^\perp, \mathcal{B}(X^\perp))$, respectively. In particular, μ_0^M is the push-forward measure of ν_0^M through $\mathcal{K}$, i.e.

$$\mu_0^M(B) = \nu_0^M(\{\xi \colon \mathcal{K}(\xi) \in B\}),$$

for all $B \in \mathcal{B}(X^M)$. Let us moreover remark that direct calculations with (1.10) show that $\mu_0^M = \mathcal{N}(0, C_0^M)$, where C_0^M is the truncated spectral decomposition of C_0, i.e.

$$C_0^M = \sum_{i=1}^{M} \lambda_i \varphi_i \varphi_i^\top. \tag{1.11}$$

We then consider the approximated posterior $\widetilde{\mu}$ on $(X, \mathcal{B}(X))$ which is given by

$$\frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\mu_0}(u) = \frac{1}{\widetilde{Z}} \exp(-\Phi(P^M u; y)),$$

where $\widetilde{Z}$ is the normalization constant

$$\widetilde{Z} = \int_X \exp(-\Phi(P^M u; y)) \, \mathrm{d}\mu_0(u).$$

Let us remark that for all $u \in X^\perp$ we have $P^M u = 0$ and therefore

$$\frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\mu_0}(u) = \left(\exp(-\Phi(0; y)) \int_X \mathrm{d}\mu_0(u) \right)^{-1} \exp(-\Phi(0; y)) = 1,$$

so that on $X^\perp$ we have $\widetilde{\mu} = \mu_0$. We can then write $\widetilde{\mu} = \mu^M \otimes \mu^\perp$, where $\mu^\perp = \mu_0^\perp$ and where μ^M is the measure on $(X^M, \mathcal{B}(X^M))$ defined by

$$\frac{\mathrm{d}\mu^M}{\mathrm{d}\mu_0^M}(u) = \frac{1}{Z^M} \exp(-\Phi(u; y)), \tag{1.12}$$

for all $u \in X^M$, where

$$Z^M = \int_{X^M} \exp(-\Phi(u; y)) \, \mathrm{d}\mu_0^M(u).$$

Let us remark that μ^M is a finite-dimensional measure, and that it is therefore amenable for computations. Moreover, reminding the notation $\nu_0^M = \mathcal{N}(0, I)$ for the prior on the coefficients of the KL expansion (1.10), the measure μ^M is the push-forward through $\mathcal{K}$ of the posterior measure ν^M on $(\mathbb{R}^M, \mathcal{B}(\mathbb{R}^M))$ given by

$$\frac{\mathrm{d}\nu^M}{\mathrm{d}\nu_0^M}(\xi) = \frac{1}{Z_\xi^M} \exp(-\Phi(\mathcal{K}(\xi); y)),$$

for $\xi \in \mathbb{R}^M$, with

$$Z_\xi^M = \int_{\mathbb{R}^M} \exp(-\Phi(\mathcal{K}(\xi); y)) \, \mathrm{d}\nu_0^M(u).$$

Indeed, since $\mathcal{K}(\mathbb{R}^M) = X^M$ a change of variable yields $Z_\xi^M = Z^M$ and for all $B \in \mathcal{B}(X^M)$

$$\begin{aligned} \mu^M(B) &= \frac{1}{Z^M} \int_B \exp(-\Phi(u; y)) \, \mathrm{d}\mu_0^M(u) \\ &= \frac{1}{Z_\xi^M} \int_{\{\xi \colon \mathcal{K}(\xi) \in B\}} \exp(-\Phi(\mathcal{K}(\xi); y)) \, \mathrm{d}\nu_0^M(\xi) = \nu^M(\{\xi \colon \mathcal{K}(\xi) \in B\}). \end{aligned}$$

In practice, we therefore compute the posterior ν^M over the M scalar coefficients of the KL expansion (1.10), and then push it forward through the operator $\mathcal{K}$ to obtain the measure μ^M for the unknown. Let us remark that results of convergence for expectations taken with respect to μ^M and μ are stated and rigorously proved in [48], where the main focus are inverse problems involving elliptic PDEs.

1.3 Markov Chain Monte Carlo Methods

In this section, we introduce the class of Markov chain Monte Carlo methods (MCMC), which can be employed to approximate the solution of Bayesian inverse problems. We point the reader to the references [71, 80, 122, 131], where these methods are introduced and extensively analyzed. In Section 1.1, we laid the theoretical basis of the Bayesian approach to inverse problems defined on a Hilbert space X. In particular, under a set of assumptions we guaranteed the well-posedness of a posterior probability measure μ on X which encapsulates the knowledge on an unknown $u \in X$ given finite-dimensional observations $y \in \mathbb{R}^L$. Moreover, in Section 1.2 we introduced a finite-dimensional approximation of the posterior measure, which makes the Bayesian inverse problem amenable in practice. The methods belonging to the MCMC class yield a methodology to finally approximate the solution of the inverse problem.

Let μ be a probability measure on X, let $\Psi\colon X \to \mathbb{R}$ be a smooth real-valued target function defined on X. We consider the problem of approximating the quantity $\mathbb{E}^\mu[\Psi]$, where $\mathbb{E}^\mu$ denotes expectation with respect to μ, i.e.

$$\mathbb{E}^\mu[\Psi] = \int_X \Psi(u)\,\mathrm{d}\mu(u). \tag{1.13}$$

Given a positive integer N_{MC} and a set of i.i.d. realizations $\{u^{(i)}\}_{i=1}^{N_{\mathrm{MC}}}$ in X such that $u^{(1)} \sim \mu$, the Monte Carlo method allows to approximate the quantity of interest $\mathbb{E}^\mu[\Psi]$ with the average $E^\mu[\Psi]$ defined as

$$E^\mu[\Psi] \coloneqq \frac{1}{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} \Psi(u^{(i)}). \tag{1.14}$$

Therefore, given a measure μ, in order to obtain an approximation of the quantity of interest $\mathbb{E}^\mu[\Psi]$ it is sufficient to generate independent samples from μ and computing the averaged functional $E^\mu[\Psi]$. Let us remark that the since the final goal is computing the high-dimensional integral (1.13), Monte Carlo methods can be interpreted as quadrature formulas. The approximation properties of the Monte Carlo method are well-known. In particular, Monte Carlo approximations do not suffer from the so-called curse of dimensionality, i.e., their quality is dimension-independent. Moreover, the Monte Carlo estimate is unbiased, i.e., we have that $\mathbb{E}[E^\mu[\Psi]] = \mathbb{E}^\mu[\Psi]$, where the outer expectation on the left-hand side is taken with respect to the sample $\{u^{(i)}\}_{i=1}^{N_{\mathrm{MC}}}$. Nevertheless, the rate of convergence of $E^\mu[\Psi]$ towards $\mathbb{E}^\mu[\Psi]$ is just of order $1/2$, in the sense

$$\sqrt{\mathrm{Var}(E^\mu[\Psi])} \leq CN^{-1/2},$$

for a constant $C > 0$ independent of the dimension d and where, again, the variance is taken with respect to the sample. Let us consider the case of μ being the posterior measure which we derived in Section 1.1. In this case, (1.3) yields

$$\begin{aligned}\mathbb{E}^\mu[\Psi] &= \int_X \Psi(u)\,\mathrm{d}\mu(u) = \frac{1}{Z}\int_X \Psi(u)\exp(-\Phi(u;y))\,\mathrm{d}\mu_0(u)\\ &= \mathbb{E}^{\mu_0}[Z^{-1}\Psi\exp(-\Phi(\cdot;y))]\end{aligned}$$

where we recall that μ_0 is a Gaussian prior measure on X, and Z the normalization constant given in (1.4). In order to approximate $\mathbb{E}^\mu[\Psi]$, one could then in principle generate i.i.d. samples from the prior μ_0, which is relatively simple in the Gaussian case, and then compute the approximation (1.14) with the function $Z^{-1}\Psi\exp(-\Phi(\cdot;y))$. The problem, evidently, is that the normalization constant Z is unknown. Moreover, approximating Z requires the computation of the high-dimensional integral (1.4), and is therefore as involved as the original problem of approximating $\mathbb{E}^\mu[\Psi]$.

The MCMC methods are designed to approximate expectations under measures for which the normalization constant is unknown, and are therefore adequate for Bayesian inverse problems. As it is suggested by their name, MCMC methods proceed by generating samples which form a Markov chain over the space X. In particular, this Markov chain is ergodic (see Section A.4) with respect to the measure μ, so that if the number of samples is sufficiently high, an average over the chain provides a good approximation of the expectation $\mathbb{E}^{\mu}[\Psi]$ due to the ergodic theorem (see Theorem A.14).

The class of MCMC methods comprises several members. We choose here to describe the random walk Metropolis–Hastings (RWMH) algorithm, first introduced in [64], and analyzed in an infinite-dimensional setting which is suitable for inverse problems in [63]. With a reference to the notation introduced in Section 1.2, let $X^M \subset X$ be the finite-dimensional subspace $X^M = \mathrm{span}\{\varphi_i\}_{i=1}^M$, and let $\mu_0^M = \mathcal{N}(0, C_0^M)$, where C_0^M is the covariance defined in (1.11). Moreover, let $Q^M = \mathcal{N}(0, C_Q^M)$ be a Gaussian distribution on X^M, which we call the proposal distribution. Given an initial guess $u^{(1)} \in X^M$, the RWMH proceeds for $i = 2, \dots, N_{\mathrm{MC}}$ as

(i) Sample $\Delta u^{(i)} \sim Q_M$ and set $\widehat{u}^{(i)} \sim u^{(i-1)} + \Delta u^{(i)}$;

(ii) Set $u^{(i)} = \widehat{u}^{(i)}$ with probability α, and $u^{(i)} = u^{(i-1)}$ with probability $1 - \alpha$, where

$$\begin{aligned} \alpha &= \min\left\{\exp(\widehat{\alpha}), 1\right\}, \\ \widehat{\alpha} &= -\Phi(\widehat{u}^{(i)}; y) + \Phi(u^{(i-1)}; y) \\ &\quad - \frac{1}{2}\left(\left(\widehat{u}^{(i)}, (C_0^M)^{-1}\widehat{u}^{(i)}\right) - \left(u^{(i-1)}, (C_0^M)^{-1}u^{(i-1)}\right)\right), \end{aligned} \tag{1.15}$$

and where $\Phi(\cdot; y)$ is the potential of (1.5).

Let us remark that the acceptance probability α above is, in practice, the ratio between the posterior densities evaluated at the previous guess. Indeed, let for the sake of clarity μ^M admit a density π^M with respect to the Lebesgue measure, so that (1.12) can be rewritten in terms of densities as

$$\pi^M(u) = \frac{1}{Z^M} \exp\left(-\Phi(u; y)\right) \pi_0^M(u),$$

where π_0^M is the prior Gaussian density

$$\pi_0^M(u) = \frac{1}{Z_0^M} \exp\left(-\frac{1}{2}(u, (C_0^M)^{-1}u)\right),$$

with Z_0^M being the normalization constant. Then, (1.15) can be rewritten as

$$\alpha = \min\left\{\frac{\pi^M(\widehat{u}^{(i)})}{\pi^M(u^{(i-1)})}, 1\right\},$$

i.e., as the ratio of the posterior density computed in the proposed value $\widehat{u}^{(i)}$ and on the previous value $u^{(i-1)}$.

Remark 1.7. We note that the proposal distribution is the only element of the RWMH algorithm which can be tuned. The easiest choice, at least for implementation, would be to fix $Q^M = \mathcal{N}(0, \sigma^2 I)$ for some user-prescribed variance σ^2. Unfortunately, the quality of the resulting Markov chain is not robust with respect to σ. In particular, if σ is too small, the probability to accept is too large and the Markov chain fails to effectively explore the posterior. At the other end of the spectrum, if σ is too large the probability of accepting a new sample reduces drastically, and the Markov chain presents a sticky behavior. One possible choice is to employ the robust adaptive Metropolis algorithm (RAM) (see [143] for details), in which the proposal covariance C_Q^M is adapted on the fly to obtain a user-specified final acceptance ratio, i.e. the ratio between the

accepted and the total number samples, which should be roughly 25% (see e.g. [143]). An option specifically tailored for high-dimensional inverse problems is the preconditioned Crank–Nicolson MCMC (pCN-MCMC), which is presented in details and analyzed in [41, 63].

2 Multiscale Ensemble Kalman Inversion

In this chapter, we introduce a methodology based on homogenization and on the Ensemble Kalman filter (EnKF) to solve multiscale inverse problems. In particular, we consider inverse problems driven by elliptic PDEs with highly oscillatory tensors, and apply homogenization and the finite element heterogeneous multiscale method (FE-HMM) to solve the forward problem, and the EnKF for the inverse problem. This chapter is based on our research article [9], from which we borrow some phrasings here, and is one of the original contributions of this thesis.

The outline of the chapter is as follows. In Section 2.1 we briefly introduce the Kalman filter and the ensemble Kalman filter (EnKF) in their natural context of state estimation for dynamical systems. Then, in Section 2.2 we introduce the technique of ensemble Kalman inversion, which allows to compute the solution to inverse problems as the ones presented in Chapter 1, in both a pointwise and a Bayesian fashions. The remainder of the chapter is dedicated to the application of ensemble Kalman inversion to problems involving multiscale elliptic PDEs. In particular, we present the problem in Section 2.3 and state our main theoretical results in Section 2.4, which are then proved in Section 2.5. In Section 2.6, we consider a standard technique to account for modeling error in the solution of the inverse problem, and prove two novel theoretical results which allow to balance precision and computational overheads. Finally, we show a numerical experiment in Section 2.7, which corroborates our analysis and showcases the potential of our new methodology.

2.1 Kalman and Ensemble Kalman Filters

In this section, we give a general introduction of the Kalman and the EnKF. The seminal work by Kalman, which first proposed the algorithm we present here, is [73]. The EnKF was introduced in [53], and we furthermore point the reader to the works [54, 55], which are standard references in the field.

Let Z be a Hilbert space, let $\Xi\colon Z \to Z$ and let $z_0 \in Z$ be an initial value. We consider the discrete recursion

$$z_n = \Xi(z_{n-1}), \quad n = 1, 2, \dots \tag{2.1}$$

We equip the dynamics above with an observation model. In particular, let $Y \subset Z$ be a Hilbert space and let $H\colon Z \to Y$ be a linear map. We then consider observations to be given by

$$y_n = Hz_n + \eta_n, \tag{2.2}$$

where we assume that $\{\eta_n\}_{n\geq 1}$ is a sequence of i.i.d. random variables such that $\eta_1 \sim \mathcal{N}(0, \Gamma)$, where Γ is a positive definite covariance matrix on Y. Kalman filters proceed by updating a

probability distribution on Z which describes the partially-observed dynamics and which accounts for both the recurrence relation (2.1) and for the observation model (2.2). In the following, we first consider the standard Kalman filter, which is an exact model for updating Gaussian distributions when Ξ is a linear map, and then proceed with the EnKF, which is a particle-based approximation for general non-linear dynamics.

2.1.1 The Kalman Filter

Kalman filters proceed recursively to estimate the state of dynamics of the form (2.1) when observations are provided by the model (2.2). At each time n, the estimation is performed in two steps. First, equation (2.1) is employed in the so-called prediction step, and then (2.2) is employed to correct the prediction in the update or analysis step. In this section, we assume for simplicity the observable space Y to be finite-dimensional.

Let Ξ be a linear map, and let us assume that at the time instant $n-1$ the state z_{n-1} is distributed as a Gaussian $z_{n-1} \sim \mathcal{N}(m_{n-1}, C_{n-1})$, where $m_{n-1} \in Z$ and C_{n-1} is a positive-definite covariance operator on Z. Employing the dynamics (2.1), we obtain an updated distribution for the next time step, which is Gaussian due to the assumption of linearity on the dynamics. At the prediction step, we indeed get the partially-updated random variable $\widehat{z}_n$ which satisfies

$$\widehat{z}_n \sim \mathcal{N}(\widehat{m}_n, \widehat{C}_n), \qquad \widehat{m}_n = \Xi m_{n-1}, \quad \widehat{C}_n = \Xi C_{n-1} \Xi^*, \tag{2.3}$$

where $\Xi^* \colon Z \to Z$ is the adjoint of Ξ. The distribution over $\widehat{z}_n$ can be interpreted as a prior knowledge on the value of the state at the next time, before observations are assimilated. We now consider the observations, and thus the update step. Due to the Gaussian assumption on the noise $\eta_n \sim \mathcal{N}(0, \Gamma)$ and since we assume Y to be finite-dimensional, we have that the likelihood of y_n given a value $\widehat{z}_n$ can be written up to a proportionality constant as

$$L(y_n \mid \widehat{z}_n) \propto \exp\left(-\frac{1}{2} \left\| \Gamma^{-1/2} (y_n - H\widehat{z}_n)^\top \right\|_2^2 \right).$$

The update step is then given by Bayes' rule. Denoting by $\widehat{\mu}_n$ the Gaussian distribution of the partially updated state $\widehat{z}_n$, taking into account the data at time n yields up to a proportionality constant

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\widehat{\mu}_n}(\widehat{z}_n \mid y_n) \propto L(y_n \mid \widehat{z}_n).$$

Let us remark that μ_n is then a Gaussian measure (see e.g. [131, Section 6.4]). In particular, one obtains by completing the square

$$\begin{aligned} &\widehat{z}_n \mid y_n \sim \mathcal{N}(m_n, C_n), \\ &m_n = \widehat{m}_n + K_n \left(y_n - H\widehat{m}_n \right), \quad C_n = (I - K_n H)\widehat{C}_n, \\ &\text{where } K_n \colon Y \to Z, \quad K_n = \widehat{C}_n H^* R_n, \\ &\text{with } R_n \colon Y \to Y, \quad R_n = \left(H\widehat{C}_n H^* + \Gamma \right)^{-1}, \end{aligned} \tag{2.4}$$

where H^* is the adjoint of H. The recursion is completed by defining the state at the next time as $z_n \coloneqq \widehat{z}_n \mid y_n$. Let us comment on the update formula above. The matrix $K_n \colon Y \to Z$, called the Kalman gain, weighs the importance of the data with respect to the dynamics. Indeed, let us consider heuristically two extreme cases:

- If the dynamics are more certain than the observations, i.e., "$\widehat{C}_n \ll \Gamma$" then

$$\begin{aligned} R_n \approx \Gamma^{-1} \implies K_n \approx \widehat{C}_n H^* \Gamma^{-1} \approx 0 \\ \implies m_n \approx \widehat{m}_n, \quad C_n \approx \widehat{C}_n, \end{aligned}$$

so that the prior prediction is not majorly influenced by the assimilation of new data and the update step is only driven by the dynamics;

- If the observations are more certain than the dynamics, i.e., "$\Gamma \ll \widehat{C}_n$" then

$$R_n \approx (H^*)^{-1}\widehat{C}_n^{-1}H^{-1} \implies K_n \approx \widehat{C}_n H^*(H^*)^{-1}\widehat{C}_n^{-1}H^{-1} = H^{-1}$$
$$\implies m_n \approx H^{-1}y_n, \quad C_n \approx 0,$$

so that the posterior mean m_n is only determined by the data, and the posterior has a high precision independently of the previous state.

In all situations which are mid-range between the two examined above, the Kalman gain yields the correct balance, in a Bayesian sense, between the prior belief and the precision with which we observe data. Resuming all the considerations above, we now express the Kalman filter in an algorithmic form. Given an initial state $z_0 \sim \mathcal{N}(m_0, C_0)$, with C_0 being possibly singular, the algorithm proceeds for $n = 1, 2, \ldots$ as

(i) Propagate the random variable $z_{n-1} \to \widehat{z}_n$ with (2.3);

(ii) Update the random variable to the next time point $\widehat{z}_n \to z_n$ with (2.4).

2.1.2 The Ensemble Kalman Filter

The EnKF is a Monte Carlo-type approximation of the Kalman filter presented above in case the dynamics are nonlinear. Let us point out that without either the Gaussian assumption on the initial state or the linearity of the dynamics, it would not be possible to find a closed-form solution to the recursive estimation of the state. The EnKF proceeds by propagating and updating an ensemble of "particles", which empirically approximate the distribution of the state. In particular, let J be a positive integer and let $\{z_{n-1}^{(j)}\}_{j=1}^J$, with $z_{n-1}^{(j)} \in Z$ for all $j = 1, \ldots, J$, be the ensemble at time $n-1$. As for the Kalman filter, we then follow a prediction-update procedure for propagating the particles. Indeed, the prediction step yields a partially-updated ensemble $\{\widehat{z}_n^{(j)}\}_{j=1}^J$ defined by

$$\widehat{z}_n^{(j)} = \Xi(\widehat{z}_{n-1}^{(j)}), \quad j = 1, \ldots, J, \quad n = 1, 2, \ldots, \tag{2.5}$$

i.e., the step in the dynamics is applied independently to each particle of the ensemble. Let us remark that in the Kalman filter the prediction mean and covariance $\widehat{m}_n$ and $\widehat{C}_n$ are necessary for the computation of the Kalman gain, and hence of the updated estimation. Therefore, we compute here and denote again by $\widehat{m}_n$ and $\widehat{C}_n$ the sample mean and covariance of the predicted ensemble, i.e.,

$$\widehat{m}_n = \frac{1}{J}\sum_{j=1}^J z_n^{(j)}, \quad \widehat{C}_n = \frac{1}{J}\sum_{j=1}^J \left(z_n^{(j)} - \widehat{m}_n\right) \otimes \left(z_n^{(j)} - \widehat{m}_n\right), \tag{2.6}$$

where $\otimes$ denotes the tensor product in Z. The update step is then given for $j = 1, \ldots, J$ by

$$\begin{aligned} &z_n^{(j)} = \widehat{z}_n^{(j)} + K_n\left(y_n^{(j)} - H\widehat{z}_n^{(j)}\right), \\ &\text{where } K_n = \widehat{C}_n H^* R_n, \\ &\text{with } R_n = \left(H\widehat{C}_n H^* + \Gamma\right)^{-1}, \end{aligned} \tag{2.7}$$

which is equivalent particle-by-particle to the update of the Kalman filter (2.4), with the empirical covariance replacing the exact predicted covariance. The only detail which is different concerns

the observations $y_n^{(j)}$, which are given by

$$y_n^{(j)} = y_n + \eta_n^{(j)}, \quad j = 1, \dots, J$$

where $\eta_n^{(j)} \sim \mathcal{N}(0, \Gamma)$ are i.i.d. random variables distributed as the noise. In practice, randomizing the data across the ensemble allows for a better exploration of the state space, and therefore for an enhanced diversity in the ensemble. Let us remark that the empirical mean and covariance of the updated ensemble serve as a point estimate and a confidence indicator in state estimation. Therefore, given an initial ensemble $\{z_0^{(j)}\}_{j=1}^J$, the EnKF proceeds for $n = 1, 2, \dots$ as

(i) Propagate the ensemble $\{z_{n-1}^{(j)}\}_{j=1}^J \to \{\widehat{z}_n^{(j)}\}_{j=1}^J$ with (2.5);

(ii) Compute the predicted mean and covariance $\widehat{m}_n$ and $\widehat{C}_n$ with (2.6);

(iii) Update the ensemble $\{\widehat{z}_n^{(j)}\}_{j=1}^J \to \{z_n^{(j)}\}_{j=1}^J$ with (2.7).

The initial ensemble $\{z_0^{(j)}\}_{j=1}^J$ can be constructed in different ways. For example, if the initial state $z_0 \in Z$ of the dynamics is known, then one can choose to have J replicas of z_0 in the initial ensemble, randomized with an appropriate source of noise. Otherwise, if the distribution of z_0 is known, then a good choice is to fix $z_0^{(j)} \sim z_0$ i.i.d. for $j = 1, \dots, J$.

Remark 2.1. Let us assume that the computational bottleneck for the execution of the EnKF is the evaluation of the map Ξ. In this case, the prediction step (2.5) is computationally dominant for a run of the EnKF. Let us suppose that we run the algorithm for N steps, and with J particle. Then, the prediction step is repeated $N \cdot J$ times, and therefore the cost is equivalent to $N \cdot J$ evaluations of Ξ. Nonetheless, let us remark that the prediction step (2.5) can be easily parallelized, since the forward operator is applied independently to each particle. Hence, for a reasonable number of particles (or a high number of computing units), we have that the overall cost is of order $\mathcal{O}(N)$.

2.2 Ensemble Kalman Inversion

In this section, we present the ensemble Kalman inversion technique for inverse problems. We refer the reader to the research articles [35, 67, 68, 114, 125], where this methodology has been developed and analyzed extensively.

Let X and Y be Hilbert spaces, with Y being finite-dimensional for simplicity, and let us consider the inverse problem

$$\text{find } u \in X \text{ given observations } y = \mathcal{G}(u) + \eta \in Y, \tag{2.8}$$

where the operator $\mathcal{G}\colon X \to Y$ is a generic forward map and the noise η follows the Gaussian distribution $\eta \sim \mathcal{N}(0, \Gamma)$ with a symmetric positive definite covariance Γ. The problem (2.8) is static, and in order to apply the Kalman filtering techniques described above it is therefore necessary to introduce dynamics artificially. For this purpose, let us consider the product space $Z = X \times Y$ and the map $\Xi\colon Z \to Z$ given by

$$\Xi(z) = \begin{pmatrix} u \\ \mathcal{G}(u) \end{pmatrix}, \quad \text{for} \quad z = \begin{pmatrix} u \\ v \end{pmatrix} \in Z.$$

The dynamics are then given by the recursion

$$z_n = \Xi(z_{n-1}), \quad n = 1, 2, \dots, \tag{2.9}$$

which, consistently with the problem (2.8), are equipped with the observation equation

$$y_n = H z_n + \eta_n, \tag{2.10}$$

where $H\colon Z \to Y$ is the projection operator defined by $H = (0 \quad I)$ and $\{\eta_n\}_{n\in\mathbb{N}}$ is an i.i.d. sequence of random variables distributed identically to the noise of the inverse problem (2.8), i.e., $\eta_n \sim \mathcal{N}(0,\Gamma)$. In fact, let us remark that combining (2.9) and (2.10) one gets $y_n = \mathcal{G}(u_n) + \eta_n$, which is in law equivalent to the equality appearing in (2.8). The EnKF described in Section 2.1.2 can then be seamlessly applied to the dynamics (2.9) and (2.10).

The initialization and termination of the EnKF are peculiar in the context of inverse problems. In particular, coherently to this framework and with a reference to Chapter 1, we assume prior knowledge is available on the parameter $u \in X$ and that it is summarized by a probability measure μ_0 on X. In this case, one can draw J i.i.d. samples $\psi^{(j)}$ from μ_0 and fix the initial ensemble as

$$z_0^{(j)} = \begin{pmatrix} \psi^{(j)} \\ \mathcal{G}(\psi^{(j)}) \end{pmatrix}.$$

Let us remark that during the run of the EnKF, the X-component of the particles never leaves the set $\mathcal{A} = \operatorname{span}\{\psi^{(j)}\}_{j=1}^J$ [68, Theorem 2.1]. Hence, another possible choice would be to choose a J-dimensional subset $\mathcal{A} \subset X$ beforehand, and then fix $\{\psi^{(j)}\}_{j=1}^J$ to be a basis for $\mathcal{A}$. For example, the set $\{\psi^{(j)}\}_{j=1}^J$ could be a basis of eigenfunctions for the covariance operator of a Gaussian prior measure μ_0, as illustrated in Section 1.2.

Concerning termination, let us fix the total number of iterations of the EnKF to a positive integer N. At the final step, we project the particles on the space X and compute a sample average of the ensemble to obtain the estimate

$$u_{\mathrm{EnKF}} = \frac{1}{J}\sum_{j=1}^J H^\perp z_N^{(j)} = \frac{1}{J}\sum_{j=1}^J u_N^{(j)},$$

where $H^\perp\colon Z \to X$ is defined by $H^\perp = (I \quad 0)$. The value $u_{\mathrm{EnKF}} \in X$ then serves as a point estimate for the solution of the inverse problem (2.8).

2.2.1 The Bayesian Interpretation

It is possible without any additional cost to recast the EnKF inversion in a Bayesian framework as the one of Chapter 1. We refer the reader to [125] for more details and further insight on the content of this section. Let us recall that in the Bayesian framework, given a prior measure μ_0 on X, then the solution to the inverse problem is expressed in terms of a posterior distribution μ whose Radon–Nykodim derivative is given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = \frac{1}{Z}\exp\left(-\Phi(u;y)\right), \tag{2.11}$$

where, due to the Gaussian assumption on the noise and since Y is finite-dimensional, the potential is given by

$$\Phi(u;y) = \frac{1}{2}\left\|\Gamma^{-1/2}(y - \mathcal{G}(u))\right\|_2^2,$$

and where Z is the normalization constant

$$Z = \int_X \exp\left(-\Phi(u;y)\right)\,\mathrm{d}\mu_0(u). \tag{2.12}$$

We now let N be a positive integer and $\Delta = 1/N$ be a "step size". The transition from the prior μ_0 to the posterior μ can be split in N steps by introducing the probability measures μ_n on X defined recursively for $n = 1, 2, \ldots, N$ as

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_{n-1}}(u) = \frac{1}{Z_n}\exp\left(-\Delta\Phi(u;y)\right), \tag{2.13}$$

where Z_n is the normalization constant, defined as above. We then have by the chain rule

$$\frac{\mathrm{d}\mu_N}{\mathrm{d}\mu_0}(u) = \prod_{n=1}^{N} \frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_{n-1}}(u) = \left(\prod_{n=1}^{N} Z_n\right)^{-1} \exp(-\Phi(u;y)).$$

Let us remark that for all $n = 1, \dots, N$ the normalizing constants satisfy

$$\begin{aligned} Z_n &= \int_X \exp(-\Delta\Phi(u;y))\,\mathrm{d}\mu_{n-1}(u) \\ &= \int_X \exp(-\Delta\Phi(u;y))\frac{\mathrm{d}\mu_{n-1}}{\mathrm{d}\mu_{n-2}}(u)\,\mathrm{d}\mu_{n-2}(u) \\ &= \frac{1}{Z_{n-1}}\int_X \exp(-2\Delta\Phi(u;y))\,\mathrm{d}\mu_{n-2}(u) \\ &= (\cdots) = \frac{1}{Z_{n-1}Z_{n-2}\cdots Z_1}\int_X \exp(-n\Delta\Phi(u;y))\,\mathrm{d}\mu_0(u). \end{aligned}$$

Therefore, it clearly holds

$$\prod_{n=1}^{N} Z_n = \int_X \exp(-\Phi(u;y))\,\mathrm{d}\mu_0(u) = Z,$$

where Z is given in (2.12). Hence, we have that $\mu_N = \mu$, and in words we have that doing N steps of the form (2.13) gradually transforms the prior μ_0 in the posterior μ.

Let us consider the prediction step (2.5) and let us modify the update step (2.7) by replacing the observation error covariance Γ by the rescaled matrix $\Delta^{-1}\Gamma$. Then, let $\{u_n^{(j)}\}_{j=1}^J$ be the set of the X components of the ensemble of particles $\{z_n^{(j)}\}_{j=1}^J$ at the n-th step of the EnKF. Moreover, let

$$\widehat{\mu}_n(du) = \frac{1}{J}\sum_{j=1}^{J} \delta_{u_n^{(j)}}(du),$$

where δ_u is the Dirac mass concentrated in $u \in X$, be the empirical distribution on X which is induced by the ensemble. It has been shown in [125] that $\widehat{\mu}_n$ is a good approximation of the measure μ_n defined in (2.13). Therefore, the evolution of the ensemble in the EnKF algorithm, and thus of the empirical measures it induces on X, mimics the step-by-step evolution of the prior distribution μ_0 to the posterior μ. Hence, after N steps the algorithm produces a set of samples from the posterior which can be employed in the same manner as other Monte Carlo algorithms for solving in practice the inverse problem and give a full uncertainty quantification of the inversion procedure (see Section 1.3).

Remark 2.2. Due to their sequential nature, the computational cost of MCMC algorithms is of the order of $\mathcal{O}(J)$ evaluations of the forward map $\mathcal{G}$ to obtain J samples approximately distributed from the posterior distribution. Conversely, in case sufficient computational power is available the EnKF requires $\mathcal{O}(N)$ evaluations of the forward map $\mathcal{G}$ to obtain the same number of samples from the posterior, as per Remark 2.1, where N is the number of steps which are necessary to reach the posterior $\mu = \mu_N$ from the prior μ_0. Therefore, choosing a large ensemble and letting it evolve over relatively few steps, i.e., if $N \ll J$, the computational advantage of the EnKF over standard MCMC algorithms is relevant.

2.3 Multiscale Ensemble Kalman Inversion

In this chapter, we consider the application of ensemble Kalman inversion to a multiscale inverse problem of the form

$$\text{find } u \in X \text{ given observations } y = \mathcal{G}^\varepsilon(u) + \eta \in Y, \tag{2.14}$$

where $\varepsilon > 0$ is the multiscale parameter, which often is $\varepsilon \ll 1$, the operator $\mathcal{G}^\varepsilon \colon X \to Y$ is the multiscale forward map and where, as above, $\eta \sim \mathcal{N}(0, \Gamma)$ for some symmetric positive definite covariance Γ on Y. Let $D \subset \mathbb{R}^d$ be an open bounded domain and let $H_0^1(D)$ denote the space of functions $v \colon D \to \mathbb{R}$ in $L^2(D)$ with first order weak derivatives in $L^2(D)$ and whose trace on ∂D vanishes. We consider the forward map $\mathcal{G}^\varepsilon$ to be the composition $\mathcal{G}^\varepsilon = \mathcal{O} \circ \mathcal{S}^\varepsilon$ of an observation operator $\mathcal{O} \colon H_0^1(D) \to Y$ and a multiscale solution operator $\mathcal{S}^\varepsilon \colon X \to H_0^1(D)$. In particular, for $u \in X$, the operator $\mathcal{S}^\varepsilon \colon u \mapsto p^\varepsilon \in H_0^1(D)$ where p^ε is the weak solution of the elliptic PDE

$$\begin{cases} -\nabla \cdot (A_u^\varepsilon \nabla p^\varepsilon) = f, & \text{in } D, \\ p^\varepsilon = 0, & \text{on } \partial D, \end{cases} \tag{2.15}$$

for a right-hand side $f \in L^2(D)$. We assume that the tensor $A_u^\varepsilon \colon D \to \mathbb{R}^{d \times d}$ is a parametrized multiscale tensor admitting explicit scale separation between slow and fast spatial variables, i.e.,

$$A_u^\varepsilon(x) = A\left(u(x), \frac{x}{\varepsilon}\right),$$

where the map $(t, x) \mapsto A(t, x/\varepsilon)$ is assumed to be known and where A is periodic in its second argument. In other words, the unknown u of the inverse problem (2.14) governs the slow-scale variations of the rapidly-oscillating tensor A_u^ε.

Let us consider now the application of ensemble Kalman inversion to the inverse problem (2.14). Since the PDE (2.15) does not in general admit a closed-form solution, one has to employ a numerical approximation to evaluate the forward map $\mathcal{G}^\varepsilon$. If ε is small and we employ the finite element method (FEM), a fine discretization is needed to resolve the smallest scale and thus evaluate the forward operator $\mathcal{G}^\varepsilon$, which clearly leads to a high computational cost. Indeed, as per Remark 2.1, a run of the EnKF algorithm would lead to $\mathcal{O}(N)$ solutions of (2.15), which is clearly unfeasible.

In order to approach the multiscale problem more efficiently we recur to the theory of homogenization (see the standard references [22,34,113]), which ensures the existence of a non-oscillating homogenized tensor A_u^0, such that for $\varepsilon \to 0$ the solution p^ε of (2.15) tends weakly in $H_0^1(D)$ to the solution p^0 of the problem

$$\begin{cases} -\nabla \cdot (A_u^0 \nabla p^0) = f, & \text{in } D, \\ p^0 = 0, & \text{on } \partial D. \end{cases} \tag{2.16}$$

Hence, the homogenized problem is a good surrogate of (2.15) when $\varepsilon \ll 1$, and its non-oscillating nature allows us to discretize it with FEM on an arbitrarily coarse mesh, whose maximum diameter is denoted by h. Therefore, denoting by $\mathcal{G}_h^0 \colon \mathcal{O} \circ \mathcal{S}_h^0$, where $\mathcal{S}_h^0 \colon u \mapsto p_h^0$, the numerical solution of (2.16), we study in this chapter the behavior of the EnKF when $\mathcal{G}^\varepsilon$ is replaced by its cheap approximation $\mathcal{G}_h^0$. Let us denote by $\{u_{n,h}^{0,(j)}\}_{j=1}^J$ the X components of the ensemble obtained after n iterations of the EnKF algorithm with the forward operators $\mathcal{G}_h^0$ in the prediction step (2.5). With this notation, given an initial ensemble $\{u_{0,h}^{0,(j)}\}_{j=1}^J$, at each step $n = 0, 1, \ldots, N-1$, our algorithm proceeds as

(i) For each $u_{n,h}^{0,(j)}$, evaluate numerically the forward map $\mathcal{G}_h^0$, thus completing the prediction step (2.5);

(ii) Perform the analyis step (2.7) to obtain the updated ensemble $\{u_{n+1,h}^{0,(j)}\}_{j=1}^J$.

We evaluate the forward map $\mathcal{G}_h^0$ employing the finite element heterogeneous multiscale method (FE-HMM) [2, 5]. The FE-HMM yields an approximation of the solution of (2.16) computed on a macro-mesh with size h with a two steps procedure. First, one computes the value of the

homogenized tensor A_u^0 on the quadrature points through the solution of appropriate elliptic equations, which are called in literature the cell problems. Then, one employs these pre-computed values to obtain the homogenized solution on the macro-mesh with standard finite elements.

Remark 2.3. Other methodologies for solving (2.14) have been developed in [3,4,100]. In particular, while the focus of [100] are one-dimensional problems, in [3] the authors solve the same inverse problem we consider here by means of Tikhonov regularization, and in [4] they employ MCMC techniques as the one presented in Section 1.3.

2.4 Statement of the Main Results

Let us first introduce some assumptions and notation which will be employed in the analysis. First, we introduce a regularity assumption on tensors which will be fulfilled by A_u^ε and A_u^0 for our analysis.

Assumption 2.4. The tensor $A_u \colon D \to \mathbb{R}^{d\times d}$ satisfies for all $u, u_1, u_2 \in X$ and $\xi \in \mathbb{R}^d$

$$\|A_{u_1} - A_{u_2}\|_{L^\infty(D;\mathbb{R}^{d\times d})} \leq M \|u_1 - u_2\|_X , \qquad A_u\xi \cdot \xi \geq \alpha_0 \|\xi\|_2^2 ,$$

where M and α_0 are positive constants.

We now introduce a regularity assumption on the observation operator.

Assumption 2.5. The observation operator $\mathcal{O} \colon H_0^1(D) \to Y$ satisfies for all $p_1, p_2 \in H_0^1(D)$

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_Y \leq C_\mathcal{O} \|p_1 - p_2\|_{L^2(D)} ,$$

where $C_\mathcal{O}$ is a positive constant.

Note that since $\mathcal{O}$ is defined on $H_0^1(D) \subset L^2(D)$, Assumption 2.5 is stronger than Lipschitz continuity. Finally, we introduce an assumption on the algorithm which will be employed in the analysis.

Assumption 2.6. All the particles in the ensemble lie at each iteration in a ball $B_R(u^*)$ for some $R > 0$ sufficiently big, where u^* is the true value of the unknown.

Remark 2.7. Let us remark that Assumption 2.6 reduces the possible outcomes of the algorithm even if R can be chosen arbitrarily big. Therefore, in the following all the expectations in the statements and in the proofs have to be intended as conditional expectations given that all particles lie in a ball $B_R(u^*)$ centered in the true value of the unknown. For example, when we write the expectation of the norm (see (2.17)) of an ensemble $u_N = \{u_N^{(j)}\}_{j=1}^J$ of particles at the N-th step of the algorithm what we mean is

$$\mathbb{E}\left[\|u_N\|\right] = \mathbb{E}\left[\|u_N\| \mid u_n^{(j)} \in B_R(u^*), \ \ \forall j = 1, \dots, J, \, n = 0, \dots, N\right].$$

This abuse of notation is repeated throughout this chapter, and expectations should be thought as above anytime Assumption 2.6 holds.

For clarity, we present the analysis in the finite-dimensional setting $X = \mathbb{R}^M$ and $Y = \mathbb{R}^L$ but claim that it can be readily generalized to the infinite-dimensional case. For an ensemble $u = \{u^{(j)}\}_{j=1}^J$ of particles in $\mathbb{R}^M$, we introduce the ensemble norm

$$\|u\| \coloneqq \frac{1}{J} \sum_{j=1}^{J} \left\|u^{(j)}\right\|_2 , \tag{2.17}$$

which is indeed a norm and where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^M$. Moreover, given a scalar α, we define the linear combination $w = u + \alpha v$ between two ensembles u and v with the same number of particles J as $\{w^{(j)} = u^{(j)} + \alpha v^{(j)}\}_{j=1}^J$.

We can now present the first main result of this work, in which we study the convergence of the ensemble obtained by the EnKF employing $\mathcal{G}_h^0$ to the one obtained employing the exact operator $\mathcal{G}^\varepsilon$ linked to the PDE (2.15).

Theorem 2.8. *Let $u_{N,h}^0 = \{u_{N,h}^{0,(j)}\}_{j=1}^J$, $u_N^\varepsilon = \{u_N^{\varepsilon,(j)}\}_{j=1}^J$ be the ensembles after N iterations of the EnKF method with forward operators $\mathcal{G}_h^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, if A_u^ε and A_u^0 satisfy Assumption 2.4 and if Assumption 2.5 and Assumption 2.6 hold, we have*

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_{N,h}^0\right\|\right] \to 0 \qquad \text{as } \varepsilon, h \to 0.$$

In particular, if the exact solution p^0 of the homogenized problem (2.16) is in $H^{q+1}(D)$ with $q \geq 1$ and we employ polynomials of degree r for the finite element basis, then

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_{N,h}^0\right\|\right] \leq C(\varepsilon + h^{s+1}),$$

where $s = \min\{r, q\}$ and $C > 0$ is a constant independent of h and ε.

The proof of this result is the main focus of Section 2.5.1. The second main theoretical result concerns the Bayesian interpretation of the EnKF methodology for inverse problems in the multiscale setting. Let μ_0 be a prior measure on X and the ensembles $u_{N,h}^0 = \{u_{N,h}^{0,(j)}\}_{j=1}^J$, $u_N^\varepsilon = \{u_N^{\varepsilon,(j)}\}_{j=1}^J$ resulting from the EnKF algorithms as in Theorem 2.8 both initialized with an i.i.d. sample from μ_0. We consider the discrete probability measures

$$\mu^\varepsilon = \frac{1}{J}\sum_{j=1}^J \delta_{u_N^{\varepsilon,(j)}} \qquad \text{and} \qquad \mu_h^0 = \frac{1}{J}\sum_{j=1}^J \delta_{u_{N,h}^{0,(j)}}, \tag{2.18}$$

i.e., the EnKF approximations of the posterior μ on u defined in (2.11). Our goal is providing a metric on how far the two measures are from each other with respect to ε and h. Let us remark that due to the randomization of the data at each step of the EnKF algorithm, both μ^ε and μ_h^0 are random probability measures. We now introduce the metric we consider for comparing the two measures, which is the equivalent to weak convergence in the context of random measures (see Section A.1).

Definition 2.9. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A sequence of random measures $\{\mu_n\}_{n\in\mathbb{N}}$ on a metric space $(E, \mathcal{B}(E))$ dependent on a random variable ξ on $(\Omega, \mathcal{F}, P)$ is said to weakly converge in $L^1(\Omega)$ to a random measure μ on the same metric space if for all bounded continuous functions $f \in C_B^0(E)$ we have

$$\mathbb{E}_\xi\left[\left|\int_E f\, d\mu_n - \int_E f\, d\mu\right|\right] \to 0.$$

In this case we write $\mu_n \xrightarrow{L^1} \mu$.

We can now state our second main result, whose proof is the main focus of Section 2.5.2.

Theorem 2.10. *Let the hypotheses of Theorem 2.8 be satisfied. Then the sequence of random measures $\{\mu^\varepsilon - \mu_h^0\}_{\varepsilon,h}$, where μ^ε and μ_h^0 are defined in (2.18), satisfies*

$$\{\mu^\varepsilon - \mu_h^0\}_{\varepsilon,h} \xrightarrow{L^1} 0 \qquad \text{as } \varepsilon, h \to 0.$$

Remark 2.11. It is possible to verify that in both Theorem 2.8 and Theorem 2.10 the limits with respect to ε and h can be interchanged.

2.5 Convergence Analysis

In this section we prove Theorem 2.8 and Theorem 2.10, the main results of this chapter. As announced above, the analysis is carried out in the finite-dimensional case $X = \mathbb{R}^M$ and $Y = \mathbb{R}^L$, but it can be generalized to the infinite-dimensional setting. For the purpose of the analysis, we introduce on top of the forward maps $\mathcal{G}^\varepsilon$ and $\mathcal{G}_h^0$, which have been introduced in Section 2.3, the operator $\mathcal{G}^0 = \mathcal{O} \circ \mathcal{S}^0$, where $\mathcal{S}^0 \colon X \to H_0^1(D)$ is the exact solution operator associated with the homogenized PDE (2.16).

2.5.1 Convergence of the Point Estimate

We now focus on Theorem 2.8. It is clear from the desired bound that the effects of homogenization and discretization can be analysed separately. In particular, we first show the convergence of the ensemble generated employing the forward operator $\mathcal{G}^\varepsilon$ to the one generated employing the exact homogenized operator $\mathcal{G}^0$ for $\varepsilon \to 0$. Then, in an analogous fashion, we prove the convergence of the ensemble generated with $\mathcal{G}_h^0$ to the ensemble generated employing $\mathcal{G}^0$. In order to introduce a compact notation, we denote by $\mathcal{U}_{J,M}$ the set of ensembles of dimension J with elements in $\mathbb{R}^M$ and we consider the homogenization error function $e \colon \mathbb{R} \times \mathcal{U}_{J,M} \to \mathbb{R}$, which is defined for a generic ensemble u as

$$e(\varepsilon, u) = \frac{1}{J} \sum_{j=1}^{J} \left\| \mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)}) \right\|_2, \tag{2.19}$$

and a discretization error function $\widetilde{e} \colon \mathbb{R} \times \mathcal{U}_{J,M} \to \mathbb{R}$ as

$$\widetilde{e}(h, u) = \frac{1}{J} \sum_{j=1}^{J} \left\| \mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)}) \right\|_2. \tag{2.20}$$

Before proving the main theorem, we introduce some preliminary results.

Let us first consider a generic forward operator involving an elliptic PDE and show that the associated forward map is Lipschitz continuous.

Lemma 2.12. *Let $\mathcal{G} \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$ be a forward operator such that $\mathcal{O} \colon H_0^1(D) \to \mathbb{R}^L$ is Lipschitz and $\mathcal{S} \colon \mathbb{R}^M \to H_0^1(D)$, $\mathcal{S} \colon u \mapsto p$ is defined by the solution of*

$$\begin{cases} -\nabla \cdot (A_u \nabla p) = f, & \text{in } D, \\ p = 0, & \text{on } \partial D, \end{cases} \tag{2.21}$$

where $D \subset \mathbb{R}^d$ is an open bounded set, the right-hand side $f \in L^2(D)$ and the tensor A_u satisfies Assumption 2.4. Then $\mathcal{G}$ is Lipschitz continuous.

The proof of Lemma 2.12 is given in Section 2.8. In the following Lemma, whose proof is also given in Section 2.8, we consider the homogenization error defined in (2.19) and show that it vanishes in the limit $\varepsilon \to 0$.

Lemma 2.13. *Let e be defined as* (2.19)*. Under Assumption 2.5, we have for all $u \in \mathcal{U}_{J,M}$*

$$e(\varepsilon, u) \to 0 \quad \text{as } \varepsilon \to 0.$$

Moreover, if the solution of the homogenized problem (2.16) *is in $H^2(D)$ independently of u, then there exists $K > 0$ independent of ε and u such that*

$$e(\varepsilon, u) \leq K\varepsilon.$$

Finally, we consider the particle empirical covariances of ensembles given by the EnKF algorithm, thus proving their boundedness and Lipschitz continuity. The proof of this Lemma can be found in Section 2.8.

Lemma 2.14. *Let $C^{up}(u) \in \mathbb{R}^{M\times L}$ and $C^{pp}(u) \in \mathbb{R}^{L\times L}$ be defined as*

$$C^{up}(u) = \frac{1}{J}\sum_{j=1}^{J}\big(u^{(j)} - \bar{u}\big)\big(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\big)^\top,$$
$$C^{pp}(u) = \frac{1}{J}\sum_{j=1}^{J}\big(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\big)\big(\mathcal{G}(u^{(j)}) - \bar{\mathcal{G}}\big)^\top,$$

where $\bar{u} \in \mathbb{R}^M$ and $\bar{\mathcal{G}} \in \mathbb{R}^L$ are the empirical averages

$$\bar{u} = \frac{1}{J}\sum_{j=1}^{J} u^{(j)}, \qquad \bar{\mathcal{G}} = \frac{1}{J}\sum_{j=1}^{J}\mathcal{G}(u^{(j)}),$$

and let $\mathcal{G}\colon \mathbb{R}^M \to \mathbb{R}^L$ be Lipschitz with constant $C_{\mathcal{G}}$. Then, there exist four constants $C_i > 0$, $i = 1,\dots,4$, such that

(i) $\|C^{up}(u)\|_2 \le C_1$,
(ii) $\|C^{pp}(u)\|_2 \le C_2$,
(iii) $\|C^{up}(u_1) - C^{up}(u_2)\|_2 \le C_3\,\|u_1 - u_2\|$,
(iv) $\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \le C_4\,\|u_1 - u_2\|$,

for all ensembles $u, u_1, u_2 \in \mathcal{U}_{J,M}$ which are stable in the sense of Assumption 2.6.

In order to clarify the exposition, we first consider the amplification of the error over one step between the EnKF algorithms employing the multiscale and the homogenized forward operators respectively, which is summarized in the following lemma.

Lemma 2.15. *For all $n = 0,\dots,N-1$, let $u_n^0 = \{u_n^{0,(j)}\}_{j=1}^J, u_n^\varepsilon = \{u_n^{\varepsilon,(j)}\}_{j=1}^J$ be the ensembles of particles at the n-th iteration of the EnKF for the forward operators $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, if A_u^ε and A_u^0 satisfy Assumption 2.4 and under Assumptions 2.5 and 2.6, there exist positive constants α and γ such that*

$$\mathbb{E}\left[\left\|u_{n+1}^\varepsilon - u_{n+1}^0\right\|\right] \le \alpha\,\mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right] + \gamma\,\mathbb{E}\left[e(\varepsilon, u_n^0)\right],$$

where $e(\varepsilon, u)$ is given in (2.19).

Proof. First, due to Assumption 2.5 and denoting by C_p the Poincaré constant we have

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_2 \le C_{\mathcal{O}}\,\|p_1 - p_2\|_{L^2(D)} \le C_{\mathcal{O}}C_p\,\|\nabla p_1 - \nabla p_2\|_{L^2(D;\mathbb{R}^d)},$$

which shows that $\mathcal{O}$ is Lipschitz with constant $C_{\mathcal{O}}C_p$. Therefore, applying Lemma 2.12, we deduce that both $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ are Lipschitz with constant $C_{\mathcal{G}}$ independent of ε. The Kalman update formulas (2.7) restricted to the u variable read (see [68])

$$u_{n+1}^{\varepsilon,(j)} = u_n^{\varepsilon,(j)} + C^{up}(u_n^\varepsilon)(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^\varepsilon(u_n^{\varepsilon,(j)})), \tag{2.22}$$
$$u_{n+1}^{0,(j)} = u_n^{0,(j)} + C^{up}(u_n^0)(C^{pp}(u_n^0) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^0(u_n^{0,(j)})). \tag{2.23}$$

Combining (2.22) and (2.23), we have

$$\mathbb{E}\left[\left\|u_{n+1}^{\varepsilon}-u_{n+1}^{0}\right\|\right] = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\Big[\Big\|u_n^{\varepsilon,(j)}+C^{up}(u_n^{\varepsilon})(C^{pp}(u_n^{\varepsilon})+\Gamma)^{-1}(y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)}))$$
$$-u_n^{0,(j)}-C^{up}(u_n^{0})(C^{pp}(u_n^{0})+\Gamma)^{-1}(y_{n+1}^{(j)}-\mathcal{G}^{0}(u_n^{0,(j)}))\Big\|_2\Big],$$

and using the triangle inequality we obtain

$$\mathbb{E}\left[\left\|u_{n+1}^{\varepsilon}-u_{n+1}^{0}\right\|\right] \le \mathbb{E}\left[\left\|u_n^{\varepsilon}-u_n^{0}\right\|\right]+S_1+S_2+S_3, \tag{2.24}$$

where

$$S_1 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^{\varepsilon})-C^{up}(u_n^{0})\right\|_2\left\|(C^{pp}(u_n^{\varepsilon})+\Gamma)^{-1}\right\|_2\left\|y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2\right],$$

$$S_2 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^{0})\right\|_2\left\|(C^{pp}(u_n^{\varepsilon})+\Gamma)^{-1}-(C^{pp}(u_n^{0})+\Gamma)^{-1}\right\|_2\right. \tag{2.25}$$
$$\left.\times\left\|y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2\right],$$

$$S_3 = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|C^{up}(u_n^{0})\right\|_2\left\|(C^{pp}(u_n^{0})+\Gamma)^{-1}\right\|_2\left\|\mathcal{G}^{0}(u_n^{0,(j)})-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2\right]. \tag{2.26}$$

Let us introduce two useful inequalities which will be employed in the following. Given A and B square invertible matrices of the same size, it holds

$$\left\|A^{-1}-B^{-1}\right\|_2 \le \left\|A^{-1}\right\|_2\left\|B^{-1}\right\|_2\left\|A-B\right\|_2. \tag{2.27}$$

Moreover, if A is positive semi-definite and B is positive definite, it holds

$$\left\|(A+B)^{-1}\right\|_2 \le \left\|B^{-1}\right\|_2. \tag{2.28}$$

Let us first consider S_1. Applying Lemma 2.14 and (2.28) to the first two factors gives

$$S_1 \le \frac{C_3}{J}\sum_{j=1}^{J}\mathbb{E}\left[\left\|u_n^{\varepsilon}-u_n^{0}\right\|\left\|\Gamma^{-1}\right\|_2\left\|y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2\right].$$

Moreover, since $y_{n+1}^{(j)} = y+\eta_{n+1}^{(j)}$ and since $y=\mathcal{G}^{\varepsilon}(u^*)+\eta$, where u^* is the true value of the unknown and η is the true realization of the noise, the triangle inequality yields

$$\left\|y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2 \le \left\|\mathcal{G}^{\varepsilon}(u^*)-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2+\left\|\eta_{n+1}^{(j)}+\eta\right\|_2,$$

which, since $\mathcal{G}^{\varepsilon}$ is Lipschitz and due to Assumption 2.6, implies

$$\left\|y_{n+1}^{(j)}-\mathcal{G}^{\varepsilon}(u_n^{\varepsilon,(j)})\right\|_2 \le C_{\mathcal{G}}R+\left\|\eta_{n+1}^{(j)}+\eta\right\|_2.$$

Hence, we get

$$S_1 \le \frac{1}{J}C_3\left\|\Gamma^{-1}\right\|_2\sum_{j=1}^{J}\mathbb{E}\left[\left\|u_n^{\varepsilon}-u_n^{0}\right\|\left(C_{\mathcal{G}}R+\left\|\eta_{n+1}^{(j)}+\eta\right\|_2\right)\right].$$

Finally, the random variables $\zeta_{n+1}^{(j)} := \eta_{n+1}^{(j)}+\eta$ are i.i.d., distributed as $\zeta\sim\mathcal{N}(0,2\Gamma)$ and independent of u_n^{ε} and u_n^{0}, which implies first

$$\mathbb{E}[\|\zeta\|_2] \le \sqrt{\mathbb{E}[\|\zeta\|_2^2]} \le \sqrt{2\mathrm{tr}(\Gamma)},$$

and second, defining $\alpha_1 \coloneqq C_3 \left\|\Gamma^{-1}\right\|_2 (C_{\mathcal{G}} R + \sqrt{2\mathrm{tr}(\Gamma)})$, yields the final bound

$$S_1 \leq \alpha_1 \, \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right]. \tag{2.29}$$

Let us now consider the second term S_2. We apply Lemma 2.14 to the norm of $C^{up}(u_n^0)$. Moreover, applying the inequalities (2.27), (2.28) and Lemma 2.14 gives

$$\left\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1} - (C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \leq C_4 \left\|\Gamma^{-1}\right\|_2^2 \left\|u_n^\varepsilon - u_n^0\right\|.$$

Reasoning as for S_1 for the third factor appearing in (2.25) finally yields

$$S_2 \leq \alpha_2 \, \mathbb{E}\left[\left\|u_n^\varepsilon - u_n^0\right\|\right], \tag{2.30}$$

where $\alpha_2 \coloneqq C_1 C_4 \left\|\Gamma^{-1}\right\|_2^2 (C_{\mathcal{G}} R + \sqrt{2\mathrm{tr}(\Gamma)})$. We now consider the last term S_3. The first factor appearing in (2.26) can be bounded by Lemma 2.14 and for the second factor we use (2.28), thus obtaining

$$\left\|(C^{pp}(u_n^0) + \Gamma)^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.$$

Regarding the third factor of (2.26), we apply the triangle inequality and the Lipschitz continuity of the forward operator $\mathcal{G}^\varepsilon$, which yield

$$\left\|\mathcal{G}^0(u_n^{0,(j)}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon,(j)})\right\|_2 \leq \left\|\mathcal{G}^0(u_n^{0,(j)}) - \mathcal{G}^\varepsilon(u_n^{0,(j)})\right\|_2 + C_{\mathcal{G}} \left\|u_n^{0,(j)} - u_n^{\varepsilon,(j)}\right\|_2.$$

Substituting back into S_3 and by definition of $e(\varepsilon, u_n^0)$ and of the ensemble norm we obtain

$$S_3 \leq C_1 \left\|\Gamma^{-1}\right\|_2 \mathbb{E}\left[e(\varepsilon, u_n^0)\right] + C_1 \left\|\Gamma^{-1}\right\|_2 C_{\mathcal{G}} \, \mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right].$$

Therefore, defining $\alpha_3 = C_1 \left\|\Gamma^{-1}\right\|_2 C_{\mathcal{G}}$ and $\gamma = C_1 \left\|\Gamma^{-1}\right\|_2$ we have the bound

$$S_3 \leq \alpha_3 \, \mathbb{E}\left[\left\|u_n^0 - u_n^\varepsilon\right\|\right] + \gamma \, \mathbb{E}\left[e(\varepsilon, u_n^0)\right]. \tag{2.31}$$

Finally, defining $\alpha \coloneqq 1 + \alpha_1 + \alpha_2 + \alpha_3$, and using the results (2.24), (2.29), (2.30) and (2.31), we obtain the desired result. □

We now present the main result about global multiscale convergence of the EnKF algorithm.

Proposition 2.16. *With the assumptions and notation of Lemma 2.15, letting $u_0^\varepsilon = u_0^0$ be the same initial ensemble, we have*

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \to 0 \qquad \text{as } \varepsilon \to 0.$$

Moreover, if the solution of the homogenized problem (2.16) is sufficiently regular, namely $p^0 \in H^2(D)$, then there exists $K_1 > 0$ independent of ε such that

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq K_1 \varepsilon.$$

Proof. Since $u_0^\varepsilon = u_0^0$, iterating the estimate of Lemma 2.15 yields

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq \gamma \sum_{i=0}^{N-1} \alpha^{N-1-i} \, \mathbb{E}\left[e(\varepsilon, u_i^0)\right].$$

Applying Lemma 2.13, we have $e(\varepsilon, u_i^0) \to 0$ for all $i = 0, \ldots, N-1$, hence as $\varepsilon \to 0$

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \to 0.$$

Moreover, if p^0 belongs to $H^2(D)$, applying Lemma 2.13 gives

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] \leq K_1 \varepsilon,$$

where $K_1 = \gamma(\alpha^N - 1)K/(\alpha - 1)$, which is the desired result. □

We now consider convergence with respect to the FEM discretization of the homogenized problem. First, we introduce a preliminary result, which plays the role of Lemma 2.13 in the context of numerical convergence and whose proof is given in Section 2.8.

Lemma 2.17. *Let $\widetilde{e}$ be defined in* (2.20) *and let Assumption 2.5 hold. If the exact solution p^0 of the homogenized problem* (2.21) *is in $H^{q+1}(D)$, the right-hand side f is in $H^{q-1}(D)$ and we employ polynomials of degree r for the finite element basis, then*

$$\widetilde{e}(h,u) \leq \widetilde{K} h^{s+1},$$

where $s = \min\{r, q\}$.

We can now state the main result concerning convergence with respect to the numerical discretization of the homogenized problem.

Proposition 2.18. *Let $u_N^0 = \{u_N^{0,(j)}\}_{j=1}^J$, $u_{N,h}^0 = \{u_{N,h}^{0,(j)}\}_{j=1}^J$ be the ensembles of particles at the last iteration of the iterative ensemble Kalman filter for the forward operators $\mathcal{G}^0$ and $\mathcal{G}_h^0$ respectively. Then, under Assumption 2.4, Assumption 2.5, Assumption 2.6 and if the exact solution p^0 of the homogenized problem* (2.21) *is in $H^{q+1}(D)$ and we use polynomials of degree r for the finite element basis, we have*

$$\mathbb{E}\left[\left\|u_{N,h}^0 - u_N^0\right\|\right] \leq K_2 h^{s+1},$$

where $s = \min\{r, q\}$ and K_2 is a positive constant independent of h.

Proof. The proof of Proposition 2.18 is identical to the proof of Proposition 2.16, except that all the ensembles $\{u_n^\varepsilon\}_{n=1}^N$ obtained by the multiscale operator $\mathcal{G}^\varepsilon$ have to be replaced by the ensembles $\{u_{n,h}^0\}_{n=1}^N$ obtained by the finite element discretization of the homogenized operator $\mathcal{G}_h^0$. Moreover Lemma 2.13 for the error e has to be replaced by Lemma 2.17 for the error $\widetilde{e}$. □

We can finally prove Theorem 2.8 and thus conclude this section.

Proof of Theorem 2.8. An application of the triangle inequality yields

$$\mathbb{E}\left[\left\|u_N^\varepsilon - u_{N,h}^0\right\|\right] \leq \mathbb{E}\left[\left\|u_N^\varepsilon - u_N^0\right\|\right] + \mathbb{E}\left[\left\|u_N^0 - u_{N,h}^0\right\|\right].$$

The two addends can be bounded applying Proposition 2.16 and Proposition 2.18, thus obtaining the desired result for $C = \max\{K_1, K_2\}$. □

2.5.2 Convergence of the Posterior Distributions

In this section, we give the proof of Theorem 2.10, i.e., the convergence of the discrete posterior measures μ^ε to μ_h^0 introduced in (2.18) as $\varepsilon, h \to 0$. Let $u^* \in \mathbb{R}^M$ and let $B_R(u^*)$ be the ball of radius R centered in u^* with respect to the norm $\|\cdot\|_s$ with $s \in [1, \infty]$. Due to the discrete nature of these distributions, we study convergence with respect to the Wasserstein metrics, for which we report its standard definition in the metric spaces $(B_R(u^*), \|\cdot\|_s)$, which can be found, e.g., in [124].

Definition 2.19. Let μ and ν be two probability measures on the metric space $(B_R(u^*), \|\cdot\|_s)$. The Wasserstein distance between μ and ν is defined for all $p \in [1, \infty)$ as

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{B_R(u^*)\times B_R(u^*)} \|u - v\|_s^p \, d\gamma(u,v)\right)^{1/p}, \tag{2.32}$$

where $\Gamma(\mu,\nu)$ denotes the collection of all joint distributions on $B_R(u^*)\times B_R(u^*)$ with marginals μ and ν on the first and second factors respectively.

Remark 2.20. If μ and ν are two discrete distributions on finite state spaces, respectively $\Omega_1=\{u_1,\dots,u_{K_1}\}$ and $\Omega_2=\{v_1,\dots,v_{K_2}\}$ included in $B_R(u^*)$, then (2.32) can be written as

$$W_{p,s}(\mu,\nu)=\left(\inf_{\gamma\in\mathbb{R}^{K_1\times K_2}_{\mu,\nu}}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\|u_i-v_j\|_s^p\,\gamma_{ij}\right)^{1/p}, \tag{2.33}$$

where $\mathbb{R}^{K_1\times K_2}_{\mu,\nu}$ is the space of $K_1\times K_2$-dimensional matrices such that

$$\sum_{j=1}^{K_2}\gamma_{ij}=\mu(u_i)\quad\text{for all } i=1,\dots K_1,\qquad \sum_{i=1}^{K_1}\gamma_{ij}=\nu(v_j)\quad\text{for all } j=1,\dots K_2.$$

We now show that the distance $W_{1,2}$ is bounded by the distance induced by the ensemble norm defined in (2.17). This result will be crucial later to prove Theorem 2.8.

Lemma 2.21. *Let $u_1=\{u_1^{(j)}\}_{j=1}^J$, $u_2=\{u_2^{(j)}\}_{j=1}^J$ be two ensembles of particles and let μ_1,μ_2 be the corresponding empirical distributions defined as sum of Dirac masses*

$$\mu_1=\frac{1}{J}\sum_{j=1}^J\delta_{u_1^{(j)}},\qquad\qquad \mu_2=\frac{1}{J}\sum_{j=1}^J\delta_{u_2^{(j)}}.$$

Then for all $s\in[1,\infty]$ and $p\in[1,\infty)$ it holds

$$W_{p,s}(\mu_1,\mu_2)\le\left(\frac{1}{J}\sum_{j=1}^J\left\|u_1^{(j)}-u_2^{(j)}\right\|_s^p\right)^{\frac{1}{p}}$$

and, in particular,

$$W_{1,2}(\mu_1,\mu_2)\le\|u_1-u_2\|.$$

Proof. Take γ^* defined as

$$\gamma^*(u_1^{(j)},u_2^{(i)})=\begin{cases}J^{-1}, & \text{if } i=j,\\ 0, & \text{if } i\neq j,\end{cases}$$

such that the matrix $(\gamma)_{ji}=\gamma^*(u_1^{(j)},u_2^{(i)})$ is in $\mathbb{R}^{J\times J}_{\mu_1,\mu_2}$, and note that

$$\sum_{j=1}^J\sum_{i=1}^J\left\|u_1^{(j)}-u_2^{(i)}\right\|_s^p\gamma_{ji}=\frac{1}{J}\sum_{j=1}^J\left\|u_1^{(j)}-u_2^{(j)}\right\|_s^p.$$

Therefore, by definition of Wasserstein distance for discrete distributions on finite spaces (2.33), we deduce that

$$W_{p,s}(\mu_1,\mu_2)\le\left(\frac{1}{J}\sum_{j=1}^J\left\|u_1^{(j)}-u_2^{(j)}\right\|_s^p\right)^{\frac{1}{p}},$$

which is the desired result. Finally, taking $p=1$ and $s=2$ and recalling the ensemble norm defined in (2.17), we obtain the second inequality. □

We now analyze the relationship between the weak L^1 convergence introduce in Definition 2.9 and the convergence with respect to the expectation of the Wasserstein distance for random probability measures. In particular, we prove that the latter implies the former, which was already proved in [124] for non-random measures. Here, we extend the result to random probability measures. The proof of the following Lemma is given in Section 2.8.

Lemma 2.22. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ and μ be random probability measures on the metric space $(B_R(u^*), \|\cdot\|_s)$ dependent on the random variable ξ on $(\Omega, \mathcal{F}, P)$. If*

$$\mathbb{E}_\xi[W_{1,s}(\mu_n, \mu)] \to 0,$$

then $\mu_n \xrightarrow{L^1} \mu$.

We can now complete the proof of Theorem 2.10.

Proof of Theorem 2.10. Applying Lemma 2.21 and due to Theorem 2.8, we deduce that for $\varepsilon, h \to 0$ it holds

$$\mathbb{E}[W_{1,2}(\mu^\varepsilon, \mu_h^0)] \to 0.$$

Note that the only difference in the update step of the EnKF when used for a point estimate and in the Bayesian framework is that Γ is replaced by $\Delta^{-1}\Gamma$ where $\Delta = 1/N$. The constants of the proof of Theorem 2.8 depend on $\left\|\Gamma^{-1}\right\|_2$, which is now replaced by $\left\|(\Delta^{-1}\Gamma)^{-1}\right\|_2$, which can be bounded by $\left\|\Gamma^{-1}\right\|_2$ as

$$\left\|(\Delta^{-1}\Gamma)^{-1}\right\|_2 = \Delta \left\|\Gamma^{-1}\right\|_2 \leq \left\|\Gamma^{-1}\right\|_2.$$

Finally, applying Lemma 2.22, we obtain the desired result. □

2.6 Modeling Error

In this section, we consider the effects of model misspecification due to the homogenization and discretization error. All the results presented in Section 2.5 concern the asymptotic case $h, \varepsilon \to 0$, which may unrealistic in applications. Let us recall that the inverse problem involves predicting the unknown u from observations originated by the model

$$y = \mathcal{G}^\varepsilon(u) + \eta, \tag{2.34}$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ is the noise. Since evaluating $\mathcal{G}^\varepsilon$ is too expensive and in many applications unfeasible, we wish to employ the cheaper forward operator $\mathcal{G}_h^0$. Hence, we rewrite (2.34) as

$$y = \mathcal{G}_h^0(u) + \mathcal{E}(u) + \eta, \tag{2.35}$$

where

$$\mathcal{E}(u) \coloneqq \mathcal{G}^\varepsilon(u) - \mathcal{G}_h^0(u).$$

The quantity $\mathcal{E}(u)$ represents the error introduced by misspecification of the forward model. Equation (2.35) shows that the observed data y can be seen as data originating by the discrete homogenized model which is affected by two sources of errors, the original noise and the modeling error. This formulation of modeling error was originally presented in [29], and then applied to multiscale inverse problems in [4]. Following [4, 29], we assume that the modeling error is a Gaussian random variable independent of the noise η, so that $\mathcal{E} \sim \mathcal{N}(m, \Sigma)$ for all u, and write

$$y = \mathcal{G}_h^0(u) + m + \zeta + \eta, \tag{2.36}$$

where $\zeta \sim \mathcal{N}(0, \Sigma)$. There is no theoretical guarantee for the modeling error to be distributed as a Gaussian in this framework. Nevertheless, it has been shown in [99] that in the one-dimensional case a Gaussian assumption can be employed effectively for the modeling error, thus partially justifying our choice. We then define

$$\widetilde{y} = y - m \qquad \text{and} \qquad \widetilde{\eta} = \eta + \zeta \sim \mathcal{N}(0, \Gamma + \Sigma)$$

and, from (2.36), we obtain

$$\widetilde{y} = \mathcal{G}_h^0(u) + \widetilde{\eta}. \tag{2.37}$$

Therefore, if the mean m and covariance Σ of the modeling error are known, a more reliable solution of the inverse problem can be obtained applying the EnKF to (2.37). The modeling error distribution, by assumption fully determined by its mean and covariance, is approximated offline. We sample $N_{\mathcal{E}}$ unknowns $\{u_i\}_{i=1}^{N_{\mathcal{E}}}$ from μ_0 and, for all $i = 1, \dots, N_{\mathcal{E}}$, we apply both the forward operators $\mathcal{G}^\varepsilon(u_i)$ and $\mathcal{G}_h^0(u_i)$. Then we compute

$$\mathcal{E}_i = \mathcal{G}^\varepsilon(u_i) - \mathcal{G}_h^0(u_i),$$

and the mean m and the covariance Σ are obtained as the empirical mean and covariance of the sample $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}}$. This procedure is computationally involved due to the multiple evaluations of $\mathcal{G}^\varepsilon$, but it has to be performed only once and can then be applied to different sets of observations and true values u^*. Let us also remark that on the one hand, due to the theory of homogenization, the modeling error can be considered negligible when ε is very small, and the expensive estimation of $\mathcal{E}$ may not be necessary. On the other hand, when ε is larger, the homogenized equation does not provide with a good approximation of the multiscale problem, and an estimation of $\mathcal{E}$ is required. One may rightfully argue that in case $\varepsilon = \mathcal{O}(1)$, it is possible to evaluate the forward operator $\mathcal{G}^\varepsilon$ without a large computational effort. Hence, the techniques presented in this section are relevant for mid-range values of ε, for which $\mathcal{E}$ is significant with respect to the noise η. Moreover, we remarked that in practice a small number $N_{\mathcal{E}}$ can be employed to obtain a satisfactory approximation of the modeling error. A theoretical justification of this practical consideration is provided by Theorems 2.23 and 2.24.

In order to obtain a more reliable approximation of the distribution of the modeling error, we can follow a dynamic approach based on the estimation of the mean m and the covariance Σ online, i.e., during the run of the EnKF algorithm. This methodology has been developed in [28], and it consists of splitting the EnKF run on $\mathcal{L}$ levels, thus obtaining a new estimation of the modeling error sequentially at the end of each level. In practice, given a prior μ_0, and initializing $\mu_0^0 \equiv \mu_0$ and $\ell = 0$, the procedure can be algorithmically summarized as:

1. approximate the distribution $\nu^\ell = \mathcal{N}(m^\ell, \Sigma^\ell)$ of the modeling error with samples of $\mathcal{E}$ obtained from μ_0^ℓ;
2. run the EnKF corrected by the modeling error ν^ℓ for N^ℓ steps and obtain the discrete approximation of the posterior
$$\mu_{N^\ell}^\ell = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n^{\ell(j)}};$$
3. set $\mu_0^{\ell+1} = \mu_{N^\ell}^\ell$, $\ell \leftarrow \ell + 1$ and if $\ell < \mathcal{L}$ return to 1.

Intuitively, this approach yields a better approximation of the modeling error. Indeed, we are sampling the modeling error from probability measures which are progressively closer to the posterior. Still, the dynamic update is performed online and is therefore computationally more expensive than approximating the modeling error fully offline.

Finally, we are interested in studying whether the simple offline method for estimating the modeling error provides indeed a good approximation. In this direction, we give in Theorems 2.23

and 2.24 a criterion on how to choose the number $N_{\mathcal{E}}$ of full multiscale problems which has to be solved in order to have a reliable approximation of the true mean m^* and covariance Σ^* of the modeling error with respect to ε and h. Before stating the theoretical results, let us recall Hoeffding's inequality, which will be employed in the proofs. Let $\{Y_i\}_{i=1}^N$ be independent random variables with values in $[a,b]$, and let $\bar{Y}$ be the sample average of $\{Y_i\}_{i=1}^N$. Then, the Hoeffding's inequality states that for all $\eta \in \mathbb{R}$ it holds

$$P\left(\left|\bar{Y} - \mathbb{E}[Y]\right| \geq \eta\right) \leq 2\exp\left(-\frac{2\eta^2 N}{(b-a)^2}\right).$$

Theorem 2.23. *Let $\alpha \in (0,1)$, $\eta > 0$ and $C_{\mathcal{E}} = \max\{K, \widetilde{K}\}$, where K and $\widetilde{K}$ are the constants of Lemma 2.13 and Lemma 2.17, respectively. Let $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}} \subset \mathbb{R}^L$ be given by*

$$\mathcal{E}_i = \mathcal{G}^{\varepsilon}(u_i) - \mathcal{G}_h^0(u_i) \qquad \text{for all } i = 1, \dots, N_{\mathcal{E}},$$

for a sample of realizations $\{u_i\}_{i=1}^{N_{\mathcal{E}}}$ drawn from a prior distribution μ_0, let m be the sample mean of $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}}$ and $m^* = \mathbb{E}[\mathcal{E}_i]$. If

$$N_{\mathcal{E}} \geq 4C_{\mathcal{E}}^2 \frac{L}{\eta^2} \log\left(\frac{2L}{\alpha}\right)\left[\varepsilon^2 + h^{2(s+1)}\right],$$

where s is given by Lemma 2.17, then

$$P\left(\|m - m^*\|_2 \leq \eta\right) \geq 1 - \alpha.$$

Proof. First, note that the modeling error is bounded, indeed by Lemma 2.13 and Lemma 2.17, we have for each $i = 1, \dots, N_{\mathcal{E}}$

$$\begin{aligned}\|\mathcal{E}_i\|_2 &= \left\|\mathcal{G}^{\varepsilon}(u_i) - \mathcal{G}_h^0(u_i)\right\|_2 \\ &\leq \left\|\mathcal{G}^{\varepsilon}(u_i) - \mathcal{G}^0(u_i)\right\|_2 + \left\|\mathcal{G}^0(u_i) - \mathcal{G}_h^0(u_i)\right\|_2 \\ &\leq K\varepsilon + \widetilde{K}h^{s+1},\end{aligned}$$

so each component $(\mathcal{E}_i)_l$, for $l = 1, \dots, L$, is bounded by the same constant

$$|(\mathcal{E}_i)_l| \leq \|\mathcal{E}_i\|_2 \leq K\varepsilon + \widetilde{K}h^{s+1} \leq C_{\mathcal{E}}(\varepsilon + h^{s+1}). \tag{2.38}$$

Observe that if

$$|m_l - m_l^*| \leq \frac{\eta}{\sqrt{L}} \qquad \text{for each } l = 1, \dots, L,$$

then

$$\|m - m^*\|_2 = \left(\sum_{l=1}^{L} |m_l - m_l^*|^2\right)^{\frac{1}{2}} \leq \eta,$$

which implies that

$$P(\|m - m^*\|_2 \leq \eta) \geq P\left(|m_l - m_l^*| \leq \frac{\eta}{\sqrt{L}}, \quad \forall\, l = 1, \dots, L\right). \tag{2.39}$$

Using (2.38) and applying Hoeffding's inequality we have

$$\begin{aligned}P\left(|m_l - m_l^*| \geq \frac{\eta}{\sqrt{L}}\right) &\leq 2\exp\left(-\frac{2\eta^2 N_{\mathcal{E}}}{4LC_{\mathcal{E}}^2(\varepsilon + h^{s+1})^2}\right) \\ &\leq 2\exp\left(-\frac{\eta^2 N_{\mathcal{E}}}{4LC_{\mathcal{E}}^2(\varepsilon^2 + h^{2(s+1)})}\right).\end{aligned} \tag{2.40}$$

Let us define the events $A_l = \left\{ |m_l - m_l^*| \leq \frac{\eta}{\sqrt{L}} \right\}$ for each $l = 1, \dots, L$. Then, we have

$$P\left(|m_l - m_l^*| \leq \frac{\eta}{\sqrt{L}} \quad \forall\, l = 1, \dots, L \right) = P\left(\bigcap_{l=1}^{L} A_l \right),$$

and, applying the De Morgan's laws and the union bound, we obtain

$$P\left(\bigcap_{l=1}^{L} A_l \right) = 1 - P\left(\left(\bigcap_{l=1}^{L} A_l \right)^C \right) = 1 - P\left(\bigcup_{l=1}^{L} A_l^C \right) \geq 1 - \sum_{l=1}^{L} P(A_l^C). \tag{2.41}$$

Therefore, thanks to (2.39), (2.40) and (2.41), we have

$$\begin{aligned} P(\|m - m^*\|_2 \leq \eta) &\geq 1 - L \max_{l=1,\dots,L} P\left(|m_l - m_l^*| \geq \frac{\eta}{\sqrt{L}} \right) \\ &\geq 1 - 2L \exp\left(-\frac{\eta^2 N_{\mathcal{E}}}{4LC_{\mathcal{E}}^2(\varepsilon^2 + h^{2(s+1)})} \right), \end{aligned} \tag{2.42}$$

and if $N_{\mathcal{E}}$ satisfies the hypothesis we obtain the desired result. □

Theorem 2.24. *Let $\alpha \in (0,1)$, $\eta > 0$ and $C_{\mathcal{E}} = \max\{K, \widetilde{K}\}$, where K and $\widetilde{K}$ are the constants of Lemma 2.13 and Lemma 2.17, respectively. Let $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}} \subset \mathbb{R}^L$ be given by*

$$\mathcal{E}_i = \mathcal{G}^{\varepsilon}(u_i) - \mathcal{G}_h^0(u_i) \qquad \text{for all } i = 1, \dots, N_{\mathcal{E}},$$

for a sample of realizations $\{u_i\}_{i=1}^{N_{\mathcal{E}}}$ drawn from a prior distribution μ_0, let m and Σ be the sample mean and covariance of $\{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}}$ and $m^* = \mathbb{E}[\mathcal{E}_i]$ and $\Sigma^* = \mathbb{E}[(\mathcal{E}_i - m^*)(\mathcal{E}_i - m^*)^\top]$. If

$$N_{\mathcal{E}} \geq \widehat{C} C_{\mathcal{E}}^4 \frac{L^2}{\eta^2} \log\left(\frac{2L(L+1)}{\alpha} \right) \left[\varepsilon^4 + h^{4(s+1)} \right],$$

where s is given by Lemma 2.17 and $\widehat{C}$ is specified in the proof, then

$$P\left(\|\Sigma - \Sigma^*\|_2 \leq \eta \right) \geq 1 - \alpha.$$

Proof. First, repeating verbatim the first part of the proof of Theorem 2.23 we have

$$|(\mathcal{E}_i)_l| \leq \|\mathcal{E}_i\|_2 \leq K\varepsilon + \widetilde{K} h^{s+1} \leq C_{\mathcal{E}}(\varepsilon + h^{s+1}). \tag{2.43}$$

Let us now rewrite the covariance matrix Σ^* and its estimator Σ as

$$\Sigma^* = \mathbb{E}[(\mathcal{E}_i - m^*)(\mathcal{E}_i - m^*)^\top] = \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] - m^*(m^*)^\top,$$

and

$$\Sigma = \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} (\mathcal{E}_i - m)(\mathcal{E}_i - m)^\top = \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - m m^\top.$$

Then by triangle inequality it holds

$$\|\Sigma - \Sigma^*\|_2 \leq \left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 + \left\| m m^\top - m^*(m^*)^\top \right\|_2,$$

and due to the elementary inequality $\left\| a b^\top \right\|_2 \leq \|a\|_2 \|b\|_2$ and the bound (2.43) we have

$$\begin{aligned} \left\| m m^\top - m^*(m^*)^\top \right\|_2 &= \left\| (m - m^*) m^\top + m^*(m - m^*)^\top \right\|_2 \\ &\leq 2C_{\mathcal{E}}(\varepsilon + h^{s+1}) \|m - m^*\|_2, \end{aligned}$$

which implies

$$\|\Sigma - \Sigma^*\|_2 \leq \left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 + 2C_{\mathcal{E}}(\varepsilon + h^{s+1}) \|m - m^*\|_2 .$$

Therefore, we obtain

$$P\left(\|\Sigma - \Sigma^*\|_2 \leq \eta\right) \geq P\left(\left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 \leq \frac{\eta}{2}, \right.$$
$$\left. \|m - m^*\|_2 \leq \frac{\eta}{4C_{\mathcal{E}}(\varepsilon + h^{s+1})} \right),$$

which, applying $P(A, B) = 1 - P(A^C \cup B^C) \geq 1 - P(A^C) - P(B^C)$, yields

$$\begin{aligned} P\left(\|\Sigma - \Sigma^*\|_2 \leq \eta\right) \geq 1 &- P\left(\left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 \geq \frac{\eta}{2} \right) \\ &- P\left(\|m - m^*\|_2 \geq \frac{\eta}{4C_{\mathcal{E}}(\varepsilon + h^{s+1})} \right). \end{aligned} \tag{2.44}$$

We then bound the two terms in the right-hand side separately. By equation (2.42) in the proof of Theorem 2.23 we first have

$$P\left(\|m - m^*\|_2 \geq \frac{\eta}{4C_{\mathcal{E}}(\varepsilon + h^{s+1})} \right) \leq 2L \exp\left(-\frac{\eta^2 N_{\mathcal{E}}}{32 L C_{\mathcal{E}}^4 (\varepsilon + h^{s+1})^4} \right). \tag{2.45}$$

Then, similarly to the last part of the proof of Theorem 2.23, since $\|\cdot\|_2 \leq \|\cdot\|_F$ where $\|\cdot\|_F$ denotes the Frobenius norm, we have

$$\begin{aligned} P\left(\left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 \geq \frac{\eta}{2} \right) \\ \leq L^2 \max_{j,k=1,\dots,L} P\left(\left| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} (\mathcal{E}_i)_j (\mathcal{E}_i)_k - \mathbb{E}[(\mathcal{E}_i)_j (\mathcal{E}_i)_k] \right| \geq \frac{\eta}{2L} \right), \end{aligned}$$

and applying the Hoeffding's inequality to the random variables $(\mathcal{E}_i)_j (\mathcal{E}_i)_k$ for all $j, k = 1, \dots, L$, which are bounded by $C_{\mathcal{E}}^2 (\varepsilon + h^{s+1})^2$ due to (2.43), we obtain

$$P\left(\left\| \frac{1}{N_{\mathcal{E}}} \sum_{i=1}^{N_{\mathcal{E}}} \mathcal{E}_i \mathcal{E}_i^\top - \mathbb{E}[\mathcal{E}_i \mathcal{E}_i^\top] \right\|_2 \geq \frac{\eta}{2} \right) \leq 2L^2 \exp\left(-\frac{\eta^2 N_{\mathcal{E}}}{8 L^2 C_{\mathcal{E}}^4 (\varepsilon + h^{s+1})^4} \right). \tag{2.46}$$

Finally, equations (2.45) and (2.46) together with (2.44) imply

$$P\left(\|\Sigma - \Sigma^*\|_2 \leq \eta\right) \geq 1 - 2L(L+1) \exp\left(-\frac{\eta^2 N_{\mathcal{E}}}{\widehat{C} L^2 C_{\mathcal{E}}^4 (\varepsilon^4 + h^{4(s+1)})} \right),$$

where $\widehat{C} = 256$ and if $N_{\mathcal{E}}$ satisfies the hypothesis we obtain the desired result. □

Remark 2.25. Note that, in Theorem 2.23 and Theorem 2.24, as expected, the number $N_{\mathcal{E}}$ of full multiscale problems tends to infinity if we require no error between the sample and the true mean and covariance ($\eta \to 0$) or certainty that the error is below a certain value ($\alpha \to 0$). Moreover, observe that for any given accuracy the number of samples required $N_{\mathcal{E}}$ is a increasing function of ε and h, so that if the model $\mathcal{G}_h^0$ is a good approximation of $\mathcal{G}$, thus computationally

expensive, then only few samples are needed. In particular, notice that in order to obtain a good approximation of the true mean, the number of full multiscale problems is

$$N_{\mathcal{E}} = \mathcal{O}\left(\eta^{-2}\log(\alpha^{-1})\left(\varepsilon^2 + h^{2(s+1)}\right)\right),$$

while to have a reliable approximation of the covariance matrix it is required that

$$N_{\mathcal{E}} = \mathcal{O}\left(\eta^{-2}\log(\alpha^{-1})\left(\varepsilon^4 + h^{4(s+1)}\right)\right).$$

2.7 Numerical Experiments

In this section, we present numerical experiments to illustrate the potential of ensemble Kalman inversion for multiscale elliptic problems. In particular, we consider the framework of the boundary problem first introduced in Calderón's seminal work [27]. We refer moreover the reader to the research article [4], from which we borrow the experimental setting.

Let us consider a class of parametrized multiscale locally periodic tensors of the type $A_u^\varepsilon(x) = A(u(x), x/\varepsilon)$, where $u\colon D \to \mathbb{R}$, $u \in X$ is the unknown of our inverse problem. In particular, we assume to know the map $(t,x) \mapsto A(t, x/\varepsilon)$ for all $x \in D$ and $t \in \mathbb{R}$ and we want to estimate the function u given measurements computed from the model

$$\begin{cases} -\nabla\cdot(A_u^\varepsilon \nabla p^\varepsilon) = 0 & \text{in } D, \\ p^\varepsilon = g & \text{on } \partial D. \end{cases} \tag{2.47}$$

Remark 2.26. Note that we presented the theory for Dirichlet homogeneous boundary conditions, i.e., in case $g \equiv 0$. Nevertheless, all our results hold for a generic smooth Dirichelet boundary condition by a standard "lifting" argument. For more details, we refer the reader to [123, Remark 8.10].

We let $u \in X$ where X is the admissible set

$$X = \{u \in L^\infty(D)\colon u^- \le u(x) \le u^+\},$$

where u^- and u^+ are two given scalars. Moreover, we let observations consist of integrals of the normal flux multiplied by a compactly-supported function on a portion of the boundary of the domain. More precisely, we consider $I \in \mathbb{N}$ disjoint portions of D, which we denote by $\Gamma_i \in \partial D$, $i = 1,\dots,I$, $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$, and I functions $\varphi_i \in H^{1/2}(\partial D)$ with compact support $\operatorname{supp}(\varphi_i) \subset \Gamma_i$ for all $i = 1,\dots,I$. Moreover, we solve (2.47) for $K \in \mathbb{N}$ Dirichlet data g_k, $k = 1,\dots,K$, and we denote by p_k^ε the solution of the problem. Let $\Lambda_{A_u^\varepsilon} : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ be the operator which maps the Dirichlet data g to the normal flux of the solution p^ε of (2.47)

$$\Lambda_{A_u^\varepsilon} g = A_u^\varepsilon \nabla p^\varepsilon \cdot \nu,$$

where ν is the exterior unit normal vector to ∂D. In literature, the function $\Lambda_{A_u^\varepsilon}$ is often referred to as the Dirichlet-to-Neumann map (see e.g. [4]). We then define the multiscale forward operator $\mathcal{G}^\varepsilon\colon X \to \mathbb{R}^L$ where $L = IK$ by components as

$$\mathcal{G}^\varepsilon(u)_{ik} = \mathcal{G}^\varepsilon(u)_l = \left\langle \Lambda_{A_u^\varepsilon} g_k, \varphi_i \right\rangle_{H^{-1/2}(\partial D), H^{1/2}(\partial D)}, \tag{2.48}$$

for $i = 1,\dots,I$ and $k = 1,\dots,K$. Let us remark that with an abuse of notation we can rewrite perhaps more intuitively (2.48) as

$$\mathcal{G}^\varepsilon(u)_{ik} = \int_{\Gamma_i} A_u^\varepsilon \nabla p_k^\varepsilon \cdot \nu \varphi_i ds.$$

The final vector of observations y is given by the sum of the operator $\mathcal{G}^\varepsilon$ and a source of Gaussian noise

$$y = \mathcal{G}^\varepsilon(u) + \eta,$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ and Γ is a given positive definite covariance matrix. We impose on the unknown u a Gaussian prior measure $\mu_0 = \mathcal{N}(0, C)$, where we choose C to be the exponential covariance operator defined by

$$C(x_1, x_2) = \delta \exp\left(-\frac{\|x_1 - x_2\|_2}{\lambda}\right), \tag{2.49}$$

where $\delta, \lambda \in \mathbb{R}^+$ and $x_1, x_2 \in D$. The parameter λ is a correlation length that describes how the values at different positions of the functions supported by the prior measure are related, while the parameter δ is an amplitude scaling factor. Let us remark that functions drawn from the prior distribution could exhibit multiple scales, thus interfering in the homogenized problem. We control this issue by tuning the parameters δ and λ, which, as illustrated by our numerical experiments, suffices in practice. Another choice for controlling the smoothness and scale-length of samples from the prior would be to a prior covariance of the Matérn kind. Let us finally remark that we solve in practice the inverse problem in a M-dimensional subspace $X^M \subset X$ by considering truncated KL expansions as in Section 1.2.

The approximated operator $\mathcal{G}_h^0$, to which we apply the EnKF as described above, is obtained via an application of the FE-HMM [2, 5]. Denoting by A_u^0 the homogenized tensor obtained via the FE-HMM, and by p_h^0 the homogenized solution computed on a coarse mesh $\mathcal{T}_h$ of D with maximum element size $h > 0$, we define the discrete homogenized operator $\mathcal{G}_h^0 \colon X \to \mathbb{R}^L$ as

$$\mathcal{G}_h^0(u)_l = \mathcal{G}_h^0(u)_{ik} = \left\langle \Lambda_{A_u^0} g_k, \varphi_i \right\rangle_{H^{-1/2}(\partial D), H^{1/2}(\partial D)}, \tag{2.50}$$

for $i = 1, \dots, I$ and $k = 1, \dots, K$, or, with the same abuse of notation as above

$$\mathcal{G}^\varepsilon(u)_{ik} = \int_{\Gamma_i} A_u^0 \nabla (p_h^0)_k \cdot \nu \varphi_i ds,$$

where $(p_h^0)_k$ is the FE-HMM solution computed with the k-th boundary condition g_k.

2.7.1 Data

We now detail the specific setting for our experiments. We let the domain D be the unit square $D = (0, 1)^2$, and consider exact tensor $A_{u^*}^\varepsilon$ given by

$$\begin{aligned}
a_{11}\left(u^*(x), \frac{x}{\varepsilon}\right) &= e^{u^*(x)}\left(\cos^2\left(\frac{2\pi x_1}{\varepsilon}\right) + 1\right) + \cos^2\left(2\pi\frac{x_2}{\varepsilon}\right), \\
a_{12}\left(u^*(x), \frac{x}{\varepsilon}\right) &= 0, \\
a_{21}\left(u^*(x), \frac{x}{\varepsilon}\right) &= 0, \\
a_{22}\left(u^*(x), \frac{x}{\varepsilon}\right) &= e^{u^*(x)}\left(\sin\left(\frac{2\pi x_2}{\varepsilon}\right) + 2\right) + \cos^2\left(2\pi\frac{x_1}{\varepsilon}\right),
\end{aligned}$$

where

$$u^*(x) = \log(1.3 + 0.3\chi_{D_1} - 0.4\chi_{D_2}),$$

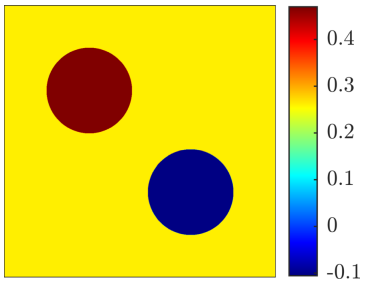


Figure 2.1 – Exact unknown u^* employed for numerical experiments.

with χ denoting the indicator function, and the sets D_1 and D_2 defined by

$$D_1 = \left\{ x = (x_1, x_2) \colon \left(x_1 - \frac{5}{16}\right)^2 + \left(x_2 - \frac{11}{16}\right)^2 \leq 0.025 \right\},$$
$$D_2 = \left\{ x = (x_1, x_2) \colon \left(x_1 - \frac{11}{16}\right)^2 + \left(x_2 - \frac{5}{16}\right)^2 \leq 0.025 \right\}.$$

We show the exact unknown u^* in Fig. 2.1. Let us remark that the tensor A_u^ε satisfies Assumption 2.4. In particular, for any $\xi \in \mathbb{R}^2$ and $u \in X$ we have

$$A_u^\varepsilon \xi \cdot \xi = a_{1,1}\left(u(x), \frac{x}{\varepsilon}\right)\xi_1^2 + a_{2,2}\left(u(x), \frac{x}{\varepsilon}\right)\xi_2^2 \geq e^{u(x)}(\xi_1^2 + \xi_2^2) \geq e^{u^-}\,\|\xi\|_2^2\,,$$

which shows that the elliptic assumption holds.

We fix the number of Dirichlet conditions $\{g_k\}_{k=1}^3$ to $K = 3$. In particular, we choose $g_k = \sqrt{\ell_k}\vartheta_k$, where $\{(\ell_k, \vartheta_k)\}_{k=1}^3$ are the three eigenvalues/eigenfunction pairs corresponding to the smallest eigenvalues of the one-dimensional Laplace operator with homogeneous Dirichlet boundary conditions on $(0, 1)$. These functions are orthonormal with respect to the scalar product in $L^2(\partial D)$, which ensures that each boundary condition yields independent information. The boundary integrals in (2.48) and (2.50) are computed on $I = 12$ portions Γ_i, three for each side of the square D. In particular, all Γ_i have length equal to 0.2 and consist of the intervals $(0.1, 0.3), (0.4, 0.6)$ and $(0.7, 0.9)$, with respect to the local coordinates of each side of D. The functions $\{\varphi_i\}_{i=1}^{12}$ are hat functions with $\mathrm{supp}(\varphi_i) = \Gamma_i$, which take value one at the midpoint of Γ_i and value 0 at the extremes of Γ_i. We refer the reader to [4] for a sensitivity analysis with respect to the number of Dirichlet conditions K.

We choose the noise covariance as $\Gamma = \gamma^2 I$, where I is the identity matrix and $\gamma = 0.01$. We fix moreover $\delta = 0.05$ and $\lambda = 0.5$ as the parameters of the prior covariance (2.49). Finally, we choose to truncate the KL expansion after $M = 100$ terms, thus computing the solution of the inverse problem on the finite-dimensional subset $X^M \subset X$ such that $\dim(X^M) = 100$. Let us remark that the true unknown u^* is discontinuous, and does not therefore belong to X^M.

In all the numerical scenarios we consider below, observations are generated by a reference solution computed with the FEM on a refined mesh with maximum element size $h_{\mathrm{obs}} = 2^{-12}$. Moreover, the macro-mesh size in the FE-HMM is fixed to $h = 2^{-5}$, again for all experiments.

2.7.2 Results

We first fix the multiscale parameter $\varepsilon = 2^{-5}$ and the ensemble size $J = 500$ and study the evolution of the EnKF estimate with respect to the number of steps N. In Fig. 2.2 we plot the

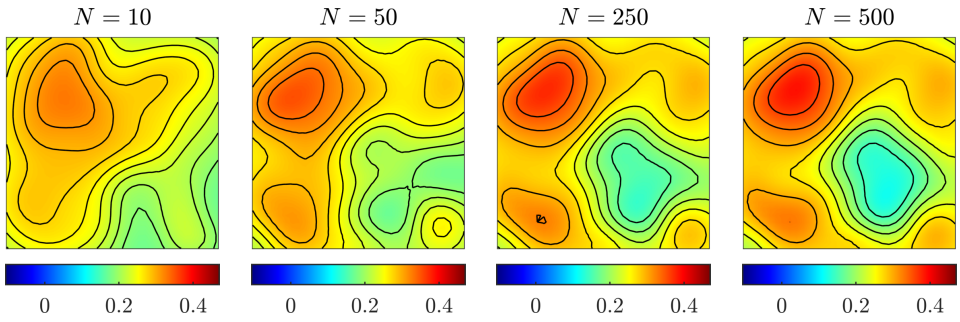


Figure 2.2 – EnKF estimation after $N = \{10, 50, 250, 500\}$ iterations.

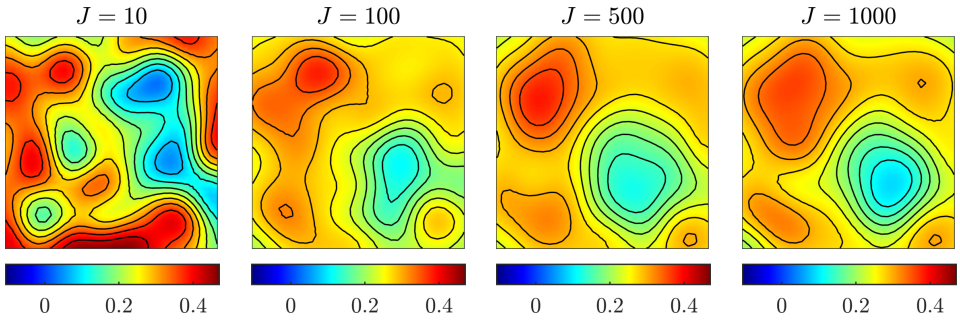


Figure 2.3 – EnKF estimation with ensemble size $J = \{10, 100, 500, 1000\}$.

estimation u_{EnKF} after $10, 50, 250$ and 500 iterations of the ensemble Kalman algorithm. We clearly see that the approximation gets better as the number of iterations increases and that convergence has been reached. In particular, already after $N = 250$ iterations the algorithm seem to have reached convergence.

Next, we perform a sensitivity analysis with respect to the ensemble size. In Fig. 2.3 we show numerical results for varying number of particles J, all computed with 500 iterations of the EnKF and for the multiscale parameter $\varepsilon = 2^{-5}$. As expected, the quality of the approximation is enhanced by taking larger ensembles. In particular, note that if the number of particles is chosen too small, e.g. $J = 10$, then the approximation of the inverse problem is not satisfying.

We then fix both the ensemble size to $J = 500$ and the number of iterations to $N = 500$ and consider different values of the multiscale parameter. Results, shown in Fig. 2.4, highlight how replacing the full model with its homogenized surrogate is a viable solution when $\varepsilon \ll 1$, whereas in case ε is not small the EnKF method fails to identify the solution of the inverse problem satisfactorily.

In order to account for the mismatch between homogenized and multiscale model, we therefore consider the modeling error techniques presented in Section 2.6. In particular, we show in Fig. 2.5 numerical results when the offline modeling error estimation is applied with $N_{\mathcal{E}} = 20$ samples. Comparing the results with Fig. 2.4, in particular for $\varepsilon = 1/4$, shows the beneficial effect of explicitly accounting for model misspecification.

Finally, in Fig. 2.6 we show the results obtained by applying the ensemble Kalman method with dynamic updating of the modeling error distribution with $\mathcal{L} = 5$ levels, $N_{\mathcal{E}}^{\ell} = 4$ samples and $N^{\ell} = 100$ iterations at each level $\ell = 1, \dots, \mathcal{L}$. For fairness of comparison with the offline approach, we consider 20 solves of the full multiscale problem and 500 iterations of the EnKF.

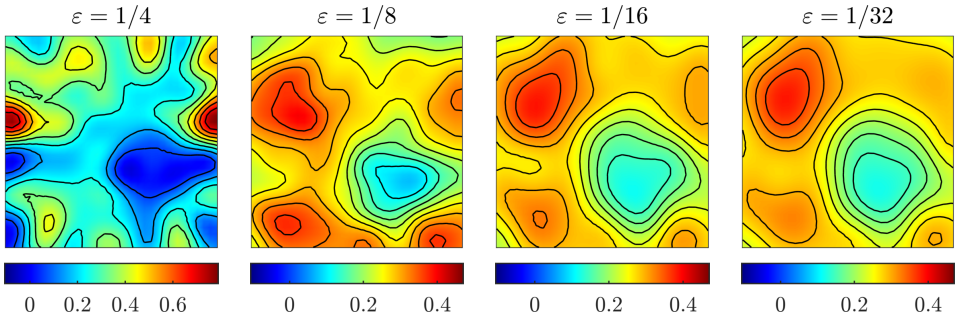


Figure 2.4 – EnKF estimation for the multiscale parameter $\varepsilon = \{1/4, 1/8, 1/16, 1/32\}$.

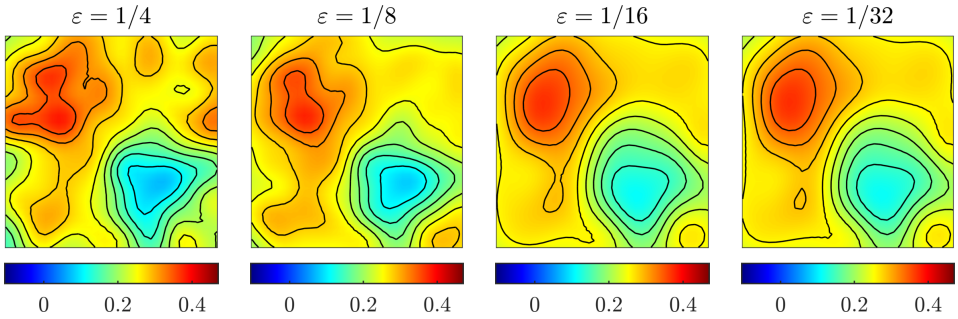


Figure 2.5 – EnKF with offline modeling error estimation for the multiscale parameter $\varepsilon = \{1/4, 1/8, 1/16, 1/32\}$.

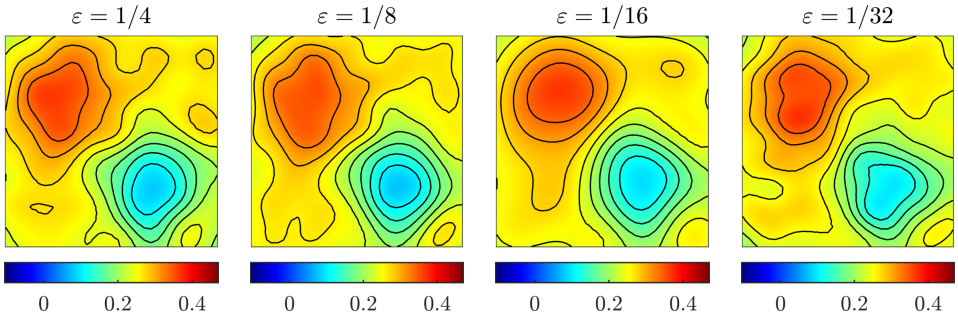


Figure 2.6 – EnKF with online iterative modeling error estimation for the multiscale parameter $\varepsilon = \{1/4, 1/8, 1/16, 1/32\}$.

Comparing these plots with the ones in Fig. 2.5, we note that updating the distribution of the modeling error dynamically still improves the results.

2.8 Proof of Technical Results

We conclude the chapter by giving the proof of some technical results, which were omitted in the text to enhance readability.

Proof of Lemma 2.12. Let $u_1, u_2 \in \mathbb{R}^M$, and $p_1 = \mathcal{S}(u_1)$, $p_2 = \mathcal{S}(u_2)$. From the weak formulations

of (2.21) we get that

$$\int_D \left(A_{u_1}\nabla p_1 - A_{u_2}\nabla p_2\right)\cdot\nabla v = 0 \qquad \text{for all } v \in H_0^1(D),$$

which yields

$$\int_D A_{u_1}(\nabla p_1 - \nabla p_2)\cdot\nabla v = -\int_D (A_{u_1} - A_{u_2})\nabla p_2\cdot\nabla v.$$

Then choosing $v = p_1 - p_2$, by the hypotheses on A_u and applying the Hölder inequality we obtain

$$\alpha_0 \left\|\nabla p_1 - \nabla p_2\right\|^2_{L^2(D;\mathbb{R}^d)} \le M \left\|u_1 - u_2\right\|_2 \left\|\nabla p_2\right\|_{L^2(D;\mathbb{R}^d)} \left\|\nabla p_1 - \nabla p_2\right\|_{L^2(D;\mathbb{R}^d)},$$

which due a standard coercivity argument implies

$$\left\|\nabla p_1 - \nabla p_2\right\|_{L^2(D;\mathbb{R}^d)} \le \frac{MC_p}{\alpha_0^2}\left\|f\right\|_{L^2(D)}\left\|u_1 - u_2\right\|_2, \tag{2.51}$$

where C_p is the Poincaré constant associated to the domain D. Hence (2.51) shows that $\mathcal{S}$ is Lipschitz with constant

$$L_{\mathcal{S}} = \frac{MC_p}{\alpha_0^2}\left\|f\right\|_{L^2(D)}.$$

Finally, since $\mathcal{G}$ is the composition of two Lipschitz operators, we deduce that it is also Lipschitz with constant $L_{\mathcal{G}} = L_{\mathcal{O}}L_{\mathcal{S}}$. □

Proof of Lemma 2.13. Let us consider an ensemble $u \in \mathcal{U}_{J,M}$ with particles $u^{(j)} \in \mathbb{R}^M$, for $j = 1, \dots, J$. For each particle we have

$$\begin{aligned}\left\|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 &= \left\|\mathcal{O}(\mathcal{S}^\varepsilon(u^{(j)})) - \mathcal{O}(\mathcal{S}^0(u^{(j)}))\right\|_2 \\ &\le C_{\mathcal{O}}\left\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)},\end{aligned}$$

where we write explicitly the dependence of the solutions p^ε and p^0 on the particle they are generated by. Due to homogenization theory (see e.g. [113, Theorem 19.1]), we have that $p^\varepsilon(u^{(j)}) \to p^0(u^{(j)})$ in $L^2(D)$ for all $j = 1, \dots, J$, which implies

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^J \left\|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 \le \frac{C_{\mathcal{O}}}{J}\sum_{j=1}^J \left\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)} \to 0.$$

Moreover, if the solution of the homogenized problem p^0 is sufficiently smooth independently of u, namely $p^0 \in H^2(D)$, letting $C > 0$ be a constant independent of ε, we have by [98] for all $j = 1, \dots, J$

$$\left\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)} \le C\varepsilon,$$

which implies

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^J \left\|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 \le \frac{C_{\mathcal{O}}}{J}\sum_{j=1}^J \left\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)} \le C_{\mathcal{O}}C\varepsilon,$$

and defining $K = C_{\mathcal{O}}C$ gives the desired result. □

Proof of Lemma 2.14. First, for all $x \in B_R(u^*)$ we have

$$\begin{aligned} \|x\|_2 &\leq \|x-u^*\|_2 + \|u^*\|_2 \leq R + \|u^*\|_2 =: m, \\ \|\mathcal{G}(x)\|_2 &\leq \|\mathcal{G}(x)-\mathcal{G}(u^*)\|_2 + \|\mathcal{G}(u^*)\|_2 \leq C_{\mathcal{G}} \|x-u^*\|_2 + \|\mathcal{G}(u^*)\|_2 \\ &\leq C_{\mathcal{G}} R + \|\mathcal{G}(u^*)\|_2 =: M. \end{aligned} \tag{2.52}$$

We can also deduce the same bounds for the mean values

$$\|\bar{u}\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \left\|u^{(j)}\right\|_2 \leq m, \qquad \text{and} \qquad \left\|\bar{\mathcal{G}}\right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \left\|\mathcal{G}(u^{(j)})\right\|_2 \leq M. \tag{2.53}$$

Then by (2.52) and (2.53) we get

$$\begin{aligned} \|C^{up}(u)\|_2 &= \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \left\| \frac{1}{J}\sum_{j=1}^{J} (u^{(j)}-\bar{u})(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}})^\top x \right\|_2 \\ &\leq \frac{1}{J}\sum_{j=1}^{J} \left(\left\|\mathcal{G}(u^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}\right\|_2\right)\left(\left\|u^{(j)}\right\|_2 + \|\bar{u}\|_2\right) \\ &\leq 4Mm, \end{aligned}$$

and defining $C_1 = 4Mm$ we get (i). The argument is similar for the matrix $C^{pp}(u)$, for which we have

$$\|C^{pp}(u)\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \left(\left\|\mathcal{G}(u^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}\right\|_2\right)^2 \leq 4M^2,$$

and defining $C_2 = 4M^2$ we get (ii). Before proving (iii) and (iv), we need the following estimates for two ensemble of particles u_1 and u_2

$$\begin{aligned} \|\bar{u}_1 - \bar{u}_2\|_2 &= \left\| \frac{1}{J}\sum_{j=1}^{J} (u_1^{(j)} - u_2^{(j)}) \right\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \left\|u_1^{(j)} - u_2^{(j)}\right\|_2 = \|u_1 - u_2\|, \\ \left\|\bar{\mathcal{G}}_1 - \bar{\mathcal{G}}_2\right\|_2 &= \left\| \frac{1}{J}\sum_{j=1}^{J} (\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})) \right\|_2 \leq \frac{C_{\mathcal{G}}}{J}\sum_{j=1}^{J} \left\|u_1^{(j)} - u_2^{(j)}\right\|_2 = C_{\mathcal{G}} \|u_1 - u_2\|. \end{aligned} \tag{2.54}$$

Then we have

$$\begin{aligned} &\|C^{up}(u_1) - C^{up}(u_2)\|_2 \\ &\quad= \sup_{x\in\mathbb{R}^L:\ \|x\|_2=1} \left\| \frac{1}{J}\sum_{j=1}^{J} \left[(u_1^{(j)}-\bar{u}_1)(\mathcal{G}(u_1^{(j)})-\bar{\mathcal{G}}_1)^\top x - (u_2^{(j)}-\bar{u}_2)(\mathcal{G}(u_2^{(j)})-\bar{\mathcal{G}}_2)^\top x\right] \right\|_2 \\ &\quad\leq \frac{1}{J}\sum_{j=1}^{J} \left(\left\|u_1^{(j)}\right\|_2 + \|\bar{u}_1\|_2\right)\left(\left\|\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1\right\|_2\right) \\ &\qquad+ \frac{1}{J}\sum_{j=1}^{J} \left(\left\|u_1^{(j)} - u_2^{(j)}\right\|_2 + \|\bar{u}_2 - \bar{u}_1\|_2\right)\left(\left\|\mathcal{G}(u_2^{(j)})\right\|_2 + \left\|\bar{\mathcal{G}}_2\right\|_2\right), \end{aligned}$$

and since $\mathcal{G}$ is Lipschitz and due to (2.52), (2.53), (2.54), we obtain

$$\begin{aligned} \|C^{up}(u_1) - C^{up}(u_2)\|_2 &\leq 2m(C_{\mathcal{G}} J \|u_1 - u_2\| + C_{\mathcal{G}} \|u_1 - u_2\|) \\ &\quad + (J\|u_1 - u_2\| + \|u_1 - u_2\|)2M \\ &\leq 2(J+1)(mC_{\mathcal{G}} + M)\|u_1 - u_2\|, \end{aligned}$$

and defining $C_3 = 2(J+1)(mC_{\mathcal{G}} + M)$ we get (iii). The argument is similar for the matrix $C^{pp}(u)$, for which we have

$$\begin{aligned}\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \leq \frac{1}{J}\sum_{j=1}^{J}\left(\left\|\mathcal{G}(u_1^{(j)})\right\| + \|\bar{\mathcal{G}}_1\|_2\right)\left(\left\|\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})\right\|_2 + \|\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1\|_2\right)\\ + \frac{1}{J}\sum_{j=1}^{J}\left(\left\|\mathcal{G}(u_1^{(j)}) - \mathcal{G}(u_2^{(j)})\right\|_2 + \|\bar{\mathcal{G}}_2 - \bar{\mathcal{G}}_1\|_2\right)\left(\left\|\mathcal{G}(u_2^{(j)})\right\| + \|\bar{\mathcal{G}}_2\|\right)\\ \leq 4(J+1)MC_{\mathcal{G}},\end{aligned}$$

and defining $C_4 = 4(J+1)MC_{\mathcal{G}}$ we get (iv), which concludes the proof. □

Proof of Lemma 2.17. Let us consider an ensemble $u \in \mathcal{U}_{J,M}$ with particles $u^{(j)} \in \mathbb{R}^M$, for $j = 1, \dots, J$. For each particle we have

$$\left\|\mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 = \left\|\mathcal{O}(\mathcal{S}_h^0(u^{(j)})) - \mathcal{O}(\mathcal{S}^0(u^{(j)}))\right\|_2 \leq C_{\mathcal{O}}\left\|p_h^0(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)},$$

where we write explicitly the dependence of the solutions p^0 and p_h^0 on the particle they are generated by. Then due to standard a priori error estimates of FEM (see e.g. [33, Theorem 3.2.5]) and higher order boundary regularity results for elliptic partial differential equations (see e.g. [52, Theorem 6.3.5]) we have for all $j = 1, \dots, J$

$$\left\|p_h^0(u^{(j)}) - p^0(u^{(j)})\right\|_{L^2(D)} \leq C\left|p^0(u^{(j)})\right|_{H^{s+1}(D)} h^{s+1} \leq C\|f\|_{H^{q-1}(D)} h^{s+1},$$

where $C > 0$ is a constant independent of h. Therefore, we obtain

$$\widetilde{e}(h,u) = \frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)})\right\|_2 \leq C_{\mathcal{O}}C\|f\|_{H^{q-1}(D)} h^{s+1},$$

and defining $\widetilde{K} = C_{\mathcal{O}}C\|f\|_{H^{q-1}(D)}$ gives the desired result. □

Proof of Lemma 2.22. We follow the same steps of the proof of Theorem 5.9 in [124]. Let us first recall the duality formula for the Wasserstein distance with $p = 1$

$$W_{1,s}(\mu_n, \mu) = \sup_{\varphi\in\Phi}\left\{\int_{B_R(u^*)}\varphi d(\mu_n - \mu)\right\},$$

where Φ is the set of all globally Lipschitz continuous functions $\varphi\colon B_R(u^*) \to \mathbb{R}$ with Lipschitz constant $C_{\text{Lip}} \leq 1$. Note that if $\varphi \in \Phi$, then also $-\varphi \in \Phi$. Hence we deduce that

$$W_{1,s}(\mu_n, \mu) = \sup_{\varphi\in\Phi}\left\{\left|\int_{B_R(u^*)}\varphi d(\mu_n - \mu)\right|\right\}.$$

Then, we have

$$\begin{aligned}\sup_{\varphi\in\Phi}\mathbb{E}_\xi\left[\left|\int_{B_R(u^*)}\varphi d(\mu_n - \mu)\right|\right] &\leq \mathbb{E}_\xi\left[\sup_{\varphi\in\Phi}\left\{\left|\int_{B_R(u^*)}\varphi d(\mu_n - \mu)\right|\right\}\right]\\ &= \mathbb{E}_\xi[W_{1,s}(\mu_n, \mu)],\end{aligned}$$

where the right hand side vanishes by hypothesis. Therefore, we obtain

$$\mathbb{E}_\xi\left[\left|\int_{B_R(u^*)}\varphi d\mu_n - \int_{B_R(u^*)}\varphi d\mu\right|\right] \to 0,$$

for all $\varphi \in \Phi$. Finally, the desired result follows by a standard density argument. □

3 Multiscale Diffusions: Homogenization and Drift Estimation

In this chapter we introduce multiscale diffusion processes, and the problem of inferring from data an effective equation that captures its slow variations. Let us remark that in this chapter we mainly present standard material in the topics of homogenization and inference of multiscale diffusion processes, and its content can be found either in classical references or in more recent research articles, which are pointed in each relevant section. Moreover, let us remark that some results and phrasings of this chapter are taken from our original paper [8], whose content is explored in more detail in Chapter 4.

Let N and d be positive integers, let $\varepsilon > 0$ denote the scale-separation parameter and let $T > 0$ be a final time. Then, let $X^\varepsilon = (X_t^\varepsilon, 0 \leq t \leq T)$ be the stochastic process with values in $\mathbb{R}^d$ solution to the stochastic differential equation (SDE)

$$\mathrm{d}X_t^\varepsilon = -\sum_{i=1}^N \alpha_i \nabla V_i(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}\nabla p\left(\frac{X_t^\varepsilon}{\varepsilon}\right)\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \tag{3.1}$$

where we denote by $V\colon \mathbb{R}^d \to \mathbb{R}^N$, $V\colon x \mapsto V(x) \coloneqq (V_1(x), V_2(x), \ldots, V_N(x))^\top$ and $V_i\colon \mathbb{R}^d \to \mathbb{R}$ the slow-scale potential and by $\alpha \coloneqq (\alpha_1, \alpha_2, \ldots, \alpha_N)^\top \in \mathbb{R}^N$ its associated drift coefficient. Moreover, we denote by $p\colon \mathbb{R}^d \to \mathbb{R}$ the fast-scale potential, which we assume to be L-periodic in all directions for a period $L > 0$. Finally, we denote by $\sigma \in \mathbb{R}^+$ the diffusion coefficient and by $W = (W_t, t \geq 0)$ a d-dimensional standard Brownian motion.

Our first goal in this chapter is to determine an effective or homogenized model which captures the slow-scale variations of (3.1). In particular, let $X = (X_t, 0 \leq t \leq T)$ be the stochastic process with values in $\mathbb{R}^d$ be the solution of the single-scale SDE

$$\mathrm{d}X_t = -\sum_{i=1}^N A_i \nabla V_i(X_t^\varepsilon)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t. \tag{3.2}$$

Then, X is a surrogate of X^ε when the scale-separation parameter ε is small. Here, the effective drift and diffusion coefficients are given by $A_i = \alpha_i K$ for $i = 1, \ldots, N$, and $\Sigma = \sigma K$, where $K \in \mathbb{R}^{d\times d}$ is the matrix defined by

$$K \coloneqq \int_{\mathcal{Y}} (I + D_y\Phi(y))\,(I + D_y\Phi(y))^\top\,\mu(\mathrm{d}y), \tag{3.3}$$

where $\mathcal{Y} \coloneqq [0, L]^d$ and where the function $\Phi\colon \mathbb{R}^d \to \mathbb{R}^d$ is the solution to the so-called cell problem

$$\mathcal{L}_0\Phi = \nabla_y p(y), \quad \mathcal{L}_0 = -\nabla_y p(y)\cdot\nabla_y + \sigma\Delta_y, \tag{3.4}$$

defined on $\mathcal{Y}$ and endowed with periodic boundary conditions. Let us remark that for a vector-valued function the operator $\mathcal{L}_0$ is applied component-wise, as in

$$\mathcal{L}_0\Phi \coloneqq (\mathcal{L}_0\Phi_1, \mathcal{L}_0\Phi_2, \cdots, \mathcal{L}_0\Phi_d)^\top.$$

The integral in (3.3) is taken with respect to the probability measure $\mu(\mathrm{d}y)$ on $\mathcal{Y}$ defined by

$$\mu(\mathrm{d}y) = \rho(y)\,\mathrm{d}y, \quad \rho(y) = \frac{1}{Z}e^{-p(y)/\sigma}, \quad Z = \int_{\mathcal{Y}} e^{-p(y)/\sigma}\,\mathrm{d}y. \tag{3.5}$$

The measure μ is connected to the fast variations of the process X_t^ε solution of (3.1), in a sense which will be made clearer in the remainder of this chapter.

Example 3.1. Let us consider the simplest case $d = N = 1$, so that $\mathcal{Y} = [0, L]$. In this case, one can compute a closed-form solution to the cell problem (3.4), whose derivative is given by

$$\Phi'(y) = \frac{L}{\widehat{Z}}e^{-p(y)/\sigma} - 1, \quad \widehat{Z} = \int_{\mathcal{Y}} e^{p(y)/\sigma}\,\mathrm{d}y.$$

Substituting into the formula for K, which is in this case a scalar, one easily gets [112, Equation (1.12)]

$$K = \frac{L^2}{Z\widehat{Z}}, \quad Z = \int_{\mathcal{Y}} e^{-p(y)/\sigma}\,\mathrm{d}y,$$

where we recall Z to be the normalization constant given in (3.5). Let us remark that the Cauchy–Schwarz inequality yields

$$\sqrt{Z\widehat{Z}} \geq \int_{\mathcal{Y}} e^{-p(y)/(2\sigma)}e^{p(y)/(2\sigma)}\,\mathrm{d}y = L,$$

so that $0 < K \leq 1$. Moreover, we notice that $K \to 1$ for $\sigma \to \infty$ and that K is an increasing function of the diffusion coefficient σ, so that for a finite value $\sigma < \infty$ it actually holds $0 < K < 1$. In Fig. 3.1 we demonstrate graphically the effects of varying σ on the homogenized model. We consider (3.1) with $\varepsilon = 0.1$, the drift coefficient $\alpha = 1$, the slow and fast-scale potentials $V(x) = x^2/2$ and $p(y) = \sin(y)$ and depict the full multiscale potential $V_\varepsilon(x) = \alpha V(x) + p(x/\varepsilon)$, along with its slow-scale component αV and with the homogenized potential AV. We vary $\sigma \in \{0.5, 1, 2\}$ and notice that for small values of the diffusion coefficient the homogenized drift is sensibly less steep than the slow-scale component of the multiscale potential, whereas for $\sigma = 2$ the two are almost indistinguishable.

Before proceeding with the homogenization result, we present in Section 3.1 a result on the geometric ergodicity of the processes X^ε and X (see Section A.4), as well as of weak convergence of their invariant measures. We then present the homogenization result in two steps. First, in Section 3.2 we show a classical formal derivation of equation (3.2), which heavily relies on the connection between SDEs and their associated backward Kolmogorov equation (BKE) introduced in Section A.3. Second, in Section 3.3, we present a rigorous proof of convergence for the multiscale process X^ε towards its single-scale surrogate X in the asymptotic limit $\varepsilon \to 0$.

The theory of homogenization, which we briefly introduced above and which is explored more thoroughly in the following, is a powerful tool to determine an effective model of the form (3.2). Nevertheless, there are severe downsides which prevent its application in a wide range of applications. In particular, for high-dimensional problems, i.e., when d is large, solving the cell problem (3.4) can be computationally expensive or even unfeasible. Moreover, the fast-scale potential p in (3.1) above must be known in order to compute the coefficients of the effective equation, which is not the case in most practical applications.

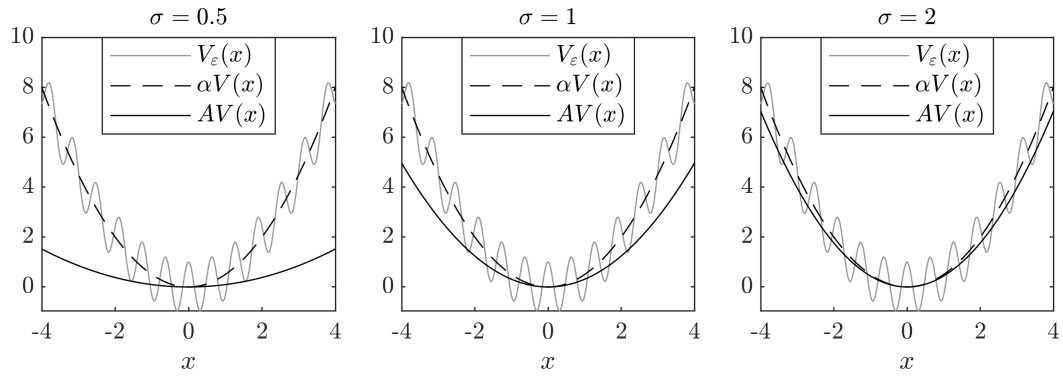


Figure 3.1 – Multiscale and homogenized confining potentials for the quadratic potential $V(x) = x^2/s$ and $p(y) = \sin(y)$, with $\varepsilon = 0.1$. From left to right, growing the diffusion coefficient σ yields for the homogenization coefficient $K \to 1$.

An alternative approach is provided by data-driven homogenization. Let us suppose that we observe a continuous-time process X^ε, whose time evolution can be modeled by (3.1). In the framework of data-driven homogenization, one determines an effective equation of the form (3.2) given these observations X^ε from (3.1). We consider a setting where the slow-scale potential V is known, but the fast scale potential p, the scale-separation parameter ε and the drift and diffusion coefficients α and σ are not. Clearly, in this case the theory of homogenization is not a viable approach to determine the effective model and one has to rely on statistical tools. The main desideratum of the resulting model is for it to be consistent with the theory of homogenization. Indeed, when an arbitrarily large amount of data is available (i.e., for $T \to \infty$) and in the homogenization regime (i.e., for $\varepsilon \to 0$) the statistically-inferred effective coefficients should coincide with the values predicted by the theory of homogenization. Unfortunately, if traditional statistical techniques are applied in this multiscale context without additional care, the resulting effective coefficient are asymptotically distant from the desired ones, due to an issue of model misspecification.

In Section 3.4 we introduce and derive from (3.2) a maximum likelihood estimator (MLE) of the drift coefficient, and show in Section 3.5 the issues that arise due to model misspecification. We then recast in Section 3.6 the problem of estimating the drift in the framework of Bayesian inference (see also Chapter 1).

3.1 Ergodic Properties

We first discuss the ergodic properties of the solutions X^ε and X of (3.1) and (3.2), respectively. Let us introduce a dissipative framework and the regularity assumptions which are fundamental for the theoretical analysis of both this chapter and for Chapter 4. The setting and results of this section can be found in [8, 94, 112, 116].

Assumption 3.2. The fast and slow-scale confining potentials p and V in (3.1) satisfy

(i) $p \in \mathcal{C}^\infty(\mathbb{R})$ and is L-periodic for some $L > 0$;

(ii) $V_i \in \mathcal{C}^\infty(\mathbb{R})$ for all $i = 1, \dots, N$ is polynomially bounded from above and bounded from below, and there exist $a, b > 0$ such that

$$-\alpha \cdot V'(x)x \leq a - bx^2;$$

(iii) V' is Lipschitz continuous, i.e. there exists a constant $C > 0$ such that

$$\|V'(x) - V'(y)\|_2 \leq C\,|x - y|\,,$$

and the components V_i' are polynomially bounded for all $i = 1, \dots, N$.

Crucially, Assumption 3.2 is sufficient for both X^ε and X to be geometrically ergodic (see Section A.4), as per, e.g., [94]. We resume this in the following result.

Proposition 3.3. *Under Assumption 3.2, the processes X^ε and X solutions of* (3.1) *are geometrically ergodic, and their invariant measures μ^ε and μ^0, respectively, satisfy*

$$\mu^\varepsilon(\mathrm{d}x) = \rho^\varepsilon(x)\,\mathrm{d}x, \quad \rho^\varepsilon(x) = \frac{1}{C_\varepsilon}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma} - \frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)\right),$$
$$\mu^0(\mathrm{d}x) = \rho^0(x)\,\mathrm{d}x, \quad \rho^0(x) = \frac{1}{C_0}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma}\right),$$

where C_ε and C_0 are the normalization constants

$$C_\varepsilon = \int_{\mathbb{R}^d}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma} - \frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)\right)\mathrm{d}x, \quad C_0 = \int_{\mathbb{R}^d}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma}\right)\mathrm{d}x.$$

Proof. For geometric ergodicity, we refer to [94]. It is moreover possible to verify through direct calculations involving the stationary Fokker–Planck equation (see Sections A.3 and A.4) that ρ^ε and ρ^0 are indeed the invariant densities of X^ε and X, respectively. □

We conclude this section with a first homogenization result, which guarantees weak convergence for the multiscale invariant measure μ^ε towards the homogenized measure μ^0 for $\varepsilon \to 0$. Let us remark that we employ the symbol $\Rightarrow$ for weak convergence of probability measures, see Section A.1. We point the reader to [112, Proposition 5.2] for a detailed proof.

Proposition 3.4. *Let μ^ε and μ^0 be the probability measures defined in Proposition 3.3. Then, it holds $\mu^\varepsilon \Rightarrow \mu^0$.*

Proof. We first notice that by definition of weak convergence of probability measures, if $\rho^\varepsilon \rightharpoonup \rho^0$ in $L^1(\mathbb{R}^d)$, where $\rightharpoonup$ denotes weak convergence in L^p spaces, then $\mu^\varepsilon \Rightarrow \mu$. Indeed, given a continuous and bounded function f, by definition of weak convergence in $L^1(\mathbb{R}^d)$ we have

$$\lim_{\varepsilon\to 0}\int_{\mathbb{R}^d} f(x)\rho^\varepsilon(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(x)\rho^0(x)\,\mathrm{d}x,$$

which shows $\mu^\varepsilon \Rightarrow \mu^0$. It remains to show $\rho^\varepsilon \rightharpoonup \rho^0$ in $L^1(\mathbb{R}^d)$, which is implied by the standard two-scale convergence result [34, Lemma 9.1], and which therefore concludes the proof. □

3.2 Derivation of the Homogenized Equation

In this section, we present a formal derivation of the homogenized model (3.2) based on an asymptotic expansion of the Backward Kolmogorov Equation (BKE), which we introduced in Section A.3. The discussion is based on [113, Chapter 11], where a more general but slightly different framework is considered, and more marginally on [22, Chapter 3] and on [112]. We remark that in [50] the authors propose numerical methods adapted to the setting we introduce here. In this section we consider for simplicity only the case $N = 1$ in (3.1).

Let us introduce the rescaled process $Y_t^\varepsilon \coloneqq X_t^\varepsilon/\varepsilon$, so that considering the evolution of X_t^ε and Y_t^ε yields the system of coupled SDEs

$$\begin{aligned} \mathrm{d}X_t^\varepsilon &= -\alpha\nabla V(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}\nabla p(Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \\ \mathrm{d}Y_t^\varepsilon &= -\frac{\alpha}{\varepsilon}\nabla V(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon^2}\nabla p(Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{\frac{2\sigma}{\varepsilon^2}}\,\mathrm{d}W_t, \end{aligned} \tag{3.6}$$

which reveals the slow/fast structure of this multiscale model. The generator $\mathcal{L}_\varepsilon$ of the joint process $(X_t^\varepsilon, Y_t^\varepsilon)^\top \in \mathbb{R}^{2d}$ reads

$$\mathcal{L}_\varepsilon \coloneqq \frac{1}{\varepsilon^2}\mathcal{L}_0 + \frac{1}{\varepsilon}\mathcal{L}_1 + \mathcal{L}_2, \tag{3.7}$$

with

$$\begin{aligned} \mathcal{L}_0 &= -\nabla_y p(y)\cdot\nabla_y + \sigma\Delta_y, \\ \mathcal{L}_1 &= -\nabla_y p(y)\cdot\nabla_x - \alpha\nabla_x V(x)\cdot\nabla_y + 2\sigma\nabla_x\cdot\nabla_y, \\ \mathcal{L}_2 &= -\alpha\nabla_x V(x)\cdot\nabla_x + \sigma\Delta_x, \end{aligned}$$

where we denote by ∇_x and ∇_y the gradient operator with respect to the variables x and y respectively, and by Δ_x and Δ_y the Laplacian with respect to the same variables.

It is possible to show that $\mathcal{L}_0$ is the generator of Y_t^ε conditioned on X_t^ε. Hence, if we consider the natural assumption that the process Y_t^ε is ergodic if the slow variable $X_t^\varepsilon = x$ is fixed, this implies that (see Section A.4)

$$\mathcal{L}_0 1(y) = 0, \tag{3.8}$$

$$\mathcal{L}_0^*\rho(y;x) = 0, \tag{3.9}$$

where $1(y)$ denotes functions which are constant with respect to y, and where $\rho(y;x)$ denotes the density of the invariant measure of Y_t^ε for any fixed $X_t^\varepsilon = x$. In particular, we have that $\rho(y;x) = \rho(y)$ is independent of x and is given by (3.5), which can directly verified to satisfy (3.9). We state explicitly the trivial equality

$$\nabla\rho(y) = -\frac{1}{\sigma}\rho(y)\nabla p(y), \quad i = 1,\dots,d, \tag{3.10}$$

since we will employ it multiple times in the following. Let us now consider the BKE for (3.6) and expand its solution $u = u(t,x,y)$ in powers of ε as

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \cdots. \tag{3.11}$$

Our goal is now deriving an effective BKE valid for the leading term u_0 only. Substituting (3.11) into the BKE and equating terms with the same powers of ε yields for the leading order terms the equations

$$\mathcal{L}_0 u_0 = 0, \tag{3.12}$$

$$\mathcal{L}_0 u_1 = -\mathcal{L}_1 u_0, \tag{3.13}$$

$$\mathcal{L}_0 u_2 = -\mathcal{L}_1 u_1 - \mathcal{L}_2 u_0 + \partial_t u_0. \tag{3.14}$$

Equation (3.12) together with (3.8) imply that u_0 is a constant with respect to y, in particular $u_0 = u_0(t,x)$. We can therefore rewrite (3.13) as

$$\mathcal{L}_0 u_1 = -\mathcal{L}_1 u_0 = \nabla_y p(y)\cdot\nabla_x u_0. \tag{3.15}$$

If we now impose

$$u_1 = \Phi\cdot\nabla_x u_0,$$

where Φ is the solution of the cell problem (3.4), then it is direct to verify the function u_1 satisfies (3.15). Let us remark that since $\mathcal{L}_0$ is a differential operator in y only, any function of the form $u_1 = \Phi \cdot \nabla_x u_0 + 1(y)$ would have been a solution, but we set the constant term to zero as it does not intervene in the following. It remains now to consider (3.14). Due to (3.8) and to the Fredholm alternative (see Theorem A.12), the right-hand side of (3.14) has to satisfy for solvability

$$0 = \int_{\mathcal{Y}} \left(-\mathcal{L}_1 u_1 - \mathcal{L}_2 u_0 + \partial_t u_0\right) \rho(y) \, \mathrm{d}y. \tag{3.16}$$

We compute the terms in the equation above singularly. Let us first remark that after some algebraic manipulations we can rewrite

$$\mathcal{L}_1 u_1 = -\alpha D_y \Phi \, \nabla_x V(x) \cdot \nabla_x u_0 + (2\sigma D_y \Phi - \nabla_y p(y) \otimes \Phi) : \nabla_x^2 u_0,$$

where $\otimes$ denotes the tensor product $v \otimes w = vw^\top$ for vectors $v, w \in \mathbb{R}^d$. Considering the second term, we can write it as

$$\mathcal{L}_2 u_0 = -\alpha \nabla_x V(x) \cdot \nabla_x u_0 + \sigma I : \nabla_x^2 u_0.$$

Hence, we have

$$\begin{aligned} \mathcal{L}_1 u_1 + \mathcal{L}_2 u_0 = &- \alpha (I + D_y \Phi) \nabla_x V(x) \cdot \nabla_x u_0 \\ &+ (\sigma I + 2\sigma D_y \Phi - \nabla_y p(y) \otimes \Phi) : \nabla_x^2 u_0. \end{aligned} \tag{3.17}$$

We now remark that it holds

$$\int_{\mathcal{Y}} \nabla_y p(y) \otimes \Phi(y) \rho(y) \, \mathrm{d}y = \sigma \int_{\mathcal{Y}} D_y \Phi(y) \rho(y) \, \mathrm{d}y,$$

which can be derived employing (3.10) and an integration by parts. Integrating (3.17), we can thus write

$$\int_{\mathcal{Y}} \left(\mathcal{L}_1 u_1 + \mathcal{L}_2 u_0\right) \rho(y) \, \mathrm{d}y = -\alpha K \nabla_x V(x) \cdot \nabla_x u_0 + \sigma K : \nabla_x^2 u_0, \tag{3.18}$$

where $K \in \mathbb{R}^{d \times d}$ is given by

$$K \coloneqq \int_{\mathcal{Y}} \left(I + D_y \Phi(y)\right) \rho(y) \, \mathrm{d}y.$$

For this matrix K and the one with identified by the same symbol given in (3.3) to actually be the same matrix, it remains to show that

$$\int_{\mathcal{Y}} D_y \Phi(y) D_y \Phi(y)^\top \rho(y) \, \mathrm{d}y = -\int_{\mathcal{Y}} D_y \Phi(y)^\top \rho(y) \, \mathrm{d}y,$$

which can be rewritten component-wise as

$$\int_{\mathcal{Y}} \nabla_y \Phi_j(y) \cdot \nabla_y \Phi_i(y) \rho(y) \, \mathrm{d}y = -\int_{\mathcal{Y}} \partial_{y_i} \Phi_j(y) \rho(y) \, \mathrm{d}y, \quad i, j = 1, \dots, d. \tag{3.19}$$

Indeed, applying an integration by parts and the identity (3.10) yields for all $i, j = 1, \dots, d$

$$\begin{aligned} \int_{\mathcal{Y}} \nabla_y \Phi_j(y) \cdot \nabla_y \Phi_i(y) \rho(y) \, \mathrm{d}y &= -\int_{\mathcal{Y}} \Phi_j(y) \nabla_y \cdot \left(\rho(y) \nabla_y \Phi_i(y)\right) \, \mathrm{d}y \\ &= -\int_{\mathcal{Y}} \Phi_j(y) \left(\nabla_y \rho(y) \cdot \nabla_y \Phi_i(y) + \rho(y) \Delta_y \Phi_i(y)\right) \, \mathrm{d}y \\ &= -\int_{\mathcal{Y}} \Phi_j(y) \left(-\frac{1}{\sigma} \nabla_y p(y) \cdot \nabla_y \Phi_i(y) + \Delta_y \Phi_i(y)\right) \rho(y) \, \mathrm{d}y. \end{aligned}$$

We now identify in the integrand on the right-hand side the quantity $\mathcal{L}_0\Phi_i$, so that by the cell problem (3.4), employing (3.10) and with a further integration by parts we obtain

$$\begin{aligned}\int_{\mathcal{Y}} \nabla_y\Phi_j(y)\cdot\nabla_y\Phi_i(y)\rho(y)\,\mathrm{d}y &= -\int_{\mathcal{Y}} \frac{1}{\sigma}\Phi_j(y)\mathcal{L}_0\Phi_i(y)\rho(y)\,\mathrm{d}y\\ &= -\int_{\mathcal{Y}} \frac{1}{\sigma}\Phi_j(y)\partial_{y_i}p(y)\rho(y)\,\mathrm{d}y\\ &= \int_{\mathcal{Y}} \Phi_j(y)\partial_{y_i}\rho(y)\,\mathrm{d}y\\ &= -\int_{\mathcal{Y}} \partial_{y_i}\Phi_j(y)\rho(y)\,\mathrm{d}y,\end{aligned}$$

which shows (3.19) and thus that the symmetric positive semi-definite matrix K is given by (3.3). Finally, we have by (3.16) and (3.18) that

$$\partial_t u_0 = -\alpha K\nabla_x V(x)\cdot\nabla_x u_0 + \sigma K : \nabla_x^2 u_0,$$

which is exactly the BKE corresponding to (3.2) by defining $A \coloneqq \alpha K$ and $\Sigma \coloneqq \sigma K$, and which therefore concludes the derivation of the homogenized model.

3.3 The Convergence Theorem

In this section, we present a rigorous proof of the homogenization result for multiscale diffusion processes. The discussion is mainly based on [113, Chapter 18], where a more general but slightly different framework is considered, and more marginally on [22, Chapter 3] and on some results of [8, 112]. As for the previous section, we consider here for simplicity only the case $N = 1$ in (3.1).

Theorem 3.5. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{H} = C^0((0,T),\mathbb{R}^d)$ and $T > 0$. Moreover, let $X^\varepsilon, X\colon \Omega \to \mathcal{H}$ be the solutions of (3.1) and (3.2), respectively, with the same initial condition $X_0^\varepsilon = X_0$ in law. Then, it holds $X^\varepsilon \Rightarrow X$ in $\mathcal{H}$ for $\varepsilon \to 0$.*

Proof. The proof is achieved with a careful term-by-term analysis of the integral expression for X_t^ε, i.e.,

$$\begin{aligned}X_t^\varepsilon - X_0^\varepsilon &= -\alpha\int_0^t \nabla_x V(X_s^\varepsilon)\,\mathrm{d}s - \frac{1}{\varepsilon}\int_0^t \nabla_y p(Y_s^\varepsilon)\,\mathrm{d}s + \sqrt{2\sigma}W_t\\ &= J_1 + J_2 + \sqrt{2\sigma}W_t,\end{aligned} \tag{3.20}$$

where we introduced the notation

$$J_1 \coloneqq -\alpha\int_0^t \nabla_x V(X_s^\varepsilon)\,\mathrm{d}s, \quad J_2 \coloneqq -\frac{1}{\varepsilon}\int_0^t \nabla_y p(Y_s^\varepsilon)\,\mathrm{d}s.$$

We first consider the term J_1 above. Let us introduce the function $\chi\colon \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}^d$ solution of the elliptic PDE

$$\begin{aligned}\mathcal{L}_0\chi(x,y) &= -\alpha(I + D_y\Phi(y))\nabla_x V(x) + \alpha K\nabla_x V(x),\\ \int_{\mathcal{Y}} \chi(x,y)\mu(\mathrm{d}y) &= 0,\end{aligned} \tag{3.21}$$

on $\mathcal{Y}$ with periodic boundary conditions, where $\mathcal{L}_0$ is introduced in (3.4), and where the measure μ and the matrix K are introduced in (3.3). Let us remark that since the right-hand side of

(3.21) is centered with respect to the measure μ by definition of K, existence and uniqueness of χ are guaranteed by the Fredholm alternative (see Theorem A.12). We now apply the Itô formula to the process $\chi_t \coloneqq \chi(X_t^\varepsilon, Y_t^\varepsilon)$, where $(X_t^\varepsilon, Y_t^\varepsilon)$ is the solution of (3.6). It holds

$$\mathrm{d}\chi_t = -\mathcal{L}_\varepsilon \chi_t \,\mathrm{d}t + \left(\sqrt{2\sigma} D_x \chi_t + \sqrt{\frac{2\sigma}{\varepsilon^2}} D_y \chi_t\right) \mathrm{d}W_t,$$

where $\mathcal{L}_\varepsilon = \varepsilon^{-2}\mathcal{L}_0 + \varepsilon^{-1}\mathcal{L}_1 + \mathcal{L}_2$ is introduced in (3.7) and is applied to χ_t component-wise. In integral form, we have the identity

$$\int_0^t \mathcal{L}_0 \chi_s \,\mathrm{d}s = \varepsilon^2(\chi_t - \chi_0) - \varepsilon \int_0^t (\mathcal{L}_1 \chi_s + \varepsilon \mathcal{L}_2 \chi_s)\,\mathrm{d}s + \varepsilon\sqrt{2\sigma} \int_0^t (\varepsilon D_x \chi_s + D_y \chi_s)\,\mathrm{d}W_s,$$

which we can rewrite in light of (3.21) as

$$J_1 = -\alpha \int_0^t (K - D_y \Phi(Y_s^\varepsilon))\, \nabla_x V(X_s^\varepsilon)\,\mathrm{d}s + J_1^{(1)} + J_1^{(2)}, \tag{3.22}$$

where

$$J_1^{(1)} = \varepsilon^2(\chi_t - \chi_0) - \varepsilon \int_0^t (\mathcal{L}_1 \chi_s + \varepsilon \mathcal{L}_2 \chi_s)\,\mathrm{d}s,$$
$$J_1^{(2)} = \varepsilon\sqrt{2\sigma} \int_0^t (\varepsilon D_x \chi_s + D_y \chi_s)\,\mathrm{d}W_s.$$

Let us now consider J_2. The Itô formula applied to $\Phi(Y_t^\varepsilon)$, where Φ is the solution of the cell problem (3.4) yields

$$\mathrm{d}\Phi(Y_t^\varepsilon) = \frac{1}{\varepsilon^2}\left(\mathcal{L}_0 \Phi(Y_t^\varepsilon) - \varepsilon\alpha D_y \Phi(Y_t^\varepsilon) \nabla_x V(X_t^\varepsilon)\right) \mathrm{d}t + \sqrt{\frac{2\sigma}{\varepsilon^2}} D_y \Phi(Y_t^\varepsilon)\,\mathrm{d}W_t.$$

Passing to the integral form and by the cell problem (3.4) we have

$$J_2 = -\alpha \int_0^t D_y \Phi(Y_s^\varepsilon)\,\mathrm{d}s + \sqrt{2\sigma} \int_0^t D_y \Phi(Y_s^\varepsilon)\,\mathrm{d}W_s + J_2^{(1)}, \tag{3.23}$$

where

$$J_2^{(1)} = -\varepsilon\left(\Phi(Y_t^\varepsilon) - \Phi(Y_0^\varepsilon)\right).$$

We now consider (3.22), (3.23) and (3.20) to obtain

$$X_t^\varepsilon - X_0^\varepsilon = -\int_0^t A \nabla_x V(X_s^\varepsilon)\,\mathrm{d}s + \sqrt{2\sigma} \int_0^t (I + D_y \Phi(Y_s^\varepsilon))\,\mathrm{d}W_s + J_1^{(1)} + J_1^{(2)} + J_2^{(1)},$$

where we replaced $A = \alpha K$ as in (3.2). Under the assumptions on V and p, it is possible to show that χ and Φ are bounded along with their derivatives [113, Lemma 18.3]. It is then direct to show that for $\varepsilon \to 0$

$$J_1^{(1)}, J_1^{(2)}, J_2^{(1)} \to 0, \quad \text{in } L^p(\Omega),$$

which implies weak convergence. Introducing the notation

$$S_t \coloneqq \sqrt{2\sigma} \int_0^t (I + D_y \Phi(Y_s^\varepsilon))\,\mathrm{d}W_s,$$

we now show employing the functional central limit theorem for martingales (see Theorem A.22) that $S \Rightarrow \sqrt{2\Sigma} W$ for $\varepsilon \to 0$ in $\mathcal{H}$, where $\Sigma = \sigma K$. Let us remark that the quadratic variation of S is given by

$$\langle S \rangle_t = 2\sigma \int_0^t (I + D_y \Phi(Y_s^\varepsilon))(I + D_y \Phi(Y_s^\varepsilon))^\top \,\mathrm{d}s.$$

Moreover, by definition of Σ (i.e., of K) it holds

$$\mathbb{E}^{\mu}\left[2\sigma(I + D_y\Phi(Y_s^\varepsilon))(I + D_y\Phi(Y_s^\varepsilon))^\top - 2\Sigma\right] = 0,$$

where μ is the measure given in Eq. (3.5). Therefore, by [112, Lemma 5.6], there exists a constant $C > 0$ independent of T and ε such that

$$\mathbb{E}^{\mu^\varepsilon}\left\|\langle S\rangle_t - 2\Sigma t\right\|^2 \leq C\left(\varepsilon^4 + \varepsilon^2 T^2 + \varepsilon^2 T\right),$$

which implies that $\langle S\rangle_t \to 2\Sigma t$ in L^2, and thus $\langle S\rangle_t \Rightarrow 2\Sigma t$ for $\varepsilon \to 0$. Hence, Theorem A.22 yields for $\varepsilon \to 0$

$$S \Rightarrow \sqrt{2\Sigma}W, \quad \text{in } \mathcal{H}.$$

We can now conclude. Indeed, we have shown that there exists a process ϑ_t such that

$$X_t^\varepsilon - X_0^\varepsilon = -\int_0^t A\nabla_x V(X_s^\varepsilon)\,\mathrm{d}s + \vartheta_t,$$

and satisfying $\vartheta \Rightarrow \sqrt{2\Sigma}W$ in $\mathcal{H}$ and for $\varepsilon \to 0$. Moreover, the mapping $\vartheta \mapsto X^\varepsilon$ is continuous [113, Lemma 18.2], and since weak convergence is preserved by continuous mappings (see Theorem A.5), we have that $X^\varepsilon \Rightarrow X$ for $\varepsilon \to 0$, where X is the solution to (3.2) with initial condition $X_0 = X_0^\varepsilon$, which concludes the proof. □

Remark 3.6. The assumption $X_0 = X_0^\varepsilon$ in law can be relaxed. In particular, let $X_0 \sim \nu^0$ and $X_0^\varepsilon \sim \nu^\varepsilon$ for some probability measure ν^0, ν^ε. If $\nu^\varepsilon \Rightarrow \nu^0$, then Theorem 3.5 holds. In particular, this holds true by Proposition 3.4 in case $X_0^\varepsilon \sim \mu^\varepsilon$ and $X_0 \sim \mu^0$, i.e., if X^ε and X are at stationarity.

3.4 Maximum Likelihood Estimation of the Drift

We now consider the problem of estimating the drift coefficient A of (3.2) when data in the form of a continuous time series are provided. For a final time $T > 0$, in this section we consider data to consist of a generic continuous process $Y \coloneqq (Y_t, 0 \leq t \leq T)$, abstracting ourselves from the multiscale setting which is the topic of this chapter. The case $Y = X^\varepsilon$, i.e., when the data originates from the multiscale model (3.1), is treated in details in Section 3.5 and in Chapter 4. We here consider for simplicity equation (3.2) in the one-dimensional case but with a multi-dimensional drift coefficient A, or in symbols we have $d = 1$ and a generic N. In this case, the slow-scale potential $V\colon \mathbb{R} \to \mathbb{R}^N$, and we denote by $V'(x) \coloneqq (V_1'(x), \ldots, V_N'(x))^\top$, and similarly for higher order derivatives.

The path-wise likelihood function of the data Y given the drift coefficient A and associated to the equation (3.2) is given by

$$L_T(Y \mid A) = \exp\left(-\frac{I_T(Y \mid A)}{2\Sigma}\right), \tag{3.24}$$

where the negative log-likelihood $I_T(Y \mid A)$ reads

$$I_T(Y \mid A) = \int_0^T A \cdot V'(Y_t)\,\mathrm{d}Y_t + \frac{1}{2}\int_0^T \left(A \cdot V'(Y_t)\right)^2\,\mathrm{d}t.$$

In the following, we drop for economy of notation the dependence of L_T and I_T on the final time, and simply write L and I. The MLE $\widehat{A}(Y,T)$ is then clearly the unique minimiser of the function $I(Y \mid A)$, which is quadratic with respect to A, and is the solution of the linear system

$$\begin{aligned} &-M(Y)\widehat{A}(Y,T) = v(Y), \\ &M(Y) \coloneqq \frac{1}{T}\int_0^T V'(Y_t) \otimes V'(Y_t)\,\mathrm{d}t, \quad v(Y) \coloneqq \frac{1}{T}\int_0^T V'(Y_t)\,\mathrm{d}Y_t. \end{aligned} \tag{3.25}$$

In the following, we omit in case of no ambiguity the argument Y from the matrix M and the right-hand side v. Let us remark that the square matrix M is by definition symmetric positive semi-definite. We introduce an assumption of positive definitess so that $\widehat{A}(Y,T)$ is well-defined.

Assumption 3.7. For all $T > 0$ and for all $Y \in C^0((0,T))$ the symmetric matrix $M(Y)$ is positive definite and there exists $\bar{\lambda} > 0$ such that $\lambda_{\min}(M(X)) \geq \bar{\lambda}$.

When data are finite-dimensional, and when the law underlying the model one wants to fit to data admits a probability density with respect to the Lebesuge measure, the likelihood function is given by the joint density of the data. Let us provide an example for clarity. Let $Y = (Y_1, Y_2, \dots, Y_n)$, with $Y_i \in \mathbb{R}$ for $i = 1, \dots, n$ be a data set consisting of independent observations, and let us compute the likelihood that the data is produced by a Gaussian distribution $\mathcal{N}(\mu, 1)$, of unknown mean μ and of unitary variance, for simplicity. The likelihood function, in this case, can be written due to the independence of the data points as

$$L(Y \mid \mu) = \frac{1}{\sqrt{2\pi}} \prod_{i=1}^{n} \exp\left(-\frac{(Y_i - \mu)^2}{2}\right),$$

and the MLE for the mean μ is clearly the sample average of Y. In the SDE context of this chapter, observations consist of a continuous path, and the model we wish to fit to data induces a measure on the infinite-dimensional space $\mathcal{H} = \mathcal{C}^0((0,T))$. In particular, denoting by $\mathcal{B}(\mathcal{H})$ the Borel σ-algebra on $\mathcal{H}$ we denote by μ_X the measure induced by the process $X \coloneqq (X_t, 0 \leq t \leq T)$ solution of (3.2) on the measurable space $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. In this framework, it is not possible to work with densities since the Lebesgue measure is not defined, and therefore the viable alternative is finding a measure on $\mathcal{H}$ with respect to which μ_X is absolutely continuous. A reasonable choice is to pick μ_W, i.e., the measure induced by Brownian motion on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, as a reference measure. Our goal for the rest of this section is therefore first showing that $\mu_X \ll \mu_W$, and then that the Radon–Nykodim derivative (see Section A.2) of μ_X with respect to μ_W reads

$$\frac{\mathrm{d}\mu_X}{\mathrm{d}\mu_W}(Y \mid A) = L(Y \mid A), \tag{3.26}$$

where L is given in (3.24).

Remark 3.8. We note that the multi-dimensional inference setting we consider here is often referred to in literature as the semi-parametric framework (see e.g. [81]). In particular the components V_i, $i = 1, \dots, N$ of the slow-scale potential can be interpreted as the known basis functions of a truncated expansion (e.g. a Taylor expansion) for an unknown confining potential $V_\alpha \colon \mathbb{R} \to \mathbb{R}$ given by

$$V_\alpha(x) = \sum_{i=1}^{N} \alpha_i V_i(x).$$

In words, one can think of inference to be performed not on a space of parameters, but on a finite-dimensional space of functions.

3.4.1 A Heuristic Derivation of the Likelihood

Before entering the details of the theoretical derivation, we now briefly present a heuristic argument which suggests that the likelihood function is indeed given by (3.24). The calculations presented here can be found in [110, Chapter 5] or [79, Chapter 6]. Let for simplicity $N = 1$, and let $V(x) = x^2/2$, so that (3.2) reads

$$\mathrm{d}X_t = -AX_t \,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t, \tag{3.27}$$

i.e., X_t is an Ornstein–Uhlenbeck process. Given a positive integer n, we consider the Euler–Maruyama approximation of (3.27), which for a sequence of time points $0 \leq t_0 < t_1 \ldots < t_n = T$, with $t_i = ih$ for a time step $h = T/n$ and an initial condition $X^h_{t_0} = x_0$ reads for all $i = 0, \ldots, n-1$

$$X^h_{t_{i+1}} = (1 - Ah)\, X^h_{t_i} + \sqrt{2\Sigma}\, \Delta W_{t_i}, \tag{3.28}$$

where $\Delta W_{t_i} \coloneqq W_{t_{i+1}} - W_{t_i}$ are the Brownian increments, which satisfy $\Delta W_{t_i} \sim \mathcal{N}(0, h)$. Hence, we clearly have that the measure of the random variable $X^h_{t_{i+1}} \mid X^h_{t_i}$ is $\mathcal{N}((1 - Ah)X^h_{t_i}, 2\Sigma h)$. Therefore, the joint distribution μ^h_X of $X^h \coloneqq (X^h_{t_1}, X^h_{t_2}, \ldots, X^h_{t_n})^\top$ admits by the Markov property a density p^h_X with respect to the Lebesgue measure which reads

$$p^h_X(x_1, x_2, \ldots, x_N \mid A) = \frac{1}{\sqrt{4\pi\Sigma h}} \prod_{i=0}^{n-1} \exp\left(-\frac{(x_{i+1} - (1 - Ah)x_i)^2}{4\Sigma h} \right),$$

where we highlight the dependence on A of the left-hand side. Similarly, the joint distribution μ^h_W of the rescaled vector of Brownian motion values $W \coloneqq \sqrt{2\Sigma}(W_{t_1}, W_{t_2}, \ldots, W_{t_n})^\top$ admits a density p^h_W with respect to the Lebesgue measure which satisfies

$$p^h_W(x_1, x_2, \ldots, x_N) = \frac{1}{\sqrt{4\pi\Sigma h}} \prod_{i=0}^{n-1} \exp\left(-\frac{(x_{i+1} - x_i)^2}{4\Sigma h} \right).$$

Taking the ratio p^h_X / p^h_W is clearly equivalent to computing the Radon–Nykodim derivative $\mathrm{d}\mu^h_X / \mathrm{d}\mu^h_W$, which after algebraic simplifications can be written as

$$\frac{\mathrm{d}\mu^h_X}{\mathrm{d}\mu^h_W}(x_1, x_2, \ldots, x_n \mid A) = \exp\left(-\frac{1}{2\Sigma} \sum_{i=0}^{n-1} A x_i \Delta x_i - \frac{1}{4\Sigma} \sum_{i=0}^{n-1} (A x_i)^2 h \right),$$

where $\Delta x_i \coloneqq x_{i+1} - x_i$. We now have an expression for the likelihood function. Assuming we are given discretized observations $Y \coloneqq (Y_{t_1}, Y_{t_2} \ldots, Y_{t_n})^\top$ and wish to compute their likelihood with respect to discretized model (3.28), we then obtain

$$L^h(Y \mid A) \coloneqq \frac{\mathrm{d}\mu^h_X}{\mathrm{d}\mu^h_W}(Y \mid A) = \exp\left(-\frac{1}{2\Sigma} \sum_{i=0}^{n-1} A Y_{t_i} \Delta Y_{t_i} - \frac{1}{4\Sigma} \sum_{i=0}^{n-1} (A Y_{t_i})^2 h \right),$$

where again $\Delta Y_{t_i} \coloneqq Y_{t_{i+1}} - Y_{t_i}$. We therefore conclude this heuristic derivation by noticing that in the limit $h \to 0$ we have that $L^h \to L$, where L is given in (3.24).

3.4.2 A Rigorous Derivation of the Likelihood

We now present the derivation of (3.24), which is based on an application of Girsanov's theorem. For the sake of simplicity, we fix in this section the diffusion coefficient so that $2\Sigma = 1$ in (3.2), and note that this choice does not affect the generality of the discussion. We refer the reader to [88, Chapters 6 and 7], where the derivation we present in this section is covered to greater extent, and to [20, 24, 110, 120] for further details and insights.

The key object in the derivation of the likelihood function (3.24) is the stochastic process $\beta \coloneqq (\beta_t, 0 \leq t \leq T)$ defined as

$$\beta_t \coloneqq \exp\left(\int_0^t A \cdot V'(X_s)\, \mathrm{d}W_s - \frac{1}{2} \int_0^t (A \cdot V'(X_s))^2\, \mathrm{d}s \right). \tag{3.29}$$

Indeed, employing the SDE (3.2) we obtain at final time

$$\begin{aligned}\beta_T &= \exp\left(\int_0^T A\cdot V'(X_t)(\mathrm{d}X_t + A\cdot V'(X_t)\,\mathrm{d}t) - \frac{1}{2}\int_0^T \left(A\cdot V'(X_t)\right)^2\,\mathrm{d}t\right)\\ &= \exp\left(\int_0^T A\cdot V'(X_t)\,\mathrm{d}X_t + \frac{1}{2}\int_0^T \left(A\cdot V'(X_t)\right)^2\,\mathrm{d}t\right),\end{aligned} \tag{3.30}$$

and notice that, formally, $\beta_T^{-1} = L(X \mid A)$, where L is the likelihood function defined in (3.24). Given the natural probability space $(\Omega, \mathcal{F}, P)$, such that W is a standard Brownian motion, let us remark that β_T can be interpreted in a twofold manner. On the one hand, we can write $\beta_T = \beta_T(\omega)$, so that $\beta_T\colon \Omega \to \mathbb{R}$ is a non-negative random variable defined on Ω. On the other hand, considering the last line of (3.30) we can see β_T to be a function of an element $x \in \mathcal{H} = \mathcal{C}^0((0,T))$ and write

$$\beta_T(x) = \exp\left(\int_0^T A\cdot V'(x_t)\,\mathrm{d}x_t + \frac{1}{2}\int_0^T \left(A\cdot V'(x_t)\right)^2\,\mathrm{d}t\right),$$

so that $\beta_T\colon \mathcal{H} \to \mathbb{R}$ is a non-negative measurable function defined on $\mathcal{H}$. Considering the second writing and recalling (3.26), our goal is then showing that

$$\frac{\mathrm{d}\mu_X}{\mathrm{d}\mu_W}(x) = \beta_T^{-1}(x),$$

where μ_X and μ_W are the measures induced by X and W, respectively, on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$.

We start by stating Girsanov's theorem, which allows to interpret the process X as a Brownian motion after a change of measure on $(\Omega, \mathcal{F})$.

Theorem 3.9 (Girsanov)**.** *Let Q be the probability measure on $(\Omega, \mathcal{F})$ with Radon–Nykodim derivative*

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) = \beta_T(\omega),$$

where β is defined in (3.29)*. Then, the process X solution of* (3.2) *is a Brownian motion with respect to Q.*

The proof of Girsanov's theorem, stated in more general frameworks, can be found in [88, Chapter 6] or [120, Chapter VIII]. We nevertheless remark that the condition $\mathbb{E}[\beta_T] = 1$, where $\mathbb{E}$ denotes expectation with respect to P, is necessary for Q to be a probability measure. Indeed, if Q is a probability measure we have

$$1 = Q(\Omega) = \int_\Omega \beta_T(\omega)\,\mathrm{d}P(\omega) = \mathbb{E}[\beta_T].$$

We verify that this necessary condition holds in our framework. An application of the Itô formula yields

$$\mathrm{d}\beta_t = \left(A\cdot V'(X_t)\right)\beta_t\,\mathrm{d}W_t,$$

or in integral form and since $\beta_0 = 1$ P-a.s.,

$$\beta_t = 1 + \int_0^t \left(A\cdot V'(X_s)\right)\beta_s\,\mathrm{d}W_s.$$

Hence, we have $\mathbb{E}[\beta_T] = 1$ and Q is indeed a probability measure on $(\Omega, \mathcal{F})$. Let us remark that in (3.26) we have a change of measure in $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, whereas Girsanov's theorem provides a change of measure on $(\Omega, \mathcal{F})$. In the following lemma we give a first characterization of the relationship between the measures μ_W and μ_X.

Lemma 3.10. *Let μ_W and μ_X be the probability measures on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ induced by the Brownian motion W and by the solution X of (3.2), respectively. Then $\mu_W \ll \mu_X$ and*

$$\frac{\mathrm{d}\mu_W}{\mathrm{d}\mu_X}(x) = \beta_T(x),$$

where β is defined in (3.29).

Proof. By definition of the induced measure μ_W on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ and Theorem 3.9, it holds for all $B \in \mathcal{B}(\mathcal{H})$

$$\mu_W(B) = P\left(\{\omega \colon W(\omega) \in B\}\right) = Q\left(\{\omega \colon X(\omega) \in B\}\right).$$

Now, since $\mathrm{d}Q = \beta_T(\omega)\,\mathrm{d}P$, we have by a change of variable

$$\mu_W(B) = \int_{\{\omega \colon X(\omega) \in B\}} \beta_T(\omega)\,\mathrm{d}P(\omega) = \int_B \beta_T(x)\,\mathrm{d}\mu_X(x),$$

which concludes the proof due to Theorem A.7. □

In the following result, we finally prove that (3.26) holds employing Lemma 3.10 and inverting the Radon–Nykodim derivative by an application of Theorem A.8.

Proposition 3.11. *Let the potential V be such that*

$$\int_0^T \left(A \cdot V'(X_t)\right)^2 \mathrm{d}t < \infty, \tag{3.31}$$

holds P-a.s., where X is the solution of (3.2). Then $\mu_X \sim \mu_W$ and

$$\frac{\mathrm{d}\mu_X}{\mathrm{d}\mu_W}(x) = \beta_T^{-1}(x).$$

Proof. We first show that $\beta_T > 0$ P-a.s., which under (3.31) is implied by

$$\int_0^T \left(A \cdot V(X_t)\right) \mathrm{d}W_t < \infty, \quad P\text{-a.s.}$$

Indeed, we have by Jensen's inequality, Itô isometry and (3.31)

$$\mathbb{E}\left[\left|\int_0^T A \cdot V(X_t)\,\mathrm{d}W_t\right|\right]^2 \leq \mathbb{E}\left[\int_0^T \left(A \cdot V(X_t)\right)^2 \mathrm{d}t\right] < \infty.$$

Then $\beta_T > 0$ follows by noticing that for any random variable $Z \colon \Omega \to \mathbb{R}$ such that $\mathbb{E}\left[|Z|\right] < \infty$, it holds $Z < \infty$ P-a.s. Theorem A.8 and Lemma 3.10 then imply

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) = \beta_T^{-1}(\omega).$$

We conclude proceeding similarly to the proof of Lemma 3.10. In particular, it holds for all $B \in \mathcal{B}(\mathcal{H})$

$$\mu_X(B) = P(\{\omega \colon X(\omega) \in B\}) = \int_{\{\omega \colon X(\omega) \in B\}} \beta_T^{-1}(\omega)\,\mathrm{d}Q(\omega) = \int_B \beta_T^{-1}(x)\,\mathrm{d}\mu_W(x),$$

which proves the desired result. □

Proposition 3.11 concludes the derivation of the likelihood function, and consequently of the MLE for the drift coefficient (3.25), and closes this section.

3.4.3 Asymptotic Consistency of the MLE

We now consider the estimator $\widehat{A}$ introduced in (3.25) and show that if we are confronted with data $(X, 0 \leq t \leq T)$ coming from the model (3.2) itself, then the estimator is asymptotically consistent, in the limit of an infinite amount of data (i.e., for $T \to \infty$). The following result is classic in the literature of statistical estimation of diffusion processes, and its proof can be found, for example, in the books [20, 24, 83, 88, 89] or in the more recent articles [115, 116]. In particular, we report here a proof based on [116].

Theorem 3.12. *Let Assumption 3.2 hold and let X be the solution of* (3.2) *with drift coefficient $A \in \mathbb{R}^N$. Then*

$$\lim_{T\to\infty} \widehat{A}(X,T) = A, \quad a.s.,$$

where $\widehat{A}$ is the MLE defined in (3.25).

Proof. By definition of the MLE and replacing (3.2) we have the decomposition

$$\widehat{A}(X,T) = -M^{-1}\frac{1}{T}\int_0^T V'(X_t)\left(-\left(A\cdot V'(X_t)\right)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t\right).$$

Let us remark that it holds

$$\frac{1}{T}\int_0^T V'(X_t)\left(-A\cdot V'(X_t)\right)\,\mathrm{d}t = -MA,$$

so that

$$\widehat{A}(X,T) = A - R(T), \quad R(T) \coloneqq M^{-1}\frac{\sqrt{2\Sigma}}{T}\int_0^T V'(X_t)\,\mathrm{d}W_t.$$

It remains to show that $R(T) \to 0$ a.s. for $T \to \infty$. The ergodic theorem (see Theorem A.14) yields

$$\lim_{T\to\infty} R(T) = \mathbb{E}^{\mu^0}[V'(X)\otimes V'(X)]^{-1}\lim_{T\to\infty}\frac{\sqrt{2\Sigma}}{T}\int_0^T V'(X_t)\,\mathrm{d}W_t.$$

Due to [116, Lemma 6.1], it holds under Assumption 3.2

$$\lim_{T\to\infty}\frac{\sqrt{2\Sigma}}{T}\int_0^T V'(X_t)\,\mathrm{d}W_t = 0, \quad \text{a.s.},$$

in $\mathbb{R}^N$, which proves $R(T) \to 0$ a.s. and the desired result. □

Let us remark that it could be further possible to prove a central limit theorem (CLT) to show that the quantity $\sqrt{T}(\widehat{A} - A)$ follows asymptotically a Gaussian distribution. This is achieved employing some form of the martingale CLT (see Section A.5). We refer the reader to [110, Chapter 5] for the proof in a specific one-dimensional case or to [20, 24, 83, 88, 89, 139] for further details.

3.5 Drift Estimation of Multiscale Diffusions

We now consider the case when the data is generated by the model (3.1), and we wish to retrieve the drift coefficient of the effective equation of the form Eq. (3.2) from the data. As we stated in

the introduction of this chapter, this problem can be interpreted as an instance of data-driven homogenization. Replacing a trajectory $X^\varepsilon = (X_t^\varepsilon, 0 \le t \le T)$ into (3.1) yields

$$-M(X^\varepsilon)\widehat{A}(X^\varepsilon, T) = v(X^\varepsilon). \tag{3.32}$$

The homogenization results Proposition 3.4 and Theorem 3.5, which show that X^ε and X, the solution of (3.2) are close if ε is small, would suggest that $\widehat{A}(X^\varepsilon, T)$ is close to $\widehat{A}(X, T)$, which in turn converges to A by Theorem 3.12 for $T \to \infty$. Hence, one would reasonably expect that $\widehat{A}(X^\varepsilon, T)$ is an asymptotically consistent estimator for A, with respect to $\varepsilon \to 0$ and $T \to \infty$. This is not the case, and the MLE tends to the drift coefficient α of the multiscale equation (3.1). We report here the proof of this negative result for completeness, and refer the reader to [112, Theorem 3.4] for the original proof. We introduce for economy of notation the quantities

$$\mathcal{M}_\varepsilon \coloneqq \mathbb{E}^{\mu^\varepsilon}[V'(X^\varepsilon) \otimes V'(X^\varepsilon)], \quad \mathcal{M}_0 \coloneqq \mathbb{E}^{\mu^0}[V'(X) \otimes V'(X)], \tag{3.33}$$

where μ^ε and μ^0 are the invariant measures of X^ε and X, given in Proposition 3.3.

Theorem 3.13. *Let Assumption 3.2 hold and let $X^\varepsilon = (X_t^\varepsilon, 0 \le t \le T)$ be the solution of* (3.1) *with drift coefficient $\alpha \in \mathbb{R}^N$. Then*

$$\lim_{T\to\infty} \widehat{A}(X^\varepsilon, T) = \alpha, \quad \textit{a.s.},$$

where $\widehat{A}$ is the MLE defined in (3.25).

Proof. Proceeding as in the proof of Theorem 3.12, we obtain the decomposition

$$\widehat{A}(X^\varepsilon, T) = \alpha + R_1^\varepsilon(T) - R_2^\varepsilon(T),$$

with

$$R_1^\varepsilon(T) \coloneqq M(X^\varepsilon)^{-1} \frac{1}{\varepsilon T} \int_0^T V'(X_t^\varepsilon) p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right) \mathrm{d}t,$$
$$R_2^\varepsilon(T) \coloneqq M(X^\varepsilon)^{-1} \frac{\sqrt{2\Sigma}}{T} \int_0^T V'(X_t^\varepsilon) \,\mathrm{d}W_t.$$

For $R_2^\varepsilon(T)$, the same reasoning as in the proof of Theorem 3.12 allows to conclude that $R_2^\varepsilon(T) \to 0$ a.s. for $T \to \infty$, uniformly in ε. It now suffices to show that $R_1^\varepsilon(T)$ vanishes asymptotically. Let us remark that the ergodic theorem yields

$$\lim_{T\to\infty} R_1^\varepsilon(T) = \frac{1}{\varepsilon} \mathcal{M}_\varepsilon^{-1} \, \mathbb{E}^{\mu^\varepsilon}\left[V'(X^\varepsilon) p'\left(\frac{X^\varepsilon}{\varepsilon}\right)\right], \quad \text{a.s.}$$

Substituting the invariant density of X^ε and integrating by parts we obtain

$$\begin{aligned}
\frac{1}{\varepsilon} \mathbb{E}^{\mu^\varepsilon}\left[V'(X^\varepsilon) p'\left(\frac{X^\varepsilon}{\varepsilon}\right)\right] &= \frac{1}{\varepsilon} \int_{\mathbb{R}} V'(x) p'\left(\frac{x}{\varepsilon}\right) \rho^\varepsilon(x) \,\mathrm{d}x \\
&= -\sigma \int_{\mathbb{R}} V'(x) \exp\left(-\frac{1}{\sigma}\alpha \cdot V(x)\right) \frac{\mathrm{d}}{\mathrm{d}x} \exp\left(-\frac{1}{\sigma} p\left(\frac{x}{\varepsilon}\right)\right) \mathrm{d}x \\
&= \sigma \, \mathbb{E}^{\mu^\varepsilon}[V''(X^\varepsilon)] - \mathcal{M}_\varepsilon \alpha,
\end{aligned}$$

so that

$$\lim_{T\to\infty} R_1^\varepsilon(T) = -\alpha + \sigma \mathcal{M}_\varepsilon^{-1} \, \mathbb{E}^{\mu^\varepsilon}[V''(X^\varepsilon)], \quad \text{a.s.}$$

Taking the limit for $\varepsilon \to 0$ and due to Proposition 3.4, we obtain

$$\lim_{\varepsilon\to 0} \lim_{T\to\infty} R_1^\varepsilon(T) = -\alpha + \sigma \mathcal{M}_0^{-1} \, \mathbb{E}^{\mu^0}[V''(X)], \quad \text{a.s.}$$

Substituting the invariant density of X and an integration by parts yield

$$\begin{aligned}\mathbb{E}^{\mu^0}[V''(X)] &= -\int_{\mathbb{R}} V'(x)\frac{\mathrm{d}}{\mathrm{d}x}\exp\left(-\frac{1}{\sigma}\alpha\cdot V(x)\right)\mathrm{d}x \\ &= \frac{1}{\sigma}\mathcal{M}_0\alpha,\end{aligned} \tag{3.34}$$

so that, finally

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} R_1^\varepsilon(T) = -\alpha+\alpha = 0, \quad \text{a.s.},$$

which proves the desired result. □

We remark that the result above can yield to arbitrarily large errors in the solution of the inference problem. Indeed, as demonstrated by Example 3.1, if the diffusion coefficient σ is small then the effective drift A can be sensibly smaller than α. In Chapter 4 we explain how the MLE (3.32) can be corrected by pre-processing the multiscale data in order to correct the negative result of Theorem 3.13. In particular, we briefly introduce the subsampling technique of [112], and then show in detail a filtering methodology, which we introduced in [8], and which is one of the original contributions of this thesis.

3.6 The Bayesian Framework

The problem of inferring the drift coefficient of diffusion processes can be seamlessly recast in a Bayesian framework, which yields a complete uncertainty quantification of the inference procedure (see Chapter 1 for details). In particular, due to the quadratic dependence on A of the log-likelihood $\log L$ (3.24), we can show that the parameter A follows a Gaussian posterior distribution. We refer the reader to our work [8] and to [115, 145] for further details.

Let $T>0$ and let us consider a generic continuous stream of data $Y=(Y_t, 0\le t\le T)$. Moreover, let us fix a Gaussian prior measure $\mu_0=\mathcal{N}(A_0,C_0)$ on the drift coefficient A of (3.2), where the mean $A_0\in\mathbb{R}^N$ and where we assume that C_0 is a non-singular covariance matrix on $\mathbb{R}^N$. Then, we can compute the posterior measure μ_T on A, which by Bayes' theorem satisfies

$$\frac{\mathrm{d}\mu_T}{\mathrm{d}\mu_0}(A\mid Y) = \frac{1}{Z}L(Y\mid A), \qquad Z=\int_{\mathbb{R}^d} L(Y\mid A)\,\mathrm{d}\mu_0(A),$$

where the likelihood $L(Y\mid A)$ is given in (3.24). Computing explicitly the posterior density π_T of μ_T with respect to the Lebesgue measure, one obtains

$$\begin{aligned}\log\pi_T(A\mid Y) = &-\log Z - \frac{T}{2\Sigma}A\cdot v(Y) - \frac{T}{4\Sigma}A\cdot M(Y)A \\ &-\frac{1}{2}(A-A_0)\cdot C_0^{-1}(A-A_0),\end{aligned}$$

where M and v are defined in (3.25). Since the log-posterior density is quadratic in A, the posterior is Gaussian, and it is therefore sufficient to determine its mean and covariance to fully characterize it. Let us denote by by $m_T(Y)$ and $C_T(Y)$ the posterior mean and covariance, respectively. Completing the squares in the log-posterior density, we formally obtain

$$C_T(Y)^{-1} = C_0^{-1}+\frac{T}{2\Sigma}M(Y), \quad m_T(Y) = C_T(Y)\left(C_0^{-1}A_0-\frac{T}{2\Sigma}v(Y)\right). \tag{3.35}$$

Under Assumptions 3.2 and 3.7, one can show that the posterior at time $T>0$ is well defined and is indeed given by $\mu_T(\cdot\mid Y)=\mathcal{N}(m_T(Y),C_T(Y))$.

Let us first state and prove an auxiliary result, which guarantees that the mean and the covariance of the posterior tend respectively to the MLE $\widehat{A}(Y,T)$ and to zero.

Lemma 3.14. *Let $m_T(Y)$ and $C_T(Y)$ be defined in* (3.35)*. Then, under Assumptions 3.2 and 3.7 it holds*

$$\lim_{T\to\infty} \left\| m_T(Y) - \widehat{A}(Y,T) \right\|_2 = 0, \quad \textit{in probability,}$$
$$\lim_{T\to\infty} \|C_T(Y)\|_2 = 0, \qquad\qquad \textit{a.s.,}$$

where $\widehat{A}(Y,T)$ is the MLE defined in (3.25)*.*

Proof. In the proof, we drop for brevity the dependence on Y from the quantities M, v, m_T and C_T. Let us first consider the covariance matrix. An algebraic identity yields

$$C_T = \frac{2\Sigma}{T}\left(M^{-1} - Q^{-1}\right), \qquad Q = M + \frac{T}{2\Sigma} M C_0 M. \tag{3.36}$$

Let us first remark that due to Assumption 3.7 and to the ergodic theorem it holds for all $T > 0$

$$\left\|M^{-1}\right\|_2 \le \frac{1}{\bar{\lambda}}, \quad \text{a.s.}$$

a.s., where $\bar{\lambda}$ is given in Assumption 3.7. We now have that for generic symmetric positive definite matrices R and S it holds

$$\left\|(R+S)^{-1}\right\|_2 \le \left\|S^{-1}\right\|_2.$$

Applying this inequality to Q^{-1}, we obtain

$$\left\|Q^{-1}\right\|_2 \le \frac{2\Sigma}{T}\left\|(MC_0M)^{-1}\right\|_2 \le \frac{2\Sigma}{T}\left\|M^{-1}\right\|_2^2 \left\|C_0^{-1}\right\|_2 \le \frac{2\Sigma}{T\bar{\lambda}^2}\left\|C_0^{-1}\right\|_2, \quad \text{a.s.,}$$

which implies

$$\lim_{T\to\infty} \left\|Q^{-1}\right\|_2 = 0, \quad \text{a.s.}$$

Finally, the triangle inequality and (3.36) yield

$$\lim_{T\to\infty} \|C_T\|_2 = 0, \quad \text{a.s.}$$

We now consider the mean. Replacing the expression of the MLE and due to the Cauchy–Schwarz and triangle inequalities, we obtain

$$\begin{aligned}\left\| m_T - \widehat{A}(Y,T) \right\|_2 &= \frac{2\Sigma}{T}\left\| M^{-1}C_0^{-1}A_0 - Q^{-1}\left(C_0^{-1}A_0 - \frac{T}{2\Sigma}v\right)\right\|_2 \\ &\le \frac{2\Sigma}{T\bar{\lambda}}\left\|C_0^{-1}\right\|_2\left(\|A_0\|_2 + \frac{1}{\bar{\lambda}}\|v\|_2 + \frac{2\Sigma}{T\bar{\lambda}}\left\|C_0^{-1}\right\|_2\|A_0\|_2\right).\end{aligned}$$

Moreover, the ergodic theorem and the weak law of large numbers for martingales (see Section A.5) guarantee that $\|v\|_2$ is bounded in probability for $T \to \infty$. Therefore

$$\lim_{T\to\infty} \left\| m_T - \widehat{A}(Y,T) \right\|_2 = 0, \quad \text{in probability,}$$

which concludes the proof. □

In the following theorem, we show that asymptotically with respect to the final time T the posterior distribution shrinks to the MLE. We characterize the contraction by verifying that the posterior measure concentrates in arbitrarily small balls. Let us finally remark that the measure μ_T is a random measure, and therefore we average the posterior distribution with respect to the measure induced by the Brownian motion W. Let us remark that the choice of the contraction measure and the proof, which we report here for completeness, are inspired by [115, Theorem 5.2]

Theorem 3.15. *Under Assumptions 3.2 and 3.7, the posterior measure $\mu_T(\cdot \mid Y)$ satisfies for all $c > 0$*

$$\lim_{T\to\infty} \mathbb{E}\left[\mu_T\left(\left\{a\colon \left\|a - \widehat{A}(Y,T)\right\|_2 \geq c\right\} \mid Y\right)\right] = 0,$$

where $\mathbb{E}$ denotes expectation with respect to the Brownian motion W and where $\widehat{A}(Y,T)$ is the MLE defined in (3.25).

Proof. The triangle inequality yields

$$\begin{aligned}\mathbb{E}\left[\mu_T\left(\left\{a\colon \left\|a - \widehat{A}(Y,T)\right\|_2 \geq c\right\} \mid Y\right)\right] \leq \mathbb{E}\left[\mu_T\left(\left\{a\colon \|a - m_T\|_2 \geq \frac{c}{2}\right\} \mid Y\right)\right] \\ + \mathbb{P}\left(\left\|m_T - \widehat{A}(Y,T)\right\|_2 \geq \frac{c}{2}\right).\end{aligned}$$

The second term vanishes due to Lemma 3.14. For the first term, Markov's inequality yields

$$\mu_T\left(\left\{a\colon \|a - m_T\|_2 \geq \frac{c}{2}\right\} \mid Y\right) \leq \frac{4}{c^2}\int_{\mathbb{R}^N} \|a - m_T\|_2^2 \, \mu_T(\mathrm{d}a \mid Y),$$

and a change of variable simply gives

$$\int_{\mathbb{R}^N} \|a - m_T\|_2^2 \, \mu_T(\mathrm{d}a \mid Y) = \mathrm{tr}(C_T).$$

Then, again by Lemma 3.14, this implies the desired result. □

We now consider the case of interest of this chapter, i.e., when data are given by the multiscale model (3.1). Theorem 3.15 implies that the Bayesian and the maximum likelihood approaches are in asymptotically equivalent. Moreover, by Theorem 3.13 we know that the MLE is biased when data are given by the multiscale model. Hence, we expect that the in the Bayesian framework, too, the inference procedure fails in this case. This is given by the following result, first appeared in [8].

Theorem 3.16. *Let $T > 0$ and $X^\varepsilon = (X_t^\varepsilon, 0 \leq t \leq T)$ be the solution of* (3.1). *Then, under the assumptions of Theorem 3.15, it holds*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \mathbb{E}\left[\mu_T\left(\{a\colon \|a - \alpha\|_2 \geq c\} \mid X^\varepsilon\right)\right] = 0,$$

where α is the drift coefficient of the multiscale equation (3.1).

Proof. Let us remark that since Lemma 3.14 holds without assumptions on the data, the covariance C_T vanishes independently of ε. Hence, by the proof of Theorem 3.15 it suffices to prove

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \|m_T(X^\varepsilon) - \alpha\|_2 = 0, \quad \text{in probability.}$$

Indeed, the triangle inequality yields

$$\|m_T(X^\varepsilon) - \alpha\|_2 \leq \left\|m_T(X^\varepsilon) - \widehat{A}(X^\varepsilon,T)\right\|_2 + \left\|\alpha - \widehat{A}(X^\varepsilon,T)\right\|_2.$$

The first term vanishes in probability due to Lemma 3.14 and independently of ε, and the second due to Theorem 3.13, which concludes the proof. □

We conclude this section and this chapter by remarking that the result above has the same consequences in the Bayesian setting as Theorem 3.13 has for the MLE. In particular, it shows that the posterior distribution obtained when data is not pre-processed concentrates asymptotically on the drift coefficient of the multiscale equation (3.1). In Chapter 4, we show how the filtering methodology proposed in our work [8] can be employed to successfully retrieve the drift of the homogenized equation in the Bayesian framework.

4 The Filtered Data Approach for Inference of Effective Diffusions

The final results of Chapter 3, in particular Theorems 3.13 and 3.16, show that in both the maximum likelihood and the Bayesian approaches standard techniques fail in the context of data-driven homogenization. In this chapter, we present a methodology based on filtered data which allows to successfully infer the effective drift coefficient given observations deriving from multiscale phenomena. The content of this chapter is based on our research article [8], and is one of the original contributions of this thesis.

The outline of this chapter is as follows. In Section 4.1 we recall from Chapter 3 the problem of interest and detail the methodology of subsampling, first introduced in [112], to obtain unbiased drift estimation of multiscale diffusions. Then, in Section 4.2 we introduce our filtering methodology, and analyze in Section 4.3 the ergodic properties that descend from the definition of the method. In Sections 4.4 and 4.5 we then prove the main results of this chapter, i.e., the unbiasedness of the filtering-based MLE in the homogenized and the multiscale regimes. In Section 4.6 we briefly consider the estimation of the effective diffusion coefficient. We then consider in Section 4.7 how to integrate our scheme in the Bayesian framework which we introduced in Chapter 1 and specialized to the context of inference of multiscale diffusions in Section 3.6. Finally, in Section 4.8 we present a series of numerical experiments which corroborate our theoretical findings and illustrate the potential and the robustness of our new methodology with respect to preexisting techniques.

4.1 The Subsampling Method for Drift Estimation

Let $\varepsilon > 0$ and let us consider the SDE (3.1), which fixing the dimension $d = 1$ for simplicity reads

$$\mathrm{d}X_t^\varepsilon = -\alpha \cdot V'(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right)\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \tag{4.1}$$

where we recall $\alpha \in \mathbb{R}^N$ and $\sigma > 0$ are the drift and diffusion coefficients, respectively, and where $V\colon \mathbb{R} \to \mathbb{R}^N$ and $p\colon \mathbb{R} \to \mathbb{R}$ are the slow and fast components of the confining potential, respectively. We report here the effective equation, which reads

$$\mathrm{d}X_t = -A \cdot V'(X_t)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t, \tag{4.2}$$

where $A = \alpha K$ and $\Sigma = K\sigma$ are the effective drift and diffusion coefficients, whose derivation is shown in detail in Section 3.2. Let us recall that given continuous-time data $Y = (Y_t, 0 \le t \le T)$, we derived in Section 3.4 the MLE for the drift coefficient, which is given by an application of

Girsanov's theorem and is the solution of the linear system

$$\begin{aligned} -M(Y)\widehat{A}(Y,T) &= v(Y), \\ M(Y) &\coloneqq \frac{1}{T}\int_0^T V'(Y_t)\otimes V'(Y_t)\,\mathrm{d}t, \quad v(Y) \coloneqq \frac{1}{T}\int_0^T V'(Y_t)\,\mathrm{d}Y_t. \end{aligned} \tag{4.3}$$

Due to Theorem 3.13, we have that asymptotically the MLE $\widehat{A}(X^\varepsilon, T)$ computed with data from (4.1) is biased, and tends to the drift coefficient of the unhomogenized model α. A technique which is widely employed in practical applications to correct this issue of model misspecification is that of subsampling the data, which has been introduced in [112] and further applied, e.g., in [14, 17, 18, 146]. In particular, let $\delta > 0$ be the subsampling rate, let $T = n\delta$ with n a positive integer and let us consider a discrete sampling $\{X^\varepsilon_{j\delta}\}_{j=0}^n$ of the continuous data X^ε. The drift estimator is then computed as the solution of the linear system

$$\begin{aligned} -M_\delta(X^\varepsilon)\widehat{A}_\delta(X^\varepsilon,T) &= v_\delta(X^\varepsilon), \\ M_\delta(X^\varepsilon) \coloneqq \frac{\delta}{T}\sum_{j=0}^{n-1} V'(X^\varepsilon_{j\delta})\otimes V'(X^\varepsilon_{j\delta}), &\quad v_\delta(X^\varepsilon) \coloneqq \frac{1}{T}\sum_{j=0}^{n-1} V'(X^\varepsilon_{j\delta})\left(X^\varepsilon_{(j+1)\delta} - X^\varepsilon_{j\delta}\right), \end{aligned}$$

which can be seen as an Euler–Maruyama discretization of $\widehat{A}(X^\varepsilon, T)$. The following result, whose statement and original proof are given in [112, Theorem 3.5], shows that by carefully choosing the subsampling rate the estimator $\widehat{A}_\delta(X^\varepsilon, T)$ is asymptotically unbiased with respect to the effective drift coefficient A.

Theorem 4.1. *Under Assumptions 3.2 and 3.7, let $0 < \zeta < 1$ and $\gamma > \zeta$. If $\delta = \varepsilon^\zeta$ and $n = \lceil \varepsilon^{-\gamma} \rceil$, then*

$$\lim_{\varepsilon\to 0} \widehat{A}_\delta(X^\varepsilon, T) = A, \quad \textit{in probability},$$

where A is the drift coefficient of (4.2).

This unbiasedness result is crucial, and shows that subsampling the data allows, asymptotically, to obtain estimators for the effective drift coefficient which are consistent with the theory of homogenization. Nevertheless, there are three main drawbacks to subsampling:

1. The scale separation parameter ε has to be known in advance in order to tune the subsampling width δ coherently with Theorem 4.1. In practical applications, the parameter ε is unknown;
2. In finite computations, i.e., for $\delta < 0$ and $T < \infty$, numerical experiments demonstrate that the estimator $\widehat{A}_\delta(X^\varepsilon, T)$ is not robust with respect to the filtering width $\delta = \varepsilon^\zeta$. An optimal value of $\zeta = 2/3$ is suggested in [112], but in applications even following this suggestion does not always result on reliable results;
3. In practice, if $\varepsilon \ll 1$ and $\delta = \varepsilon^\zeta$, then subsampling leads to disregarding a high percentage of the data, which sometimes could be as high as 99%. First, this leads to an inevitable increase of the variance of the estimator with respect to data. Second, in modern applications accumulating and exploiting a large amount of data is fundamental, and disregarding information seems unreasonable, or even unjustifiable.

The novel methodology based on filtered data we present in this chapter allows to bypass each of these three issues. Moreover, our novel approach can be naturally employed in the Bayesian framework of Section 3.6, and thus correct the faulty behavior highlighted by Theorem 3.16.

4.2 The Filtered Data Approach

We now introduce our methodology based on filtered data to address the issue that the MLE estimator, when confronted with multiscale data, is biased. Let $\beta, \delta > 0$ and let us consider a family of exponential kernel functions $k\colon \mathbb{R}^+ \to \mathbb{R}$ defined as

$$k(r) = C_\beta \delta^{-1/\beta} e^{-r^\beta/\delta}, \quad C_\beta = \beta\,\Gamma(1/\beta)^{-1}, \tag{4.4}$$

where C_β is chosen so that the kernel is normalized, in the sense

$$\int_0^\infty k(r)\,\mathrm{d}r = 1,$$

and where $\Gamma(\cdot)$ is the gamma function. We then consider the filtered process $Z^\varepsilon \coloneqq (Z_t^\varepsilon, 0 \le t \le T)$ defined by the truncated convolution

$$Z_t^\varepsilon \coloneqq \int_0^t k(t-s) X_s^\varepsilon\,\mathrm{d}s.$$

The process Z^ε can be interpreted as a smoothed version of the original trajectory X^ε. In fact, in the field of signal processing the kernel (4.4) belongs to the class of low-pass linear time-invariant filters, which cut the high frequencies in a signal to highlight its slowest components. In the following, rigorous analysis is conducted only when $\beta = 1$. Nonetheless, numerical experiments show that for higher values of β the performances of estimators computed employing the filter are more robust and qualitatively better.

Remark 4.2. Given a trajectory X^ε, it is relatively inexpensive to compute Z^ε from a computational standpoint. In particular, the process Z^ε is the truncated convolution of the kernel with the process X^ε. Hence, computational tools based on the Fast Fourier Transform (FFT) exist and allow to compute Z^ε fast component-wise. Moreover, the process Z^ε can be computed, in case $\beta = 1$, in a recursive manner and therefore "online".

Given a trajectory X^ε and the filtered data Z^ε, the estimator of the drift coefficient we propose is given by the solution of the linear system

$$\begin{aligned} &-\widetilde{M}^{-1}(X^\varepsilon)\widehat{A}_k(X^\varepsilon, T) = \widetilde{v}(X^\varepsilon) \\ &\widetilde{M}(X^\varepsilon) = \frac{1}{T}\int_0^T V'(Z_t^\varepsilon) \otimes V'(X_t^\varepsilon)\,\mathrm{d}t, \quad \widetilde{v}(X^\varepsilon) = \frac{1}{T}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}X_t^\varepsilon. \end{aligned} \tag{4.5}$$

For economy of notation we drop explicit reference to the dependence of $\widetilde{M}$ and $\widetilde{v}$ on X^ε. Let us remark that the formula above is obtained by replacing in the matrix and in the right-hand side only one instance of X_t^ε with Z_t^ε. In particular, it is fundamental for proving unbiasedness to keep in the definition of $\widetilde{v}$ the differential of the original process $\mathrm{d}X_t^\varepsilon$ (see Remark 4.10). Let us furthermore remark that $\widehat{A}_k(X^\varepsilon, T)$ need not be the minimizer of some likelihood function based on filtered data, but just as a perturbation of $\widehat{A}(X^\varepsilon, T)$. Indeed, likewise the subsampling estimator $\widehat{A}_\delta(X^\varepsilon, T)$, we work directly at the level of estimators and after the likelihood maximization. The only theoretical guarantee which is still needed for the well-posedness of $\widehat{A}_k(X^\varepsilon, T)$ is for $\widetilde{M}$ to be invertible.

Given this considerations, we now set the boundaries of our analysis to the dissipative case which is considered in Chapter 3, too. In particular, the assumption below is repeatedly evoked in the rest of the chapter.

Assumption 4.3. Assumptions 3.2 and 3.7 hold, and the matrix $\widetilde{M}$ defined in (4.5) is invertible.

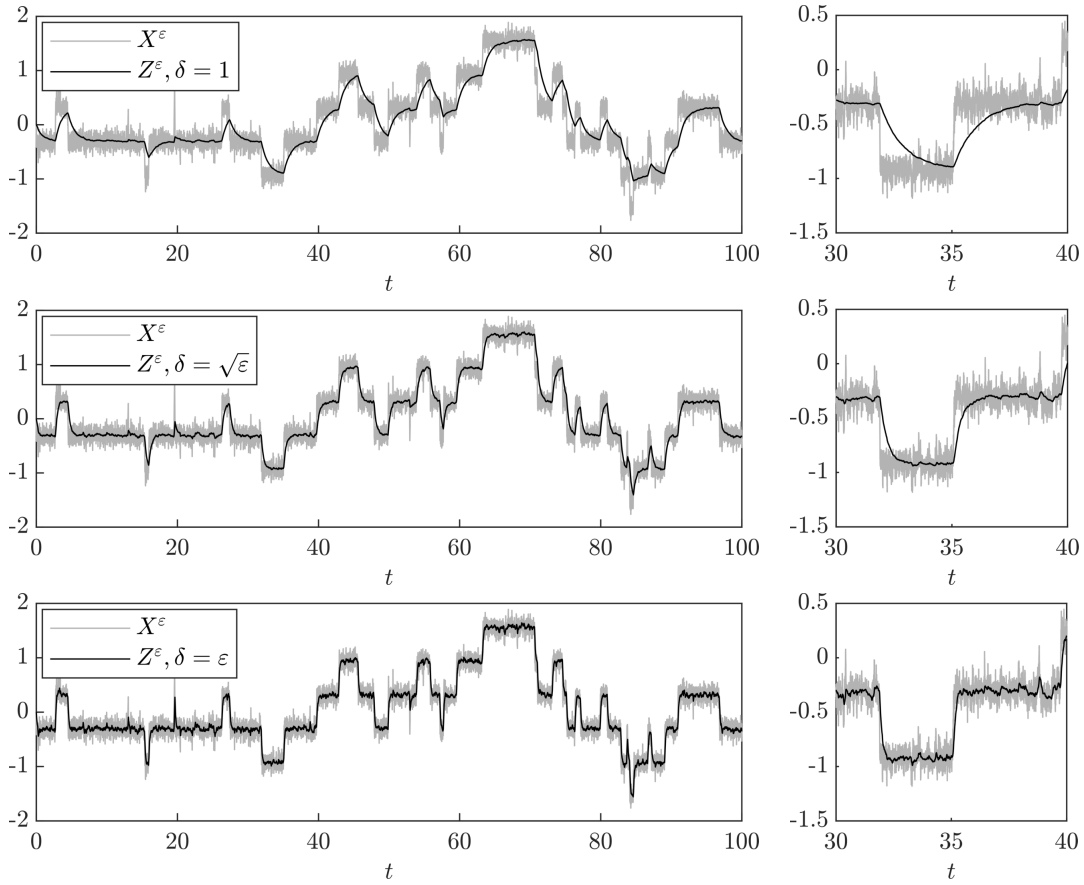


Figure 4.1 – Filtering a trajectory X^ε obtained with $V(x) = x^2/2$, $p(y) = \cos(y)$, $\alpha = 1$, $\sigma = 0.5$ and $\varepsilon = 0.1$. The filtering width is $\delta = \{1, \sqrt{\varepsilon}, \varepsilon\}$ from top to bottom, respectively, and $\beta = 1$.

Remark 4.4. In the following, in particular in the proof of Lemma 4.6, we will employ Assumption 3.2(ii) for the whole drift of the SDE (4.1), i.e., the function

$$V^\varepsilon(x) \coloneqq \alpha \cdot V(x) + p\left(\frac{x}{\varepsilon}\right).$$

Since $p \in C^\infty(\mathbb{R})$ and is periodic, all derivatives of p are in $L^\infty(\mathbb{R})$. Therefore, the assumption above is sufficient for V^ε to satisfy Assumption 3.2(ii) with different values for a and b. In particular, assume Assumption 3.2(ii) holds for V. Then, we have for all $\gamma > 0$ by Young's inequality

$$\begin{aligned} -(V^\varepsilon)'(x)x &\leq a - bx^2 - \frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)x \\ &\leq \left(a + \frac{1}{2\varepsilon^2\gamma}\left\|p'\right\|^2_{L^\infty(\mathbb{R})}\right) - \left(b - \frac{\gamma}{2}\right)x^2. \end{aligned}$$

Hence, Assumption 3.2(ii) holds for V^ε with a coefficient b which is arbitrarily close to the coefficient for V, alone.

Let us from now on consider $\beta = 1$. For this value of β, the parameter δ appearing in (4.4) regulates the width of the filtering window. In practice, larger values of δ will lead to trajectories which are smoother and for which fast-scale oscillations are practically canceled. Let us remark that the filtering width resembles the subsampling step employed for the estimator $\widehat{A}_\delta(X^\varepsilon, T)$

introduced and analyzed in [112]. For subsampling, the choice guaranteeing asymptotically unbiased results is $\delta = \varepsilon^\zeta$ with $\zeta \in (0,1)$, and a similar analysis is due for our technique. For visualization purposes, we depict in Figure 4.1 the filtered trajectory Z^ε for three different values of δ, namely $\delta = \{1, \sqrt{\varepsilon}, \varepsilon\}$. With $\delta = 1$, all oscillations at the fast scale are canceled and the filtered trajectory Z^ε presents only slow-scale variations. Reducing the value of δ, fast-scale oscillations are progressively taken into account.

In the following, we first focus on the ergodic properties of the process Z^ε when it is coupled with the process X^ε. This analysis is practically independent of the choice of δ, and is therefore presented on its own. Then, we focus on two different cases which depend on the choice of the width δ of the filter. First, in Section 4.4, we consider δ to be independent of ε, and therefore we filter at the speed of the homogenized process. In this case, we are able to prove that our estimator of the drift coefficient of the homogenized equation is asymptotically unbiased almost surely. This result will be presented in Theorem 4.15. We then move on in Section 4.5 to the case $\delta \propto \varepsilon^\zeta$, which corresponds to filtering the data at the speed of the multiscale process. In this case, we show that under some conditions on the exponent ζ, we can still obtain estimators which are asymptotically unbiased in probability. This result is proved in Theorem 4.21. For this second case, we widely employ techniques and estimates from [112].

4.3 Ergodic Properties

Let us consider the filtering kernel (4.4) with $\beta = 1$, i.e.,

$$k(r) = \frac{1}{\delta} e^{-r/\delta}.$$

In this case, Leibniz integral rule yields the equality

$$\mathrm{d}Z_t^\varepsilon = k(0) X_t^\varepsilon \,\mathrm{d}t + \int_0^t k'(t-s) X_s^\varepsilon \,\mathrm{d}s \,\mathrm{d}t = \frac{1}{\delta} \left(X_t^\varepsilon - Z_t^\varepsilon \right) \mathrm{d}t,$$

which can be interpreted as an ordinary differential equation for Z_t^ε driven by the stochastic signal X^ε. Considering the processes X^ε and Z^ε together, we obtain the system of two one-dimensional SDEs

$$\begin{aligned} \mathrm{d}X_t^\varepsilon &= -\alpha \cdot V'(X_t^\varepsilon) \,\mathrm{d}t - \frac{1}{\varepsilon} p'\left(\frac{X_t^\varepsilon}{\varepsilon} \right) \mathrm{d}t + \sqrt{2\sigma} \,\mathrm{d}W_t, \\ \mathrm{d}Z_t^\varepsilon &= \frac{1}{\delta} \left(X_t^\varepsilon - Z_t^\varepsilon \right) \mathrm{d}t. \end{aligned} \tag{4.6}$$

The first ingredient for studying the ergodic properties of the two-dimensional process $(X^\varepsilon, Z^\varepsilon)^\top := ((X_t^\varepsilon, Z_t^\varepsilon)^\top, 0 \le t \le T)$ is verifying that the measure induced by the stochastic process admits a smooth density with respect to the Lebesgue measure. Since noise is present only on the first component, this is a consequence of the theory of hypo-ellipticity, as summarized in the following Lemma, whose proof is given in Section 4.9.1.

Lemma 4.5. *Let $(X^\varepsilon, Z^\varepsilon)^\top$ be the solution of (4.6) and let $\widetilde{\mu}_t^\varepsilon$ be the measure induced by the joint process at time t. Then, the measure $\widetilde{\mu}_t^\varepsilon$ admits a smooth density $\widetilde{\rho}_t^\varepsilon$ with respect to the Lebesgue measure.*

Once it is established that the law of the process admits a smooth density for all times $t > 0$, which satisfies a time-dependent Fokker–Planck equation (see Section A.3), we are interested in the limiting properties of this law. In particular, Proposition 3.3 guarantees that the process X^ε is geometrically ergodic (see Section A.4), and we wish the couple $(X^\varepsilon, Z^\varepsilon)^\top$ to inherit the same property. The following Lemma guarantees that the couple is indeed geometrically ergodic, and its proof is given in Section 4.9.1.

Lemma 4.6. *Let Assumption 4.3 hold and let $b > 0$ be given in Assumption 3.2(ii). Then, if $\delta > 1/(4b)$, the process $(X^\varepsilon, Z^\varepsilon)^\top$ solution of (4.6) is geometrically ergodic. Moreover, denoting by $\widetilde{\mu}^\varepsilon$ the invariant measure of the couple $(X^\varepsilon, Z^\varepsilon)^\top$, its density with respect to the Lebesgue measure $\widetilde{\rho}^\varepsilon$ is the solution to the stationary Fokker–Planck equation*

$$\sigma\partial^2_{xx}\widetilde{\rho}^\varepsilon(x,z) + \partial_x\left(\left(\alpha\cdot V'(x) + \frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)\right)\widetilde{\rho}^\varepsilon(x,z)\right) + \frac{1}{\delta}\partial_z\left((z-x)\widetilde{\rho}^\varepsilon(x,z)\right) = 0. \tag{4.7}$$

Remark 4.7. The condition $\delta > 1/(4b)$ is not very restrictive. Let the parameter dimension $N = 1$ and let $V(x) \propto x^{2r}$ for an integer $r > 1$. Then, Assumption 3.2(ii) holds for an arbitrarily large $b > 0$. Therefore, the parameter of the filter δ can be chosen along the entire positive real axis. A similar argument can be employed for higher dimensions $N > 1$.

In a general case, it is not possible to find an explicit solution to (4.7). Nevertheless, it is possible to show some relevant properties of the solution itself, which are summarized in the following Lemma, whose proof is given in Section 4.9.1.

Lemma 4.8. *Under the assumptions of Lemma 4.6, let $\widetilde{\rho}^\varepsilon$ be the solution of (4.7) and let us write*

$$\widetilde{\rho}^\varepsilon(x,z) = \rho^\varepsilon(x)\psi^\varepsilon(z)R^\varepsilon(x,z), \tag{4.8}$$

where ρ^ε and ψ^ε are the marginal densities of X^ε and Z^ε respectively, i.e.,

$$\rho^\varepsilon(x) = \int_{\mathbb{R}} \widetilde{\rho}^\varepsilon(x,z)\,\mathrm{d}z, \quad \psi^\varepsilon(z) = \int_{\mathbb{R}} \widetilde{\rho}^\varepsilon(x,z)\,\mathrm{d}x.$$

Then, it holds

$$\sigma\delta\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z = \mathbb{E}^{\widetilde{\mu}^\varepsilon}[(X^\varepsilon - Z^\varepsilon)^2 V''(Z^\varepsilon)]. \tag{4.9}$$

Remark 4.9. Lemma 4.8, and in particular the equality (4.9), plays a fundamental role in the proof of unbiasedness of the estimator based on filtered data. In particular, this equality allows to bypass the explicit knowledge of the function $R^\varepsilon(x,z)$, which governs the correlation between the processes X^ε and Z^ε at stationarity, for which a closed-form expression is not available in the general case. Moreover, let us remark that the marginal invariant density ρ^ε of X^ε is known explicitly by Proposition 3.3.

Remark 4.10. Let us return to the definition of $\widehat{A}_k$ and replace the differential $\mathrm{d}X^\varepsilon_t$ with $\mathrm{d}Z^\varepsilon_t$ in $\widetilde{v}$. In this case, it holds

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T V'(Z^\varepsilon_t)\,\mathrm{d}Z^\varepsilon_t = \lim_{T\to\infty}\frac{1}{\delta T}\int_0^T V'(Z^\varepsilon_t)(X^\varepsilon_t - Z^\varepsilon_t)\,\mathrm{d}t = \frac{1}{\delta}\,\mathbb{E}^{\mu^\varepsilon}\left[V'(Z^\varepsilon)(X^\varepsilon - Z^\varepsilon)\right] = 0,$$

where the last equality is obtained as in the proof of Lemma 4.8, with the choice $f(x,z) = V(z)$ at the last line. Therefore, we stress again that it is indeed necessary to employ the original differential $\mathrm{d}X^\varepsilon_t$ in the vector $\widetilde{v}$ in the definition (4.5) of $\widehat{A}^\varepsilon_k$.

Remark 4.11. Let us consider the kernel (4.4) with $\beta > 1$. In this case, the steps leading to the system (4.6) do not yield a system of Itô SDEs, but of stochastic delay differential equations. The analysis of the estimator in case $\beta > 1$ is therefore based on different arguments than the one we present in this work.

4.4 Filtered Data in the Homogenized Regime

In this section, we analyze the behavior of the estimator $\widehat{A}_k(X^\varepsilon, T)$ based on filtered data given in (4.5) when the filtering width δ is independent of ε. The analysis in this case is based on

the convergence of the couple $(X^\varepsilon, Z^\varepsilon)^\top$ with respect to the multiscale parameter $\varepsilon \to 0$. In particular, Proposition 3.4 guarantees that the invariant measure of X^ε converges weakly to the invariant measure of X, the solution of the homogenized equation (4.2). The following result guarantees the same kind of convergence for the couple $(X^\varepsilon, Z^\varepsilon)^\top$.

Lemma 4.12. *Under Assumption 4.3, let $\widetilde{\mu}^\varepsilon$ be the invariant measure of the couple $(X^\varepsilon, Z^\varepsilon)^\top$. If δ is independent of ε, then the measure $\widetilde{\mu}^\varepsilon$ converges weakly to the measure $\widetilde{\mu}^0(\mathrm{d}x, \mathrm{d}z) = \widetilde{\rho}^0(x,z)\,\mathrm{d}x\,\mathrm{d}z$, whose density $\widetilde{\rho}^0$ is the unique solution of the stationary Fokker–Planck equation*

$$\Sigma\partial^2_{xx}\widetilde{\rho}^0(x,z) + \partial_x\left(A\cdot V'(x)\widetilde{\rho}^0(x,z)\right) + \frac{1}{\delta}\partial_z\left((z-x)\widetilde{\rho}^0(x,z)\right) = 0, \tag{4.10}$$

where A and Σ are the coefficients of the homogenized equation (4.2).

Proof. Let $(X,Z)^\top := \left((X_t, Z_t)^\top, 0 \le t \le T\right)$ be the solution of

$$\begin{aligned}\mathrm{d}X_t &= -A\cdot V'(X_t)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t,\\ \mathrm{d}Z_t &= \frac{1}{\delta}\left(X_t - Z_t\right)\,\mathrm{d}t,\end{aligned}$$

with $(X_0, Z_0)^\top \sim \widetilde{\mu}^0$. The arguments of Section 4.3 can be repeated to conclude that the invariant measure of $(X,Z)^\top$ admits a smooth density $\widetilde{\rho}^0$ which satisfies (4.10). Moreover, the proof of Theorem 3.5 can be readily adapted for the couple $(X^\varepsilon, Z^\varepsilon)^\top$ to show that $(X^\varepsilon, Z^\varepsilon)^\top \Rightarrow (X,Z)^\top$ in $\mathcal{C}^0([0,T];\mathbb{R}^2)$, provided that $(X_0^\varepsilon, Z_0^\varepsilon)^\top \sim \widetilde{\mu}^\varepsilon$. This implies that $\widetilde{\mu}^\varepsilon \Rightarrow \widetilde{\mu}^0$ for $\varepsilon \to 0$, and therefore concludes the proof. □

Example 4.13. A closed form solution of (4.10) can be obtained in a simple case. Let the dimension of the parameter $N = 1$ and let $V(x) = x^2/2$. Then, the analytical solution is given by

$$\widetilde{\rho}^0(x,z) = \frac{1}{C_0}\exp\left(-\frac{A}{\Sigma}\frac{x^2}{2} - \frac{1}{\delta\Sigma}\frac{(x-(1+A\delta)z)^2}{2}\right),$$

where

$$C_0 = \int_{\mathbb{R}}\int_{\mathbb{R}}\exp\left(-\frac{A}{\Sigma}\frac{x^2}{2} - \frac{1}{\delta\Sigma}\frac{(x-(1+A\delta)z)^2}{2}\right)\mathrm{d}x\,\mathrm{d}z = \frac{2\pi\Sigma\sqrt{\delta}}{(1+A\delta)\sqrt{A}}.$$

This is the density of a multivariate normal distribution $\mathcal{N}(0,\Gamma)$, where the covariance matrix is given by

$$\Gamma = \frac{\Sigma}{A(1+A\delta)}\begin{pmatrix}1+A\delta & 1\\ 1 & 1\end{pmatrix}.$$

Let us remark that this distribution can be obtained from direct computations involving Gaussian processes. In particular, we have that X is in this case an Ornstein–Uhlenbeck process and it is therefore known that $X \sim \mathcal{GP}(m_t, \mathcal{C}(t,s))$, where at stationarity $m_t = 0$ and

$$\mathcal{C}(t,s) = \frac{\Sigma}{A}e^{-A|t-s|}.$$

The basic properties of Gaussian processes imply that Z is a Gaussian process, and that the couple $(X,Z)^\top$ is a Gaussian process, too, whose mean and covariance are computable explicitly.

We now present an analogous result to Lemma 4.8 for the limit distribution.

Corollary 4.14. *Let $\widetilde{\rho}^0$ be the solution of* (4.10) *and let us write*

$$\widetilde{\rho}^0(x,z) = \rho^0(x)\psi^0(z)R^0(x,z),$$

where ρ^0 and ψ^0 are the marginal densities, i.e.,

$$\rho^0(x) = \int_{\mathbb{R}} \widetilde{\rho}^0(x,z)\,\mathrm{d}z, \quad \psi^0(z) = \int_{\mathbb{R}} \widetilde{\rho}^0(x,z)\,\mathrm{d}x.$$

Then, if A and Σ are the coefficients of the homogenized equation (4.2)*, it holds*

$$\Sigma\delta \int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^0(x)\psi^0(z)\partial_x R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z = \mathbb{E}^{\widetilde{\mu}^0}[(X-Z)^2 V''(Z)].$$

Proof. The proof is directly obtained from Lemma 4.8 setting $p(y) = 0$ and replacing α, σ by A, Σ respectively. □

Let us recall the notation

$$\mathcal{M}_\varepsilon \coloneqq \mathbb{E}^{\mu^\varepsilon}[V'(X^\varepsilon)\otimes V'(X^\varepsilon)], \quad \mathcal{M}_0 \coloneqq \mathbb{E}^{\mu^0}[V'(X)\otimes V'(X)], \tag{4.11}$$

which we introduced in (3.33), and let us moreover denote by

$$\widetilde{\mathcal{M}}_\varepsilon \coloneqq \mathbb{E}^{\widetilde{\mu}^\varepsilon}[V'(Z^\varepsilon)\otimes V'(X^\varepsilon)], \quad \widetilde{\mathcal{M}}_0 \coloneqq \mathbb{E}^{\widetilde{\mu}^0}[V'(Z)\otimes V'(X)], \tag{4.12}$$

the equivalent quantities in the context of the estimator (4.5). We can now state and prove the main result of this section, namely the convergence of the estimator based on filtered data of the drift coefficient of the homogenized equation.

Theorem 4.15. *Under Assumption 4.3, let $\widehat{A}_k(X^\varepsilon, T)$ be defined in* (4.5) *with δ independent of ε. Then*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widehat{A}_k(X^\varepsilon, T) = A, \quad \textit{a.s.},$$

where A is the drift coefficient of the homogenized equation (4.1)*.*

Proof. Replacing the expression of $\mathrm{d}X_t^\varepsilon$ into (4.5), we get for $\widetilde{v}$

$$\widetilde{v} = -\widetilde{M}\alpha - \frac{1}{T}\int_0^T \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right) V'(Z_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2\sigma}}{T}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}W_t.$$

Therefore, we have

$$\begin{aligned} &\widehat{A}_k(X^\varepsilon, T) = \alpha + I_1^\varepsilon(T) - I_2^\varepsilon(T), \\ &I_1^\varepsilon(T) \coloneqq \frac{1}{T}\widetilde{M}^{-1}\int_0^T \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right) V'(Z_t^\varepsilon)\,\mathrm{d}t, \quad I_2^\varepsilon(T) \coloneqq \frac{\sqrt{2\sigma}}{T}\widetilde{M}^{-1}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}W_t \end{aligned} \tag{4.13}$$

We study the terms $I_1^\varepsilon(T)$ and $I_2^\varepsilon(T)$ separately. First, the ergodic theorem (Theorem A.14) applied to $I_1^\varepsilon(T)$ yields

$$\lim_{T\to\infty} I_1^\varepsilon(T) = \widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right) V'(Z^\varepsilon)\right], \quad \text{a.s.} \tag{4.14}$$

Replacing the decomposition (4.8), the expression of the density ρ^ε of the marginal measure of X^ε and integrating by parts, we have

$$\begin{aligned} \mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right) V'(Z^\varepsilon)\right] &= -\frac{\sigma}{C_\varepsilon}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-\frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)}\right) e^{-\frac{1}{\sigma}\alpha\cdot V(x)} V'(z)\psi^\varepsilon(z) R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \frac{\sigma}{C_\varepsilon}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)}\partial_x\left(e^{-\frac{1}{\sigma}\alpha\cdot V(x)} R^\varepsilon(x,z)\right) V'(z)\psi^\varepsilon(z)\,\mathrm{d}x\,\mathrm{d}z, \end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(Z^\varepsilon)\right] &= -\left(\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\otimes V'(x)\widetilde{\rho}^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z\right)\alpha \\ &\quad + \sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= -\widetilde{\mathcal{M}}_\varepsilon\alpha + \sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z.\end{aligned}$$

Replacing the equality above into (4.14), we obtain

$$\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \widetilde{\mathcal{M}}_\varepsilon^{-1}\sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z, \quad \text{a.s.}$$

We now remark that the right-hand side is exactly the quantity appearing in Lemma 4.8. Therefore, we have

$$\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \frac{1}{\delta}\widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}[(X^\varepsilon - Z^\varepsilon)^2 V''(Z^\varepsilon)], \quad \text{a.s.} \tag{4.15}$$

Since δ is independent of ε, we can pass to the limit as ε goes to zero and Lemma 4.12 yields

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \frac{1}{\delta}\widetilde{\mathcal{M}}_0^{-1}\,\mathbb{E}^{\widetilde{\mu}^0}[(X - Z)^2 V''(Z)], \quad \text{a.s.} \tag{4.16}$$

Due to Corollary 4.14, we have

$$\frac{1}{\delta}\,\mathbb{E}^{\widetilde{\mu}^0}[(X - Z)^2 V''(Z)] = \Sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\rho^0(x)\psi^0(z)\partial_x R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z,$$

and moreover, an integration by parts yields

$$\begin{aligned}\frac{1}{\delta}\,\mathbb{E}^{\widetilde{\mu}^0}[(X - Z)^2 V''(Z)] &= -\Sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)(\rho^0)'(x)\psi^0(z)R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= -\frac{\Sigma}{C_0}\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-\frac{1}{\Sigma}A\cdot V(x)}\right)\psi^0(z)R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \left(\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\otimes V'(x)\widetilde{\rho}^0(x,z)\,\mathrm{d}x\,\mathrm{d}z\right)A \\ &= \widetilde{\mathcal{M}}_0 A.\end{aligned}$$

We can therefore conclude that

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + A, \quad \text{a.s.} \tag{4.17}$$

We now consider the second term $I_2^\varepsilon(T)$, and rewrite it as

$$I_2^\varepsilon(T) = \sqrt{2\sigma}I_{2,1}^\varepsilon(T)I_{2,2}^\varepsilon(T),$$

where

$$\begin{aligned}I_{2,1}^\varepsilon(T) &\coloneqq \left(\frac{1}{T}\int_0^T V'(Z_t^\varepsilon)\otimes V'(X_t^\varepsilon)\,\mathrm{d}t\right)^{-1}\left(\frac{1}{T}\int_0^T V'(Z_t^\varepsilon)\otimes V'(Z_t^\varepsilon)\,\mathrm{d}t\right), \\ I_{2,2}^\varepsilon(T) &\coloneqq \left(\frac{1}{T}\int_0^T V'(Z_t^\varepsilon)\otimes V'(Z_t^\varepsilon)\,\mathrm{d}t\right)^{-1}\left(\frac{1}{T}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}W_t\right).\end{aligned}$$

The ergodic theorem yields

$$\lim_{T\to\infty} I_{2,1}^\varepsilon(T) = \widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[V'(Z^\varepsilon)\otimes V'(Z^\varepsilon)\right] =: \gamma^\varepsilon,$$

where γ^ε is bounded uniformly in ε due to the theory of homogenization, Assumption 3.2(iii)-3.7 and Lemma 4.30. Moreover, always due to Lemma 4.30 and Assumption 3.2(iii) we have that $V'(Z^\varepsilon)$ is square integrable, and hence [116, Lemma 6.1] yields

$$\lim_{T\to\infty} I^\varepsilon_{2,2}(T) = 0, \quad \text{a.s.},$$

independently of ε. Therefore

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I^\varepsilon_2(T) = 0, \quad \text{a.s.},$$

which, together with (4.17) and (4.13), proves the desired result. □

Remark 4.16. Let us remark that the assumption that δ is independent of ε is necessary to pass from (4.15) to (4.16) but is not needed before (4.15). Moreover, the term $I^\varepsilon_2(t)$ in the proof vanishes a.s. independently of ε. Therefore, in the analysis of the case $\delta = \mathcal{O}(\varepsilon^\zeta)$ it will be sufficient for unbiasedness to show that

$$\lim_{\varepsilon\to 0} \frac{1}{\delta}\widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}[(X^\varepsilon - Z^\varepsilon)^2 V''(Z^\varepsilon)] = A,$$

which is a non-trivial limit since $\delta \to 0$ for $\varepsilon \to 0$.

4.5 Filtered Data in the Multiscale Regime

We now consider the case of the filtering width $\delta = \mathcal{O}(\varepsilon^\zeta)$, where $\zeta > 0$ will be specified in the following. In this case, the filtered process resembles more the original process X^ε, as can be noted in Figure 4.1. Moreover, the techniques employed for proving Theorem 4.15 can only be partly exploited, as highlighted by Remark 4.16. In fact, in order to prove unbiasedness it is necessary to characterize precisely the difference between the processes Z^ε and X^ε. A first characterization is given by the following Proposition, whose proof can be found in Section 4.9.2.

Proposition 4.17. *Let Assumption 4.3 hold and $\varepsilon, \delta > 0$ be sufficiently small. Then, it holds for every $t > 0$*

$$X^\varepsilon_t - Z^\varepsilon_t = \delta B^\varepsilon_t + R(\varepsilon, \delta),$$

where the stochastic process B^ε_t is defined as

$$B^\varepsilon_t \coloneqq \sqrt{2\sigma}\int_0^t k(t-s)(1+\Phi'(Y^\varepsilon_s))\,\mathrm{d}W_s, \tag{4.18}$$

where Φ is the solution of the cell problem (3.4)*, W_s is the Brownian motion appearing in* (4.1) *and $Y^\varepsilon_t = X^\varepsilon_t/\varepsilon$. Moreover, B^ε_t and the remainder $R(\varepsilon,\delta)$ satisfy for every $p \geq 1$ the estimates*

$$\mathbb{E}^{\mu^\varepsilon}\left[|B^\varepsilon_t|^p\right]^{1/p} \leq C\delta^{-1/2}, \tag{4.19}$$

and

$$\left(\mathbb{E}^{\varphi^\varepsilon}\,|R(\varepsilon,\delta)|^p\right)^{1/p} \leq C\left(\delta + \varepsilon + \max\{1,t\}e^{-t/\delta}\right), \tag{4.20}$$

where C is independent of ε, δ and t, and μ^ε is the invariant measure of X^ε.

It is clear from the Proposition above that understanding the properties of the process B^ε_t is key to understanding the behavior of the difference between X^ε and Z^ε. In particular, we can write the dynamics of B^ε_t with an application of the Itô formula (or the stochastic version of Leibniz integral rule) and due to the properties of the smoothing kernel k as

$$\mathrm{d}B^\varepsilon_t = -\frac{1}{\delta}B^\varepsilon_t\,\mathrm{d}t + \frac{\sqrt{2\sigma}}{\delta}(1+\Phi'(Y^\varepsilon_t))\,\mathrm{d}W_t.$$

This equation can be coupled with the dynamics of the processes X_t^ε, Y_t^ε and Z_t^ε, thus describing the evolution of the quadruple $(X^\varepsilon, Y^\varepsilon, Z^\varepsilon, B^\varepsilon)$ together. In particular, it is possible to show that the results of Section 4.3 hold for the quadruple, and the properties of the invariant measure of the quadruple can be exploited to prove the unbiasedness of the estimator in the case $\delta = \mathcal{O}(\varepsilon^\zeta)$ in the same way as in the case δ independent of ε. In this context, a further assumption on the potential V is necessary.

Assumption 4.18. The derivatives V'' and V''' of the potential $V\colon \mathbb{R} \to \mathbb{R}^N$ are component-wise polynomially bounded, and the second derivative is Lipschitz, i.e., there exists a constant $L > 0$ such that

$$\|V''(x) - V''(y)\| \leq L\,|x - y|\,,$$

for all $x, y \in \mathbb{R}$.

In light of Remark 4.16, it is fundamental to understand the behavior of the quantity

$$\frac{1}{\delta}(X_t^\varepsilon - Z_t^\varepsilon)^2 V''(Z_t^\varepsilon),$$

as well as its limit for $t \to \infty$ and for $\varepsilon \to 0$. Let us remark that due to Proposition 4.17 we have

$$\frac{1}{\delta}(X_t^\varepsilon - Z_t^\varepsilon)^2 V''(Z_t^\varepsilon) \approx \delta (B_t^\varepsilon)^2 V''(Z_t^\varepsilon),$$

and therefore studying the right hand side of the approximate equality above is the goal of the upcoming discussion. The following result, whose proof is in Section 4.9.3, gives a first characterization.

Lemma 4.19. *Under Assumptions 4.3 and 4.18, let η^ε be the invariant measure of the quadruple $(X^\varepsilon, Y^\varepsilon, Z^\varepsilon, B^\varepsilon)$. Then it holds*

$$\delta\, \mathbb{E}^{\eta^\varepsilon}\left[(B^\varepsilon)^2 V''(Z^\varepsilon)\right] = \sigma\, \mathbb{E}^{\eta^\varepsilon}[(1 + \Phi'(Y^\varepsilon))^2 V''(Z^\varepsilon)] + \widetilde{R}(\varepsilon, \delta),$$

where the remainder $\widetilde{R}(\varepsilon, \delta)$ satisfies

$$\left|\widetilde{R}(\varepsilon, \delta)\right| \leq C\left(\delta^{1/2} + \varepsilon\right).$$

Let us remark that the quantity appearing above hints towards the theory of homogenization. In fact, we recall from Chapter 3 that the homogenization coefficient K is given by

$$K = \int_0^L (1 + \Phi'(y))^2\, \mu(\mathrm{d}y),$$

where μ is the marginal measure of the process Y^ε when coupled with X^ε. Therefore, the next step is the homogenization limit, i.e., the limit of vanishing ε, which is considered in the following Lemma, and whose proof is given in Section 4.9.3.

Lemma 4.20. *Let the assumptions of Lemma 4.19 hold, and let $\delta = \varepsilon^\zeta$ with $\zeta > 0$. Then, it holds*

$$\lim_{\varepsilon \to 0} \sigma\, \mathbb{E}^{\eta^\varepsilon}[(1 + \Phi'(Y^\varepsilon))^2 V''(Z^\varepsilon)] = \mathcal{M}_0 A,$$

where A is the drift coefficient of the homogenized equation (4.2)*, and $\mathcal{M}_0$ is the matrix defined in* (4.11)*.*

Provided with the results presented above, we can prove the following Theorem, stating that the estimator $\widehat{A}_k(X^\varepsilon, T)$ is asymptotically unbiased even in the case of the filtering width δ vanishing with respect to the multiscale parameter ε.

Theorem 4.21. *Let Assumption 4.3 and the assumptions of Lemmas 4.6 and 4.20 hold. Let* $\widehat{A}_k(X^\varepsilon, T)$ *be defined in* (4.5) *and* $\delta = \varepsilon^\zeta$ *with* $\zeta \in (0,2)$*. Then*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widehat{A}_k(X^\varepsilon, T) = A, \quad \textit{in probability,}$$

where A *is the drift coefficient of the homogenized equation* (4.2)*.*

Proof. Let us introduce the notation

$$\mathcal{A}^\varepsilon(\delta) \coloneqq \frac{1}{\delta}\widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}[(X^\varepsilon - Z^\varepsilon)^2 V''(Z^\varepsilon)],$$

where $\widetilde{\mathcal{M}}_\varepsilon$ is defined in (4.12). Then following the proof of Theorem 4.15 and in light of Remark 4.16, we only need to show that if $\delta = \varepsilon^\zeta$ with $\zeta \in (0,2)$ we have

$$\lim_{\varepsilon\to 0} \mathcal{A}^\varepsilon(\delta) = A, \quad \text{in probability.}$$

Using Proposition 4.17 and geometric ergodicity for taking the limit for $t\to\infty$ (Lemma 4.6), we have the following equality

$$\begin{aligned}
\mathcal{A}^\varepsilon(\delta) &= \widetilde{\mathcal{M}}_\varepsilon^{-1}\frac{1}{\delta}\lim_{t\to\infty}\mathbb{E}[(X_t^\varepsilon - Z_t^\varepsilon)^2 V''(Z_t^\varepsilon)] \\
&= \widetilde{\mathcal{M}}_\varepsilon^{-1}\frac{1}{\delta}\lim_{t\to\infty}\mathbb{E}\left[(\delta B_t^\varepsilon + R(\varepsilon,\delta))^2\, V''(Z_t^\varepsilon)\right] \\
&=: \widetilde{\mathcal{M}}_\varepsilon^{-1}\lim_{t\to\infty}\left(J_1^\varepsilon(t) + J_2^\varepsilon(t) + J_3^\varepsilon(t)\right),
\end{aligned}$$

where $R(\varepsilon,\delta)$ is given in Proposition 4.17, $\mathbb{E}$ denotes the expectation with respect to the Wiener measure and

$$\begin{aligned}
J_1^\varepsilon(t) &= \delta\,\mathbb{E}\left[(B_t^\varepsilon)^2 V''(Z_t^\varepsilon)\right], \\
J_2^\varepsilon(t) &= 2\,\mathbb{E}\left[B_t^\varepsilon R(\varepsilon,\delta) V''(Z_t^\varepsilon)\right], \\
J_3^\varepsilon(t) &= \frac{1}{\delta}\,\mathbb{E}\left[R(\varepsilon,\delta)^2 V''(Z_t^\varepsilon)\right].
\end{aligned}$$

Let us consider the three terms separately. First, by geometric ergodicity and applying Lemma 4.19 and Lemma 4.20 we get

$$\begin{aligned}
\lim_{\varepsilon\to 0}\lim_{t\to\infty} J_1^\varepsilon(t) &= \lim_{\varepsilon\to 0}\delta\,\mathbb{E}^{\eta^\varepsilon}\left[(B^\varepsilon)^2 V''(Z^\varepsilon)\right] \\
&= \lim_{\varepsilon\to 0}\left(\sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(Z^\varepsilon)(1+\Phi'(Y^\varepsilon))^2] + \widetilde{R}(\varepsilon,\delta)\right) \\
&= \mathcal{M}_0 A.
\end{aligned}$$

Let us now consider $J_2^\varepsilon(t)$. Considering Hölder conjugates p, q, r the Hölder inequality yields

$$|J_2^\varepsilon(t)| \le \mathbb{E}[(B_t^\varepsilon)^p]^{1/p}\,\mathbb{E}[R(\varepsilon,\delta)^q]^{1/q}\,\mathbb{E}[V''(Z^\varepsilon)^r]^{1/r}.$$

Now, we can bound the first two terms with (4.19) and (4.20), respectively. The third term is bounded due to Assumption 4.18 and Lemma 4.30. Hence, we have for t sufficiently large

$$|J_2^\varepsilon(t)| \le C\left(\delta^{1/2} + \varepsilon\delta^{-1/2}\right).$$

We consider now $J_3^\varepsilon(t)$. The Hölder inequality yields for conjugates p and q

$$|J_3^\varepsilon(t)| \le \mathbb{E}[R(\varepsilon,\delta)^{2p}]^{1/p}\,\mathbb{E}[V''(Z_t^\varepsilon)^q]^{1/q},$$

which, similarly as above, yields for t sufficiently large

$$|J_3^\varepsilon(t)| \leq C\left(\delta + \varepsilon^2\delta^{-1}\right).$$

Therefore, since $\delta = \mathcal{O}(\varepsilon^\zeta)$ for $\zeta \in (0,2)$, the terms $J_2^\varepsilon(t)$ and $J_3^\varepsilon(t)$ vanish in the limit for $t \to \infty$ and $\varepsilon \to 0$. Furthermore, by Lemma 4.33 and by weak convergence of the invariant measure μ^ε to μ^0, we have

$$\lim_{\varepsilon\to 0} \widetilde{\mathcal{M}}_\varepsilon = \mathcal{M}_0.$$

Therefore

$$\lim_{\varepsilon\to 0} \mathcal{A}^\varepsilon(\delta) = \mathcal{M}_0^{-1}\mathcal{M}_0 A = A,$$

which implies the desired result. □

We conclude this section with a negative convergence result, i.e., that if $\delta = \varepsilon^\zeta$ with $\zeta > 2$, the estimator based on filtered data converges to the coefficient α of the unhomogenized equation. This result is relevant for two reasons. First, it shows the sharpness of the bound on ζ in the assumptions of Theorem 4.21. Second, it shows an interesting switch between two completely different regimes at $\zeta = 2$, which happens arbitrarily fast in the limit $\varepsilon \to 0$.

Theorem 4.22. *Let Assumption 4.3 and the assumptions of Lemmas 4.6 and 4.20 hold. Let* $\widehat{A}_k(X^\varepsilon, T)$ *be defined in* (4.5) *and* $\delta = \varepsilon^\zeta$ *with* $\zeta > 2$. *Then*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widehat{A}_k(X^\varepsilon, T) = \alpha, \quad \textit{in probability},$$

where α *is the drift coefficient of the multiscale equation* (4.1).

The proof of Theorem 4.22 is given in Section 4.9.3.

4.6 The Diffusion Coefficient

Estimating the effective diffusion coefficient Σ of the homogenized SDE (4.2) is as well a relevant problem. Indeed, knowing Σ besides the drift coefficient A gives a complete estimation of the model (4.2), which is effective for the multiscale data generated by (4.1) in the sense of homogenization theory. The standard approach for estimating the diffusion coefficient is to approximate the quadratic variation of the path. In [112, Theorem 3.4], the authors show that this approach fails in case the data is not pre-processed, meaning that the quadratic variation of X^ε equals the diffusion coefficient σ of (4.1), even in the limit for $\varepsilon \to 0$. They propose therefore the estimator $\widehat{\Sigma}_\delta$ based on subsampling and defined as

$$\widehat{\Sigma}_\delta \coloneqq \frac{1}{2T}\sum_{i=1}^{n}\left(X_{i\delta}^\varepsilon - X_{(i-1)\delta}^\varepsilon\right)^2,$$

where δ is the subsampling width and where $T = n\delta$. It is possible to show that $\widehat{\Sigma}_\delta$ indeed estimates the effective diffusion coefficient Σ asymptotically for $\varepsilon \to 0$ and $T \to \infty$ [112, Theorem 3.6].

Despite the focus of this chapter being mainly the estimation of the effective drift coefficient, we propose here the estimator for Σ in (4.2) based on filtered data and given by

$$\widehat{\Sigma}_k(X^\varepsilon, T) \coloneqq \frac{1}{\delta T}\int_0^T \left(X_t^\varepsilon - Z_t^\varepsilon\right)^2 \mathrm{d}t, \tag{4.21}$$

where again we employ the subscript k for reference to the kernel (4.4) of the filter. The estimator $\widehat{\Sigma}$ is unbiased for the effective diffusion coefficient Σ in case $\beta = 1$ and when we filter data at the multiscale regime, i.e., when δ is a vanishing function of ε. In particular, the following result holds.

Theorem 4.23. *Let the assumptions of Theorem 4.21 hold. Then, if $\delta = \varepsilon^\zeta$, with $\zeta \in (0,2)$, it holds*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widehat{\Sigma}_k(X^\varepsilon, T) = \Sigma, \quad \textit{in probability,}$$

where Σ is the diffusion coefficient of the homogenized equation (4.2).

Proof. First, the ergodic theorem yields

$$\lim_{T\to\infty} \widehat{\Sigma}_k = \frac{1}{\delta}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[(X^\varepsilon - Z^\varepsilon)^2\right],$$

then applying Proposition 4.17 at stationarity we obtain

$$\begin{aligned}\lim_{T\to\infty} \widehat{\Sigma}_k &= \delta\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[(B^\varepsilon)^2\right] + 2\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[B^\varepsilon R(\varepsilon,\delta)\right] + \frac{1}{\delta}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[R(\varepsilon,\delta)^2\right]\\ &=: I_1^\varepsilon + I_2^\varepsilon + I_3^\varepsilon,\end{aligned}$$

and due to the Cauchy-Schwarz inequality and estimates (4.19) and (4.20) we have

$$|I_2^\varepsilon| \le C\left(\delta^{1/2} + \varepsilon\delta^{-1/2}\right) \qquad \text{and} \qquad |I_3^\varepsilon| \le C\left(\delta + \varepsilon^2\delta^{-1}\right), \tag{4.22}$$

for a constant $C > 0$ independent of ε and δ. Let us now consider I_1^ε. Employing equation (4.35) with the function $f(z,b) = 1/2b^2$ gives

$$\mathbb{E}^{\eta^\varepsilon}\left[(B^\varepsilon)^2\right] = \frac{\sigma}{\delta}\,\mathbb{E}^{\eta^\varepsilon}\left[1 + \Phi'(Y^\varepsilon)\right] = \frac{\sigma K}{\delta} = \frac{\Sigma}{\delta},$$

which together with bounds (4.22) and the hypothesis on δ implies

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widehat{\Sigma} = \Sigma, \quad \text{in probability,}$$

which is the desired result. □

4.7 Filtering the Data in The Bayesian Framework

In Section 3.6 we have presented how the problem of estimating the drift coefficient for multiscale diffusion processes can be recast in a Bayesian framework. In particular, Theorem 3.16 highlights at the posterior level the biasedness issue which occurs if data from the multiscale model are employed without additional care. In this section, we present how to correct this faulty behavior by employing filtered data.

We recall that without pre-processing the data, the posterior distribution is given by the Gaussian $\mu(\cdot \mid X^\varepsilon) = \mathcal{N}(m_T(X^\varepsilon), C_T(X^\varepsilon))$, where

$$C_T(X^\varepsilon)^{-1} = C_0^{-1} + \frac{T}{2\Sigma}M(X^\varepsilon), \quad m_T(X^\varepsilon) = C_T(X^\varepsilon)\left(C_0^{-1}A_0 - \frac{T}{2\Sigma}v(X^\varepsilon)\right), \tag{4.23}$$

and where $M(X^\varepsilon)$ and $v(X^\varepsilon)$ are given in (4.3). Let us consider the modified likelihood function

$$\widetilde{L}(X^\varepsilon \mid A) = \exp\left(-\frac{\widetilde{I}(X^\varepsilon \mid A)}{2\Sigma}\right),$$

where

$$\begin{aligned}\widetilde{I}(X^\varepsilon \mid A) &= \int_0^T A \cdot V'(Z_t^\varepsilon)\,\mathrm{d}X_t^\varepsilon + \frac{1}{2}\int_0^T \left(A \cdot V'(X_t^\varepsilon)\right)^2 \mathrm{d}t \\ &= \widetilde{v}(X^\varepsilon)\cdot A + \frac{1}{2} A \cdot M(X^\varepsilon) A,\end{aligned}$$

and where $\widetilde{v}(X^\varepsilon)$ is given in (4.5). Since $M(X^\varepsilon)$ is symmetric positive definite by Assumption 4.3, the function $\widetilde{L}(X^\varepsilon \mid A)$ is indeed a valid Gaussian likelihood function. Proceeding as in Section 3.6 and fixing a prior $\mu_0 = \mathcal{N}(A_0, C_0)$ on the parameter, we then obtain the modified posterior measure $\widetilde{\mu}_T(\cdot \mid X^\varepsilon) = \mathcal{N}(\widetilde{m}_T(X^\varepsilon), C_T(X^\varepsilon))$, whose mean and covariance are given by

$$C_T(X^\varepsilon)^{-1} = C_0^{-1} + \frac{T}{2\Sigma} M(X^\varepsilon), \quad \widetilde{m}_T(X^\varepsilon) = C_T(X^\varepsilon)\left(C_0^{-1} A_0 - \frac{T}{2\Sigma}\widetilde{v}(X^\varepsilon)\right)$$

Let us remark that the posterior $\widetilde{\mu}_T$ has the same covariance as μ_T given in (4.23) and that therefore it is indeed a valid Gaussian posterior distribution. Nevertheless, in order to employ the tool of convergence introduced in Theorem 3.15, we need to study the properties of the MLE based on the likelihood $\widetilde{L}(X^\varepsilon \mid A)$, i.e., the solution of the linear system

$$-M(X^\varepsilon)\widetilde{A}_k(X^\varepsilon, T) = \widetilde{v}(X^\varepsilon). \tag{4.24}$$

The following theorem guarantees the unbiasedness of this estimator under a condition on the parameter δ of the filter.

Theorem 4.24. *Let the assumptions of Theorem 4.21 hold. Then, if $\delta = \varepsilon^\zeta$, with $\zeta \in (0,2)$, it holds*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widetilde{A}_k(X^\varepsilon, T) = A, \quad \textit{in probability,}$$

for $\widetilde{A}_k(X^\varepsilon, T)$ defined in (4.24).

Proof. We first consider the difference between the two estimators $\widetilde{A}_k(X^\varepsilon, T)$ and $\widehat{A}_k(X^\varepsilon, T)$. In particular, the ergodic theorem and an algebraic equality imply

$$\begin{aligned}\lim_{T\to\infty}\left(\widetilde{A}_k(X^\varepsilon, T) - \widehat{A}_k(X^\varepsilon, T)\right) &= \left(\mathcal{M}_\varepsilon^{-1} - \widetilde{\mathcal{M}}_\varepsilon^{-1}\right)\lim_{T\to\infty}\widetilde{v} \\ &= -\mathcal{M}_\varepsilon^{-1}\left(\mathcal{M}_\varepsilon - \widetilde{\mathcal{M}}_\varepsilon\right)\widetilde{\mathcal{M}}_\varepsilon^{-1}\lim_{T\to\infty}\widetilde{v} \\ &= \mathcal{M}_\varepsilon^{-1}\left(\mathcal{M}_\varepsilon - \widetilde{\mathcal{M}}_\varepsilon\right)\lim_{T\to\infty}\widehat{A}_k(X^\varepsilon, T),\end{aligned}$$

almost surely, where $\mathcal{M}_\varepsilon$ and $\widetilde{\mathcal{M}}_\varepsilon$ are defined in (4.11) and (4.12), respectively. Therefore, due to Assumption 4.3 which allows controlling the norm of $\mathcal{M}_\varepsilon^{-1}$ and due to Lemma 4.33 we have for a constant $C > 0$

$$\lim_{T\to\infty}\left\|\widetilde{A}_k(X^\varepsilon, T) - \widehat{A}_k(X^\varepsilon, T)\right\|_2 \le C\left(\varepsilon + \delta^{1/2}\right), \tag{4.25}$$

where we remark that $\widehat{A}_k(X^\varepsilon, T)$ has a bounded norm for ε sufficiently small due to Theorem 4.21. Now, the triangle inequality yields

$$\left\|\widetilde{A}_k(X^\varepsilon, T) - A\right\|_2 \le \left\|\widetilde{A}_k(X^\varepsilon, T) - \widehat{A}_k(X^\varepsilon, T)\right\|_2 + \left\|\widehat{A}_k(X^\varepsilon, T) - A\right\|_2.$$

Therefore, due to Theorem 4.21, the inequality (4.25) and since $\delta = \varepsilon^\zeta$, the desired result holds. □

Remark 4.25. One could argue that we could have carried on the whole analysis for the estimator $\widetilde{A}_k(X^\varepsilon, T)$ instead of the estimator $\widehat{A}_k(X^\varepsilon, T)$. Nevertheless, the latter guarantees the strong result of almost sure convergence in case δ is independent of ε, which is false for the former. Conversely, analysing the properties of the estimator $\widetilde{A}_k(X^\varepsilon, T)$ is fundamental for the Bayesian setting, in which the matrix $\widetilde{M}$ cannot be employed as its symmetric part is not positive definite in general.

In light of Theorem 3.15, the result above guarantees that the mean of the posterior distribution $\widetilde{\mu}_T(\cdot \mid X^\varepsilon)$ converges to the drift coefficient of the homogenized equation. Since the covariance matrix is the same for $\mu_T(\cdot \mid X^\varepsilon)$ and $\widetilde{\mu}_T(\cdot \mid X^\varepsilon)$, it is possible to prove a positive convergence result for $\widetilde{\mu}_T(\cdot \mid X^\varepsilon)$, which is given by the following Theorem.

Theorem 4.26. *Let the assumptions of Theorem 4.24 hold. Then, the modified posterior measure* $\widetilde{\mu}_T(\cdot \mid X^\varepsilon) = \mathcal{N}(\widetilde{m}_T(X^\varepsilon), C_T(X^\varepsilon))$ *satisfies*

$$\lim_{\varepsilon \to 0} \lim_{T \to \infty} \mathbb{E}\left[\widetilde{\mu}_T\left(\{a\colon \|a - A\|_2 \geq c\} \mid X^\varepsilon\right)\right] = 0,$$

where $\mathbb{E}$ *denotes expectation with respect to the Wiener measure and* A *is the drift coefficient of the homogenized equation* (4.2).

Proof. The proof follows from the proof of Theorem 3.16 and from Theorem 4.24. □

4.8 Numerical Experiments

In this section we show numerical experiments confirming our theoretical findings and showcasing the potential of the filtered data approach to overcome model misspecification arising when multiscale data is used to fit homogenized models.

Remark 4.27. In practice, we consider for numerical experiment the data to be in the form of a high-frequency discrete time series from the solution X^ε of (4.1). Let n be a positive integer, $\tau = T/n$ be the time step at which data is observed, and let $X^\varepsilon \coloneqq (X^\varepsilon_0, X^\varepsilon_\tau, \dots, X^\varepsilon_{n\tau})$. We then compute the estimator $\widehat{A}_k$ as the solution of the linear system

$$-\widetilde{M}_\tau^{-1}(X^\varepsilon)\widehat{A}_{k,\tau}(X^\varepsilon, T) = \widetilde{v}_\tau(X^\varepsilon),$$

where

$$\widetilde{M}_\tau(X^\varepsilon) = \frac{\tau}{T}\sum_{j=0}^{n-1} V'(Z^\varepsilon_{j\tau}) \otimes V'(X^\varepsilon_{j\tau}), \qquad \widetilde{v}_\tau(X^\varepsilon) = \frac{1}{T}\sum_{j=0}^{n-1} V'(Z^\varepsilon_{j\tau})(X^\varepsilon_{(j+1)\tau} - X^\varepsilon_{j\tau}).$$

We take in all experiments $\tau \ll \varepsilon^2$, so that the discretization of the data has negligible effects and does not compromise the validity of our theoretical results.

4.8.1 Parameters of the Filter

For the first preliminary experiments, we consider $N = 1$ and the quadratic potential $V(x) = x^2/2$. In this case, the solution of the homogenized equation is an Ornstein–Uhlenbeck process. Moreover, we set the the fast potential in the multiscale equation (4.1) as $p(y) = \cos(y)$. In all experiments, data is generated employing the Euler–Maruyama method with a fine time step.

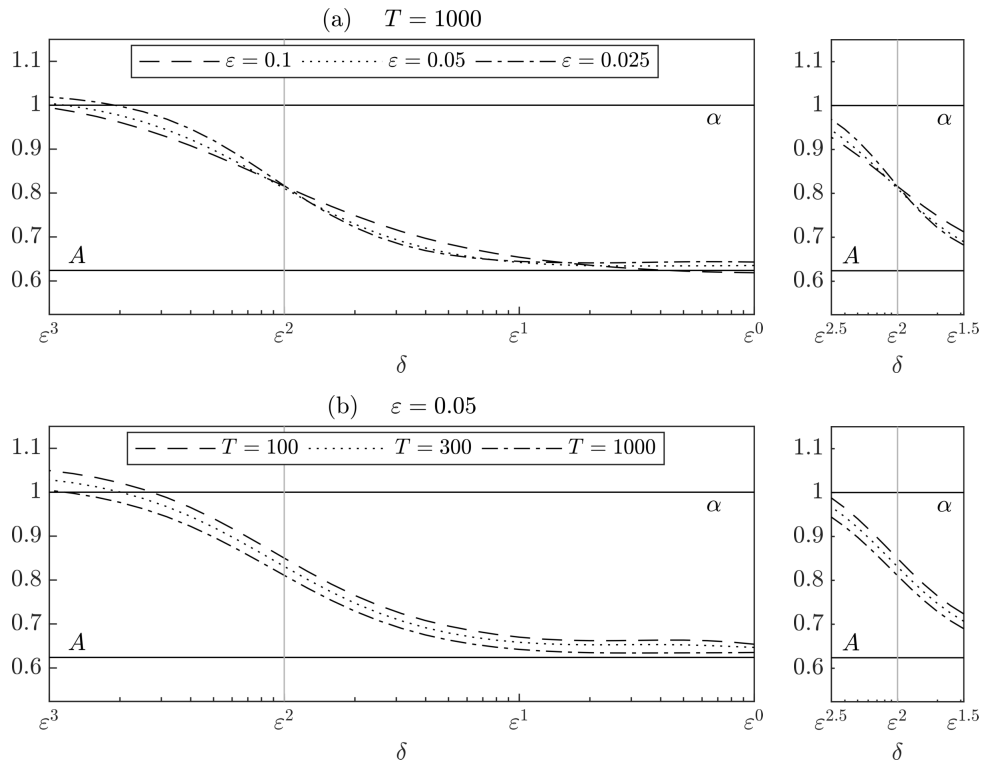


Figure 4.2 – Results for Section 4.8.1. On both figures, horizontal lines represent α and A, the drift coefficients of the unhomogenized and homogenized equations, and the grey vertical line represents the lower bound for the validity of Theorem 4.21. The curved lines (dashed, dotted and dash-dotted) represent on figure (a) the values of $\widehat{A}_k(X^\varepsilon, T)$ for $\varepsilon = \{0.1, 0.05, 0.025\}$, respectively, computed with $T = 10^3$. On figure (b), they correspond to the values of $\widehat{A}_k(X^\varepsilon, T)$ at $T = \{100, 300, 1000\}$, respectively, computed with $\varepsilon = 0.05$. We plot next to both figures (a) and (b) a zoom on a neighbourhood of ε^2 to show the transition between the two regimes highlighted by the theoretical results. Note that the δ-axis is in logarithmic scale and is normalized with respect to ε.

Verification of Theoretical Results

We first demonstrate numerically the validity of Theorem 4.15, Theorem 4.21 and Theorem 4.22, i.e., the unbiasedness of $\widehat{A}_k(X^\varepsilon, T)$ for $\delta = \varepsilon^\zeta$ with $\zeta \in [0, 2)$ and biasedness for $\zeta > 2$. Let us recall that for $\zeta = 0$ the analysis and the theoretical result are fundamentally different than for $\zeta \in (0, 2)$. We consider $\varepsilon \in \{0.1, 0.05, 0.025\}$, the diffusion coefficient $\sigma = 1$ and generate data X_t^ε for $0 \leq t \leq T$ with $T = 10^3$. Then we filter the data by choosing $\delta = \varepsilon^\zeta$, and $\zeta = 0, 0.1, 0.2, \ldots, 3$, and compute $\widehat{A}_k(X^\varepsilon, T)$. Results are displayed in Figure 4.2, and show that for $\zeta > 2$, i.e., $\delta = o(\varepsilon^2)$, the estimator tends to the drift coefficient α of the unhomogenized equation. Conversely, as predicted by the theory, for $\zeta \in [0, 2)$ the estimator tends to A, the drift coefficient of the homogenized equation. Therefore, the point $\delta = \varepsilon^2$ acts asymptotically as a switch between two completely different regimes, which is theoretically sharp in the limit for $T \to \infty$ and $\varepsilon \to 0$. Let us remark that the results displayed in Figure 4.2.(a) demonstrate that the transition occurs more rapidly for the smallest values of ε. Moreover, in Figure 4.2.(b), one can see how with bigger final times T the estimator is closer both to A when $\zeta \in [0, 2]$ and to α when $\zeta > 2$. Still, we observe that in finite computations the switch between A and α is smoother than what we expect from the theory, which suggests to fix, if possible, $\delta = 1$.

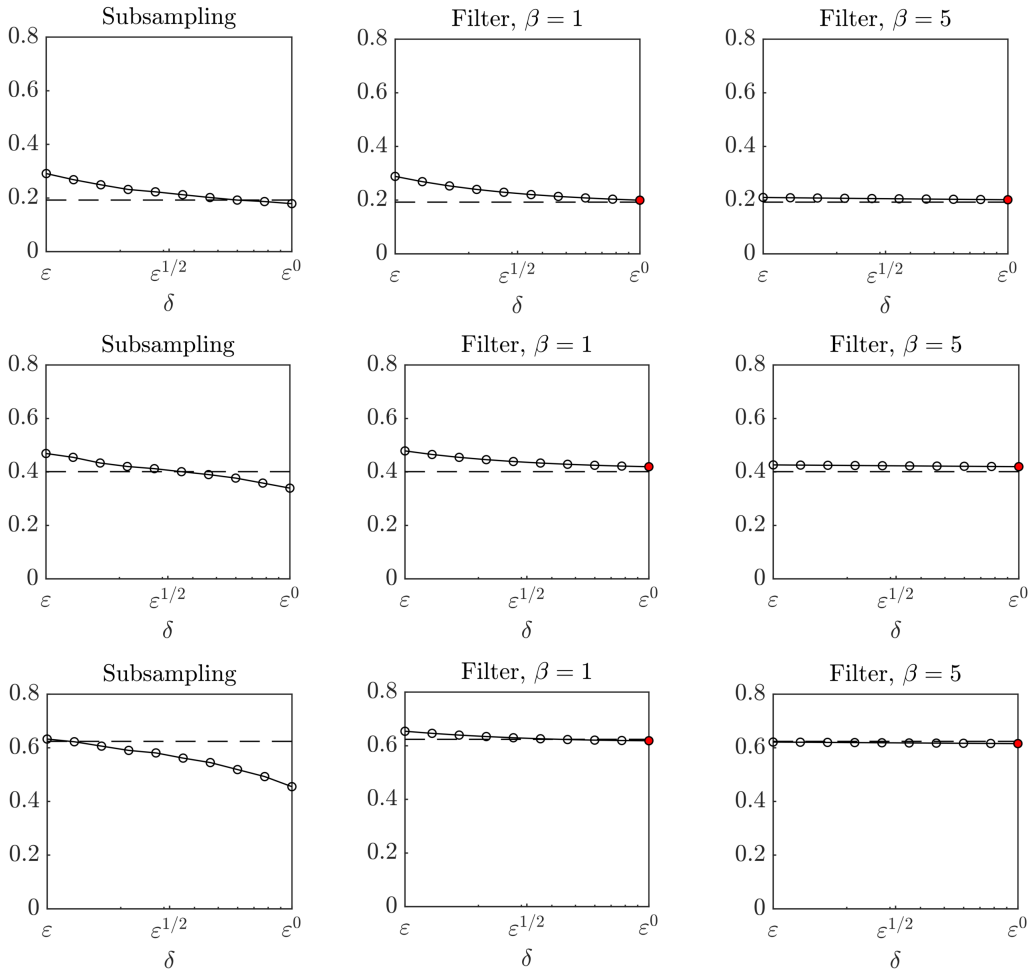


Figure 4.3 – Results for Section 4.8.1. The case of $\delta = 1$ is highlighted as a solid dot for the filtered data technique, as the analysis and theoretical result is different in this case. The three rows correspond to $\sigma = 0.5, 0.7, 1.0$ from top to bottom, and the dashed line corresponds to the true value of A.

Comparison with Subsampling

We now compare the results given by the filtered data technique with the results given by subsampling the data, i.e., the difference between the estimators $\widehat{A}_k(X^\varepsilon, T)$ and $\widehat{A}_\delta(X^\varepsilon, T)$. We fix the diffusion coefficient $\sigma = 0.5$, the multiscale parameter $\varepsilon = 0.1$ and generate data for $0 \leq t \leq T$ with $T = 10^3$. We choose $\delta = \varepsilon^\zeta$ and vary $\zeta \in [0, 1]$, where δ is the filtering and the subsampling width, respectively. Moreover, for the filtered data approach we consider both $\beta = 1$ and $\beta = 5$. We report in Figure 4.3 the experimental results. Let us remark that:

(i) for $\sigma = 0.5$ the results given by subsampling and by the filter with $\beta = 1$ are similar, while for higher values of σ the filtered data approach seems better than subsampling;

(ii) in general, choosing a higher value of β seems beneficial for the quality of the estimator;

(iii) the dependence on δ of numerical results given by the filter seems relevant only in case $\beta = 1$ and for small values of σ. For $\beta = 1$ and higher values of σ, the estimator is stable with respect to this parameter. This can be observed for a higher value of β but we have

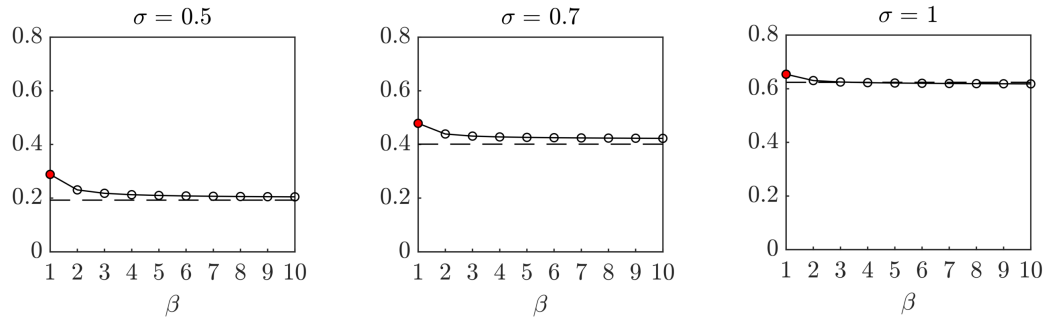


Figure 4.4 – Results for the estimator based on filter data with respect to the parameter β (Section 4.8.1). The result for $\beta = 1$, for which there are theoretical guarantees given by Theorem 4.21, is highlighted as a solid dot. From left to right we consider different values of σ, and the dashed line corresponds to the true value of A.

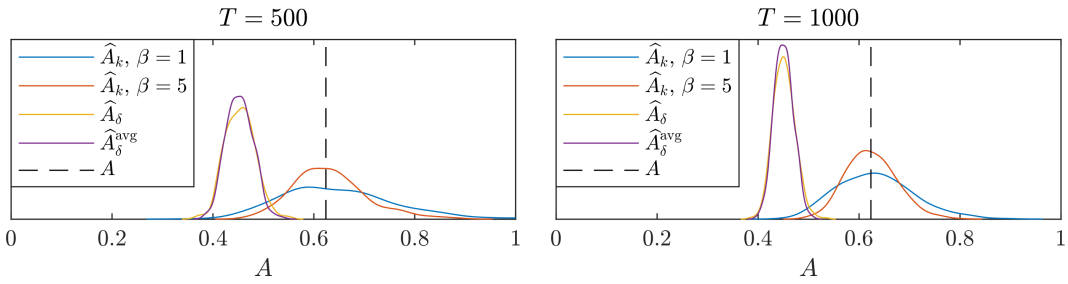


Figure 4.5 – Numerical results for Section 4.8.2. Comparison between the density of the estimator of the drift based on filtered data with $\beta = \{1, 5\}$, the estimator based on subsampling and the estimator based on shift-subsampling and averaging of (4.27). On the left and on the right, the final time is $T = \{500, 1000\}$, respectively.

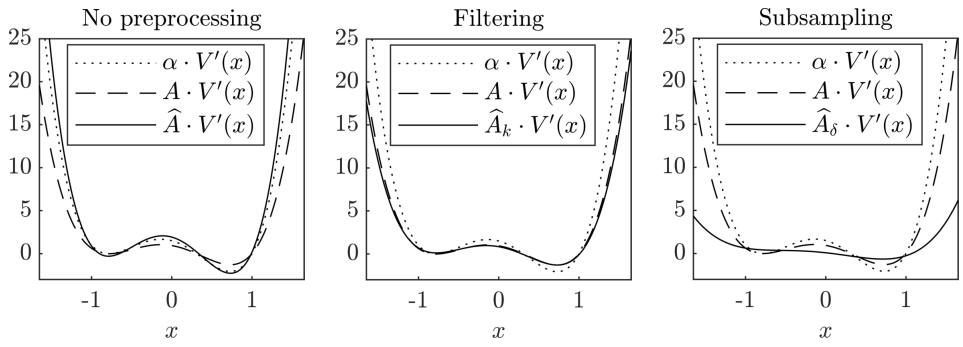
no theoretical guarantee in this case.

The Influence of β

We finally test the variability of the estimator with respect to β in (4.4). We consider $\delta = \varepsilon$, which corresponds to $\zeta = 1$ and seems to be the worst-case scenario for the filter, at least for $\beta = 1$. We consider again $\sigma = 0.5, 0.7, 1$ and vary $\beta = 1, 2, \ldots, 10$. Results, given in Figure 4.4, show empirically that the estimator stabilizes fast with respect to β. Nevertheless, there is no theoretical guarantee supporting this empirical observation.

4.8.2 Variance of the Estimators

We now compare the estimators $\widehat{A}_k$ based on filtered data and $\widehat{A}_\delta$ based on subsampling in terms of variance. We consider for this experiment the SDE (4.1) with $N = 1$, the bistable potential $V(x) = x^4/4 - x^2/2$, the multiscale drift coefficient $\alpha = 1$, the diffusion coefficient $\sigma = 1$ and with $\varepsilon = 0.1$. We then let $X^\varepsilon = (X_t, 0 \le t \le T)$ be the solution of (4.1) and generate $N_{\text{s}} = 500$ i.i.d. samples of X^ε. We then compute the estimators $\widehat{A}_k$ and $\widehat{A}_\delta$ on each of the realizations of X^ε, thus obtaining N_{s} replicas $\{\widehat{A}_k^{(i)}\}_{i=1}^{N_{\text{s}}}$ and $\{\widehat{A}_\delta^{(i)}\}_{i=1}^{N_{\text{s}}}$. For the estimator $\widehat{A}_k$, we consider the kernel (4.4) with $\beta = \{1, 5\}$ and with $\delta = 1$. For the estimator $\widehat{A}_\delta$, we employ the subsampling width $\delta = \varepsilon^{2/3}$,



Coefficient	Multiscale α	Homogenized A	No preprocessing $\widehat{A}$	Filtering $\widehat{A}_k$	Subsampling $\widehat{A}_\delta$
1	-1	-0.62	-0.92	-0.70	-0.59
2	-0.5	-0.31	-0.70	-0.27	0.05
3	0.5	0.31	0.55	0.31	0.14
4	1	0.62	1.22	0.57	0.13

Figure 4.6 – Results for Section 4.8.3. In the figure, from left to right the potential function estimated with the data itself, the filter, subsampled data. In the table, numerical results for the single components of the true and estimated drift coefficients.

which is heuristically optimal following [112]. It could be argued that another estimator based on subsampling and shifting could be employed to reduce the variance. In particular, we let $\tau > 0$ be the time step at which the data is observed. Indeed, in practice we work with high-frequency discrete data, and observe $X^\varepsilon := (X_0^\varepsilon, X_\tau^\varepsilon, \dots, X_{n\tau}^\varepsilon)$, with $n\tau = T$. We assume for simplicity that the subsampling width δ is a multiple of τ and compute for all $k = 0, 1, \dots, \delta/\tau - 1$

$$\widehat{A}_{\delta,k}(X^\varepsilon, T) = -\frac{\sum_{j=0}^{n-1} V'(X_{j\delta+k}^\varepsilon)(X_{(j+1)\delta+k}^\varepsilon - X_{j\delta+k}^\varepsilon)}{\delta \sum_{j=0}^{n-1} V'(X_{j\delta+k}^\varepsilon)^2}, \tag{4.26}$$

i.e. the subsampling estimator obtained by shifting the origin by $k\tau$. We then average over the index k and obtain the new estimator

$$\widehat{A}_\delta^{\text{avg}}(X^\varepsilon, T) = \frac{\tau}{\delta} \sum_{k=0}^{\delta/\tau - 1} \widehat{A}_{\delta,k}(X^\varepsilon, T). \tag{4.27}$$

We include this estimator in the numerical study for completeness, and compute N_s replicas of $\widehat{A}_\delta^{\text{avg}}$ on all the realizations of X^ε. Results, given in Figure 4.5 for the final times $T = \{500, 1000\}$, show that our novel approach does not outperform subsampling in terms of variance, but clearly does in terms of bias. Moreover, we notice numerically that the shifted-averaged estimator $\widehat{A}_\delta^{\text{avg}}$ does not reduce sensibly the variance in this case with respect to $\widehat{A}_\delta$. In fact, this is only partly surprising, since the estimators $\widehat{A}_{\delta,k}$ of (4.26) are highly correlated. Finally, we notice that the filtering estimator $\widehat{A}_k$ with $\beta = 5$ has a lower variance with respect to the same estimator with $\beta = 1$. This confirms that choosing a higher value of β improves the estimation of the effective drift coefficient.

4.8.3 Multidimensional Drift Coefficient

Let us consider the Chebyshev polynomials of the first kind, i.e., the polynomials $T_i\colon \mathbb{R} \to \mathbb{R}$, $i = 0, 1, \ldots$, defined by the recurrence relation

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x).$$

We consider the potential function $V(x) = (V_1(x), V_2(x), V_3(x), V_4(x))^\top$, with

$$V_i(x) = T_i(x), \quad i = 1, \ldots, 4,$$

thus considering the semi-parametric framework (see Remark 3.8). The potential V then satisfies Assumption 4.3 whenever N is even and if the leading coefficient α_N is positive. We set $N = 4$ and the drift coefficient $\alpha = (-1, -1/2, 1/2, 1)$. With this drift coefficient, the potential function is of the bistable kind. Moreover, we set $\varepsilon = 0.05$, the diffusion coefficient $\sigma = 1$, the fast potential $p(y) = \cos(y)$ and simulate a trajectory of X^ε for $0 \le t \le T$ with $T = 10^3$ employing the Euler–Maruyama method with time step $\Delta t = \varepsilon^3$. We estimate the drift coefficient $A \in \mathbb{R}^4$ with the estimators:

(i) $\widehat{A}(X^\varepsilon, T)$ based on the data X^ε itself;

(ii) $\widehat{A}_\delta(X^\varepsilon, T)$ based on subsampled data with subsampling parameter $\delta = \varepsilon^{2/3}$;

(iii) $\widehat{A}_k(X^\varepsilon, T)$ based on filtered data Z^ε computed with $\beta = 1$ and $\delta = 1$.

In particular, we pick this specific value of δ for the subsampling following the optimality criterion given in [112]. Results, given in Figure 4.6, show that the filter-based estimation captures well the homogenized potential as well as the coefficient A. Moreover, it is possible to remark the negative result given by Theorem 3.13 holds in practice, i.e., with no pre-processing the estimator $\widehat{A}(X^\varepsilon, T)$ tends to the drift coefficient α of the unhomogenized equation. Finally, we can observe that the subsampling-based estimator fails to capture the homogenized coefficients. Indeed, the estimator strongly depends on the sampling rate and on the diffusion coefficient, as shown in the numerical experiments of [112]. Even though the authors suggest the choice of $\delta = \varepsilon^{2/3}$, this is just an heuristic and is not guaranteed to be the optimal value in all cases. In the asymptotic limit of $\varepsilon \to 0$ and $T \to \infty$, any valid choice of the subsampling rate is guaranteed theoretically to work, but not in the pre-asymptotic regime. Moreover, in [112] the authors suggest that different subsampling rates may be employed for different components of the drift coefficient, which renders the subsampling approach complex to apply in practice. Our estimator, conversely, seems to perform better with no particular tuning of the parameters even in this multi-dimensional case, which demonstrates the robustness of our novel approach.

4.8.4 The Bayesian Approach: Bistable Potential

In this numerical experiment we consider $N = 2$ and the bistable potential, i.e., the function V defined as

$$V(x) = \left(\frac{x^4}{4} \quad -\frac{x^2}{2} \right)^\top,$$

with coefficients $\alpha_1 = 1$ and $\alpha_2 = 2$. We then consider the multiscale equation with $\sigma = 0.7$, the fast potential $p(y) = \cos(y)$ and $\varepsilon = 0.05$, thus simulating a trajectory X^ε. We adopt here a Bayesian approach and compute the posterior distribution $\widetilde{\mu}_T$ obtained with the filtered data approach introduced in Section 4.7. The parameters of the filter are set to $\beta = 1$ and $\delta = \varepsilon$ in (4.4). We choose the non-informative prior $\mu_0 = \mathcal{N}(0, I)$, where I is the identity matrix in $\mathbb{R}^2$. Let us remark that in order to compute the posterior covariance the diffusion coefficient Σ of the homogenized equation has to be known. In this case, we pre-compute the value of Σ via

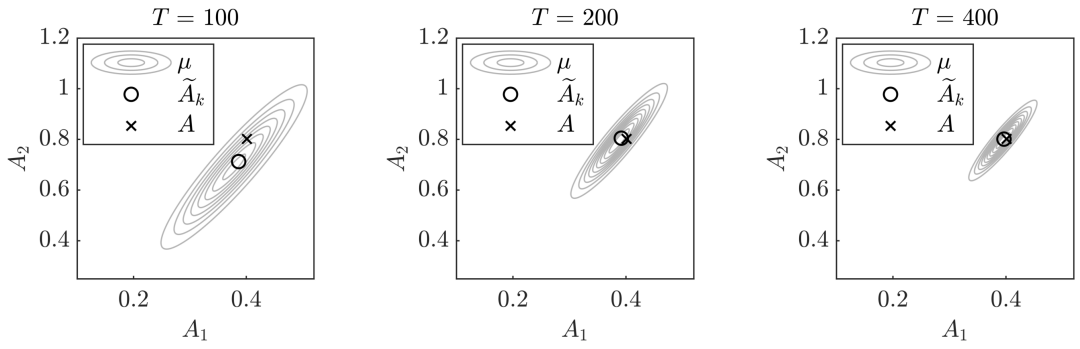


Figure 4.7 – Results for Section 4.8.4. Posterior distributions over the parameter $A = (A_1, A_2)^\top$ for the bistable potential obtained with the filtered data approach. The figures refer to final time $T = 100, 200, 400$ from left to right, respectively. The MLE $\widetilde{A}_k(X^\varepsilon, t)$ is represented with a circle, while the true value A of the drift coefficient of the homogenized equation is represented with a cross.

the coefficient K and the theory of homogenization, but notice that Σ could be estimated either employing the subsampling technique of [112] or using the estimator $\widehat{\Sigma}_k$ based on filtered data defined in (4.21). In particular, in this case $\Sigma \approx 0.2807$, and we compute numerically

$$\widehat{\Sigma}_k(X^\varepsilon, 100) = 0.2901, \quad \widehat{\Sigma}_k(X^\varepsilon, 200) = 0.2835, \quad \widehat{\Sigma}_k(X^\varepsilon, 400) = 0.2813,$$

so that employing the estimator $\widehat{\Sigma}_k$ instead of the true value would have negligible effects on the computation of the posterior over the effective drift coefficient. We stop computations at times $T = \{100, 200, 400\}$ in order to observe the shrinkage of the Gaussian posterior towards the MLE $\widetilde{A}_k(X^\varepsilon, T)$ with respect to time. In Figure 4.7, we observe that the posterior does indeed shrink towards the MLE, which in turn gets progressively closer to the true value of the drift coefficient A of the homogenized equation.

4.9 Proof of Technical Results

We conclude the chapter by giving the proof of some technical results, which were omitted in the text to enhance readability.

4.9.1 Proofs of Sections 4.3

Proof of Lemma 4.5. We have to show that the joint process solution to (4.6) is hypo-elliptic. Denoting as $f\colon \mathbb{R} \to \mathbb{R}$ the function

$$f(x) = -\alpha \cdot V'(x) - \frac{1}{\varepsilon} p'\left(\frac{x}{\varepsilon}\right),$$

the generator of the process $(X^\varepsilon, Z^\varepsilon)^\top$ is given by

$$\mathcal{L} = f\partial_x + \sigma\partial^2_{xx} + \frac{1}{\delta}(x - z)\partial_z =: \mathcal{X}_0 + \sigma\mathcal{X}_1^2,$$

where

$$\mathcal{X}_0 = f\partial_x + \frac{1}{\delta}(x - z)\partial_z, \quad \mathcal{X}_1 = \partial_x.$$

The commutator $[\mathcal{X}_0, \mathcal{X}_1]$ applied to a test function v then gives

$$\begin{aligned}[\mathcal{X}_0, \mathcal{X}_1]v &= f\partial_x^2 v + \frac{1}{\delta}(x-z)\partial_x\partial_z v - \partial_x\left(f\partial_x v + \frac{1}{\delta}(x-z)\partial_z v\right)\\ &= -\partial_x f\partial_x v - \frac{1}{\delta}\partial_z v.\end{aligned}$$

Consequently,

$$\operatorname{Lie}(\mathcal{X}_1, [\mathcal{X}_0, \mathcal{X}_1]) = \operatorname{Lie}\left(\partial_x, -\partial_x f\partial_x - \frac{1}{\delta}\partial_z\right),$$

which spans the tangent space of $\mathbb{R}^2$ at (x, z), denoted $T_{x,z}\mathbb{R}^2$. The desired result then follows from Hörmander's theorem (see e.g. [110, Chapter 6]). □

Proof of Lemma 4.6. Lemma 4.5 guarantees that the Fokker–Planck equation can be written directly from the system (4.6). For geometric ergodicity, let

$$\mathcal{S}(x,z) \coloneqq \begin{pmatrix} -\alpha\cdot V'(x) - \frac{1}{\varepsilon}p'(\frac{x}{\varepsilon}) \\ \frac{1}{\delta}(x-z)\end{pmatrix}\cdot\begin{pmatrix}x\\z\end{pmatrix} = -\left(\alpha\cdot V'(x) + \frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)\right)x + \frac{1}{\delta}(xz - z^2).$$

Due to Assumption 4.3, Remark 4.4 and Young's inequality, we then have for all $\gamma > 0$

$$\mathcal{S}(x,z) \le a + \left(\frac{1}{2\gamma\delta} - b\right)x^2 + \frac{1}{\delta}\left(\frac{\gamma}{2} - 1\right)z^2.$$

We choose $\gamma = \gamma^* \coloneqq 1 - b\delta + \sqrt{1 + (1-b\delta)^2} > 0$ so that

$$C(\gamma^*) \coloneqq -\frac{1}{2\gamma^*\delta} + b = -\frac{1}{\delta}\left(\frac{\gamma^*}{2} - 1\right),$$

and we notice that $C(\gamma^*) > 0$ if $\delta > 1/(4b)$. In this case, we have

$$\mathcal{S}(x,z) \le a - C(\gamma^*)\left\|\begin{pmatrix}x & z\end{pmatrix}^\top\right\|^2,$$

and problem (4.6) is dissipative. It remains to prove the irreducibility condition [94, Condition 4.3]. We remark that the system (4.6) fits the framework of the example the end of [94, Page 199], and therefore [94, Condition 4.3] is satisfied. The result then follows from [94, Theorem 4.4]. □

Proof of Lemma 4.8. Let us remark that the marginal density ρ^ε of X^ε satisfies the stationary FPE

$$\sigma(\rho^\varepsilon)''(x) + \frac{\mathrm{d}}{\mathrm{d}x}\left(\left(\alpha\cdot V'(x) + \frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)\right)\rho^\varepsilon(x)\right) = 0, \tag{4.28}$$

In view of (4.8) and integrating (4.7) with respect to x, we then obtain in light of (4.28)

$$\partial_x\left(\sigma\rho^\varepsilon\psi^\varepsilon\partial_x R^\varepsilon\right) + \partial_z\left(\frac{1}{\delta}(z-x)\rho^\varepsilon\psi^\varepsilon R^\varepsilon\right) = 0.$$

We now multiply the equation above by a continuous differentiable function $f\colon \mathbb{R}^2 \to \mathbb{R}^N$, $f = f(x,z)$, and integrate with respect to x and z. An integration by parts yields

$$\begin{aligned}\sigma\int_{\mathbb{R}}\int_{\mathbb{R}} &\partial_x f(x,z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z\\ &= \frac{1}{\delta}\int_{\mathbb{R}}\int_{\mathbb{R}}\partial_z f(x,z)(x-z)\rho^\varepsilon(x)\psi^\varepsilon(z)R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z,\end{aligned}$$

which implies the following identity in $\mathbb{R}^N$

$$\sigma\delta \int_{\mathbb{R}}\int_{\mathbb{R}} \partial_x f(x,z)\rho^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z = \mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[\partial_z f(X^\varepsilon, Z^\varepsilon)(X^\varepsilon - Z^\varepsilon)\right].$$

Finally, choosing

$$f(x,z) = (x-z)V'(z) + V(z),$$

we obtain the desired result. □

4.9.2 Proof of Proposition 4.17

Preliminary estimates

In order to prove the characterization provided by Proposition 4.17, we need to prove two additional results on the filter. First, we prove a Jensen-like inequality for the kernel.

Lemma 4.28. *Let $\delta > 0$ and $k(r)$ be defined as*

$$k(r) = \frac{1}{\delta}e^{-r/\delta}.$$

Then, for any $t > 0$, $p \geq 1$ and any function $g \in \mathcal{C}^0([0,t])$ it holds

$$\left|\int_0^t k(t-s)g(s)\,\mathrm{d}s\right|^p \leq \int_0^t k(t-s)\,|g(s)|^p\,\mathrm{d}s.$$

Proof. Let us first note that

$$\int_0^t k(t-s)\,\mathrm{d}s = 1 - e^{-t/\delta}.$$

Therefore, the measure $\kappa_t(\mathrm{d}s)$ on $[0,t]$ defined as

$$\kappa_t(\mathrm{d}s) \coloneqq \frac{k(t-s)}{1-e^{-t/\delta}}\,\mathrm{d}s,$$

is a probability measure. An application of Jensen's inequality therefore yields

$$\begin{aligned}\left|\int_0^t k(t-s)g(s)\,\mathrm{d}s\right|^p &\leq (1-e^{-t/\delta})^p \int_0^t |g(s)|^p\,\kappa_t(\mathrm{d}s)\\ &= (1-e^{-t/\delta})^{p-1}\int_0^t k(t-s)\,|g(s)|^p\,\mathrm{d}s.\end{aligned}$$

Finally since $0 < (1-e^{-t/\delta}) < 1$ and $p \geq 1$, this implies the desired result. □

The following lemma characterizes the action of the filter when it is applied to polynomials.

Lemma 4.29. *With the notation of Lemma 4.28, it holds for all $p \geq 0$*

$$\int_0^t k(t-s)(t-s)^p\,\mathrm{d}s \leq C\delta^p,$$

where $C > 0$ is a positive constant independent of δ.

Proof. The change of variable $u = (t-s)/\delta$ yields

$$\int_0^t k(t-s)(t-s)^p\,\mathrm{d}s = \delta^p\int_0^{t/\delta} u^p e^{-u}\,\mathrm{d}u = \delta^p\gamma\left(p+1, \frac{t}{\delta}\right),$$

where γ is the lower incomplete Gamma function, which is bounded by the complete Gamma function $\Gamma(p+1)$ independently of the second argument. □

Proof of Proposition 4.17

Denoting $Y_t^\varepsilon \coloneqq X_t^\varepsilon/\varepsilon$, we will make use of the decomposition [112, Formula 5.8]

$$\begin{aligned} X_t^\varepsilon - X_s^\varepsilon = &- \int_s^t (\alpha \cdot V'(X_r^\varepsilon))(1 + \Phi'(Y_r^\varepsilon))\,\mathrm{d}r \\ &+ \sqrt{2\sigma} \int_s^t (1 + \Phi'(Y_r^\varepsilon))\,\mathrm{d}W_r - \varepsilon(\Phi(Y_t^\varepsilon) - \Phi(Y_s^\varepsilon)), \end{aligned} \tag{4.29}$$

which is obtained applying the Itô formula to Φ, the solution of the cell problem (3.4) Recall that by definition of Z_t^ε we have

$$X_t^\varepsilon - Z_t^\varepsilon = \int_0^t k(t-s)(X_t^\varepsilon - X_s^\varepsilon)\,\mathrm{d}s + e^{-t/\delta} X_t^\varepsilon.$$

Plugging the decomposition (4.29) into the equation above, we obtain

$$X_t^\varepsilon - Z_t^\varepsilon = I_1^\varepsilon(t) + I_2^\varepsilon(t) + I_3^\varepsilon(t) + I_4^\varepsilon(t),$$

where

$$\begin{aligned} I_1^\varepsilon(t) &\coloneqq - \int_0^t k(t-s) \int_s^t (\alpha \cdot V'(X_r^\varepsilon))(1 + \Phi'(Y_r^\varepsilon))\,\mathrm{d}r\,\mathrm{d}s, \\ I_2^\varepsilon(t) &\coloneqq \sqrt{2\sigma} \int_0^t k(t-s) \int_s^t (1 + \Phi'(Y_r^\varepsilon))\,\mathrm{d}W_r\,\mathrm{d}s, \\ I_3^\varepsilon(t) &\coloneqq -\varepsilon \int_0^t k(t-s)(\Phi(Y_t^\varepsilon) - \Phi(Y_s^\varepsilon))\,\mathrm{d}s, \\ I_4^\varepsilon(t) &= e^{-t/\delta} X_t. \end{aligned}$$

Let us analyze the terms above singularly. For $I_1^\varepsilon(t)$, one can show [112, Proposition 5.8]

$$\int_s^t (\alpha \cdot V'(X_r^\varepsilon))(1 + \Phi'(Y_r^\varepsilon))\,\mathrm{d}r = (t-s)(A \cdot V'(X_t^\varepsilon)) + R_1^\varepsilon(t-s),$$

where the remainder R_1^ε satisfies

$$\mathbb{E}^{\mu^\varepsilon}\left[|R_1^\varepsilon(t-s)|^p\right]^{1/p} \leq C(\varepsilon^2 + \varepsilon(t-s)^{1/2} + (t-s)^{3/2}). \tag{4.30}$$

Therefore, it holds

$$\begin{aligned} I_1^\varepsilon(t) &= -(A \cdot V'(X_t^\varepsilon)) \int_0^t k(t-s)(t-s)\,\mathrm{d}s + \int_0^t k(t-s) R_1^\varepsilon(t-s)\,\mathrm{d}s \\ &= -\delta(A \cdot V'(X_t^\varepsilon)) + e^{-t/\delta}(t+\delta)(A \cdot V'(X_t^\varepsilon)) + \widetilde{R}_1^\varepsilon(t), \end{aligned}$$

where we exploited the equality

$$\int_0^t k(t-s)(t-s)\,\mathrm{d}s = \delta - e^{-t/\delta}(t+\delta),$$

and where

$$\widetilde{R}_1^\varepsilon(t) \coloneqq \int_0^t k(t-s) R_1^\varepsilon(t-s)\,\mathrm{d}s.$$

Now, Lemma 4.28, the inequality (4.30) and Lemma 4.29 yield for all $p \geq 1$

$$\begin{aligned} \mathbb{E}^{\mu^\varepsilon}\left[\left|\widetilde{R}_1^\varepsilon(t)\right|^p\right] &\leq C \int_0^t k(t-s)\,\mathbb{E}^{\mu^\varepsilon}\left|R_1^\varepsilon(t-s)\right|^p\,\mathrm{d}s \\ &\leq C \int_0^t k(t-s)(\varepsilon^{2p} + \varepsilon^p (t-s)^{p/2} + (t-s)^{3p/2})\,\mathrm{d}s \\ &\leq C\left(\varepsilon^{2p} + \varepsilon^p \delta^{p/2} + \delta^{3p/2}\right), \end{aligned}$$

where C is a positive constant independent of ε and δ. Therefore, for δ sufficiently small, we get

$$\left(\mathbb{E}^{\varphi^\varepsilon}\,|I_1^\varepsilon(t)|^p\right)^{1/p} \leq C\left(\delta+\varepsilon^2+\varepsilon\delta^{1/2}+te^{-t/\delta}\right).$$

We now consider the second term. Let us introduce the notation

$$Q_t^\varepsilon := \int_0^t (1+\Phi'(Y_r^\varepsilon))\,\mathrm{d}W_r,$$

and therefore rewrite

$$I_2^\varepsilon(t) = \sqrt{2\sigma}\int_0^t k(t-s)(Q_t^\varepsilon - Q_s^\varepsilon)\,\mathrm{d}s.$$

An application of the Itô formula to $u(s,Q_s^\varepsilon)$ where $u(s,x)=k(t-s)x$ yields

$$\begin{aligned} I_2^\varepsilon(t) &= \sqrt{2\sigma}\left(Q_t^\varepsilon\int_0^t k(t-s)\,\mathrm{d}s - Q_t^\varepsilon + \delta\int_0^t k(t-s)\,(1+\Phi'(Y_s^\varepsilon))\,\mathrm{d}W_s\right) \\ &= \delta B_t^\varepsilon - \sqrt{2\sigma}e^{-t/\delta}Q_t^\varepsilon =: \delta B_t^\varepsilon - R_2^\varepsilon(t). \end{aligned} \tag{4.31}$$

where B_t^ε is defined in (4.18). For the remainder $R_2^\varepsilon(t)$, let us remark that for all $p\geq 1$ it holds

$$\mathbb{E}\left[|Q_t^\varepsilon|^p\right]^2 \leq \mathbb{E}\left[|Q_t^\varepsilon|^{2p}\right] \leq Ct^{p-1}\int_0^t \mathbb{E}\,|1+\Phi'(Y_r^\varepsilon)|^{2p}\,\mathrm{d}r \leq Ct^p$$

where we applied Jensen's inequality, an estimate for the moments of stochastic integrals [74, Formula (3.25), p. 163] and the boundedness of Φ. Therefore we have

$$\mathbb{E}^{\mu^\varepsilon}\left[|R_2^\varepsilon(t)|^p\right]^{1/p} \leq C\sqrt{t}e^{-t/\delta}. \tag{4.32}$$

In order to obtain the bound (4.19) on B_t^ε, let us remark that from (4.31) it holds for a constant $C>0$ depending only on p

$$\mathbb{E}\left[|B_t^\varepsilon|^p\right]^{1/p} \leq C\delta^{-1}\left(\mathbb{E}\,|I_2^\varepsilon(t)|^p\right)^{1/p} + C\delta^{-1}\left(\mathbb{E}\,|R_2^\varepsilon(t)|^p\right)^{1/p}.$$

The second term is bounded exponentially fast with respect to t and δ due to (4.32). For the first term, applying Lemma 4.28, the inequality [74, Formula (3.25), p. 163] and Lemma 4.29 we obtain for a constant $C>0$ independent of δ and t

$$\begin{aligned} \mathbb{E}\left[|I_2^\varepsilon(t)|^p\right] &\leq C\int_0^t k(t-s)\,\mathbb{E}\,|Q_t-Q_s|^p\,\mathrm{d}s \\ &\leq C\int_0^t k(t-s)(t-s)^{p/2}\,\mathrm{d}s \leq C\delta^{p/2}. \end{aligned}$$

Therefore, it holds for δ sufficiently small

$$\mathbb{E}\left[|B_t^\varepsilon|^p\right]^{1/p} \leq C\delta^{-1/2},$$

which proves the bound (4.19). Let us now consider $I_3^\varepsilon(t)$. Since Φ is bounded, we simply have

$$|I_3^\varepsilon(t)| \leq C\varepsilon,$$

almost surely. Finally, due to [112, Corollary 5.4], we know that X_t^ε has bounded moments of all orders and therefore

$$\mathbb{E}^{\mu^\varepsilon}\left[|I_4^\varepsilon(t)|^p\right]^{1/p} \leq Ce^{-t/\delta},$$

which concludes the proof. □

4.9.3 Proofs of Section 4.5

Preliminary estimates

The following lemma shows that Z^ε has bounded moments of all orders.

Lemma 4.30. *Under Assumption 4.3, let Z^ε be distributed as the invariant measure $\widetilde{\mu}^\varepsilon$ of the couple $(X^\varepsilon, Z^\varepsilon)^\top$. Then for any $p \geq 1$ there exists a constant $C > 0$ uniform in ε such that*

$$\mathbb{E}^{\widetilde{\mu}^\varepsilon} |Z^\varepsilon|^p \leq C.$$

Proof. Let X_t^ε be at stationarity with respect to its invariant measure μ^ε. Let Z_t^ε be the corresponding filtered process. By definition of Z_t^ε and applying Lemma 4.28 we have

$$\begin{aligned}\mathbb{E}^{\mu^\varepsilon} |Z_t^\varepsilon|^p &= \mathbb{E}^{\mu^\varepsilon} \left| \int_0^t k(t-s) X_s^\varepsilon \, \mathrm{d}s \right|^p \\ &\leq \int_0^t k(t-s) \, \mathbb{E}^{\mu^\varepsilon} |X_s^\varepsilon|^p \, \mathrm{d}s,\end{aligned}$$

which, together with the definition of k and the fact that X_s^ε has bounded moments of all orders [112, Corollary 5.4], implies for a constant $C > 0$

$$\mathbb{E}^{\mu^\varepsilon} |Z_t^\varepsilon|^p \leq C.$$

In order to conclude, we remark that due to Lemma 4.6 we have for all $t \geq 0$

$$\mathbb{E}^{\widetilde{\mu}^\varepsilon} |Z^\varepsilon|^p \leq \mathbb{E}^{\mu^\varepsilon} |Z_t^\varepsilon|^p + Ce^{-\lambda t},$$

which, for t sufficiently big, yields the desired result. □

Corollary 4.31 is a direct consequence of Proposition 4.17 and provides a rough estimate of the difference between the trajectories X_t^ε and Z_t^ε when they are at stationarity.

Corollary 4.31. *Under Assumption 4.3, let the couple $(X^\varepsilon, Z^\varepsilon)^\top$ be distributed as its invariant measure $\widetilde{\mu}^\varepsilon$. Then, if $\delta \leq 1$, it holds for any $p \geq 1$*

$$\left(\mathbb{E}^{\widetilde{\mu}^\varepsilon} |X^\varepsilon - Z^\varepsilon|^p \right)^{1/p} \leq C \left(\varepsilon + \delta^{1/2} \right),$$

for a constant $C > 0$ independent of ε and δ.

Proof. Let $p \geq 1$, then due to Proposition 4.17 there exists a constant $C > 0$ depending only on p such that

$$\mathbb{E}^{\mu^\varepsilon} |X_t^\varepsilon - Z_t^\varepsilon|^p \leq C \left(\varepsilon^p + \delta^{p/2} \right).$$

Let us now remark that this result holds for X_t^ε being at stationarity and for Z_t^ε being its filtered process, and not for a couple $(X^\varepsilon, Z^\varepsilon)^\top \sim \widetilde{\mu}^\varepsilon$. In order to conclude, we remark that due to Lemma 4.6 we have for all $t \geq 0$

$$\mathbb{E}^{\widetilde{\mu}^\varepsilon} |X^\varepsilon - Z^\varepsilon|^p \leq \mathbb{E}^{\widetilde{\mu}^\varepsilon} |X_t^\varepsilon - Z_t^\varepsilon|^p + Ce^{-\lambda t},$$

which, for t sufficiently big, yields the desired result. □

The result above can be in some sense rather counter-intuitive. Indeed, for a fixed $\varepsilon > 0$ and for $\delta \to 0$ independently of ε, one expects the filtered trajectory Z^ε to approach X^ε. This is provided by the following Lemma.

Lemma 4.32. *Under Assumption 4.3, let the couple $(X^\varepsilon, Z^\varepsilon)^\top$ be distributed as its invariant measure $\widetilde{\mu}^\varepsilon$. Then, if $\delta \leq 1$, it holds for any $p \geq 1$*

$$\left(\mathbb{E}^{\widetilde{\mu}^\varepsilon} |X^\varepsilon - Z^\varepsilon|^p\right)^{1/p} \leq C\left(\delta\varepsilon^{-1} + \delta^{1/2}\right),$$

for a constant $C > 0$ independent of ε and δ.

Proof. By equation (4.1) we have for all $0 \leq s < t$

$$X_t^\varepsilon - X_s^\varepsilon = -\alpha \int_s^t V'(X_r^\varepsilon)\,\mathrm{d}r - \frac{1}{\varepsilon}\int_s^t p'\left(\frac{X_r^\varepsilon}{\varepsilon}\right)\mathrm{d}r + \sqrt{2\sigma}(W_t - W_s).$$

Therefore, by Assumption 4.3 and since X_t^ε has bounded moments of all orders at stationarity [112, Corollary 5.4], it holds for any $p \geq 1$ and a constant $C > 0$

$$\mathbb{E}^{\mu^\varepsilon} |X_t^\varepsilon - X_s^\varepsilon|^p \leq C\left((t-s)^p + (t-s)^p\varepsilon^{-p} + (t-s)^{p/2}\right), \tag{4.33}$$

where μ^ε is the invariant measure of X^ε. By definition of Z_t^ε we have

$$X_t^\varepsilon - Z_t^\varepsilon = \int_0^t k(t-s)(X_t^\varepsilon - X_s^\varepsilon)\,\mathrm{d}s + e^{-t/\delta}X_t^\varepsilon,$$

which, applying Lemma 4.28, the inequality (4.33) and Lemma 4.29, implies

$$\begin{aligned}\mathbb{E}^{\mu^\varepsilon} |X_t^\varepsilon - Z_t^\varepsilon|^p &\leq C\left(\int_0^t k(t-s)\,\mathbb{E}^{\mu^\varepsilon} |X_t^\varepsilon - X_s^\varepsilon|^p\,\mathrm{d}s + e^{-pt/\delta}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon} |X_t^\varepsilon|^p\right)\\ &\leq C\left(\delta^p + \delta^p\varepsilon^{-p} + \delta^{p/2} + e^{-pt/\delta}\right).\end{aligned}$$

Geometric ergodicity (Lemma 4.6) then implies for $\widetilde{\mu}^\varepsilon$ the measure of the couple $(X^\varepsilon, Z^\varepsilon)^\top$

$$\mathbb{E}^{\widetilde{\mu}^\varepsilon} |X^\varepsilon - Z^\varepsilon|^p \leq \mathbb{E}^{\mu^\varepsilon} |X_t^\varepsilon - Z_t^\varepsilon|^p + Ce^{-\lambda t},$$

which, for t sufficiently big and since $\delta \leq 1$ yields the desired result. □

Let us conclude with a last preliminary estimate concerning the matrices $\widetilde{\mathcal{M}}_\varepsilon$ and $\mathcal{M}_\varepsilon$ defined in (4.12) and (4.11), respectively.

Lemma 4.33. *Let the assumptions of Corollary 4.31 hold. Then the matrices $\mathcal{M}_\varepsilon$ and $\widetilde{\mathcal{M}}_\varepsilon$ satisfy*

$$\left\|\mathcal{M}_\varepsilon - \widetilde{\mathcal{M}}_\varepsilon\right\|_2 \leq C\left(\varepsilon + \delta^{1/2}\right),$$

for a constant $C > 0$ independent of ε and δ.

Proof. Applying Jensen's and Cauchy–Schwarz inequalities we have

$$\begin{aligned}\left\|\mathcal{M}_\varepsilon - \widetilde{\mathcal{M}}_\varepsilon\right\|_2 &\leq \mathbb{E}^{\widetilde{\mu}^\varepsilon} \left\|(V'(Z^\varepsilon) - V'(X^\varepsilon)) \otimes V'(X^\varepsilon)\right\|_2\\ &\leq \left(\mathbb{E}^{\widetilde{\mu}^\varepsilon} \|V'(Z^\varepsilon) - V'(X^\varepsilon)\|_2^2\right)^{1/2}\left(\mathbb{E}^{\widetilde{\mu}^\varepsilon} \|V'(X^\varepsilon)\|_2^2\right)^{1/2}.\end{aligned}$$

The Lipschitz condition on V' together with the boundedness of the moments of X^ε and Corollary 4.31 yield for a constant $C > 0$

$$\left\| \mathcal{M}_\varepsilon - \widetilde{\mathcal{M}}_\varepsilon \right\|_2 \leq C \left(\mathbb{E}^{\widetilde{\mu}^\varepsilon} |Z^\varepsilon - X^\varepsilon|^2 \right)^{1/2} \leq C \left(\varepsilon + \delta^{1/2} \right),$$

which is the desired result. □

Proof of Lemma 4.19

Let us consider the following system of stochastic differential equations for the processes $X_t^\varepsilon, Z_t^\varepsilon, B_t^\varepsilon, Y_t^\varepsilon$

$$\begin{aligned}
\mathrm{d}X_t^\varepsilon &= -\alpha \cdot V'(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon} p'(Y_t^\varepsilon)\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \\
\mathrm{d}Z_t^\varepsilon &= \frac{1}{\delta}\left(X_t^\varepsilon - Z_t^\varepsilon\right)\,\mathrm{d}t, \\
\mathrm{d}B_t^\varepsilon &= -\frac{1}{\delta} B_t^\varepsilon\,\mathrm{d}t + \frac{\sqrt{2\sigma}}{\delta}(1 + \Phi'(Y_t^\varepsilon))\,\mathrm{d}W_t, \\
\mathrm{d}Y_t^\varepsilon &= -\frac{1}{\varepsilon}\alpha \cdot V'(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon^2} p'(Y_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2\sigma}}{\varepsilon}\,\mathrm{d}W_t,
\end{aligned}$$

whose generator $\widetilde{\mathcal{L}}_\varepsilon$ is given by

$$\begin{aligned}
\widetilde{\mathcal{L}}_\varepsilon = &- \left(\alpha \cdot V'(x) + \frac{1}{\varepsilon} p'(y) \right) \partial_x + \frac{1}{\delta}(x - z)\partial_z - \frac{1}{\delta} b \partial_b - \left(\frac{1}{\varepsilon}\alpha \cdot V'(x) + \frac{1}{\varepsilon^2} p'(y) \right) \partial_y \\
&+ \sigma \left(\partial_{xx}^2 + \frac{2}{\varepsilon}\partial_{xy}^2 + \frac{1}{\varepsilon^2}\partial_{yy}^2 \right) \\
&+ \sigma \left(\frac{2(1 + \Phi'(y))}{\delta}\partial_{xb}^2 + \frac{2(1 + \Phi'(y))}{\varepsilon\delta}\partial_{yb}^2 + \frac{(1 + \Phi'(y))^2}{\delta^2}\partial_{bb}^2 \right).
\end{aligned}$$

Let us denote by $\varphi^\varepsilon \colon \mathbb{R}^3 \times [0, L] \to \mathbb{R}$, $\varphi^\varepsilon = \varphi^\varepsilon(x, z, b, y)$, the density of the invariant measure η^ε of the quadruple $(X_t^\varepsilon, Z_t^\varepsilon, B_t^\varepsilon, Y_t^\varepsilon)$. Then φ^ε solves the stationary FPE $\widetilde{\mathcal{L}}_\varepsilon^* \varphi^\varepsilon = 0$ (see Section A.3), i.e., explicitly

$$\begin{aligned}
0 = &\partial_x \left(\left(\alpha \cdot V'(x) + \frac{1}{\varepsilon} p'(y) \right) \varphi^\varepsilon \right) + \frac{1}{\delta}\partial_z \left((z - x)\varphi^\varepsilon \right) \\
&+ \frac{1}{\delta}\partial_b (b\varphi^\varepsilon) + \partial_y \left(\left(\frac{1}{\varepsilon}\alpha \cdot V'(x) + \frac{1}{\varepsilon^2} p'(y) \right) \varphi^\varepsilon \right) \\
&+ \sigma \left(\partial_{xx}^2 \varphi^\varepsilon + \frac{2}{\varepsilon}\partial_{xy}^2 \varphi^\varepsilon + \frac{1}{\varepsilon^2}\partial_{yy}^2 \varphi^\varepsilon \right) \\
&+ \sigma \left(\frac{2}{\delta}\partial_{xb}^2 \left((1 + \Phi'(y))\varphi^\varepsilon \right) + \frac{2}{\varepsilon\delta}\partial_{yb}^2 \left((1 + \Phi'(y))\varphi^\varepsilon \right) \right) \\
&+ \sigma \left(\frac{1}{\delta^2}\partial_{bb}^2 \left((1 + \Phi'(y))^2 \varphi^\varepsilon \right) \right).
\end{aligned} \tag{4.34}$$

We now multiply the equation above by a continuous differentiable function $f \colon \mathbb{R}^2 \to \mathbb{R}^N$, $f = f(z, b)$, and integrate with respect to x, z, b and y. Then an integration by parts yields

$$\begin{aligned}
&\int_{\mathbb{R}^3 \times [0,L]} \partial_b f(z, b) b \varphi^\varepsilon \\
&\qquad = \int_{\mathbb{R}^3 \times [0,L]} \partial_z f(z, b)(x - z)\varphi^\varepsilon + \frac{\sigma}{\delta} \int_{\mathbb{R}^3 \times [0,L]} \partial_{bb}^2 f(z, b)(1 + \Phi'(y))^2 \varphi^\varepsilon,
\end{aligned}$$

which implies the following identity in $\mathbb{R}^N$

$$\begin{aligned}\delta\,\mathbb{E}^{\eta^\varepsilon}\left[\partial_b f(Z^\varepsilon,B^\varepsilon)B^\varepsilon\right] &= \sigma\,\mathbb{E}^{\eta^\varepsilon}\left[\partial^2_{bb} f(Z^\varepsilon,B^\varepsilon)(1+\Phi'(Y^\varepsilon))\right] \\ &\quad + \delta\,\mathbb{E}^{\eta^\varepsilon}\left[\partial_z f(Z^\varepsilon,B^\varepsilon)(X^\varepsilon - Z^\varepsilon)\right].\end{aligned} \tag{4.35}$$

Choosing

$$f(z,b) = \frac{1}{2}b^2 V''(z),$$

we obtain

$$\begin{aligned}\delta\,\mathbb{E}^{\eta^\varepsilon}\left[(B^\varepsilon)^2 V''(Z^\varepsilon)\right] &= \sigma\,\mathbb{E}^{\eta^\varepsilon}\left[V''(Z^\varepsilon)(1+\Phi'(Y^\varepsilon))\right] + \frac{\delta}{2}\,\mathbb{E}^{\eta^\varepsilon}\left[(B^\varepsilon)^2 V'''(Z^\varepsilon)(X^\varepsilon - Z^\varepsilon)\right] \\ &=: \sigma\,\mathbb{E}^{\eta^\varepsilon}\left[V''(Z^\varepsilon)(1+\Phi'(Y^\varepsilon))\right] + \widetilde{R}(\varepsilon,\delta).\end{aligned}$$

We now consider the remainder and, applying Hölder's inequality, Corollary 4.31, Lemma 4.30, Assumption 4.18 and (4.19), we get for p,q,r such that $1/p+1/q+1/r=1$

$$\left|\widetilde{R}(\varepsilon,\delta)\right| \le C\delta\left(\mathbb{E}^{\eta^\varepsilon}\left|B^\varepsilon\right|^{2p}\right)^{1/p}\left(\mathbb{E}^{\eta^\varepsilon}\left|V'''(Z^\varepsilon)\right|^q\right)^{1/q}\left(\mathbb{E}^{\eta^\varepsilon}\left|X^\varepsilon - Z^\varepsilon\right|^r\right)^{1/r} \le C(\delta^{1/2}+\varepsilon),$$

which completes the proof. □

Proof of Lemma 4.20

Let us introduce the notation

$$\Delta(\varepsilon) = \left|\sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(Z^\varepsilon)(1+\Phi'(Y^\varepsilon))^2] - \mathcal{M}_0 A\right|,$$

and note that the aim is to show that $\lim_{\varepsilon\to 0}\Delta(\varepsilon)=0$. By the triangle inequality we get

$$\begin{aligned}\Delta(\varepsilon) \le &\left|\sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(Z^\varepsilon)(1+\Phi'(Y^\varepsilon))^2] - \sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(X^\varepsilon)(1+\Phi'(Y^\varepsilon))^2]\right| \\ &+ \left|\sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(X^\varepsilon)(1+\Phi'(Y^\varepsilon))^2] - \Sigma\,\mathbb{E}^{\varphi^0}[V''(X)]\right| \\ =:&\Delta_1(\varepsilon)+\Delta_2(\varepsilon).\end{aligned}$$

We first study $\Delta_1(\varepsilon)$ and due to the boundedness of Φ', Assumption 4.18 and Corollary 4.31 we have

$$\Delta_1(\varepsilon) \le C\,\mathbb{E}^{\eta^\varepsilon}\left|X^\varepsilon - Z^\varepsilon\right| \le C(\delta^{1/2}+\varepsilon) = C(\varepsilon^{\zeta/2}+\varepsilon),$$

which implies

$$\lim_{\varepsilon\to 0}\Delta_1(\varepsilon) = 0.$$

We now consider $\Delta_2(\varepsilon)$. Integrating equation (4.34) with respect to z and b we obtain the Fokker–Planck equation for the stationary marginal density $\lambda\colon \mathbb{R}\times[0,L]$, $\lambda=\lambda(x,y)$, of the couple $(X^\varepsilon,Y^\varepsilon)$

$$\begin{aligned}\partial_x\left(\left(\alpha\cdot V'(x)+\frac{1}{\varepsilon}p'(y)\right)\lambda\right) + \partial_y\left(\left(\frac{1}{\varepsilon}\alpha\cdot V'(x)+\frac{1}{\varepsilon^2}p'(y)\right)\lambda\right) \\ +\sigma\left(\partial^2_{xx}\lambda + \partial^2_{xy}\left(\frac{2}{\varepsilon}\lambda\right) + \partial^2_{yy}\left(\frac{1}{\varepsilon^2}\lambda\right)\right) = 0,\end{aligned}$$

whose solution is given by

$$\lambda(x,y) = \frac{1}{C_\lambda}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma} - \frac{1}{\sigma}p(y)\right),$$

where

$$\begin{aligned} C_\lambda &= \int_{\mathbb{R}}\int_0^L \exp\left(-\frac{\alpha\cdot V(x)}{\sigma} - \frac{1}{\sigma}p(y)\right)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int_{\mathbb{R}} \exp\left(-\frac{\alpha\cdot V(x)}{\sigma}\right)\,\mathrm{d}x\right)\left(\int_0^L \exp\left(-\frac{1}{\sigma}p(y)\right)\,\mathrm{d}y\right) \\ &=: C_{\lambda_x} C_{\lambda_y}. \end{aligned}$$

Therefore, since $\Sigma = K\sigma$ and by (3.3) and Proposition 3.3 we have

$$\begin{aligned} \sigma\,\mathbb{E}^{\eta^\varepsilon}[V''(X^\varepsilon)(1+\Phi'(Y^\varepsilon))^2] &= \sigma\left(\int_{\mathbb{R}} V''(x)\frac{1}{C_{\lambda_x}}\exp\left(-\frac{\alpha\cdot V(x)}{\sigma}\right)\,\mathrm{d}x\right) \\ &\quad\times\left(\int_0^L (1+\Phi'(y))^2\frac{1}{C_{\lambda_y}}\exp\left(-\frac{1}{\sigma}p(y)\right)\,\mathrm{d}y\right) \\ &= \sigma K\,\mathbb{E}^{\mu^0}[V''(X)] = \Sigma\,\mathbb{E}^{\mu^0}[V''(X)]. \end{aligned}$$

Moreover, by (3.34) we have $\Sigma\,\mathbb{E}^{\mu^0}[V''(X)] = \mathcal{M}_0 A$, which shows that $\Delta_2(\varepsilon) = 0$ and completes the proof. □

Proof of Theorem 4.22

Let us consider the decomposition (4.13), i.e.,

$$\widehat{A}_k(X^\varepsilon, T) = \alpha + I_1^\varepsilon(T) - I_2^\varepsilon(T),$$

where $I_1^\varepsilon(T)$ is defined in (4.13) and satisfies

$$\lim_{T\to\infty} I_1^\varepsilon(T) = \widetilde{\mathcal{M}}_\varepsilon^{-1}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(Z^\varepsilon)\right], \quad \text{a.s.}$$

and, by the proof of Theorem 4.15 we have independently of ε

$$\lim_{T\to\infty} I_2^\varepsilon(T) = 0, \quad \text{a.s.}$$

A Taylor expansion of the first order of V' yields

$$V'(Z^\varepsilon) = V'(X^\varepsilon) + V''(\widetilde{X}^\varepsilon)(Z^\varepsilon - X^\varepsilon),$$

where $\widetilde{X}^\varepsilon$ is a random variable which assumes values between X^ε and Z^ε. We can therefore write

$$\begin{aligned} \lim_{T\to\infty} I_1^\varepsilon(T) &= \widetilde{\mathcal{M}}_\varepsilon^{-1}\left(\frac{1}{\varepsilon}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(X^\varepsilon)\right] + \frac{1}{\varepsilon}\,\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left[p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V''(\widetilde{X}^\varepsilon)(Z^\varepsilon - X^\varepsilon)\right]\right) \\ &=: \widetilde{\mathcal{M}}_\varepsilon^{-1}\left(J_1^\varepsilon + J_2^\varepsilon\right). \end{aligned}$$

We now consider the two terms separately and show they vanish for $\varepsilon \to 0$. Proceeding as in the proof of Theorem 3.13, we directly obtain that

$$\lim_{\varepsilon\to 0} \widetilde{\mathcal{M}}_\varepsilon^{-1} J_1^\varepsilon = 0. \tag{4.36}$$

We now turn to J_2^ε. The Hölder's inequality with conjugate exponents p and q and the assumptions on p and V yield

$$|J_2^\varepsilon| \le C\varepsilon^{-1}\left(\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left|\widetilde{X}^\varepsilon\right|^q\right)^{1/q}\left(\mathbb{E}^{\widetilde{\mu}^\varepsilon}\left|Z^\varepsilon - X^\varepsilon\right|^p\right)^{1/p}.$$

Since $\widetilde{X}^\varepsilon$ assumes values between X^ε and Z^ε, it has bounded moments by [112, Corollary 5.4] and Lemma 4.30. Hence, applying Lemma 4.32 we have

$$|J_2^\varepsilon| \leq C\left(\delta\varepsilon^{-2} + \delta^{1/2}\varepsilon^{-1}\right),$$

which, since $\delta = \varepsilon^\zeta$ with $\zeta > 2$, implies

$$\lim_{\varepsilon\to 0} |J_2^\varepsilon| = 0. \tag{4.37}$$

Finally, Lemma 4.33 and the weak convergence of the invariant measure μ^ε to μ^0 imply

$$\lim_{\varepsilon\to 0} \widetilde{\mathcal{M}}_\varepsilon = \mathcal{M}_0,$$

which, together with (4.36), (4.37) implies that $I_1^\varepsilon(T) \to 0$ for $T \to \infty$ and $\varepsilon \to 0$, which implies the desired result. □

5 Conclusion of Part I

In this first part of the thesis we introduced two novel methods to solve inverse problems in the context of multiscale elliptic PDEs and multiscale SDEs of the Langevin type. Despite both inverse problems being amenable to the theory of homogenization, the misspecification due to the replacement of a multiscale equation with a single-scale effective model has to be treated in a substantially different manner in the PDE and the SDE cases. In particular, for elliptic PDEs the theory of homogenization translates naturally to inverse problems, and multiscale data are assimilated seamlessly by a coarse-grained model. Conversely, if data arising from a multiscale diffusion process are directly employed to infer the parameters of an effective SDE, the resulting coefficients are biased with respect to the one predicted by homogenization theory. In this case, it is necessary to employ additional techniques of pre-processing of the data in order to correct this issue of unbiasedness.

In Chapter 1 we gave a brief but general introduction to Bayesian inverse problems, with which we lay the basis for our notation and for some concepts which are repeatedly employed throughout the thesis. In particular, we focused on finite-dimensional approximations of otherwise infinite-dimensional inverse problems and on practical techniques to compute their solution. We remark that more complete and deeper introductions to the topic of inverse problems and their Bayesian interpretation have been presented in the literature (e.g. [49, 71, 131]).

In Chapter 2 we presented a methodology based on numerical homogenization and on the EnKF for inverse problems involving multiscale elliptic PDEs with tensors highly oscillatory at a scale $\varepsilon \ll 1$. By combining the EnKF and the FE-HMM we managed to significantly reduce the computational cost for elliptic inverse problems which would be otherwise computationally involved or unfeasible. Our main theoretical result was showing that the ensemble of particles approximating the unknown parameter generated by our multiscale algorithm converges to the ensemble generated by the true model as the small scale parameter ε and the numerical discretization parameter h go to zero. In particular, Theorem 2.8 gives guarantees in the asymptotic regime of pointwise estimations derived from the EnKF, while Theorem 2.10 enables to deduce convergence when the EnKF is recast into a Bayesian framework. Hence when $\varepsilon \ll 1$ and the full model is expensive to solve, the multiscale numerical method we propose is both accurate and efficient to solve inverse problems involving in multiscale elliptic PDEs.

If the scale-separation parameter ε is not in the asymptotic regime, or the discretization is not refined, then our method may fail to retrieve the unknown due to the misspecification of the model. In order to alleviate this issue, we equipped our method with a technique which allows to account for the discrepancy between the artificial homogenized surrogate forward model and the true multiscale data. This technique requires additional offline or online computations involving the

numerical solution of the full multiscale problem. The optimal number of such additional solves is quantified in Theorem 2.23 and Theorem 2.24. In particular, we have proved that the number of solves needed to reach any required accuracy tends to zero when the small scale parameter ε and the numerical discretization parameter h vanish. Hence, we were able to conclude that accounting for model misspecification is particularly beneficial for mid-range values of ε, when a small number of full solves should be computationally affordable.

In Chapter 3 we introduced multiscale diffusion processes of the overdamped Langevin kind, and the problem of inferring their drift coefficient. In particular, we focused on the ergodic properties of such stochastic processes, and we have shown a traditional homogenization result, which guarantees that there exists an effective SDE that captures the slow variation of the multiscale model. Then, we considered the inference problem for the drift coefficient and we showed a derivation based on Girsanov's change of measure formula of the MLE for the drift of the effective model. We then considered the case of interest, i.e., when multiscale data are observed and inserted in the estimation procedure. In this case, we have shown that the MLE is asymptotically biased, and that therefore the theory of homogenization does not translate from the forward to the inverse problem. We concluded by showing how this issue of model misspecification also affects the Bayesian setting, which naturally arises in this scenario.

In Chapter 4 we presented a novel methodology based on filtered data for the estimation of the drift of the homogenized diffusion process, which allows to overcome the issue of model misspecification and the asymptotic biasedness of estimators when one is confronted with multiscale data. In particular, we proved asymptotic unbiasedness of estimators drawn from our methodology. Moreover, we were able to derive a modified Bayesian approach which guarantees robust uncertainty quantification and posterior contraction, based on the same filtered data approach. Numerical experiments demonstrate how the estimator based on filtered data requires less knowledge of the characteristic time-scales of the multiscale equation with respect to subsampling, and how it can be employed as a black-box tool for parameter estimation on a range of academic examples.

The topics of Bayesian inverse problems and parameter inference for deterministic and stochastic multiscale models have been widely explored in the literature. Nevertheless, we believe that several research directions could be interesting and relevant. In particular, we think it would be worth to:

(i) Adapt the EnKF/FE-HMM approach to a wider class of multiscale inverse problems, involving parabolic or hyperbolic differential equations;

(ii) Apply the EnKF/FE-HMM methodology to real-world data to demonstrate their practical usefulness;

(iii) Analyze the filtered data approach when the coefficient $\beta > 1$ in (4.4), which seems to give more robust results in practice. Other implementations of low-pass filters could be of interest, too;

(iv) Extend the analysis of Chapter 4 to the non-parametric framework, i.e., adapt the methodology to infer the drift function. This could be most likely achieved by means of Bayesian regularization techniques;

Part II
Probabilistic Methods for Differential Equations

The second part of this thesis is devoted to probabilistic numerical methods for differential equations.

Probabilistic numerics (PN) is a relatively new and rapidly expanding field of numerical analysis, whose main aim is introducing appropriate probability measures over approximate solutions of otherwise deterministic problems. The underlying goal is obtaining richer information from a numerical method by equipping its pointwise output with a full quantification of the uncertainty due to approximate computations. Characterizing the error in a probabilistic fashion is particularly helpful in case the numerical method is one component of a more complex pipeline of computations, such as inverse problems.

In Chapter 6 we introduce the field of PN, setting an abstract framework for probabilistic numerical schemes and introducing the notation which we employ in the subsequent chapters. Specifically, we consider in Chapter 6 the two sub-fields of perturbation-based and measure-valued probabilistic schemes, and define some of their properties. We focus primarily on perturbation-based methods, for which we prove a novel result on Monte Carlo estimators, which implies that any family of draws from the probabilistic solution has a good quality regardless of the sample size. We then explain how it is in practice possible to combine PN schemes with the Bayesian paradigm in order to enhance the solution of inverse problems. We finally conclude by detailing a probabilistic numerical scheme for ODEs for each of the two sub-fields of PN, arguing about their respective advantages and disadvantages.

In Chapter 7 we introduce the random time step Runge–Kutta method (RTS-RK), a perturbation-based probabilistic method for geometric and chaotic ODEs. The conception and analysis of the RTS-RK is one of the original contributions of this thesis. After analyzing in detail the convergence properties of the RTS-RK, we focus on its geometric features, such as the exact conservation of polynomial invariants and, most importantly, its symplecticity for Hamiltonian equations. In particular, the RTS-RK is to our knowledge the first geometry-aware method to be proposed in the field of PN. The potential of the RTS-RK in the context of geometric ODEs and inference problems is illustrated by an exhaustive series of numerical experiments.

In Chapter 8 we present the random mesh finite element method (RM-FEM), a perturbation-based probabilistic numerical scheme for elliptic PDEs, which is one of the original contributions of this thesis. The RM-FEM shares its fundamental idea with the RTS-RK, i.e., to randomly perturb the discretization instead of the solution itself, which is instead the main strategy in the literature for deriving perturbation-based probabilistic solvers for differential equations. Our main contribution in this chapter is deriving a posteriori error estimators which are entirely based on probabilistic information and which are rigorously justified in the one-dimensional case. Numerical experiments illustrate the correctness of these estimators in several academic examples, and the potential of the RM-FEM when employed together with the Bayesian approach to solve inverse problems.

Finally, in Chapter 9 we draw our conclusions and give suggestion for possible future developments.

We suggest a reader who is mainly interested in PN solvers for elliptic PDEs and therefore wishes to skip Chapter 7, to read Chapter 6 to get acquainted with our notation and to have a reference frame for the field of PN.

6 An Introduction to Probabilistic Numerics

In this chapter we give an introduction to probabilistic numerics (PN). The main idea associating all contributions to the field of PN is to introduce a probability measure over the solution of traditional numerical methods. The underlying rationale is to quantify the uncertainty due to numerical errors in a probabilistic manner, rather than with standard error estimates. Indeed, a probability measure over approximate solutions can be readily pushed through a pipeline of computations, thus justifying the need of probabilistic methods especially when the solution of the problem at hand is employed as the input of a subsequent analysis. Let us refer the reader to the review papers [38, 65, 102], which both give an historical framework for the field of PN and summarize the most recent developments, and which partially inspired us in the writing of this chapter. Some notation and results, especially concerning the application of probabilistic methods to Bayesian inverse problems, are taken from [87]. Moreover, let us remark that some phrasings employed here are borrowed from our articles [6, 7], and that Theorem 6.7 is a generalization of our result [6, Theorem 3].

The outline of this chapter is as follows. In Section 6.1 we define some fundamental properties and desiderata of probabilistic methods, with a particular focus on convergence and on Monte Carlo estimators. We proceed by describing in Section 6.2 how probabilistic methods can be employed in the framework of Bayesian inverse problems (see Chapter 1 for an introduction), which has been identified in several contributions to the literature of PN as one of the most successful applications of this class of numerical schemes. We conclude this chapter with Section 6.3, in which we describe two probabilistic methods for ODEs belonging to two different sub-classes of PN.

6.1 Definition and Properties of Probabilistic Methods

In this section, we first introduce an abstract and generic framework for probabilistic numerical methods, and then define their convergence properties.

Let X and Y be possibly infinite-dimensional Banach spaces, which we call respectively the input and output spaces. One can think of the space X as a container for all the relevant data of a specific problem, whose solution (or a quantity of interest derived from the solution) lies in the space Y. Let moreover $\mathcal{G}\colon X \to Y$ be a function, which we call the forward map. We consider scenarios where it is not possible to evaluate $\mathcal{G}$ exactly as, for example, it is the case when evaluating $\mathcal{G}$ involves the solution of a differential problem. We then let $h > 0$ be a discretization parameter and call a map $\mathcal{G}_h\colon X \to Y$ a numerical method for $\mathcal{G}$ if $\mathcal{G}_h$ approximates $\mathcal{G}$ on Y.

The field of PN can be roughly split in two classes of methods, whose purpose is common but whose definition is intrinsically different. In particular, the shared goal is accounting in a probabilistic manner for the uncertainty due to discretization of the forward map $\mathcal{G}$. The first class is the one of perturbation-based probabilistic methods, which are constructed on existing numerical schemes, then randomized by appropriate perturbations. The second class is the one of the measure-valued probabilistic methods, which build a probability measure on the output space Y, often adopting a Bayesian approach. We now give a definition of the two classes.

Definition 6.1. Let $h > 0$ be a discretization parameter and let $(\Omega, \mathcal{F}, P)$ be a probability space. Given a map $\mathcal{G}\colon X \to Y$, we call a random variable $\widetilde{\mathcal{G}}_h\colon \Omega \times X \to Y$ a perturbation-based numerical method if for all $\omega \in \Omega$ the map $\widetilde{\mathcal{G}}_h(\omega, \cdot)\colon X \to Y$ is a numerical method for $\mathcal{G}$.

Definition 6.2. Let $h > 0$ be a discretization parameter, let $\mathcal{F}(Y)$ be a σ-algebra on Y and let $\mathcal{M}(Y)$ be the space of probability measures on $(Y, \mathcal{F}(Y))$. Given a map $\mathcal{G}$, we call a measure-valued probabilistic method a map $\widetilde{\mathcal{G}}_h\colon X \to \mathcal{M}(Y)$ such that for all $u \in X$ the measure $\widetilde{\mathcal{G}}_h(u) \in \mathcal{M}(Y)$ contains information on $\mathcal{G}$.

Remark 6.3. The definitions above are abstract and do not agree with other definitions which could be found in literature. For example, in the review article [38] the authors restrict the scope to methods which are, in some sense, Bayesian.

Remark 6.4. In practice, the output of measure-valued probabilistic methods is often restricted to subspaces of $\mathcal{M}(Y)$ of families of measures which are fully determined by a finite number of parameters. In particular, many probabilistic methods which have been proposed in the literature are tailored to ouptut Gaussian measures on Y, which are chosen due to their versatility and natural aptitude to the Bayesian framework.

In the following, we mainly focus on sampling-based probabilistic methods and their properties. Indeed, the two probabilistic schemes for ODEs and PDEs which we present in Chapters 7 and 8, respectively, can be ascribed to this class. Nevertheless, we describe for completeness in Section 6.3.2 a numerical scheme which belongs to the class of measure-valued probabilistic methods.

6.1.1 Convergence

In this section, we define the notion of convergence which can be employed for perturbation-based probabilistic numerical methods. Let us recall that for a deterministic numerical method $\mathcal{G}_h\colon X \to Y$, it is customary to say that it has order of convergence r with respect to h if it holds for all $u \in X$

$$\|\mathcal{G}_h(u) - \mathcal{G}(u)\|_Y \leq Ch^r,$$

for a constant $C > 0$ which is independent of h, but which possibly depends on $u \in X$.

Let us introduce some notation. For a sampling-based numerical method, we denote by $\nu_h \in \mathcal{M}(Y)$ the measure on $(Y, \mathcal{F}(Y))$ which is induced by $\widetilde{\mathcal{G}}_h$, i.e.

$$\nu_h(F) = P(\widetilde{\mathcal{G}}_h(\cdot, u) \in F),$$

for all $F \in \mathcal{F}(Y)$. The measure ν_h clearly depends on the data $u \in X$, but we omit the dependence of ν_h on u for economy of notation. For the same reason, we omit in the following the dependence of $\widetilde{\mathcal{G}}_h$ on the event $\omega \in \Omega$ and simply write $\widetilde{\mathcal{G}}_h(u)$. We moreover denote by $\mathbb{E}^{\nu_h}$ the expectation with respect to ν_h. Endowed with the probability measure ν_h, we can employ for probabilistic methods the standard definitions of weak and mean-square convergence which are familiar, for example, in the literature on numerical methods for SDEs (see e.g. the standard references [79, 95, 96]).

Definition 6.5. The probabilistic method $\widetilde{\mathcal{G}}_h$ has weak order of convergence $r > 0$ if for any sufficiently smooth function $\Psi\colon \mathbb{R}^d \to \mathbb{R}$ there exists a constant $C > 0$ independent of h such that

$$\left|\mathbb{E}^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}(u))\right| \leq Ch^r,$$

holds for all $u \in X$.

Definition 6.6. The probabilistic method $\widetilde{\mathcal{G}}_h$ has mean-square order of convergence $r > 0$ if there exists a constant $C > 0$ independent of h such that

$$\mathbb{E}^{\nu_h}\left[\left\|\widetilde{\mathcal{G}}_h(u) - \mathcal{G}(u)\right\|_Y^2\right]^{1/2} \leq Ch^r,$$

holds for all $u \in X$.

Let us comment on the two definitions above. First, it is clear by Jensen's inequality that mean-square convergence is stronger than weak convergence, and a method with mean-square order r has weak order r, too. Second, we can in both cases separate the effects due discretization and those due to the randomization of the method. In particular, for weak convergence we have by the triangle inequality

$$\left|\mathbb{E}^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}(u))\right| \leq \left|\Psi(\mathcal{G}_h(u)) - \Psi(\mathcal{G}(u))\right| + \left|\mathbb{E}^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}_h(u))\right|,$$

so that the first term can be bounded by considering the convergence of the numerical method, and the second term by analyzing the random perturbation introduced by the probabilistic scheme. For the mean-square convergence, we have

$$\mathbb{E}^{\nu_h}\left[\left\|\widetilde{\mathcal{G}}_h(u) - \mathcal{G}(u)\right\|_Y^2\right] \leq 2\left\|\mathcal{G}_h(u) - \mathcal{G}(u)\right\|_Y^2 + 2\,\mathbb{E}^{\nu_h}\left[\left\|\widetilde{\mathcal{G}}_h(u) - \mathcal{G}_h(u)\right\|_Y^2\right],$$

and again the two terms are due to discretization and randomization, respectively. One of the goals in probabilistic methods is to find the correct tuning of the random perturbations so that the two terms above are balanced with respect to the discretization parameter h, and therefore the randomization of the scheme is consistent in accounting for numerical errors in a probabilistic manner.

A further goal, which has been little explored in the literature of PN, is the one of employing the probabilistic map $\widetilde{\mathcal{G}}_h$ to derive a posteriori estimators for the numerical error. This is the main topic of Chapter 8, where we derive an a posteriori error estimators based on the finite element method (FEM) for an elliptic PDE.

6.1.2 A Result on Monte Carlo Estimators

In order to obtain an approximation of the map $\mathcal{G}$ with a perturbation-based numerical method, one has to recur to a Monte Carlo approximation. Indeed, the expectation $\mathbb{E}^{\nu_h}$ is not computable, as these methods often involve one or more high-dimensional support random variables in their definition. This has been identified as a weakness for perturbation-based methods with respect to measure-valued probabilistic schemes, since the latter generate a probability measure on the output deterministically and without the practical need of any sampling strategy. An exhaustive discussion about this issue can be found in [76, Section 2]. In this section, we present a result on Monte Carlo estimators drawn from the measure ν_h on Y induced by a convergent perturbation-based probabilistic method. In particular, we show that the quality of Monte Carlo estimators

is in this context driven by the discretization parameter h, and can be thought of as being independent of the dimension of the sample.

Let us remark that the following result and its proof can be found in our article [6, Theorem 3], where it is specialized to the probabilistic method of ODEs, which is the focus of Chapter 7

Theorem 6.7. *Let $\mathcal{G}\colon X \to Y$ be a forward map and let $h > 0$ be a discretization parameter. Moreover, let $\widetilde{\mathcal{G}}_h\colon \Omega \times X \to Y$ be a perturbation-based probabilistic method in the sense of Definition 6.1 with weak and mean-square order of convergence p_w and p_s, respectively. Let moreover $\Psi\colon Y \to \mathbb{R}$ be a Lipschitz continuous function, let $u \in X$ and let for a positive integer M*

$$E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] \coloneqq \frac{1}{M}\sum_{i=1}^{M}\Psi\left(\widetilde{\mathcal{G}}_h^{(i)}(u)\right),$$

where $\{\widetilde{\mathcal{G}}_h^{(i)}\}_{i=1}^M$ are i.i.d. realizations of the probabilistic method. Then, it holds

$$\mathbb{E}^M\left[\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}(u))\right)^2\right]^{1/2} \leq C\left(h^{p_w} + \frac{h^{p_s}}{\sqrt{M}}\right),$$

where $\mathbb{E}^M$ denotes expectation with respect to the sample, and where C is a positive constant independent of h and M, but possibly depending on u.

Proof. We drop in the proof for economy of notation the dependence on $u \in X$, and we introduce the mean-square error (MSE)

$$\mathrm{MSE} \coloneqq \mathbb{E}^M\left[\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right] - \Psi(\mathcal{G})\right)^2\right]. \tag{6.1}$$

The variance-bias decomposition of the MSE yields

$$\mathrm{MSE} = \mathrm{Var}^M\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right]\right) + \left(\mathbb{E}^M\left[E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right]\right] - \Psi(\mathcal{G})\right)^2,$$

where Var^M denotes variance with respect to the sample. For the second term, due to the unbiasedness of Monte Carlo estimators it holds

$$\mathbb{E}^M\left[E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right]\right] - \Psi(\mathcal{G}) = \mathbb{E}^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right] - \Psi(\mathcal{G}),$$

and hence, since the method has weak order p_w, we obtain

$$\mathrm{MSE} \leq \mathrm{Var}^M\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right]\right) + Ch^{2p_w}. \tag{6.2}$$

Moreover, since the samples are i.i.d. and distributed as $\widetilde{\mathcal{G}}_h$, the variance satisfies

$$\mathrm{Var}^M\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h)\right]\right) = \frac{1}{M}\mathrm{Var}^{\nu_h}\left(\Psi(\widetilde{\mathcal{G}}_h)\right), \tag{6.3}$$

where Var^{ν_h} denotes variance with respect to the measure ν_h. Since $\mathcal{G}$ is deterministic, and hence independent of ν_h, the variance of the estimator can then be bounded by exploiting the Lipschitz continuity of Ψ and the independence of the samples as

$$\begin{aligned}\mathrm{Var}^{\nu_h}\left(\Psi(\widetilde{\mathcal{G}}_h)\right) &= \mathrm{Var}^{\nu_h}\left(\Psi(\widetilde{\mathcal{G}}_h) - \Psi(\mathcal{G})\right)\\ &\leq \mathbb{E}^{\nu_h}\left[\left(\Psi(\widetilde{\mathcal{G}}_h) - \Psi(\mathcal{G})\right)^2\right]\\ &\leq C\,\mathbb{E}^{\nu_h}\left[\left\|\widetilde{\mathcal{G}}_h - \mathcal{G}\right\|_Y^2\right],\end{aligned}$$

where $C > 0$ is independent of h. Finally, since $\widetilde{\mathcal{G}}_h$ has mean-square order p_s, we get

$$\mathrm{Var}^{\nu_h}\left(\Psi(\widetilde{\mathcal{G}}_h)\right) \leq Ch^{2p_s},$$

which, together with (6.1), (6.2) and (6.3), proves the desired result. □

Remark 6.8. The first clear implication of Theorem 6.7 is that

$$\lim_{h\to 0} \mathbb{E}^M\left[\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}(u))\right)^2\right]^{1/2} = 0,$$

so that if the probabilistic discretization $\widetilde{\mathcal{G}}_h$ of the forward map $\mathcal{G}$ is refined, then the Monte Carlo estimator is close to the true value regardless of the number of samples M. For a general $h > 0$, it is customary to choose the number of samples M so that the sources of bias and variance are balanced. In this case, a reasonable choice is

$$M = \mathcal{O}\left(h^{2(p_s - p_w)}\right), \tag{6.4}$$

so that it holds

$$\mathbb{E}^M\left[\left(E_M^{\nu_h}\left[\Psi(\widetilde{\mathcal{G}}_h(u))\right] - \Psi(\mathcal{G}(u))\right)^2\right]^{1/2} \leq Ch^{p_w}.$$

Let us remark that it is often possible to tune probabilistic methods (see e.g. [6, 7, 39]) so that $p_s = p_w$. In this case, the indication provided by (6.4) is $M = \mathcal{O}(1)$ with respect to h, i.e., one can fix the number of samples to be independent of the discretization parameter.

6.2 Probabilistic Numerics and Bayesian Inverse Problems

We consider in this section the application of probabilistic numerical methods for the solution of Bayesian inverse problems, which we have introduced in Chapter 1. We restrict the scope of the discussion here to problems where the output space Y is finite-dimensional, and in particular $Y \equiv \mathbb{R}^L$. Given a forward map $\mathcal{G}\colon X \to Y$, we recall that inverse problems can be stated as

$$\text{find } u \in X \text{ given observations } y = \mathcal{G}(u) + \eta,$$

where we assume $\eta \sim \mathcal{N}(0, \Gamma)$ for a positive-definite covariance matrix Γ on Y. Given a prior measure $\mu_0 = \mathcal{N}(m_0, C_0)$ on X, we then have that the solution to the inverse problem is given under Assumption 1.1 by the posterior distribution μ whose Radon–Nykodim derivative with respect to μ_0 is given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(u) = \frac{1}{Z}\exp\left(-\Phi(u; y)\right), \tag{6.5}$$

for all $u \in X$, where Z is the normalization constant

$$Z = \int_X \exp(-\Phi(u; y))\,\mathrm{d}\mu_0(u),$$

and where the potential Φ is defined as

$$\Phi(u; y) = \frac{1}{2}\left\|\Gamma^{-1/2}\left(\mathcal{G}(u) - y\right)\right\|_2^2.$$

We recall moreover that if we replace the exact forward map $\mathcal{G}$ by a numerical surrogate $\mathcal{G}_h$, and compute the posterior measure μ_h associated to the approximate forward map, then by Proposition 1.3 the measure μ_h is a good approximation of the true posterior μ. In particular,

if $\mathcal{G}_h$ has order of convergence r, then the same order of convergence holds for the posterior measures with respect to the Hellinger distance.

The convergence result provided by Proposition 1.3 is fundamental and guarantees that the solution of the inverse problem is correct in the asymptotic limit of infinite computations. Nevertheless, it has been demonstrated heuristically in diverse settings that the approximate posterior measure μ_h can be overly confident on the parameter if $h > 0$ is a finite value [6, 7, 32, 36, 37, 39, 87, 101]. In particular, this issue is amplified in case the observation model is much more precise than the approximation quality of the forward model, i.e., heuristically, in case $\|\Gamma\| \ll h^r$. It is nonetheless useful in applications to have a cheap surrogate which can be evaluated quickly, without renouncing to a complete uncertainty quantification of the solution to the inverse problem. Probabilistic numerical methods can be employed for this purpose.

Let us introduce for all $y \in Y$ the random potential $\widetilde{\Phi}_h \colon \Omega \times X \to \mathbb{R}$ which is defined as

$$\widetilde{\Phi}_h(\omega, u; y) = \frac{1}{2} \left\| \Gamma^{-1/2} \left(\widetilde{\mathcal{G}}_h(\omega, u) - y \right) \right\|_2^2,$$

where $\widetilde{\mathcal{G}}_h$ is a perturbation-based probabilistic method for $\mathcal{G}$ as in Definition 6.1. Letting $\mathcal{M}(X)$ denote the space of probability measures on X, we then consider the random measure $\widetilde{\mu}_h \colon \Omega \to \mathcal{M}(X)$ whose Radon–Nykodim derivative with respect to the prior is given by

$$\frac{\mathrm{d}\widetilde{\mu}_h}{\mathrm{d}\mu_0}(\omega, u) = \frac{1}{\widetilde{Z}_h(\omega)} \exp\left(-\widetilde{\Phi}_h(\omega, u; y) \right), \tag{6.6}$$

where $\widetilde{Z}_h \colon \Omega \to \mathbb{R}$ is the random normalization constant

$$\widetilde{Z}_h(\omega) = \int_X \exp\left(-\widetilde{\Phi}_h(\omega, u; y) \right) \mathrm{d}\mu_0(u),$$

so that $\widetilde{\mu}_h \in \mathcal{M}(X)$ for all $\omega \in \Omega$. In the following, we drop for economy of notation the dependence on the event $\omega \in \Omega$ on the quantities introduced above. Similarly to the Monte Carlo approximation introduced in Section 6.1.2, we now describe a methodology to compute in practice a deterministic approximation of the random measure $\widetilde{\mu}_h$. We refer the reader to [87], where this topic is treated in depth, and from which we borrow some notation and results here. There are two possibilities to compute a deterministic posterior measure from (6.6). The first, which is called the marginal approximation, is to compute the measure $\widetilde{\mu}_{h,\mathrm{mar}}$ defined as

$$\frac{\mathrm{d}\widetilde{\mu}_{h,\mathrm{mar}}}{\mathrm{d}\mu_0}(u) = \frac{1}{\mathbb{E}^{\nu_h}\left[\widetilde{Z}_h\right]} \mathbb{E}^{\nu_h}\left[\exp\left(-\widetilde{\Phi}_h(u; y) \right) \right], \tag{6.7}$$

where we recall that $\mathbb{E}^{\nu_h}$ denotes expectation with respect to the measure induced by the probabilistic map on Y. Under some mild technical assumptions, it is possible to prove for a forward map $\widetilde{\mathcal{G}}_h$ with mean-square order of convergence p_s that the measures $\widetilde{\mu}_h$ and $\widetilde{\mu}_{h,\mathrm{mar}}$ satisfy respectively

$$\begin{aligned} \mathbb{E}^{\nu_h}\left[d_{\mathrm{Hell}}(\widetilde{\mu}_h, \mu)^2 \right]^{1/2} &\leq C h^{p_s}, \\ d_{\mathrm{Hell}}(\widetilde{\mu}_{h,\mathrm{mar}}, \mu) &\leq C h^{p_s}, \end{aligned} \tag{6.8}$$

where C is a positive constant independent of h and where μ is the true posterior given in (6.5) [87, Theorems 3.1, 3.2].

The second possibility to obtain a deterministic posterior measure from the random measure $\widetilde{\mu}_h$ is to proceed with a Monte Carlo estimation. Indeed, letting M be a positive integer, one can

generate M i.i.d. realizations of the probabilistic map, i.e., of the random potential $\widetilde{\Phi}_h$, which we call $\{\widetilde{\Phi}_h^{(i)}\}_{i=1}^M$. Then, an approximation of $\widetilde{\mu}_h$ is given by the measure $\widetilde{\mu}_{h,\mathrm{MC}}$ defined as

$$\frac{\mathrm{d}\widetilde{\mu}_{h,\mathrm{MC}}}{\mathrm{d}\mu_0}(u) = \frac{1}{M}\sum_{i=1}^{M}\frac{\mathrm{d}\widetilde{\mu}_h^{(i)}}{\mathrm{d}\mu_0}(u) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{\widetilde{Z}_h^{(i)}}\exp\left(-\widetilde{\Phi}_h^{(i)}(u;y)\right), \tag{6.9}$$

where $\{\widetilde{\mu}_h^{(i)}\}_{i=1}^M$ are M realizations of the random posterior (6.6) and where for all $i = 1, \dots, M$

$$Z_h^{(i)} = \int_X \exp\left(-\widetilde{\Phi}_h^{(i)}(u;y)\right)\,\mathrm{d}\mu_0(u).$$

There is no work in the literature, to our knowledge, in which the posterior $\widetilde{\mu}_{h,\mathrm{MC}}$ is considered and analyzed. Numerical experiments and a partial analysis lead us to conjecture that the quality of the approximation of μ by $\widetilde{\mu}_{h,\mathrm{MC}}$ can be described similarly to Theorem 6.7, and proving such an approximation property could be an interesting line for future work.

Remark 6.9. There exist other approaches to factor the effects of discretization into Bayesian inverse problems. In particular, numerical error can be treated as modeling discrepancies, and the techniques introduced in Section 2.6 can therefore be applied.

6.2.1 Sampling from the Posterior

In Chapter 1 we have introduced a methodology to draw samples from the posterior distribution and thus approximate the solution of Bayesian inverse problems. In particular, we considered in Section 1.2 a procedure based on the Karhunen–Loève (KL) expansion to obtain finite-dimensional approximations of inverse problem. Then, in Section 1.3 we presented the Metropolis–Hastings (MH) algorithm, and more in general Markov chain Monte Carlo methods (MCMC), which allow to sample from the posterior distribution. Here, we consider the random posterior measure $\widetilde{\mu}_h$ of (6.6) and detail how to obtain samples and thus solve the inverse problem. In particular, given a smooth function $\Psi\colon X \to \mathbb{R}$, we are interested in approximating the quantity of interest Q defined as

$$Q \coloneqq \mathbb{E}^{\nu_h}\left[\mathbb{E}^{\widetilde{\mu}_h}\left[\Psi\right]\right].$$

We first consider the measure $\widetilde{\mu}_{h,\mathrm{MC}}$ of (6.9), for which the application of standard MH is direct. Let in particular the measure ν_h be independent on the prior μ_0, so that one can evaluate on multiple values $u \in X$ for any fixed realization of $\widetilde{\mathcal{G}}_h$. Then, letting M_{MC} be a positive integer, we generate M_{MC} i.i.d. realizations of the probabilistic method, which correspond to M_{MC} posterior measures $\{\widetilde{\mu}_h^{(i)}\}_{i=1}^{M_{\mathrm{MC}}}$. For each of these posteriors, we generate M_{chain} samples employing the MH algorithm, where M_{chain} is a positive integer. Schematically, the algorithm to generate $M_{\mathrm{MC}} \cdot M_{\mathrm{chain}}$ samples approximately distributed as $\widetilde{\mu}_h$ proceeds for $i = 1, \dots, M_{\mathrm{MC}}$ as

(i) Generate independently a realization $\widetilde{\mathcal{G}}_h^{(i)}$ of the random forward map;

(ii) Generate M_{chain} samples $\{u^{(i,j)}\}_{j=1}^{M_{\mathrm{chain}}}$ from the posterior $\widetilde{\mu}_h^{(i)}$ associated to $\widetilde{\mathcal{G}}_h^{(i)}$ with the MH algorithm.

The quantity of interest Q is finally approximated by the quantity Q_{MC} computed as

$$Q_{\mathrm{MC}} = \frac{1}{M_{\mathrm{MC}} \cdot M_{\mathrm{chain}}}\sum_{i=1}^{M_{\mathrm{MC}}}\sum_{j=1}^{M_{\mathrm{chain}}}\Psi\left(u^{(i,j)}\right).$$

A different approach has to be adopted for the marginal approximation $\widetilde{\mu}_{h,\mathrm{mar}}$ of (6.7). In particular, due to its definition it is natural to apply the pseudo-marginal Metropolis–Hastings

(PMMH) of [16] to generate samples from $\widetilde{\mu}_{h,\mathrm{mar}}$. As above, let M_{chain} and M_{MC} be positive integers. Then, the PMMH simply generates M_{chain} samples with a MH algorithm, where in the acceptance ratio (1.15) the likelihood function on the proposed value is replaced by a Monte Carlo estimator on M_{MC} samples. Even though the acceptance ratio is not computed exactly, the unbiasedness of the Monte Carlo estimator of the likelihood is sufficient ot guarantee that the resulting Markov chain is ergodic with respect to $\widetilde{\mu}_{h,\mathrm{mar}}$ (see e.g. [16]). Following this approach, we obtain M_{chain} samples $\{u^{(i)}\}_{i=1}^{M_{\mathrm{chain}}}$ distributed as $\widetilde{\mu}_{h,\mathrm{mar}}$ and approximate the quantity of interest Q with the quantity Q_{mar} given by

$$Q_{\mathrm{mar}} = \frac{1}{M_{\mathrm{chain}}} \sum_{j=1}^{M_{\mathrm{chain}}} \Psi\left(u^{(i)}\right).$$

Remark 6.10. The two approaches described above are substantially different in terms of computational cost. In particular, let M_{tot} be a positive integer and let us suppose that we wish to sample M_{tot} values approximately distributed from the posterior $\widetilde{\mu}_h$ of (6.6). If we choose the approximation given by $\widetilde{\mu}_{h,\mathrm{MC}}$, it is sufficient to choose M_{MC} and M_{chain} such that $M_{\mathrm{tot}} = M_{\mathrm{MC}} \cdot M_{\mathrm{chain}}$, and thus we have

$$\mathrm{cost}_{\mathrm{MC}} = M_{\mathrm{tot}},$$

where cost is measured in terms of evaluations of the approximated forward map. Conversely, if we choose the marginal approximation provided by $\widetilde{\mu}_{h,\mathrm{mar}}$ we need to run the PM-MCMC algorithm for M_{tot} iterations, each of which involves M_{MC} evaluations of the forward map. Hence, the cost in this case is

$$\mathrm{cost}_{\mathrm{mar}} = M_{\mathrm{tot}} \cdot M_{\mathrm{MC}}.$$

Moreover, in this second case it is known [15,16] that choosing a small value for M_{MC} might result in a "sticky" behavior of the resulting Markov chain, i.e., it can be difficult due to the variance of the Monte Carlo approximation to escape states with a relatively low probability. Nevertheless, employing the marginal approximation $\widetilde{\mu}_{h,\mathrm{marg}}$ has the advantage of being theoretically justified by (6.8), i.e., by [87, Theorem 3.1].

6.3 Probabilistic Solvers for ODEs

We conclude this chapter by introducing two methods for ordinary differential equations (ODE) which are contributions to the field of PN. In particular, we present first the additive-noise Runge–Kutta integrator of [39], which is a perturbation-based probabilistic method in the sense of Definition 6.1, and then the numerical method based on the Kalman filter of [76], which, conversely, is a measure-valued probabilistic method in the sense of Definition 6.2.

Let d be a positive integer and let us consider the ODE on $\mathbb{R}^d$

$$y'(t) = f(y(t)), \quad y(0) = y_0, \tag{6.10}$$

where we assume that the right-hand side $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is globally Lipschitz continuous, and where $y_0 \in \mathbb{R}^d$ is a given initial condition. We consider for simplicity the ODE to be autonomous, and remark that any right-hand side which depends explicitly on the independent time variable t can be rewritten in the form (6.10) by augmenting the state space.

Let us frame the ODE (6.10) into the setting of Section 6.1. If we consider a final time $T > 0$, the data of the problem are the right-hand side f and the initial condition y_0. Therefore, we have that the input space X is given by $X = \mathcal{C}^{0,1}((0,T);\mathbb{R}^d) \times \mathbb{R}^d$, where we denote by $\mathcal{C}^{0,1}((0,T);\mathbb{R}^d)$ the space of $\mathbb{R}^d$-valued Lipschitz continuous functions. Let us remark that X is a Banach space. The forward operator $\mathcal{G}$ maps in this framework the input into the solution $y(t)$ for $0 \le t \le T$. With

the assumption $f \in \mathcal{C}^{0,1}((0,T);\mathbb{R}^d)$, we have that the unique solution y of (6.10) is continuously differentiable, i.e., the output space $Y = \mathcal{C}^1((0,T),\mathbb{R}^d)$. Let us remark that in the autonomous case the solution $y(t)$ can be conveniently written as

$$y(t) = \varphi_t(y_0; f), \tag{6.11}$$

where for any $f \in \mathcal{C}^{0,1}((0,T);\mathbb{R}^d)$ the function $\varphi_t(\cdot; f)\colon \mathbb{R}^d \to \mathbb{R}^d$ is called the flow map of the ODE (6.10). Therefore, defining

$$\begin{aligned} \mathcal{G}\colon \quad X \quad &\to \quad Y, \\ (f, y_0) &\mapsto \{\varphi_t(y_0; f)\}_{0\le t\le T}, \end{aligned}$$

we can write the problem of solving the ODE (6.10) into the abstract framework we introduced in Section 6.1.

6.3.1 The Additive Noise Runge–Kutta Method

We now present the probabilistic Runge–Kutta method based on additive noise, first introduced in [39] and further analyzed in [86]. We refer the reader to the standard references [60–62] for an exhaustive introduction on Runge–Kutta methods and their properties.

Let $T > 0$ be a final time, let N be a positive integer and let $0 = t_0 < t_1 < \ldots < t_N = T$ be a time grid. Moreover, let $h_n \coloneqq t_n - t_{n-1}$ be a sequence of time steps, and call $h \coloneqq \max_i h_i$. Runge–Kutta methods approximate the solution y of (6.10) on the points of the time grid by mimicking the flow (6.11) through the recursion

$$y_n = \psi_{h_n}(y_{n-1}; f), \quad n = 1, \ldots, N, \tag{6.12}$$

where y_0 is the initial condition of the ODE and where for any $t > 0$ and $f \in \mathcal{C}^{0,1}((0,T);\mathbb{R}^d)$ the function $\psi_t(\cdot; f)\colon \mathbb{R}^d \to \mathbb{R}^d$ is called the numerical flow. In particular, for all $n = 1, \ldots, N$ the value y_n then serves as an approximation of the exact solution $y(t_n)$. In the following, we drop for economy of notation the dependence of ψ on the right-hand side f, and simply write $\psi_t(y)$. Let us remark that since the value y_n is uniquely determined by the previous iterate y_{n-1}, Runge–Kutta schemes belong to the class of one-step methods. Given a positive integer s, a time step $h > 0$ and a vector $y \in \mathbb{R}^d$, the numerical flow map ψ_h applied to y is uniquely determined by a set of real coefficients $\{b_i\}_{i=1}^s$ and $\{a_{ij}\}_{i,j=1}^s$ and by the relation

$$\begin{aligned} k_i &= f\left(y + h\sum_{j=1}^{s} a_{ij}k_j\right), \quad i = 1, \ldots, s, \\ \psi_h(y) &= y + h\sum_{i=1}^{s} b_i k_i, \end{aligned}$$

where the vectors $\{k_i\}_{i=1}^s$ of $\mathbb{R}^d$ are called the internal stages. Carefully setting the number of stages and the coefficients, one can obtain Runge–Kutta methods which are tailored for a wide and diverse range of time-dependent problems. In particular, if the coefficients satisfy the so-called order conditions (see e.g. [61, Chapter III.2]), it is possible to build a numerical flow map such that for all $h > 0$ it holds

$$\sup_{y\in\mathbb{R}^d} \|\psi_h(y) - \varphi_h(y)\| \le Ch^{q+1}, \tag{6.13}$$

where the exponent $q \ge 1$ is called the local order of the Runge–Kutta method, and where C is a positive constant independent of h. In this case, the one-step error propagates through the dynamics induced by the numerical flow map on $\mathbb{R}^d$ as described by the following result (see e.g. [61, Theorem 3.4]).

Theorem 6.11. *Let $f \in \mathcal{C}^{0,1}((0,T);\mathbb{R}^d)$ and let the Runge–Kutta method have local order $q \geq 1$. Then, it holds*

$$\sup_{n=1,\dots,N} \|y(t_n) - y_n\|_2 \leq Ch^q,$$

where $C > 0$ is a constant independent of the time step h.

Let us frame Runge–Kutta methods in the notation of Section 6.1. Fixed a set of coefficients $\{b_i\}_{i=1}^s$ and $\{a_{ij}\}_{i,j=1}^s$, as well as a sequence of time steps $\{h_i\}_{i=1}^N$, the method takes as an input the same couple (f, y_0) and outputs an approximation of the solution y computed with the recursion (6.12) on the points of the time grid. Hence, we can write the Runge–Kutta method as a function $\mathcal{G}_h \colon X \to Y^N \equiv (\mathbb{R}^+ \times \mathbb{R}^d)^N$ such that

$$\begin{aligned} \mathcal{G}_h \colon \quad X \quad &\to \quad Y^N, \\ (f, y_0) &\mapsto \{(t_n, y_n)\}_{n=1}^N. \end{aligned}$$

Convergence is then measured following Theorem 6.11 by comparing values on the points of the time grid, i.e., we equip the space Y^N with the norm

$$\|y\|_{Y^N} = \sup_{n=1,\dots,N} \|y_i\|_2 . \tag{6.14}$$

Let us consider for simplicity an uniformly-spaced time grid, i.e., $t_i = hi$ for a fixed time step $h = T/N$. The additive-noise Runge–Kutta scheme (AN-RK) constructs a sequence of random variables $\{Y_n\}_{n=1}^N$ with values in $\mathbb{R}^d$ by adding at each step of the recursion (6.12) random contributions to the numerical flow, i.e.,

$$Y_n = \psi_h(Y_{n-1}) + \xi_n, \quad n = 1, \dots, N, \tag{6.15}$$

where $Y_0 = y_0$ is given by the initial condition of (6.10), and where the random variable $\xi \coloneqq \{\xi_n\}_{n=1}^N$ has i.i.d. Gaussian components with values in $\mathbb{R}^d$. The following assumption on ξ is crucial to prove results of weak and mean-square convergence.

Assumption 6.12. There exists $p \geq 1/2$ and a symmetric positive definite matrix $Q \in \mathbb{R}^{d \times d}$ such that $\xi_1 \sim \mathcal{N}(0, Qh^{2p+1})$.

In practice, the assumption above allows to calibrate the randomness introduced by the probabilistic method with the error due to discretization. Indeed, if $h \ll 1$ then the error is small, and the random contributions tend to vanish. Let us remark that the Gaussian requirement of Assumption 6.12 has been weakened in [86].

The probabilistic method given by the recursion (6.15) can be framed into the setting of Definition 6.1 by considering the random forward map $\widetilde{\mathcal{G}}_h \colon \Omega \times X \to Y^N$ defined as

$$\begin{aligned} \widetilde{\mathcal{G}}_h \colon \Omega \times X \quad &\to \quad Y^N, \\ \omega \times (f, y_0) &\mapsto \{(t_n, Y_n(\omega))\}_{n=1}^N. \end{aligned} \tag{6.16}$$

The following weak and mean-square convergence results hold, in the sense of Definitions 6.5 and 6.6 [39, Theorems 2.2 and 2.4].

Theorem 6.13. *Let $T > 0$, N be a positive integer and $h = T/N$. Moreover, let Assumption 6.12 hold and let the Runge–Kutta method identified by the flow ψ_h have order of convergence q, as defined in* (6.13)*. Then, the AN-RK method $\widetilde{\mathcal{G}}_h \colon X \times \Omega \to Y^N$ of* (6.16) *has weak order of convergence $p_w = \min\{2p, q\}$ and mean-square order of convergence $p_s = \min\{p, q\}$ with respect to the norm* (6.14)*.*

Remark 6.14. In [39], the authors rightfully argue that the correct scaling for the user-prescribed random perturbations is therefore $p = q$, where p is given in Assumption 6.12 and q is the order of the Runge–Kutta method identified by ψ_t. In this case, we have the natural consequence that both the weak and strong order of convergence are $p_w = p_s = q$. Moreover, by Remark 6.8 it suffices to choose $M = \mathcal{O}(1)$ samples to obtain accurate Monte Carlo estimations of quantities based on the probabilistic solution.

6.3.2 A Probabilistic ODE Solver Based on Filtering

We now present a filtering-based Bayesian probabilistic ODE solver, which was introduced in [126] and further developed in [76, 77, 137]. In particular, we focus on the version given in [76] and analyzed in [77], which is based on the Kalman filter. Let us recall the reader that the formulation of the Kalman filer in a general framework is reported in Chapter 2. In this section we consider for simplicity the ODE (6.10) in the one-dimensional case, i.e., we fix $d = 1$. A discussion on generalization to higher dimensions can be found in [76, 77].

The main idea underlying filtering-based probabilistic solvers for ODEs is introducing a prior model for the q-th derivative of the solution y of (6.10), with q being an integer such that $q \geq 1$. In particular, the choice that is explored in [76, 77] is to assume that, a priori, the quantity $y^{(q)}$ acts as Brownian motion, so that the prior model for the solution y itself is that of an iterated Brownian integral (q times). Hence, the prior model reads

$$\begin{aligned} \mathrm{d}y_t^{(k)} &= y_t^{(k+1)}\,\mathrm{d}t, \quad k = 0, 1, \dots, q-1, \\ \mathrm{d}y_t^{(q)} &= \sigma^2\,\mathrm{d}W_t, \end{aligned} \tag{6.17}$$

where W_t is a one-dimensional Brownian motion. Denoting $Z_t = (y_t, y_t', y_t^{(2)}, \dots, y_t^{(q)})^\top$, so that Z_t is a stochastic process with values in $\mathbb{R}^{q+1}$, we can rewrite (6.17) more compactly as the Itô SDE

$$\mathrm{d}Z_t = F Z_t\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t, \tag{6.18}$$

where, this time, W_t is a $(q+1)$-dimensional Brownian motion and where the matrices $F, \Sigma \in \mathbb{R}^{(q+1)\times(q+1)}$ are given by

$$F = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ & 0 & 1 & \cdots & 0 \\ & & \ddots & \ddots & \vdots \\ \vdots & & & 0 & 1 \\ 0 & & & \cdots & 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix}.$$

The properties of the solution of (6.18) are known. In particular, let us remark that (6.18) admits the explicit solution

$$Z_{t+h} = \exp(Fh) Z_t + \int_t^{t+h} \exp\left(F(t+h-s)\right) \Sigma\,\mathrm{d}W_s,$$

where $h > 0$ is a given time step. Moreover, the solution Z_t is distributed as a Gaussian for all $t \geq 0$. Therefore, introducing the matrices

$$Q(h) \coloneqq \mathrm{Cov}\left(\int_t^{t+h} \exp\left(F(t+h-s)\right) \Sigma\,\mathrm{d}W_s \right), \qquad A(h) \coloneqq \exp(Fh),$$

we can completely define the distribution of Z_{t+h} with the relations

$$\begin{aligned}\mathbb{E}[Z_{t+h}] &= A(h)\,\mathbb{E}[Z_t],\\ \mathrm{Cov}(Z_{t+h}) &= A(h)\mathrm{Cov}(Z_t)A(h)^\top + Q(h),\end{aligned}$$

where for a random vector X we write $\mathrm{Cov}(X) \coloneqq \mathbb{E}[XX^\top] - \mathbb{E}[X]\,\mathbb{E}[X]^\top$. For the sake of completeness, let us remark that due to the Itô isometry it holds

$$Q(h) = \int_t^{t+h} \exp\left(F(t+h-s)\right)\Sigma\Sigma^\top \exp\left(F(t+h-s)\right)^\top \,\mathrm{d}s.$$

Let us moreover remark that in this simple scenario both $A(h)$ and $Q(h)$ can be computed explicitly, as shown in [76, 77] and references therein. The prior model we introduced above is employed in the prediction step of the Kalman filter (see Chapter 2). In particular, let us assume that $Z_t \sim \mathcal{N}(m_t, C_t)$, for some mean $m_t \in \mathbb{R}^{q+1}$ and a positive definite covariance $C_t \in \mathbb{R}^{(q+1)\times(q+1)}$. Then, the prediction step yields the random variable $\widehat{Z}_{t+h} \sim \mathcal{N}(\widehat{m}_{t+h}, \widehat{C}_{t+h})$ where

$$\widehat{m}_{t+h} = A(h)m_t, \qquad \widehat{C}_{t+h} = A(h)C_tA(h)^\top + Q(h). \tag{6.19}$$

In the spirit of Kalman filtering, the random variable $\widehat{Z}_{t+h}$ has to be confronted with observations in order to obtain with Bayes' rule the final distribution at time $t+h$, but in the framework of ODEs no "real" observations are provided. Nevertheless, the right-hand side f of (6.10) yields information on the first derivative y', i.e., on the second component of the vector Z_{t+h}. Therefore, an observation model is given by the equation

$$y_{t+h} = HZ_{t+h}, \qquad H \coloneqq \begin{pmatrix}0 & 1 & 0 & \cdots & 0\end{pmatrix}.$$

The value of Z_{t+h} is unknown beforehand and a surrogate has to be computed. In [76], the authors make the strong modeling assumption

$$y = \widehat{f} + \eta, \text{ in law, with } \eta \sim N\left(0, \Gamma\right), \tag{6.20}$$

where denoting $\widehat{\mu}_{t+h} = \mathcal{N}(\widehat{m}_{t+h}, \widehat{C}_{t+h})$ the predicted distribution of $\widehat{Z}_{t+h}$ we define

$$\widehat{f} = \mathbb{E}^{\widehat{\mu}_{t+h}}[f], \quad \Gamma = \mathrm{Var}^{\widehat{\mu}_{t+h}}(f).$$

Let us remark that in [76] the authors propose a series of methodologies to approximate the otherwise unknown quantities $\widehat{f}$ and Γ, e.g. by means of a Monte Carlo simulation or by Bayesian quadrature. Endowed with the observation model (6.20), we can finally perform the update step, which reads

$$\begin{aligned}&Z_{t+h} \coloneqq \widehat{Z}_{t+h} \mid Y_{t+h} \sim \mathcal{N}(m_{t+h}, C_{t+h}),\\ &m_{t+h} = \widehat{m}_{t+h} + K_{t+h}\left(y_{t+h} - H\widehat{m}_{t+h}\right), \quad C_{t+h} = (I - K_{t+h}H)\widehat{C}_{t+h},\\ &\text{where } K_{t+h} = \widehat{C}_{t+h}H^\top R_{t+h},\\ &\text{with } R_{t+h} = \left(H\widehat{C}_{t+h}H^\top + \Gamma\right)^{-1},\end{aligned} \tag{6.21}$$

and where we recall that K_{t+h} is the Kalman gain. The two steps of prediction and update are repeated until the final time T is reached. We refer the reader to Fig. 6.1 for a schematic representation of this filtering methodology.

The method we described above naturally falls in the category of the measure-valued probabilistic methods of Definition 6.2. In particular, the whole recursion can be easily written as $\widetilde{\mathcal{G}}_h \colon X \to \mathcal{N}(Y)$, where we denote by $\mathcal{N}(Y)$ the space of Gaussian measures on Y. We remark moreover

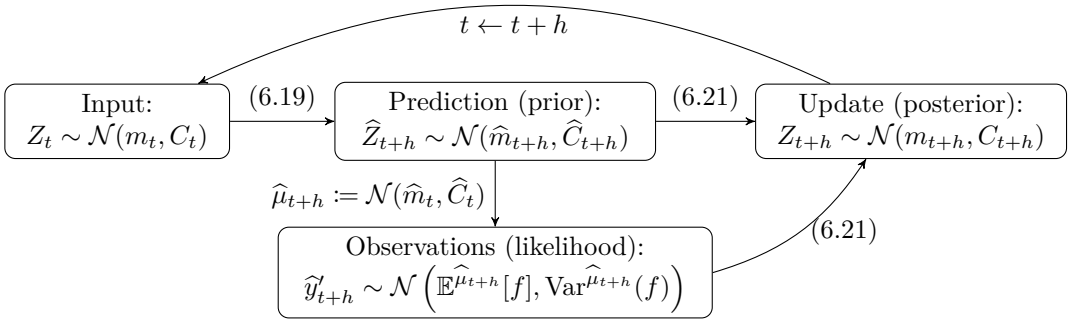


Figure 6.1 – Flow diagram for the filtering method for ODEs of [76, 77].

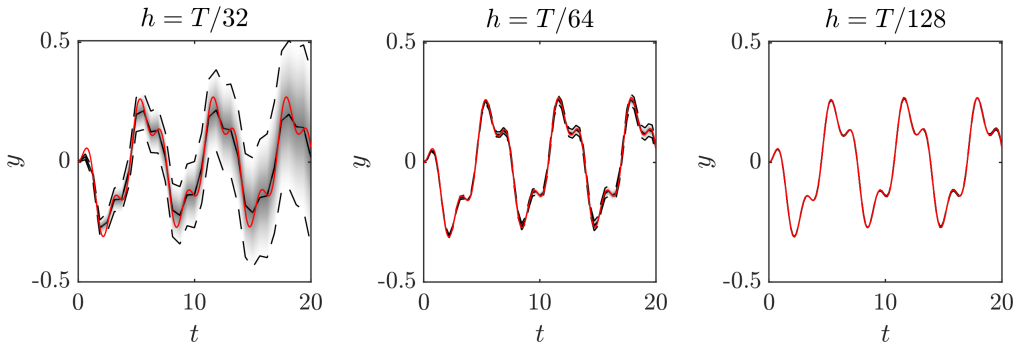


Figure 6.2 – Numerical results for the filtering probabilistic method for ODEs. The final time T is fixed to $T = 20$ and the right-hand side of (6.10) is $f(t, y) = -(y - \sin(t)\cos(2t))/2$. Results for $q = 2$ in the prior model (6.18), for $\sigma^2 = 0.1$ in (6.17) and for time steps $h = \{T/32, T/64, T/128\}$. A reference solution is given in red, the mean estimate of the filter in a solid black line and the boundaries of a 95% confidence interval in dashed black lines. The confidence interval itself is represented by a shaded gray area.

that since the update step of Kalman filtering is motivated by Bayes' rule, this methodology can be described as being Bayesian.

Let us consider a numerical example on a simple test case. In particular, let (6.10) hold with a non-homogeneous right-hand side given by

$$f(t, y) = -\frac{1}{2}(y - \sin(t)\cos(2t)), \tag{6.22}$$

and with initial condition $y_0 = 0$. We set the final time to $T = 20$, and consider the prior model (6.18) with $q = 2$ and the pre-multiplying factor for the Brownian noise $\sigma^2 = 0.1$. Moreover, we generate the observations in (6.20) with a standard Monte Carlo simulation. Finally, we consider the time step to be given by $h = \{T/32, T/64, T/128\}$ in order to verify the influence of h on the numerical output. In Fig. 6.2 we notice heuristically that for $h \to 0$ the output mean converges to the exact solution while the uncertainty shrinks to zero. In fact, the convergence properties of this probabilistic method have been analyzed in a simple setting in [77], and it has been proved that the mean indeed converges to the exact solution for $h \to 0$.

We believe there are a series of advantages and disadvantages of the filtering approach for ODEs illustrated above. In particular, we identified the following advantages:

- The output measure on the solution is Gaussian, and can therefore be readily pushed

through pipelines of computations, which can be practical in real-world applications;

- The recursion described by the diagram in Fig. 6.1 is fully deterministic and computationally rather inexpensive, conversely to the sampling-based method of Section 6.3.1.

In our opinion, though, the downsides of this methodology which should be accounted for are:

- The method seems to us too rigid in its implementation. Runge–Kutta methods, for example, have been demonstrated to work theoretically and in practice on a diverse and large class of problems. Squandering the versatility of Runge–Kutta methods for the sake of maintaining a Bayesian/Gaussian approach seems unreasonable, especially in light of Theorem 6.7;
- The procedure to generate data and thus perform the update step seems unnatural and difficult to justify from a theoretical standpoint. In particular, as per [102], the methodology introduced here cannot be labeled as Bayesian due to this specific misstep in the algorithm;
- The theoretical analysis is industrious and justifies convergence theoretically only for a restrained set of possible implementations of the method. In particular, in [77] the authors show linear convergence with respect to h for $q = 2$ in (6.18), and suggest that a higher order holds for $q > 2$;
- Some numerical experiments we performed seem to show that the method is prone to instabilities on stiff problems (see e.g. [62]) in case the coefficient q of (6.18) is chosen to be relatively large.

7 Probabilistic Geometric Integration of ODEs

In this chapter, we introduce a perturbation-based probabilistic numerical method for quantifying the uncertainty induced by the time integration of ordinary differential equations (ODEs). The method is based on a randomization of the time discretization and on Runge–Kutta integrators, and we therefore call it random time step Runge–Kutta method (RTS-RK). A source of inspiration for the RTS-RK is undeniably given by the additive noise Runge–Kutta method (AN-RK) of [39,86], which we summarized in Section 6.3.1, and which is fundamental in the field of perturbation-based solvers for ODEs. Nevertheless, the AN-RK fails in some instances to reproduce the favorable properties of the Runge–Kutta method it is built on. This degradation is particularly accentuated in the context of ODEs with certain geometric properties, such as the conservation of polynomials invariants and Hamiltonian systems. This motivates the RTS-RK, whose intrinsic randomization, opposed to the extrinsic source of additive noise of the AN-RK, allows to create a family of probabilistic solutions which all possess the geometric properties of the underlying Runge–Kutta integrator. The content of this chapter is based on our article [6], and is one of the original contribution of this thesis.

The outline of this chapter is as follows. In Section 7.1 we introduce the setting for probabilistic numerics and present our novel numerical scheme. We then show in Section 7.2 and Section 7.3 the properties of weak and mean-square convergence of the numerical solution towards the exact solution of the ODE. The geometric properties of the numerical scheme are presented in Section 7.4 and Section 7.5, while in Section 7.6 we introduce Bayesian inverse problems in the ODE setting, and show how our method can be integrated in existing sampling strategies. Finally, we show a series of numerical experiments which validate our analysis and illustrate the potential of our method in Section 7.7.

7.1 Random Time Step Runge–Kutta Method

Let us consider the setting of Section 6.3, i.e., let $f\colon \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz continuous function let us consider the ODE

$$y'(t) = f(y(t)), \quad y(0) = y_0 \in \mathbb{R}^d. \tag{7.1}$$

We recall from Section 6.3 that the solution $y(t)$ can be written for simplicity by employing the flow $\varphi_t\colon \mathbb{R}^d \to \mathbb{R}^d$, which satisfies

$$y(t) = \varphi_t(y_0).$$

Moreover, given a time step h, we recall from Section 6.3 that a Runge–Kutta method can be written in terms of a numerical flow $\psi_t\colon \mathbb{R}^d \to \mathbb{R}^d$, which is uniquely determined by the coefficients

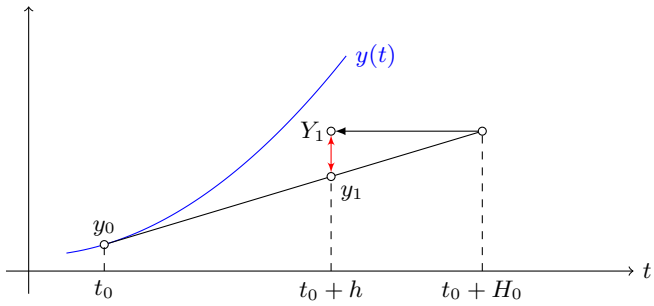


Figure 7.1 – Graphical representation of one step of the RTS-RK method with $\psi_h(y) = y + hf(y)$. The red arrow is the stochastic contribution due to random time-stepping.

of the method, as

$$y_{k+1} = \psi_h(y_k), \quad k = 0, 1, \dots.$$

Maintaining the same notation as in Section 6.3.1, we present and analyze in this chapter the random time-stepping Runge–Kutta method (RTS-RK), i.e., the scheme defined by the recurrence relation

$$Y_{k+1} = \psi_{H_k}(Y_k), \quad k = 0, 1, \dots, \tag{7.2}$$

where Y_k is still a random variable approximating $y(t_k)$ and the time steps H_k are locally given by a sequence of i.i.d. random variables with values in $\mathbb{R}^+$. A graphical representation of one step of the RTS-RK method is given in Fig. 7.1. Let us finally remark that the sequence Y_k, $k = 0, 1, \dots$, form a homogeneous Markov chain, as the transition probability is independent of the index k.

Remark 7.1. We note that simulating the AN-RK method of Section 6.3.1 and the RTS-RK method is equivalent in terms of computational cost.

7.1.1 Assumptions and Notation

We now present the main assumptions and notations which are used throughout the rest of this work. Firstly, we have to consider the possible values taken by the random step sizes, which have to satisfy restrictions that are necessary not to spoil the properties of deterministic methods.

Assumption 7.2. The i.i.d. random variables H_k satisfy for all $k = 0, 1, \dots$

(i) $H_k > 0$ a.s.,

(ii) there exists $h > 0$ such that $\mathbb{E}[H_k] = h$,

(iii) there exist $p \geq 1/2$ and $C > 0$ independent of k such that $\mathrm{Var}(H_k) = Ch^{2p+1}$.

The class of random variables satisfying the hypotheses above is general. However, it is practical for an implementation point of view to have examples of these variables.

Example 7.3. Let us consider the random variables $\{H_k\}_{k\geq 0}$ such that

$$H_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(h - h^{p+1/2}, h + h^{p+1/2}), \quad 0 < h < 1, \quad p \geq 1/2.$$

We easily verify that the assumptions (i) and (ii) are verified as $h < 1$, and that (iii) is verified with $C = 1/3$. Another choice of random variables could simply be

$$H_k \overset{\text{i.i.d.}}{\sim} \log \mathcal{N}\left(\log h - \log\sqrt{1 + h^{2p}}, \log(1 + h^{2p})\right), \tag{7.3}$$

for which the properties above are trivially verified (with $C = 1$), provided $p > 1/2$.

We secondly introduce an assumption on the deterministic method underlying the RTS-RK scheme, identified by its numerical flow ψ_h.

Assumption 7.4. The Runge–Kutta method defined by the numerical flow $\{\psi_t\}_{t\geq 0}$ is of order q in the sense of (6.13), i.e., for h small enough, there exists a constant $C > 0$ such that

$$\|\psi_h(y) - \varphi_h(y)\| \leq Ch^{q+1}, \quad \forall y \in \mathbb{R}^d.$$

Remark 7.5. Depending on the domain of definition of the vector field f, the choice of an unbounded distribution for the time step could give rise to two critical issues. In particular,

(i) if $f\colon D \to \mathbb{R}^d$, where $D \in \mathbb{R}^d$ is a bounded open subset of $\mathbb{R}^d$, allowing the time step to assume unbounded values as, e.g., in case of the log-normal distribution (7.3), may force the solution outside D,

(ii) if ψ_h is the numerical flow of an implicit method, the solution could be ill-posed.

In both the two cases above, we suggest to employ uniform time steps as in Example 7.3, which allow the time steps to be small enough almost surely. For the former issue, more sophisticated techniques of path rejection could be employed [97], but the mean-square convergence properties which will be examined in Section 7.3 would not hold.

In order to avoid the second issue presented in Remark 7.5, we introduce a further assumption.

Assumption 7.6. If the map ψ_t is implicit, the time steps H_k satisfy $H_k \leq M < \infty$ almost surely, where M is small enough for the scheme to be well-posed.

Let us finally remark that the choice of the distribution of the time steps is artificial and therefore arbitrary. Hence, choosing a bounded distribution does not represent a limitation to the numerical scheme.

7.2 Weak Convergence Analysis

In this section, we analyze the RTS-RK in terms of its weak convergence, in the sense of Definition 6.5. In the following, we denote by $\mathcal{C}_b^l(\mathbb{R}^d, \mathbb{R})$ for any integer l the functions in $\mathcal{C}^l(\mathbb{R}^d, \mathbb{R})$ with all derivatives up to order l bounded uniformly in $\mathbb{R}^d$. Moreover, we consider the integration of (7.1) over the finite length domain $[0, T]$, where $T > 0$ is the final time.

Let us introduce the Lie derivative of the flow $\mathcal{L} = f \cdot \nabla$, which allows us to adopt the semi-group notation for the exact solution of (7.1) (see e.g. [60, Section III.5.1] or [113, Section 4.3]) and write for any smooth function Ψ

$$\Psi(\varphi_h(y)) = e^{h\mathcal{L}}\Psi(y). \tag{7.4}$$

Moreover, let us recall that the probabilistic numerical solution $\{Y_k\}_{k\geq 0}$ forms a homogeneous Markov chain. Therefore, given $h > 0$ there exists an operator $\mathcal{P}_h$, the generator [110, Section 2.3], such that

$$\mathbb{E}\left[\Psi(Y_{k+1}) \mid Y_k = y\right] = (\mathcal{P}_h\Psi)(y).$$

In order to have an analogy with the notation (7.4), we adopt the exponential form of the infinitesimal generator and denote in the following $\mathcal{P}_h = e^{h\mathcal{L}_h}$, where we explicitly write the dependence of the Markov generator on the step size h. Furthermore, due to the homogeneity of the Markov chain, we can write

$$\mathbb{E}\left[\Psi(Y_{k+1}) \mid Y_0 = y\right] = e^{h\mathcal{L}_h}\,\mathbb{E}\left[\Psi(Y_k) \mid Y_0 = y\right]. \tag{7.5}$$

We can now prove that the RTS-RK converges weakly after one step.

Lemma 7.7. *Let Assumption 7.2, Assumption 7.4 and Assumption 7.6 hold and let f in (7.1) be sufficiently smooth. If $\mathbb{E}\left[H_0^4\right] < \infty$, there exists a constant $C > 0$ independent of h and y such that for any function $\Psi \in \mathcal{C}_b^l(\mathbb{R}^d, \mathbb{R})$, with $l = \max\{q, 3\}$*

$$|\mathbb{E}\left[\Psi(Y_1) \mid Y_0 = y\right] - \Psi(\varphi_h(y))| \leq Ch^{\min\{2p+1, q+1\}}.$$

Proof. Since f is sufficiently smooth, the map $t \mapsto \psi_t(y)$ is of class $C^2(\mathbb{R}^+, \mathbb{R}^d)$ and Lipschitz continuous with constant L_ψ independent of y. Let us expand the functional Ψ computed on the numerical solution as

$$\begin{aligned}
\Psi(Y_1) &= \Psi(\psi_{H_0}(Y_0)) \\
&= \Psi\left(\psi_h(Y_0) + (H_0 - h)\partial_t\psi_h(Y_0) + \frac{1}{2}(H_0 - h)^2\partial_{tt}\psi_h(Y_0) + \mathcal{O}(|H_0 - h|^3)\right) \\
&= \Psi(\psi_h(Y_0)) + \left((H_0 - h)\partial_t\psi_h(Y_0) + \frac{1}{2}(H_0 - h)^2\partial_{tt}\psi_h(Y_0)\right) \cdot \nabla\Psi(\psi_h(Y_0)) \\
&\quad + \frac{1}{2}(H_0 - h)^2\partial_t\psi_h(Y_0)\partial_t\psi_h(Y_0)^\top : \nabla^2\Psi(\psi_h(Y_0)) + \mathcal{O}(|H_0 - h|^3),
\end{aligned} \tag{7.6}$$

where we denote by $\nabla^2\Psi$ the Hessian matrix of Ψ, and by : the inner product on matrices induced by the Frobenius norm on $\mathbb{R}^d$, i.e., $A : B = \mathrm{tr}(A^\top B)$. Taking the conditional expectation with respect to $Y_0 = y$ and applying Assumption 7.2 we get

$$\begin{aligned}
e^{h\mathcal{L}_h}\Psi(y) - \Psi(\psi_h(y)) &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h(y) \cdot \nabla\Psi(\psi_h(y)) \\
&\quad + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h(y)\partial_t\psi_h(y)^\top : \nabla^2\Psi(\psi_h(y)) + \mathcal{O}(h^{3p+3/2}),
\end{aligned} \tag{7.7}$$

where we exploited Hölder's inequality for the last term. Moreover, expanding Ψ around y we get

$$\begin{aligned}
\Psi(\psi_h(y)) &= \Psi\left(\psi_0(y) + h\partial_t\psi_0(y) + \mathcal{O}(h^2)\right) \\
&= \Psi(y) + \mathcal{O}(h),
\end{aligned}$$

which implies

$$\begin{aligned}
e^{h\mathcal{L}_h}\Psi(y) - \Psi(\psi_h(y)) &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h(y) \cdot \nabla\Psi(y) \\
&\quad + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h(y)\partial_t\psi_h(y)^\top : \nabla^2\Psi(y) + \mathcal{O}(h^{2p+1}).
\end{aligned} \tag{7.8}$$

Let us remark that due to the smoothness of the flow we have

$$e^{h\mathcal{L}}\Psi(y) - \Psi(\psi_h(y)) = \mathcal{O}(h^{q+1}). \tag{7.9}$$

Combining (7.9) and (7.8) we have the one-step weak error of the probabilistic method on the original ODE, i.e.,

$$e^{h\mathcal{L}}\Psi(y) - e^{h\mathcal{L}_h}\Psi(y) = \mathcal{O}(h^{\min\{2p+1, q+1\}}),$$

which proves the desired result. □

Remark 7.8. Let us remark that rigorously if $\partial_{tt}\psi_h(y)$ is bounded independently of y then the equality (7.6) holds. In fact, as it can be noticed in (7.7), a weaker and sufficient requirement is that $h^{p+1/2}\partial_{tt}\psi_h(y)$ is bounded independently of h.

In order to obtain a result on the global order of convergence we need a further stability assumption, which is the same as Assumption 3 in [39].

Assumption 7.9. The function f and the distribution of the random time steps H_k, $k = 0, 1, \ldots$, are such that the operator $e^{h\mathcal{L}_h}$ satisfies for all functions $g \in \mathcal{C}_b^q(\mathbb{R}^d, \mathbb{R})$ and a positive constant L,

$$\sup_{u\in\mathbb{R}^d} \left|e^{h\mathcal{L}_h} g(u)\right| \leq (1 + Lh) \sup_{u\in\mathbb{R}^d} |g(u)|, \tag{7.10}$$

where L may depend on f and on the distribution of the random time steps, but not on g or h.

Remark 7.10. Let us remark that in order for ψ_h to satisfy Assumption 7.4, i.e., for ψ_h to be of order q, the right hand side f must be of class $\mathcal{C}_b^q(\mathbb{R}^d, \mathbb{R}^d)$ (see, e.g. [61, Theorem II.3.1]). Therefore, in order to apply the bound (7.10) to composite functions $\Psi \circ \varphi_h \colon \mathbb{R}^d \to \mathbb{R}$ where $\Psi \in \mathcal{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$, by the chain rule we need Assumption 7.9 to hold for functions in $C_b^q(\mathbb{R}^d, \mathbb{R})$. This fact will be exploited in the proof of Theorem 7.12 below.

We now give a lemma useful for bounding discrete sequences, which is taken from [96, Lemma 1.6].

Lemma 7.11. *Suppose that for arbitrary N and $k = 0, \ldots, N$ we have*

$$e_k \leq (1 + Ah)e_{k-1} + Bh^r,$$

where $h = T/N$, $A > 0$, $B \geq 0$, $r \geq 1$ and $e_k \geq 0$, $k = 0, \ldots, N$. Then

$$e_k \leq e^{AT} e_0 + \frac{B}{A}(e^{AT} - 1)h^{r-1}.$$

The proof of Lemma 7.11 follows from the discrete Grönwall inequality. We can now state the main result on weak convergence.

Theorem 7.12. *Let the assumptions of Lemma 7.7 and Assumption 7.9 hold. Then, there exists a constant $C > 0$ independent of h and of the initial condition such that for all functions $\Psi \in \mathcal{C}_b^l(\mathbb{R}^d, \mathbb{R})$, with $l = \max\{q, 3\}$*

$$|\mathbb{E}\left[\Psi(Y_k)\right] - \Psi(y(kh)))| \leq Ch^{\min\{2p,q\}}, \tag{7.11}$$

for all $k = 1, 2, \ldots, N$ and $T = Nh$.

Proof. Let us introduce the following notation

$$w_k(u) = \Psi(\varphi_{t_k}(u)),$$
$$W_k(u) = \mathbb{E}\left[\Psi(Y_k) \mid Y_0 = u\right].$$

By the triangle inequality and the Markov property (7.5), we have

$$\begin{aligned}\sup_{u\in\mathbb{R}^d} |W_k(u) - w_k(u)| \leq & \sup_{u\in\mathbb{R}^d} \left|e^{h\mathcal{L}} w_{k-1}(u) - e^{h\mathcal{L}_h} w_{k-1}(u)\right| \\ & + \sup_{u\in\mathbb{R}^d} \left|e^{h\mathcal{L}_h} w_{k-1}(u) - e^{h\mathcal{L}_h} W_{k-1}(u)\right|.\end{aligned}$$

We then apply Lemma 7.7 to the first term and Assumption 7.9 to the second and denote $e_k \coloneqq \sup_{u\in\mathbb{R}^d} |W_k(u) - w_k(u)|$, thus obtaining

$$e_k \leq Ch^{\min\{2p+1,q+1\}} + (1 + Lh)e_{k-1}.$$

We can therefore apply Lemma 7.11 with $A = L$ and $r = \min\{2p + 1, q + 1\}$, and therefore get for a constant $C > 0$

$$\sup_{u\in\mathbb{R}^d} |w_k(u) - W_k(u)| \leq Ch^{\min\{2p,q\}},$$

which is the desired result. □

Remark 7.13. In [39], Conrad et al. define ordinary and stochastic modified equations in order to prove a result of weak convergence applying techniques of backward error analysis. In particular, they show that their probabilistic solver approximates in the weak sense a stochastic differential equation (SDE) where the deterministic part is given by the original ODE. For our probabilistic solver, it is possible to prove that the numerical solutions approximates in the weak sense the solution of an SDE which depends on the derivative of the map $t \mapsto \psi_t(y)$. Such a construction is shown in Section 7.8.

Remark 7.14. Let us recall that the random variable Y_k given by RTS-RK is thought of as an approximation of $y(kh)$ regardless of the value of the sum of the random time steps. Hence, the comparison in (7.11) is legitimate and does not induce time misalignment between true and numerical solutions. This basic property applies to all results in the following.

7.3 Mean-square Convergence Analysis

The second property of the RTS-RK method we analyze is its mean-square order of convergence, in the sense of Definition 6.6. We start by analysing how the method converges with respect to the mean step size h in the local sense, i.e., after one step of the numerical integration.

Lemma 7.15. *Under Assumption 7.2, Assumption 7.4 and Assumption 7.6 the numerical solution Y_1 given by one step of the RTS-RK method* (7.2) *satisfies*

$$\mathbb{E}\left[\|Y_1 - y(h)\|^2\right]^{1/2} \leq Ch^{\min\{p+1/2,q+1\}}, \tag{7.12}$$

where C is a real positive constant independent of h and of the initial condition y_0 and the coefficients p, q are given in the assumptions.

Proof. By the triangle and Young's inequalities we have for all $y \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\psi_{H_0}(y) - \varphi_h(y)\|^2\right] \leq 2\,\mathbb{E}\left[\|\psi_{H_0}(y) - \psi_h(y)\|^2\right] + 2\,\|\psi_h(y) - \varphi_h(y)\|^2 .$$

We now consider Assumption 7.4 and Assumption 7.2, thus getting

$$\begin{aligned}\mathbb{E}\left[\|\psi_{H_0}(y) - \varphi_h(y)\|^2\right] &\leq 2L_\psi^2\,\mathbb{E}\left[|H_0 - h|^2\right] + 2C_1h^{2(q+1)} \\ &= 2L_\psi^2 C_2 h^{2p+1} + 2C_1h^{2(q+1)} \\ &\leq C^2 h^{2\min\{p+1/2,q+1\}},\end{aligned}$$

where C_1 and C_2 are the constants given in Assumption 7.4 and Assumption 7.2 respectively. This is the desired result with $C = \max\{2L_\psi^2 C_2, 2C_1\}^{1/2}$. □

As a consequence of the one-step convergence, we can prove a result of global mean-square convergence.

Theorem 7.16. *Let f be globally Lipschitz and $t_k = kh$ for $k = 1, 2, \ldots, N$, where $Nh = T$. Then, under the assumptions of Lemma 7.15 the numerical solution given by* (7.2) *satisfies*

$$\sup_{k=1,2,\ldots,N} \mathbb{E}\left[\|Y_k - y(t_k)\|^2\right]^{1/2} \leq Ch^{\min\{p,q\}}, \tag{7.13}$$

where C is a real positive constant independent of h and of the initial condition.

In order to prove this result, let us introduce the following lemma.

Lemma 7.17. *Given the ODE* (7.1) *with f globally Lipschitz, then for any y and w in $\mathbb{R}^d$ and $0 < h < 1$ we have*

$$\|\varphi_h(y) - \varphi_h(w)\| \leq (1 + Ch)\,\|y - w\|\,, \tag{7.14}$$
$$\|\varphi_h(y) - \varphi_h(w) - (y - w)\| \leq Ch\,\|y - w\|\,, \tag{7.15}$$

where C is a positive constant independent of h and of the initial condition y_0.

The proof of Lemma 7.17 follows from the global Lipschitz continuity of f and the Grönwall inequality. We can now prove the main result on mean-square convergence.

Proof of Theorem 7.16. In the following, we denote by C a constant that does not depend on h and on the initial condition y_0 whose value may change from line to line. Let us define $e_k^2 \coloneqq \mathbb{E}\left[\|Y_k - y(t_k)\|^2\right]$. Adding and subtracting the exact flow applied to the numerical solution, we obtain

$$\begin{aligned} e_{k+1}^2 = \mathbb{E}\left[\|\psi_{H_k}(Y_k) - \varphi_h(Y_k)\|^2\right] + \mathbb{E}\left[\|\varphi_h(Y_k) - \varphi_h(y(t_k))\|^2\right] \\ + 2\,\mathbb{E}\left[\Big((\varphi_h(Y_k) - \varphi_h(y(t_k)))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\Big)\right]. \end{aligned} \tag{7.16}$$

Let us consider the three terms in (7.16) separately. For the first term, we have by Lemma 7.15

$$\mathbb{E}\left[\|\psi_{H_k}(Y_k) - \varphi_h(Y_k)\|^2\right] \leq Ch^{\min\{2p+1,2(q+1)\}}. \tag{7.17}$$

For the second term, due to (7.14), we have

$$\mathbb{E}\left[\|\varphi_h(Y_k) - \varphi_h(y(t_k))\|^2\right] \leq (1 + Ch)^2 e_k^2. \tag{7.18}$$

Let us now define $Z = \varphi_h(Y_k) - \varphi_h(y(t_k)) - (Y_k - y(t_k))$. We can rewrite the scalar product as

$$\begin{aligned} \mathbb{E}\left[(\varphi_h(Y_k) - \varphi_h(y(t_k)))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right] \\ = \mathbb{E}\left[(Y_k - y(t_k))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right] \\ + \mathbb{E}\left[Z^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right]. \end{aligned} \tag{7.19}$$

We bound the two terms in (7.19) separately. For the first term, by the law of total expectation, we have

$$\begin{aligned} \mathbb{E}\left[(Y_k - y(t_k))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y_k - y(t_k))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k)) \mid Y_k\right]\right] \\ &= \mathbb{E}\left[(Y_k - y(t_k))^\top \mathbb{E}\left[\psi_{H_k}(Y_k) - \varphi_h(Y_k) \mid Y_k\right]\right]. \end{aligned}$$

Applying the Cauchy–Schwarz inequality to the outer expectation we get

$$\begin{aligned} \mathbb{E}\left[(Y_k - y(t_k))^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right] &\leq \mathbb{E}\left[\|\mathbb{E}\left[\psi_{H_k}(Y_k) - \varphi_h(Y_k) \mid Y_k\right]\|^2\right]^{1/2} e_k \\ &\leq Ch^{\min\{2p+1,q+1\}} e_k, \end{aligned}$$

where we applied Lemma 7.7. We now consider the second term in (7.19). By the Cauchy–Schwarz inequality we have

$$\mathbb{E}\left[Z^\top (\psi_{H_k}(Y_k) - \varphi_h(Y_k))\right] \leq \mathbb{E}\left[\|Z\|^2\right]^{1/2} \mathbb{E}\left[\|\psi_{H_k}(Y_k) - \varphi_h(Y_k)\|^2\right]^{1/2}.$$

We now apply (7.15) and Lemma 7.15 to obtain

$$\mathbb{E}\left[Z^\top \left(\psi_{H_k}(Y_k) - \varphi_h(Y_k)\right)\right] \leq Ch^{\min\{p+3/2,q+2\}} e_k.$$

We can hence bound the scalar product in (7.19) with Young's inequality and assuming $h < 1$ as

$$\begin{aligned}\mathbb{E}\left[\left(\varphi_h(Y_k) - \varphi_h(y(t_k))\right)^\top \left(\psi_{H_k}(Y_k) - \varphi_h(Y_k)\right)\right] &\leq Ch^{\min\{p+3/2,q+1\}} e_k \\ &\leq \frac{he_k^2}{2} + C\frac{h^{\min\{2p+2,2q+1\}}}{2}.\end{aligned} \tag{7.20}$$

Combining (7.17), (7.18) and (7.20), we have

$$e_{k+1}^2 \leq Ch^{\min\{2p+1,2q+1\}} + (1+Ch)e_k^2,$$

which implies the desired result by Lemma 7.11 and since $e_0 = 0$. □

Remark 7.18. The difference between global and local orders of convergence, i.e., between (7.12) and (7.13), is not exactly one, as it usually is in the purely deterministic case. In fact, due to the independence of the random variables there is only a 1/2 loss in the random part of the exponent, while the natural loss of one order is verified in the deterministic component.

Remark 7.19. As for the additive noise method proposed in [39], the result of mean-square convergence suggests that a reasonable choice for the noise scale p is to fix $p = q$, where q is the order of the Runge–Kutta method ψ_h. In this way, the properties of convergence of the underlying deterministic method are preserved, while yielding a probabilistic interpretation of the numerical solution.

Remark 7.20. Let $\Psi\colon \mathbb{R}^d \to \mathbb{R}$ be a function in $\mathcal{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$, let $T > 0$ and let us consider the problem of approximating the quantity $\Psi(y(T))$. As highlighted in Section 6.1.2, in perturbation-based methods such as the RTS-RK one has to recur to Monte Carlo approximations. In particular, let M be a positive integer and let

$$\widehat{\Psi}_{M,N} \coloneqq \frac{1}{M}\sum_{i=1}^{M} \Psi(Y_N^{(i)}),$$

where $\{Y_N^{(i)}\}_{i=1}^M$ are i.i.d. samples from the RTS-RK approximating the solution $y(T)$ at final time. Due to Theorem 7.16 and Theorem 7.12, Theorem 6.7 applies and it therefore holds

$$\mathbb{E}\left[\left(\widehat{\Psi}_{M,N} - \Psi(y(T))\right)^2\right] \leq C\left(h^{2\min\{2p,q\}} + \frac{h^{2\min\{p,q\}}}{M}\right). \tag{7.21}$$

Balancing the two terms for the different values that p can take yields the optimal choice of the number of trajectories

$$M = \begin{cases} \mathcal{O}(1), & \text{if } p \geq q, \\ \mathcal{O}(h^{2(p-q)}), & \text{if } p < q \leq 2p, \\ \mathcal{O}(h^{-2p}), & \text{if } 2p < q. \end{cases}$$

Hence, if the scaling $p = q$ is chosen following the indication of Remark 7.19, then it is sufficient to sample $\mathcal{O}(1)$ trajectories to obtain convergent and optimal Monte Carlo estimators. Let us finally remark that in order to have uncertainty quantification for a fixed value $h > 0$ it is still necessary to draw $M > 1$ samples. Indeed, Theorem 6.7 does not provide an indication of how the value of M should be chosen in order to have a good empirical description of the probability measure induced by the RTS-RK method, but still ensures quantitatively that the Monte Carlo estimators drawn from this distribution have a good quality.

7.4 Conservation of First Integrals

There exist numerical methods for ODEs endowed with certain geometric properties which are particularly useful when integrating ODEs with a similar geometric structure [60]. Wwe investigate here whether the random choice of time steps in (7.2) spoils the properties of the underlying deterministic Runge–Kutta method. We consider here the conservation of first integrals of motion, and start by recalling the definition of first integral for an ODE.

Definition 7.21. Given a function $I\colon \mathbb{R}^d \to \mathbb{R}$, then $I(y)$ is a first integral of (7.1) if $I'(y)f(y) = 0$ for all $y \in \mathbb{R}^d$.

If this property of the ODE is conserved by a numerical integrator, i.e., if for the any $y \in \mathbb{R}^d$ it is true that $I(\psi_h(y)) = I(y)$, then we say that the numerical method conserves the first integral. In particular, this implies that the first integral I is conserved along the trajectory of the numerical solution, i.e., $I(y_k) = I(y_0)$ for all $k \geq 0$.

Example 7.22. To illustrate this concept we first discuss the case of linear first integrals, which can be seen as a general case of the conservation of mass in physical systems. Let us consider a linear first integral $I(y) = v^\top y$ and any Runge–Kutta method with coefficients $\{b_i\}_{i=1}^s$, $\{a_{ij}\}_{i,j=1}^s$. Then, we have for a time step $H_0 > 0$

$$I(Y_1) = v^\top y_0 + H_0 \sum_{i=1}^{s} b_i v^\top f\left(y_0 + H_0 \sum_{j=1}^{s} a_{ij} K_j\right),$$

where $\{K_i\}_{i=1}^s$ are the internal stages of the Runge–Kutta method. Since $I(y)$ is a first integral, $v^\top f(y) = 0$ for any $y \in \mathbb{R}^d$. Hence $I(Y_1) = I(y_0)$ and iteratively $I(Y_k) = I(y_0)$ for all $k \geq 0$ along the numerical trajectory. The equality above shows that any RTS-RK method conserves linear first integrals path-wise, or in the strong sense.

It is known that no Runge–Kutta method can conserve any polynomial invariant of order $n \geq 3$ [60, Theorem IV.3.3]. Nonetheless, for some particular problems there exist tailored Runge–Kutta methods which can conserve polynomial invariants of higher order. We therefore can state the following general result.

Theorem 7.23. *Let $I(y)$ be a first integral for* (7.1) *and ψ_h be the numerical flow of a Runge–Kutta scheme for* (7.1)*. If the scheme defined by ψ_h conserves $I(y)$ for any $h > 0$, then the RTS-RK method given in* (7.2) *conserves $I(y)$ almost surely.*

Proof. If $I(\psi_h(y)) = I(y)$ for any h, then $I(\psi_{H_0}(y)) = I(y)$ almost surely for any value that H_0 can assume. □

We now consider quadratic first integrals, i.e., first integrals of the form $I(y) = y^\top S y$ with S a symmetric matrix, which are conserved by Runge–Kutta methods that satisfy the hypotheses of Cooper's theorem [60, Theorem IV.2.2]. The conservation of quadratic first invariants is of the utmost importance, e.g., for Hamiltonian systems, as it implies the symplecticity of the scheme. It is known [60, Theorem IV.2.1] that all Gauss methods conserve quadratic first integrals. The simplest member of this class of methods is the implicit midpoint rule, which is a one-stage method defined by coefficients $b_1 = 1$ and $a_{11} = 1/2$.

Corollary 7.24. *If the Runge–Kutta scheme defined by ψ_h conserves quadratic first integrals then the numerical method* (7.2) *conserves quadratic first integrals almost surely.*

Proof. This result is a direct consequence of Theorem 7.23. □

The properties above for the RTS-RK method are not satisfied by the additive noise method presented in [39]. In particular, let us remark that the conservation of first integrals is exact for any trajectory of the RTS-RK method, and is not an average property. In other words, we can say that (7.2) conserves linear first integrals in the strong sense. For the AN-RK method of Section 6.3.1, we have

$$\begin{aligned} I(Y_1) &= v^\top y_0 + h\sum_{i=1}^{s} b_i v^\top f\left(y_0 + h\sum_{j=1}^{s} a_{ij}K_j\right) + v^\top \xi_0(h), \\ &= v^\top (y_0 + \xi_0(h)). \end{aligned}$$

If the random variable ξ_0 is zero-mean, then $\mathbb{E}\left[I(Y_1)\right] = I(y_0)$ and iteratively along the solution $\mathbb{E}\left[I(Y_k)\right] = I(y_0)$. Linear first integrals are therefore conserved in average, but not in a path-wise fashion. For quadratic first integrals, we have instead that the additive noise method does not conserve them neither path-wise nor in the weak sense, as we have

$$\begin{aligned} I(Y_1) &= (\psi_h(y_0) + \xi_0(h))^\top S(\psi_h(y_0) + \xi_0(h)) \\ &= I(y_0) + 2\xi_0(h)^\top S\psi_h(y_0) + \xi_0(h)^\top S\xi_0(h). \end{aligned}$$

Under Assumption 6.12, i.e., if the random variables are zero-mean and if there exists a matrix Q such that $\mathbb{E}\left[\xi_0(h)\xi_0(h)^\top\right] = Qh^{2p+1}$ for some $p \geq 1$ (see [39, Assumption 1]) we then have

$$\mathbb{E}\left[I(Y_1)\right] = I(y_0) + Q : Sh^{2p+1}. \tag{7.22}$$

Hence, along the trajectories of the solution a bias is introduced in the first integral which persists even in the mean sense. In general, Theorem 7.23 is not valid for the additive noise method, as the random contribution drives the first integral far from its true value at each time step. In practice, this could produce large deviations of the numerical approximation from the true solution, especially in the long time regime.

7.5 Hamiltonian systems

A class of dynamical systems of particular interest for their geometric properties is the class of Hamiltonian systems. Given a function $Q\colon \mathbb{R}^{2d} \to \mathbb{R}$, called the Hamiltonian, Hamiltonian systems can be written as

$$y' = J^{-1}\nabla Q(y), \quad y(0) = y_0 \in \mathbb{R}^{2d}, \tag{7.23}$$

where the matrix $J \in \mathbb{R}^{2d\times 2d}$ is defined as

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix},$$

and where I is the identity matrix in $\mathbb{R}^{d\times d}$. The Hamiltonian Q is a first integral for (7.23), hence we require numerical integrators to conserve the energy, or at least not to deviate from its true value in an uncontrolled fashion. As it was shown in the previous section, when Q is a polynomial it is possible to obtain exact conservation with deterministic integrators and with their probabilistic counterparts obtained with the RTS-RK method. If Q is not a polynomial, exact conservation is in general not achievable, but a good approximation of the energy over long time spans is achievable through the notion of symplectic differentiable maps.

Definition 7.25 (Definition VI.2.2 in [60])**.** Let $U \subset \mathbb{R}^{2d}$ be a non-empty open set. A differentiable map $g\colon U \to \mathbb{R}^{2d}$ is called symplectic if the Jacobian matrix g' is everywhere symplectic, i.e., if

$$(g')^\top J g' = J.$$

It is well-known that the flow $\varphi_t \colon \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ of any system of the form (7.23) is symplectic. In a natural manner, a numerical integrator is called symplectic if its numerical flow ψ_h is a symplectic map whenever it is applied to a smooth Hamiltonian system [60, Definition VI.3.1]. In the following, we will analyse both the local and global properties of the RTS-RK method built on symplectic integrators and applied to (7.23).

7.5.1 Symplecticity of the RTS-RK Method

It has been pointed out [60, Section VIII.1] that applying an adaptive step size technique to a symplectic method can destroy its symplecticity. Therefore, Skeel and Gear [128] write any adaptive technique in terms of a map $\tau(y, h)$ such that the k-th time step h_k is selected as $h_k = \tau(y_k, h)$, where h is a base value for the time step. Hence, in order to have again a symplectic method for variable time steps, the new condition to be satisfied is

$$V^\top J V = J, \quad V = \partial_y \psi_{\tau(y,h)}(y) + \partial_t \psi_{\tau(y,h)}(y) \partial_y \tau(y, h)^\top.$$

Let us now consider the RTS-RK method based on a symplectic deterministic integrator. We have the following lemma.

Lemma 7.26. *If the flow ψ_h of the deterministic integrator is symplectic, then the flow of the random time-stepping probabilistic method* (7.2) *is symplectic.*

Proof. For the RTS-RK scheme, the k-th time step H_k is generated by a random mapping as $H_k = \tau(y, h) = \tau(h) = h\Theta_k$, where Θ_k are appropriately scaled random variables such that H_k satisfies Assumption 7.2. Hence, τ is independent of y, i.e., $\partial_y \tau(y, h) = 0$, and with the notation introduced above

$$V = \partial_y \psi_{\tau(h)}(y).$$

Therefore, by the symplecticity of ψ_t the condition $V^\top J V = J$ is satisfied and the flow map of the RTS-RK method is symplectic. □

Let us remark that the local symplecticity of the flow map is not sufficient for good conservation of the Hamiltonian for the numerical solution. Global properties of approximation of the energy are therefore presented below.

7.5.2 Long-time Conservation of Hamiltonians

We now wish to study the mean conservation of the Hamiltonian along the trajectories of the RTS-RK method based on symplectic integrators. Our goal is obtaining a bound on the quantity $\mathbb{E}\left[|Q(Y_n) - Q(y_0)|\right]$ that holds over long times. Showing theoretically long time conservation of the energy function in Hamiltonian systems requires backward error analysis. In the following, we will introduce the basis of this technique and show how they apply to our probabilistic integrator. For further details, a comprehensive treatment of backward error analysis ought to be found in [60, Chapter IX].

The first ingredient needed to perform a rigorous backward error analysis is a rather strong assumption on the regularity of the ODE, see e.g. [60, Section IX.7].

Assumption 7.27. The function f is analytic in a neighbourhood of the initial condition y_0 and there exist constants $C, R > 0$ such that $\|f(y)\| \leq C$ for $\|y - y_0\| \leq 2R$.

In general, backward error analysis is based on determining a modified equation $y' = \widetilde{f}(y)$ such that the numerical approximation is its exact solution. Hence, the function $\widetilde{f}$ will both depend on the original ODE and on the numerical flow map ψ_h. In particular, for an integrator of order q the modified equation is given by a function $\widetilde{f}$ defined as

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \dots,$$

where the functions $\{f_i\}_{i>q}$ are uniquely determined by f, its derivatives and by the coefficients of the Runge–Kutta method. The exactness of the numerical solution for the modified equation is nonetheless only formal, as the infinite sum defining $\widetilde{f}$ is not guaranteed to converge. Thus, it is necessary to truncate the sum in order to perform a rigorous analysis, i.e.,

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \ldots + h^{N-1} f_N(y). \tag{7.24}$$

where $q < N < \infty$ is the truncation index. Let us remark that in the following we will always refer to the truncated function above when using the symbol $\widetilde{f}$. The truncation of the infinite sum implies that the numerical solution is not exact for the modified equation anymore. In particular, the error committed over one step on the modified equation is given by (see e.g. [60, Theorem IX.7.6])

$$\|\widetilde{\varphi}_h(y) - \psi_h(y)\| \leq Che^{-\kappa/h}, \tag{7.25}$$

where $\widetilde{\varphi}$ is the exact flow of the modified equation and κ and C are constants depending on the coefficients of the method and on the regularity of f.

It is possible to prove (see e.g. [60, Section IX.8]) that for a Hamiltonian system (7.23) and a symplectic integrator the modified equation is still a Hamiltonian system, i.e., there exists a modified Hamiltonian $\widetilde{Q}$ defined as

$$\widetilde{Q}(y) = Q(y) + h^q Q_{q+1}(y) + \ldots + h^{N-1} Q_N(y), \tag{7.26}$$

such that $\widetilde{f} = J^{-1}\nabla\widetilde{Q}$. The estimate (7.25) implies that the modified Hamiltonian is almost conserved by the symplectic integrator. In particular, if Q is Lipschitz, we have

$$\left|\widetilde{Q}(\psi_h(y)) - \widetilde{Q}(y)\right| \leq Che^{-\kappa/h}. \tag{7.27}$$

The bound above guarantees that the modified Hamiltonian is well approximated for a long time, and as a consequence that the original Hamiltonian is almost conserved for the same time span. In particular, the following result is valid, see e.g. [60, Theorem IX.8.1.] or [21].

Theorem 7.28. *Under Assumption 7.27 and for h sufficiently small, if the numerical solution y_n given by a symplectic method of order q applied to an Hamiltonian system is close enough to the initial condition y_0, then*

$$\widetilde{Q}(y_n) = \widetilde{Q}(y_0) + \mathcal{O}(e^{-\kappa/2h}),$$
$$Q(y_n) = Q(y_0) + \mathcal{O}(h^q).$$

over exponentially long time intervals $nh \leq e^{\kappa/2h}$.

The randomisation of the time steps implies that a general modified equation does not exist. Nonetheless, due to Lemma 7.26, it is possible to construct locally a random Hamiltonian modified equation at each time step. We thus define at each step the random modified Hamiltonian as

$$\widehat{Q}_j(y) = Q(y) + H_j^q Q_{q+1}(y) + \ldots + H_j^{N-1} Q_N(y). \tag{7.28}$$

As for the deterministic case, the random modified Hamiltonian $\widehat{Q}$ will be almost conserved by the numerical flow. In particular, we define the random local truncation error as

$$\eta_j \coloneqq \widehat{Q}_j(\psi_{H_j}(y)) - \widehat{Q}_j(y), \tag{7.29}$$

which, in light of (7.27), satisfy

$$|\eta_j| \leq C H_j e^{-\kappa/H_j}, \tag{7.30}$$

almost surely. In order to prove the conservation of the Hamiltonian over long time for the RTS-RK method, it is necessary to introduce a technical assumption on the higher moments of the random time steps.

Assumption 7.29. There exists $\bar{r} > 1$ such that for any $1 < r < \bar{r}$, the random time steps $\{H_j\}_{j\geq 0}$ satisfy

$$\mathbb{E}\left[H_j^r\right] = h^r + C_r h^{2p+r-1},$$

where p is defined in Assumption 7.2 and $C_r > 0$ satisfies $C_{2r} > 2C_r$ and is independent of h. Moreover, there exists $m, M > 0$ with $M > m$ such that $mh \leq H_j \leq Mh$ almost surely for all $j \geq 0$.

This assumption guarantees that the higher moments of the random time steps are close to the corresponding powers of h in the mean and mean-square sense. In particular, it is possible to verify that

$$\begin{aligned}
&\mathbb{E}\left[H_j^r - h^r\right] = C_r h^{2p+r-1},\\
&\mathbb{E}\left[(H_j^r - h^r)^2\right] = (C_{2r} - 2C_r) h^{2p+2r-1}.
\end{aligned}$$

Then, for any $r, s > 1$ such that $r + s < R$, it holds

$$\begin{aligned}
&\mathbb{E}\left[H_j^{r+s} - h^{r+s}\right] = \widehat{C}_{r,s} h^s \,\mathbb{E}\left[H_j^r - h^r\right],\\
&\mathbb{E}\left[(H_j^{r+s} - h^{r+s})^2\right] = \widetilde{C}_{r,s} h^{2s} \,\mathbb{E}\left[(H_j^r - h^r)^2\right],
\end{aligned}$$

where $\widehat{C}_{r,s} = C_{r+s}/C_r$ and $\widetilde{C}_{r,s} = (C_{2(r+s)} - 2C_{r+s})/(C_{2r} - 2C_r)$. Finally, let us remark that Assumption 7.29 is satisfied for the uniform random time steps $H_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(h - h^{p+1/2}, h + h^{p+1/2})$ introduced in Example 7.3. Let us now give an explicit which holds for the random variables η_j defined in (7.29).

Lemma 7.30. *Suppose that Assumption 7.2, Assumption 7.6 and Assumption 7.29 hold true, and suppose that $0 < h \leq 1$. Then the random variables η_j defined in* (7.29) *satisfy*

$$\mathbb{E}\left[|\eta_j|^r\right] \leq C h^{\min\{r, p+r-3/2\}} e^{-r\kappa/(Mh)},$$

where $C > 0$ is independent of h and for all $r \in \mathbb{N}$ with $r \geq 1$.

The proof of Lemma 7.30 is given in Section 7.8. Let us furthermore introduce two lemmas, which will be employed for proving long-time conservation of Hamiltonians. Let us remark that Lemma 7.31 holds for generic positive integers n, q, N.

Lemma 7.31. *Let n, q, N be positive integers with $N > q$, and let us define the sets of real numbers $a = a_{n,q,N} \coloneqq \{a_{jk}, j = 0, \dots, n-1,\ k = q, \dots, N-1\}$ and $b = b_n \coloneqq \{b_j, j = 0, \dots, n-1\}$. Then*

$$\left(\sum_{j=0}^{n-1}\left(\sum_{k=q}^{N-1} a_{jk} + b_j\right)\right)^2 = \sum_{j=0}^{n-1} a_{jq}^2 + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} a_{jq}a_{iq} + R(a) + S(a,b),$$

where the remainder $R(a)$ can be written as $R = R_1 + R_2 + R_3$, with

$$\begin{aligned}
R_1(a) &= \sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1} a_{jk}^2, \qquad & R_2(a) &= 2\sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl},\\
R_3(a) &= 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\sum_{k=q}^{N-1}\sum_{\substack{l=q\\ l+k>2q}}^{N-1} a_{jk}a_{il},
\end{aligned}$$

and the remainder $S(a,b)$ can be written as $S = S_1 + S_2 + S_3 + S_4$, with

$$S_1(a,b) = \sum_{j=0}^{n-1} b_j^2, \qquad S_2(a,b) = 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} b_i b_j,$$
$$S_3(a,b) = 2\sum_{j=1}^{n-1}\sum_{k=q}^{N-1} b_j a_{jk}, \qquad S_4(a,b) = 2\sum_{j=1}^{n-1}\sum_{i=0}^{n-1}\left(b_j \sum_{k=q}^{N-1} a_{ik} + b_i \sum_{k=q}^{N-1} a_{jk}\right).$$

The proof of Lemma 7.31 is given in Section 7.8.

Lemma 7.32. *Let Assumption 7.2 hold with $p \geq 3/2$ and $h < 1$, and let Assumption 7.4, Assumption 7.27 and Assumption 7.29 hold. Moreover, let q be specified in Assumption 7.4 and N be the truncation index of the modified right hand side* (7.24)*. Let us consider the sets of real-valued random variables $\Delta \coloneqq \{\Delta_{j,k}(H_j^k - h^k), j = 0,\dots,n-1,\ k = q,\dots,N-1\}$, where $\Delta_{j,k} \coloneqq Q_{k+1}(Y_j) - Q_{k+1}(Y_{j+1})$ and $\eta \coloneqq \{\eta_j, j = 0,\dots,n-1\}$. Then, with the notation of Lemma 7.31, there exist positive constants C_1, C_2 independent of h and n, but possibly dependent on q and N, such that*

$$\mathbb{E}\left[R(\Delta)\right] \leq C_1\left(t_n h^{2(p+q+1/2)} + t_n^2 h^{2(2p+q-1/2)}\right),$$
$$\mathbb{E}\left[S(\Delta,\eta)\right] \leq C_2\left((t_n h + t_n^2)e^{-2\kappa/(Mh)} + (t_n h^{p+q+1/2} + t_n^2 h^{2p+q-1})e^{-\kappa/(Mh)}\right),$$

where $t_n = nh$.

The proof of Lemma 7.32 is also given in Section 7.8. It is now possible to prove a result of long conservation of the Hamiltonian for symplectic RTS-RK methods.

Theorem 7.33. *Let $0 < h \leq 1$. Suppose that Assumption 7.2 holds for $p \geq 3/2$, that Assumption 7.6 and Assumption 7.27 hold, and that Assumption 7.29 holds with $\bar{r}$ sufficiently large. Moreover, let Y_n be the solution given by the RTS-RK method built on a symplectic integrator of order q applied to a Hamiltonian system with Hamiltonian Q. If $Y_0 = y_0$ and the numerical solution Y_n is close enough to the initial condition y_0 almost surely, then there exist a constant $C > 0$ independent of h and n such that*

$$\mathbb{E}\left[|Q(Y_n) - Q(y_0)|\right] \leq Ch^q,$$

for time intervals of length

$$t_n = \mathcal{O}\left(\min\left\{h^{1-2p}, e^{\kappa/(4Mh)}h^{-(2p+2q-1)/4}, e^{\kappa/(2Mh)}\right\}\right)$$

where p is given in Assumption 7.2 and M in Assumption 7.29.

Proof. In the following proof, we denote by C a positive constant independent of h and n which can possibly change value from line to line. Let us first consider the modified Hamiltonian $\widetilde{Q}$ and expand the difference $\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)$ in a telescopic sum as

$$\widetilde{Q}(Y_n) - \widetilde{Q}(y_0) = \sum_{j=0}^{n-1}\left(\widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j)\right). \tag{7.31}$$

We then consider each element of the sum, add and subtract the random modified Hamiltonian $\widehat{Q}_j$ computed in Y_{j+1} thus obtaining

$$\begin{aligned}\widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j) &= \widetilde{Q}(Y_{j+1}) - \widehat{Q}_j(Y_{j+1}) + \widehat{Q}_j(Y_{j+1}) - \widetilde{Q}(Y_j)\\ &= \widetilde{Q}(Y_{j+1}) - \widehat{Q}_j(Y_{j+1}) + \widehat{Q}_j(Y_j) - \widetilde{Q}(Y_j) + \eta_j.\end{aligned}$$

Hence, by applying the definition (7.26) of $\widetilde{Q}$ and (7.28) of $\widehat{Q}_j$, we get

$$\widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j) = \sum_{k=q}^{N-1} (H_j^k - h^k)\Delta_{j,k} + \eta_j,$$

where $\Delta_{j,k}$ is defined in Lemma 7.32. Going back to (7.31), applying Jensen's inequality and Lemma 7.31 we obtain

$$\begin{aligned}
\mathbb{E}\left[\left|\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)\right|\right]^2 &\leq \mathbb{E}\left[\left(\sum_{j=0}^{n-1}\left(\sum_{k=q}^{N-1}(H_j^k - h^k)\Delta_{j,k} + \eta_j\right)\right)^2\right] \\
&= \sum_{j=0}^{n-1}\mathbb{E}\left[(H_j^q - h^q)^2\Delta_{j,q}^2\right] \\
&\quad + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\mathbb{E}\left[(H_j^q - h^q)\Delta_{j,q}(H_i^q - h^q)\Delta_{i,q}\right] \\
&\quad + \mathbb{E}\left[R(\Delta)\right] + \mathbb{E}\left[S(\Delta,\eta)\right].
\end{aligned} \tag{7.32}$$

The first term in (7.32) satisfies

$$\left(\sum_{j=0}^{n-1}\mathbb{E}\left[(H_j^q - h^q)^2\Delta_{j,q}^2\right]\right)^{1/2} \leq C\sqrt{t_n}h^{p+q}, \tag{7.33}$$

due to (7.52). Now, considering (7.54), we obtain that the second term in (7.32) satisfies

$$\left(2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\mathbb{E}\left[(H_j^q - h^q)\Delta_{j,q}(H_i^q - h^q)\Delta_{i,q}\right]\right)^{1/2} \leq Ct_nh^{2p+q-1}.$$

For the remainder term $\mathbb{E}\left[R(\Delta)\right]$, due to Lemma 7.32 we get

$$\mathbb{E}\left[R(\Delta)\right]^{1/2} \leq C\left(\sqrt{t_n}h^{p+q+1/2} + t_nh^{2p+q-1/2}\right).$$

For the remainder term $\mathbb{E}\left[S(\Delta,\eta)\right]$, due to Lemma 7.32 and since $h \leq 1$ and $p \geq 3/2$ by assumption, we get

$$\begin{aligned}
\mathbb{E}\left[S(\Delta,\eta)\right]^{1/2} &\leq C\left(t_n^2\left(e^{-2\kappa/(Mh)} + h^{p+q+1/2}e^{-\kappa/(Mh)}\right)\right)^{1/2} \\
&\leq Ct_n\left(e^{-\kappa/(Mh)} + h^{(2p+2q+1)/4}e^{-\kappa/(2Mh)}\right).
\end{aligned} \tag{7.34}$$

Finally, taking the square root of both sides of (7.32), replacing the expressions we obtained above and since $h \leq 1$, we get that the modified Hamiltonian satisfies

$$\begin{aligned}
\mathbb{E}\left[\left|\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)\right|\right] \leq C\big(&\sqrt{t_n}h^{p+q} + t_nh^{2p+q-1} \\
&+ t_n\left(e^{-\kappa/(Mh)} + h^{(2p+2q+1)/4}e^{-\kappa/(2Mh)}\right)\big).
\end{aligned}$$

Hence, imposing for a constant $C > 0$

$$t_n \leq C\min\{h^{1-2p}, e^{\kappa/(4Mh)}h^{-(2p+2q-1)/4}, e^{\kappa/(2Mh)}\},$$

and since exponential terms are dominated by polynomial terms (see e.g. [60, Theorem IX.8.1]), we obtain

$$\mathbb{E}\left[\left|\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)\right|\right] \leq Ch^q. \tag{7.35}$$

Finally, applying the triangle inequality, since for all $y \in \mathbb{R}^d$ it holds $\left|Q(y) - \widetilde{Q}(y)\right| \leq Ch^q$ by definition of the modified Hamiltonian $\widetilde{Q}$ and due to (7.35) we get

$$\begin{aligned}\mathbb{E}\left[|Q(Y_n) - Q(y_0)|\right] &\leq \mathbb{E}\left[\left|Q(Y_n) - \widetilde{Q}(Y_n)\right|\right] + \mathbb{E}\left[\left|Q(y_0) - \widetilde{Q}(y_0)\right|\right] + \mathbb{E}\left[\left|\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)\right|\right] \\ &\leq Ch^q,\end{aligned}$$

which is the desired result. □

Remark 7.34. The result of Theorem 7.33 is consistent with the theory of deterministic symplectic integrators. In fact, in the limit $p \to \infty$, one can choose the coefficient M in Assumption 7.29 arbitrarily close to 1 and we have

$$\mathbb{E}\left[|Q(Y_n) - Q(y_0)|\right] \overset{=}{} \mathcal{O}(h^q),$$

for exponentially long time spans $t_n = \mathcal{O}\left(e^{\kappa/(2h)}\right)$, which is consistent with the theory of deterministic symplectic integrators summarised by Theorem 7.28.

Remark 7.35. It has been observed (see for example [59, 60]) that adopting variable step sizes in symplectic integration destroys the good properties of conservation of the Hamiltonian. In particular, the error on the Hamiltonian has a linear drift in time, i.e., the approximation has the same quality as the one given by a standard non-symplectic algorithm. Conversely, Theorem 7.33 proves that random step sizes do not spoil, under the assumptions specified above, the good long time properties of symplectic integrators with fixed step size.

Remark 7.36. As it can be noticed in the proof of Lemma 7.32, we introduce the assumption $p \geq 3/2$ in order to simplify the terms composing the remainder $S(\Delta, \eta)$. In case $1 \leq p < 3/2$, e.g. when the symplectic Euler method is employed ($q = 1$) and the natural scaling $p = q$ is chosen, the $\mathcal{O}(h^q)$ approximation of the Hamiltonian still holds but with a slight reduction in the exponential terms appearing in the time span of validity.

Remark 7.37. Let us remark that in order for (7.34) to hold we implicitly assumed $t_n \geq 1$ to bound $\sqrt{t_n} \leq t_n$. If $t_n < 1$, we can bound every appearance of t_n from (7.33) to (7.34) as $t_n \leq 1$, and the desired result would still hold.

7.6 Bayesian Inference

In this section, we introduce a Bayesian inverse problem in the ODE setting and illustrate how the RTS-RK method can be employed in this framework. Let us recall that Bayesian inverse problems are introduced in Chapter 1 and the application of probabilistic methods in this framework in Section 6.2. Here, we give only the details which are specific to the ODE case.

Let us consider a function $f_\vartheta \colon \mathbb{R}^d \to \mathbb{R}^d$ which depends on a real parameter $\vartheta \in X$, where $X \equiv \mathbb{R}^n$, and the ODE

$$y'_\vartheta = f_\vartheta(y), \quad y(0) = y_0 \in \mathbb{R}^d.$$

We write in this case that the exact solution is given by $y_\vartheta(t) = \varphi_t(y_0; \vartheta)$, where we write explicitly the dependence of the flow on the parameter. Let us remark that the initial condition y_0 could be as well subject to inference, and it can be seamlessly included in the inversion procedure.

Let us remark that we could as well consider a non-parametric setting such as the one introduced in Chapter 1, thus approaching the problem with a finite-dimensional approximation as in

Section 1.2. Nevertheless, we consider here for simplicity X to be finite-dimensional, and note that this suffices for demonstrating the potential of probabilistic methods in Bayesian inversion.

We let in this case the forward model $\mathcal{G}\colon X \to \mathbb{R}^m$ be given by $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$, where the solution operator $\mathcal{S}\colon X \to \mathcal{C}^0([0,T])$ maps the parameter ϑ into the flow $\{\varphi_t(y_0;\vartheta)\}_{0\leq t\leq T}$, and where $\mathcal{O}\colon \mathcal{C}^0([0,T]) \to \mathbb{R}^m$ outputs pointwise observations of the solution, i.e.

$$\mathcal{O}\left(\{\varphi_t(y_0;\vartheta)\}_{0\leq t\leq T}\right) = \begin{pmatrix} y(t_1^*) & y(t_2^*) & \cdots & y(t_m^*) \end{pmatrix}^\top,$$

for observation points $0 < t_1^* < t_2^* < \ldots < t_m^* \leq T$. We then consider the observation model

$$z = \mathcal{G}(\vartheta) + \beta,$$

where $\beta \sim \mathcal{N}(0,\Gamma)$ is a Gaussian source of noise, and the inverse problem

$$\text{find } \vartheta \in X \text{ given observations } z^* = \mathcal{G}(\vartheta^*) + \beta,$$

where $\vartheta^* \in X$ is the true value of the parameter. Given a Gaussian prior $\mu_0 = \mathcal{N}(0,\Gamma_0)$, where Γ_0 is a positive-definite covariance matrix on X, the inverse problem is well-posed in the Bayesian sense (see e.g. [87]), and the posterior μ is given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{Z}\exp\left(-\Phi(\vartheta;z)\right),$$

where for any $z \in \mathbb{R}^m$ the potential $\Phi(\cdot;z)\colon X \to \mathbb{R}$ is given by

$$\Phi(\vartheta;z) = \frac{1}{2}\left\|\Gamma^{-1/2}\left(\mathcal{G}(\vartheta) - z\right)\right\|_2^2,$$

where we recall that Γ is the covariance of the Gaussian noise, and where Z is the normalization constant

$$Z = \int_X \exp\left(-\Phi(\vartheta;z)\right)\,\mathrm{d}\mu_0(\vartheta).$$

It is then possible to consider the deterministic and the probabilistic approximations μ_h and $\widetilde{\mu}_h$ of the posterior, respectively, as described in detail in Section 6.2. In particular, the measures μ_h and $\widetilde{\mu}_h$ are obtained respectively by discretizing the solution operator $\mathcal{S}$ with a Runge–Kutta integrator and with the RTS-RK method based on the same deterministic scheme. For the RTS-RK, we then have the random forward operator $\widetilde{\mathcal{G}}_h = \mathcal{O} \circ \widetilde{\mathcal{S}}_h$, where $\widetilde{\mathcal{S}}_h$ associates the parameter ϑ to the random approximate RTS-RK solution. Let us recall that $\widetilde{\mu}_h$ is thus a random measure, and should be approximated deterministically. We choose in this setting to consider the marginal approximation $\widetilde{\mu}_{h,\mathrm{mar}}$ defined in Section 6.2, which we recall to be given by

$$\frac{\mathrm{d}\widetilde{\mu}_{h,\mathrm{mar}}}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{\mathbb{E}\left[\widetilde{Z}_h\right]}\,\mathbb{E}\left[\exp\left(-\widetilde{\Phi}_h(\vartheta;z)\right)\right],$$

where expectation is with respect to the measure induced by the random time steps, where the random potential $\widetilde{\Phi}_h$ is given by

$$\widetilde{\Phi}_h(\vartheta;z) = \frac{1}{2}\left\|\Gamma^{-1/2}\left(\widetilde{\mathcal{G}}_h(\vartheta) - z\right)\right\|_2^2,$$

and where $\widetilde{Z}_h$ is the random normalization constant

$$\widetilde{Z}_h = \int_X \exp\left(-\widetilde{\Phi}_h(\vartheta;z)\right)\,\mathrm{d}\mu_0(\vartheta).$$

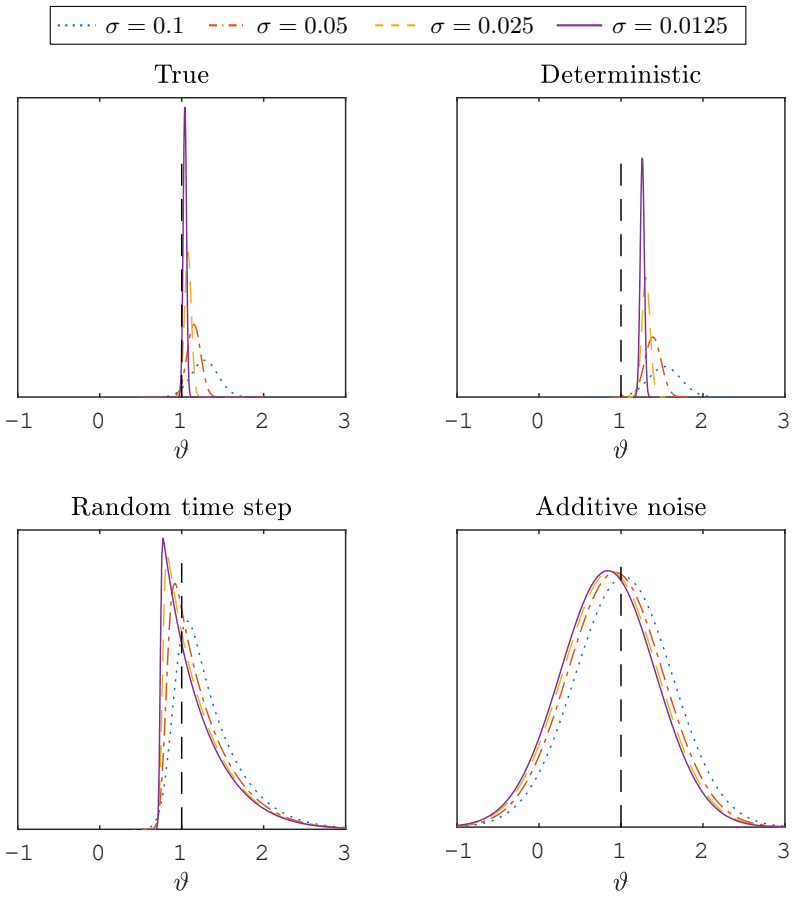


Figure 7.2 – True analytical posterior distributions in the linear case of Section 7.6.1 and its approximations with the deterministic explicit Euler method and with the AN-RK and the RTS-RK both based on the explicit Euler method. In this case, $h = 0.5$ and the variance σ^2 of the observation error is reduced progressively. The true value of the initial condition $\vartheta^* = 1$ is shown with a vertical black dashed line.

Sampling from $\widetilde{\mu}_h$ is performed employing the pseudo-marginal Metropolis–Hastings (PMMH) of [16], as described in Section 6.2. We remark that employing the deterministic approximation $\widetilde{\mu}_{h,\mathrm{mar}}$ and the PMMH instead of the alternative $\widetilde{\mu}_{h,\mathrm{MC}}$ entails a higher computational cost by Remark 6.10. Nevertheless, practical numerical experiments led us to employ the marginal approximation, since in this ODE setting we consider examples where the dimension of the parameter is much smaller than the dimension of the support random variable (i.e., of the sequence of random time steps).

7.6.1 Closed-form Posteriors for a Linear Problem

We consider here a very simple example for which it is possible to compute explicitly the posterior distributions introduced above, and which illustrates the application of probabilistic methods for ODEs to inverse problems. In particular, let us consider the following one dimensional ODE

$$y'(t) = -y(t), \quad y(0) = \vartheta,$$

where we consider the initial condition $\vartheta \in \mathbb{R}$ to be the parameter of interest, and where in this case the right-hand side is known. Given a fixed $T > 0$, we consider the problem of inferring the initial condition ϑ from a single observation $z = \varphi_T(\vartheta^*) + \beta$, where ϑ^* is a true value for the initial condition and where $\beta \sim \mathcal{N}(0, \sigma^2)$. As a prior, we fix $\mu_0 = \mathcal{N}(0, 1)$.

In this simple scenario, the exact forward map admits a closed-form expression. In particular, it holds

$$\mathcal{G}(\vartheta) = \exp(-T)\vartheta.$$

The forward map is therefore linear with respect to ϑ, and simple manipulations with Gaussian densities yield the exact posterior μ given by

$$\mu = \mathcal{N}\left(\frac{z\exp(-T)}{\sigma^2 + \exp(-2T)}, \frac{\sigma^2}{\sigma^2 + \exp(-2T)}\right). \tag{7.36}$$

Consistently, in case the observational noise σ^2 vanishes we have that the observation $z \to \vartheta^* \exp(-T)$, and therefore the posterior μ shrinks to the true value ϑ^*.

Let us now consider the numerical approximation of this simple inverse problem by means of the explicit Euler method. In particular, we assume T being sufficiently small so that we perform only one step of the method, i.e., we fix $h = T$. Hence, the numerical forward map is in this case given by

$$\mathcal{G}_h(\vartheta) = (1 - h)\vartheta.$$

The numerical forward map is also linear with respect to the parameter ϑ, and therefore the posterior μ_h is Gaussian. In particular, simple calculations with Gaussian densities give

$$\mu_h = \mathcal{N}\left(\frac{(1-T)z}{\sigma^2 + (1-T)^2}, \frac{\sigma^2}{\sigma^2 + (1-T)^2}\right), \tag{7.37}$$

where we recall that $T = h$. In the same limit of $\sigma^2 \to 0$ as above, we get in this case that the posterior distribution shrinks to a value ϑ_{lim} which is given by

$$\vartheta_{\text{lim}} = \frac{\exp(-T)}{1-T}\vartheta^*.$$

Let us remark that in the limit for $T \to 0$ (i.e., for $h \to 0$), we have $\vartheta_{\text{lim}} \to \vartheta^*$, but for a fixed positive value $T > 0$ the posterior μ_h presents a bias in the inference of the initial condition.

Let us consider the AN-RK of Section 6.3.1 built on one step of the explicit Euler method, i.e., the random approximation $y(T) \approx Y_1$, where $Y_1 = (1-h)\vartheta + \xi$. We consider Assumption 6.12 to hold with $Q \equiv 1$ and $p = q = 1$, so that $\xi \sim \mathcal{N}(0, h^3)$. The random forward map $\widetilde{\mathcal{G}}_h$ is then defined as

$$\widetilde{\mathcal{G}}_h(\vartheta) = (1-h)\vartheta + \xi.$$

The marginal posterior distribution $\widetilde{\mu}_{h,\text{mar}}$ admits in this simple case a closed-form expression, and manipulations with Gaussian densities yield

$$\widetilde{\mu}_{h,\text{mar}} = \mathcal{N}\left(\frac{(1-T)z}{\widetilde{\sigma}^2 + (1-T)^2}, \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}^2 + (1-T)^2}\right).$$

where $\widetilde{\sigma}^2 = \sigma^2 + T^3$ and where we remind that $T = h$. Therefore, in this case the distribution $\widetilde{\mu}_{h,\text{mar}}$ is not degenerate in the limit for $\sigma^2 \to 0$, but we actually have a limiting distribution $\widetilde{\mu}^{\text{lim}}_{h,\text{mar}}$ given by

$$\widetilde{\mu}^{\text{lim}}_{h,\text{mar}} = \mathcal{N}\left(\frac{(1-T)\exp(-T)\vartheta^*}{T^3 + (1-T)^2}, \frac{T^3}{T^3 + (1-T)^2}\right). \tag{7.38}$$

Let us remark that consistently the posterior $\widetilde{\mu}^{\lim}_{h,\mathrm{mar}}$ shrinks to the true value ϑ^* in case T (i.e., h). Nevertheless, the relevant feature of $\widetilde{\mu}^{\lim}_{h,\mathrm{mar}}$ is that while the mean of the limiting distribution is still biased from ϑ^*, a positive variance accounts for the uncertainty in the solution of the inverse problem.

Let us now consider the RTS-RK based on one step of the explicit Euler with step size distribution $H \sim \mathcal{U}(T - T^{p+1/2}, T + T^{p+1/2})$, where we recall again that $T = h$ and where we fix $p = q = 1$ following Remark 7.19. In this case, the random forward model $\widetilde{\mathcal{G}}_h$ is given by

$$\widetilde{\mathcal{G}}_h(\vartheta) = (1 - H)\vartheta.$$

Let us remark that the random perturbation introduced by the probabilistic method does not follow a Gaussian distribution, and hence the marginal posterior $\widetilde{\mu}_{h,\mathrm{mar}}$ is not Gaussian even though the forward map is linear with respect to ϑ. Nevertheless, it is possible to compute a closed-form expression for the density of $\widetilde{\mu}_{h,\mathrm{mar}}$ with respect to the Lebesgue measure, i.e., the function $\widetilde{\pi}_{h,\mathrm{mar}} \colon \mathbb{R} \to \mathbb{R}$ such that $\widetilde{\mu}_{h,\mathrm{mar}}(\mathrm{d}\vartheta) = \widetilde{\pi}_{h,\mathrm{mar}}(\vartheta)\,\mathrm{d}\vartheta$. In particular, it holds

$$\begin{aligned}\widetilde{\pi}_{h,\mathrm{mar}}(\vartheta) \propto \exp\left(-\frac{\vartheta^2}{2}\right)\frac{1}{\vartheta}\left(F_{\mathcal{N}}\left(\frac{((1-T)+T^{3/2})\vartheta - z}{\sigma}\right)\right. \\ \left. -F_{\mathcal{N}}\left(\frac{((1-T)-T^{3/2})\vartheta - z}{\sigma}\right)\right),\end{aligned} \tag{7.39}$$

where the symbol $\propto$ denotes equality up to multiplicative constants independent of ϑ and such that $\widetilde{\pi}_{h,\mathrm{mar}}$ integrates to one over $\mathbb{R}$, and where $F_{\mathcal{N}}$ is the distribution function of a $\mathcal{N}(0,1)$ random variable. We compute also in this case the limit for $\sigma^2 \to 0$, which yields the limiting distribution $\widetilde{\mu}^{\lim}_{h,\mathrm{mar}}(\mathrm{d}\vartheta) = \widetilde{\pi}^{\lim}_{h,\mathrm{mar}}(\vartheta)\,\mathrm{d}\vartheta$, where the limiting density satisfies

$$\widetilde{\pi}^{\lim}_{h,\mathrm{mar}}(\vartheta) \propto \exp\left(-\frac{\vartheta^2}{2}\right)\frac{1}{\vartheta}\chi_{\{\vartheta_{\min} \le \vartheta \le \vartheta_{\max}\}},$$

with χ being the indicator function and where $\vartheta_{\min}$ and $\vartheta_{\max}$ are given by

$$\vartheta_{\min} = \frac{\exp(-T)\vartheta^*}{((1-T)+T^{3/2})}, \quad \vartheta_{\max} = \frac{\exp(-T)\vartheta^*}{((1-T)-T^{3/2})}.$$

We remark that in this case, too, the limit for vanishing observational noise yields a posterior with a positive variance. Moreover, in the limit for $T \to 0$ (i.e., $h \to 0$), we have that $\vartheta_{\min}$ and $\vartheta_{\max}$ both tend to ϑ^*, and therefore the posterior $\widetilde{\mu}^{\lim}_{h,\mathrm{mar}}$ consistently collapses to the true value ϑ^*.

In order to represent graphically the findings above, we fix $T = h = 0.5$ and consider $\sigma = \{0.1, 0.05, 0.025, 0.0125\}$, thus generating four observational noises η_i as $\eta_i = \sigma_i Z$ for a random variable $Z \sim \mathcal{N}(0,1)$. In Fig. 7.2 we show the posteriors (7.36), (7.37), (7.38) and (7.39), which confirm our claim, i.e., that probabilistic methods take into account the uncertainty in the forward model due to numerical approximation and transfer it to the posterior belief.

7.7 Numerical Experiments

In this section, we present a series of numerical experiments which validate our analysis, and which illustrate the potential of the RTS-RK method on equations with geometric properties and Bayesian inference problems.

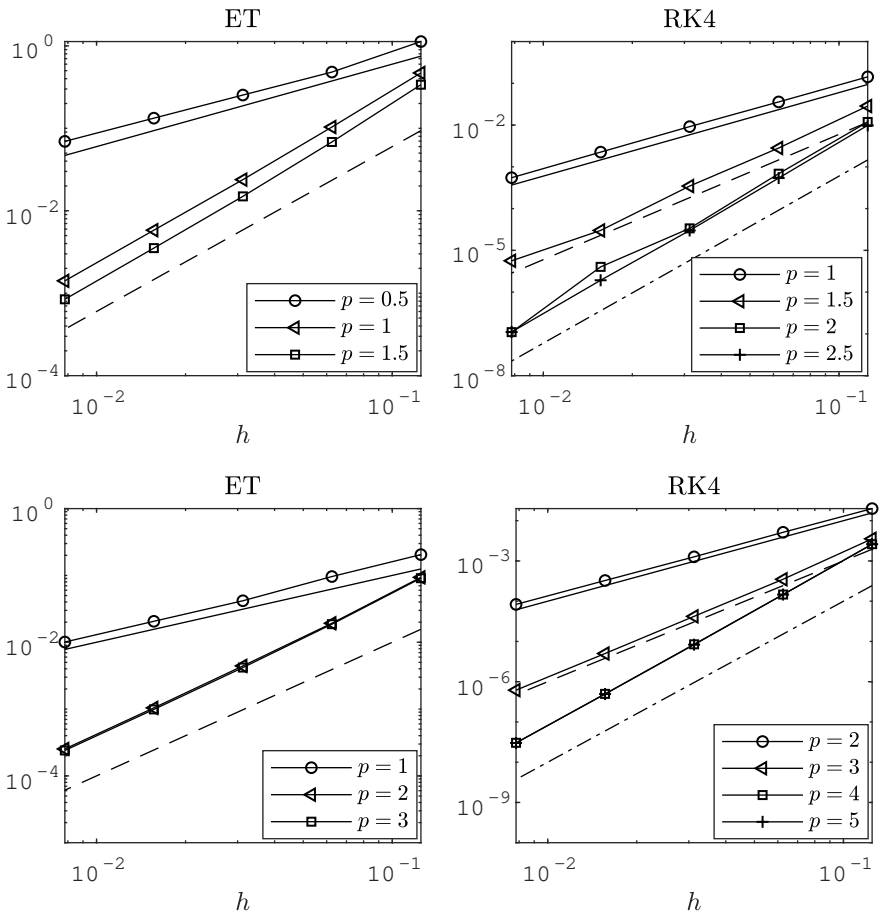


Figure 7.3 – Convergence for the RTS-RK built on the explicit trapezoidal (ET) and fourth-order Runge–Kutta (RK4) as a function of the value of p of Assumption 7.2. First row: weak order of convergence (Theorem 7.12). Second row: mean-square order of convergence (Theorem 7.16). In the left column, the reference slopes 1 and 2 are displayed (solid and dashed lines), while in the right column reference slopes 2, 3 and 4 are displayed (solid, dashed and dash-dotted lines).

7.7.1 Convergence

In order to verify the result predicted by Theorems 7.12 and 7.16, we consider the FitzHugh–Nagumo equation, which is defined as

$$\begin{aligned} y_1' &= c\left(y_1 - \frac{y_1^3}{3} + y_2\right), \quad y_1(0) = -1, \\ y_2' &= -\frac{1}{c}(y_1 - a + by_2), \quad y_2(0) = 1, \end{aligned} \tag{7.40}$$

where a, b, c are real parameters with values $a = 0.2$, $b = 0.2$, $c = 3$. We integrate the equation from time $t_0 = 0$ to final time $T = 1$. The reference solution is generated with a high-order method on a fine time scale. The deterministic integrators we choose in this experiment are the explicit trapezoidal rule (ET) and the classic fourth-order Runge–Kutta method (RK4), which verify Assumption 7.4 with $q = 2$ and $q = 4$, respectively. The random steps are uniform as in Example 7.3. We vary their mean in the range $h_i = 0.125 \cdot 2^{-i}$ with $i = 0, 1, \ldots, 4$, and we vary the value of p in Assumption 7.2 in order to verify the theoretical convergence results.

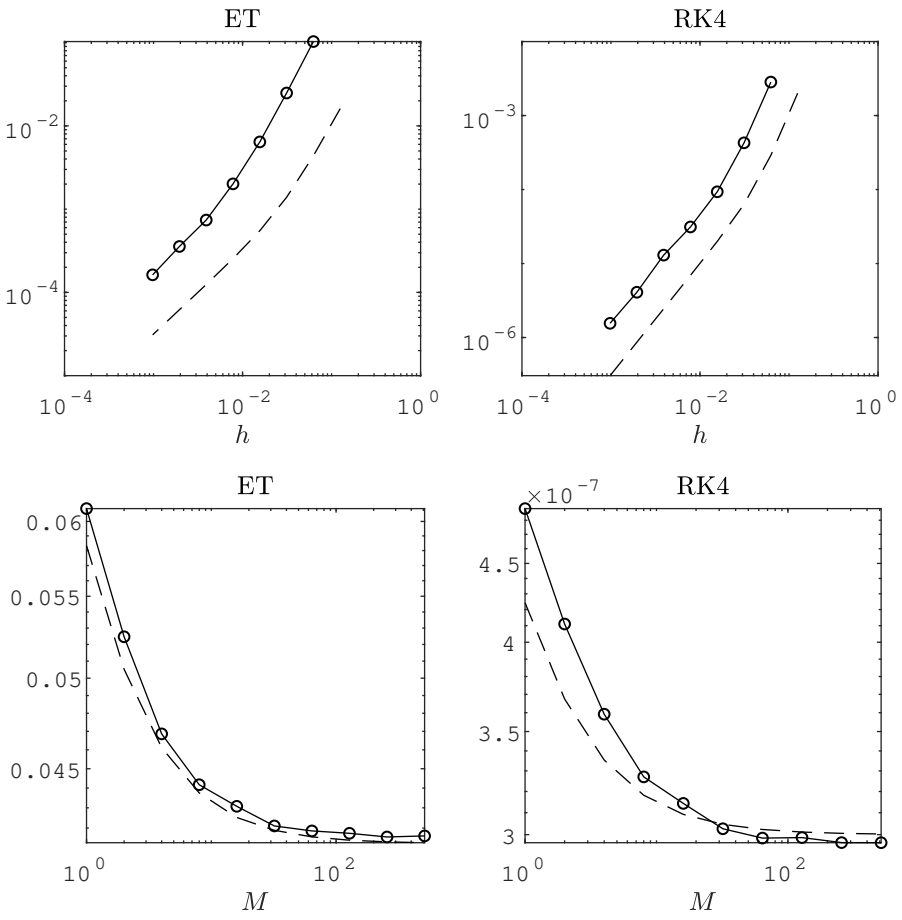


Figure 7.4 – Convergence of the square root of the MSE of the Monte Carlo estimator for the random time-stepping explicit trapezoidal (ET) (left column) and fourth-order Runge–Kutta (RK4) (right column) with respect to the time step h (first row) and the number of trajectories M (second row). The dashed line corresponds to the prediction of Theorem 6.7.

For weak convergence, we consider $p \in \{0.5, 1, 1.5\}$ for the ET and $p \in \{1, 1.5, 2, 2.5\}$ for RK4. Theorem 7.12 then predicts weak order $p_w = \{1, 2, 2\}$ for the ET and $p_w = \{2, 3, 4, 4\}$ for the RK4. The functional $\Psi\colon \mathbb{R}^d \to \mathbb{R}$ of the solution we consider is defined as $\Psi(x) \coloneqq x^\top x$. Finally, we consider 10^6 trajectories of the numerical solution in order to approximate the expectation with a Monte Carlo sum. Results (Fig. 7.3) show that the order of convergence predicted by Theorem 7.12 is confirmed by numerical experiments.

For mean-square convergence, we consider $p \in \{1, 2, 3\}$ for the ET and $p \in \{2, 3, 4, 5\}$ for the RK4. Theorem 7.16 then predicts strong order $p_s = \{1, 2, 2\}$ for the ET and $p_s = \{2, 3, 4, 4\}$ for the RK4. We simulate 10^3 realizations of the numerical solution and compute the approximate mean-square order of convergence for each value of h with a Monte Carlo mean. Results (Fig. 7.3) show that the orders predicted by Theorem 7.16 are confirmed numerically.

7.7.2 Mean-square Convergence of Monte Carlo Estimators

We shall now verify numerically the validity of Theorem 6.7 for the RTS-RK method. We consider the ODE (7.40), with final time $T = 1$ and the same parameters as above. In this case as well, we consider the ET rule and the RK4 with uniform random time steps having mean $h_i = 0.125 \cdot 2^{-i}$ with $i = 0, 1, \dots, 7$. For the explicit trapezoidal rule, we fix $M = 10^3$ and $p = 1$, so that for bigger values of h the first term in the bound (7.21) dominates, while in the regime of small h, the higher order of the first term makes the second term larger in magnitude. This behavior results in the change of slope in the convergence plot which can be observed in Fig. 7.4, both in the theoretical estimate and in the numerical results. We perform the same experiment using the RK4, fixing $M = 10^4$ and $p = 1.5$, thus obtaining a numerical confirmation of the theoretical result.

As a second experiment, we consider the same setup as above but wish to verify the dependence of the MSE on the number of samples M, which we vary as $M = 2^i$, with $i = 0, 1, \dots, 9$. For the explicit trapezoidal rule, we consider $p = q = 2$, which is the optimal choice for the intrinsic variability of the RTS-RK method. Moreover, we fix $h = 0.05$. In this case, the bound (7.21) reduces to

$$\mathrm{MSE}(\widehat{\Psi}_{N,M}) \leq Ch^{2q}\left(1 + \frac{1}{M}\right).$$

In Fig. 7.4 we show that the convergence of the MSE depends on M as predicted by the theoretical bound. We repeat the same experiment using the fourth order explicit Runge–Kutta method, for which we take $h = 0.01$ and $p = q = 4$, thus confirming numerically our theoretical result.

7.7.3 Robustness

In this numerical experiment we verify the robustness of RTS-RK when applied to chemical reactions. Let us consider the Peroxide-Oxide chemical reaction, which is macroscopically defined by the following balance equation

$$\mathrm{O_2} + 2\mathrm{NADH} + 2\mathrm{H^+} \rightarrow 2\mathrm{H_2O} + 2\mathrm{NAD^+},$$

where NADH and $\mathrm{NAD^+}$ are the oxidized and reduced form of the nicotinamide adenine dinucleotide (NAD) respectively. This reaction has to be catalyzed by an enzyme to take place, which reacts with the reagents to create intermediate products of the reaction. A successful model [105] to describe the time-evolution of the chemical system is the following

$$\begin{aligned} &\mathrm{B} + \mathrm{X} \xrightarrow{k_1} 2\mathrm{X}, \quad 2\mathrm{X} \xrightarrow{k_2} 2\mathrm{Y}, \quad \mathrm{A} + \mathrm{B} + \mathrm{Y} \xrightarrow{k_3} 3\mathrm{X}, \\ &\mathrm{X} \xrightarrow{k_4} \mathrm{P}, \quad \mathrm{Y} \xrightarrow{k_5} \mathrm{Q}, \quad \mathrm{X_0} \xrightarrow{k_6} \mathrm{X}, \\ &\mathrm{A_0} \xleftrightarrow{k_7} \mathrm{A}, \quad \mathrm{B_0} \xrightarrow{k_8} \mathrm{B}. \end{aligned}$$

Here, A and B are respectively $[\mathrm{O_2}]$ and [NADH], P, Q are the products and X, Y are intermediate results of the reaction process. It is therefore possible to model the time evolution of the reaction with the following system of nonlinear ODEs

$$\begin{aligned} \mathrm{A}' &= k_7(\mathrm{A_0} - \mathrm{A}) - k_3\mathrm{ABY}, & \mathrm{A}(0) &= 6, \\ \mathrm{B}' &= k_8\mathrm{B_0} - k_1\mathrm{BX} - k_3\mathrm{ABY}, & \mathrm{B}(0) &= 58, \\ \mathrm{X}' &= k_1\mathrm{BX} - 2k_2\mathrm{X}^2 + 3k_3\mathrm{ABY} - k_4\mathrm{X} + k_6\mathrm{X_0}, & \mathrm{X}(0) &= 0, \\ \mathrm{Y}' &= 2k_2\mathrm{X}^2 - k_5\mathrm{Y} - k_3\mathrm{ABY}, & Y(0) &= 0, \end{aligned} \tag{7.41}$$

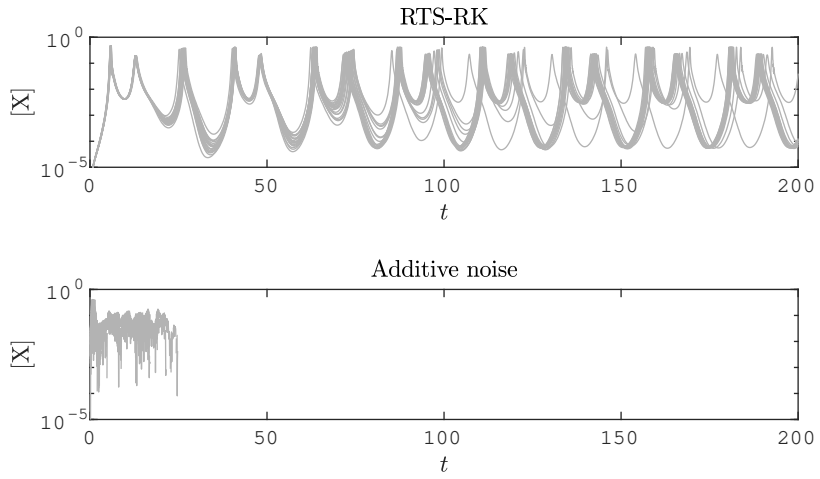


Figure 7.5 – Fifty trajectories of the numerical value of the concentration of the X species for the random time-stepping and additive noise methods (above and below respectively).

where $A_0 = 8$, $B_0 = 1$, $X_0 = 1$ and the real parameters k_i, $i = 1, \dots, 8$ representing the reaction rates take values

$$k_1 = 0.35, \quad k_2 = 250, \quad k_3 = 0.035, \quad k_4 = 20,$$
$$k_5 = 5.35, \quad k_6 = 10^{-5}, \quad k_7 = 0.1, \quad k_8 = 0.825.$$

It has been shown [105] that for these values of the parameters the system exhibits a chaotic behavior. In particular, at long time the trajectories lie in a strange attractor, and the system shows a strong sensitivity to perturbations on the initial condition.

Since the components of the solution represent the concentration of chemicals, we require the numerical solution to be positive. Apart from physical considerations, we observe numerically that if one of the components takes negative values, the solution shows strong instabilities. For the RTS-RK method, the distribution of the random time steps can be selected so that the probability of obtaining a negative solution is zero, see e.g. Example 7.3. In contrast, for the additive noise method we can have disruptive effects even for h small if the solution has a small magnitude, as the probability for negative populations will never be zero. Hence, in this case employing the additive noise method likely produces instabilities regardless of the chosen time step.

Let us apply the AN-RK and the RTS-RK methods to (7.41). We choose $h = 0.05$ as the mean of uniformly distributed time steps for the RTS-RK and as the time step for the AN-RK, while we employ the Runge–Kutta–Chebyshev method (RKC) (see [138]) as deterministic integrator. Since the RKC has order 1, we fix $p = q = 1$. As the problem is stiff, stabilized methods prevent a step size restriction while remaining explicit. We note that the RKC method is a stabilized numerical integrator of first order and that higher order explicit stabilized methods such as ROCK2 or ROCK4 [1, 10] could also be used as deterministic solvers for the RTS-RK method. It can be seen in Fig. 7.5 that the RTS-RK method maintains the numerical solution positive and captures the chaotic nature of the chemical reaction. In contrast, the additive noise scheme produces negative values, thus showing strong instabilities in the long-time behavior. In particular, all the numerical trajectories turn negative or diverge before approximately $t = 25$, which is the reason why after this time they are not displayed in Fig. 7.5.

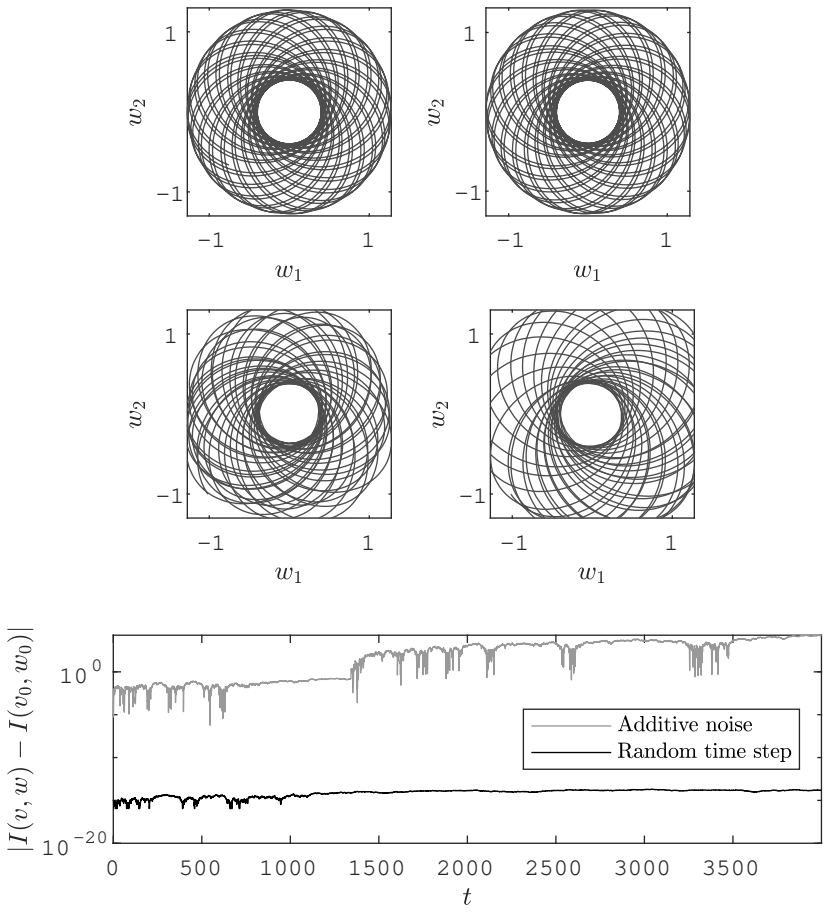


Figure 7.6 – Trajectories of (7.42) given by the RTS-RK method (7.2) for $0 \leq t \leq 200$ and $3800 \leq t \leq 4000$ (first row), and by AN-RK method (see Section 6.3.1) for $0 \leq t \leq 200$ and $200 \leq t \leq 400$ (second row). Error on the angular momentum I defined in (7.43) for $0 \leq t \leq 4000$ given by the two methods.

7.7.4 Conservation of Quadratic First Integrals

A simple model for the two-body problem in celestial mechanics is the Kepler system with a perturbation, which reads

$$\begin{aligned} w_1' &= v_1, \quad v_1' = -\frac{w_1}{\|q\|^3} - \frac{\delta w_1}{\|q\|^5}, \\ w_2' &= v_2, \quad v_2' = -\frac{w_2}{\|q\|^3} - \frac{\delta w_2}{\|q\|^5}, \end{aligned} \tag{7.42}$$

where v_1, v_2 are the two components of the velocity and w_1, w_2 are the two components of the position. We set the perturbation parameter δ to be equal to 0.015 and the initial condition to be

$$w_1(0) = 1 - e, \quad w_2(0) = 0, \quad v_1(0) = 0, \quad v_2(0) = \sqrt{(1+e)/(1-e)},$$

where $e = 0.6$ is the eccentricity. It is well-known that this equation has the Hamiltonian and the angular momentum as quadratic first integrals. In particular, we focus here on the angular momentum, which reads

$$I(v, w) = w_1 v_2 - w_2 v_1. \tag{7.43}$$

We consider the simplest Gauss collocation method, namely the implicit midpoint rule, as the deterministic Runge–Kutta method. It is known that Gauss collocation methods conserve quadratic first integrals. According to Theorem 7.23, we expect therefore that the RTS-RK implemented on the implicit midpoint rule also conserves quadratic first integrals. We integrate (7.42) with uniformly distributed random time steps with mean $h = 0.01$ from time $t = 0$ to time $t = 4000$ which corresponds to approximately 636 revolutions of the system (long-time behavior). Since the implicit midpoint rule is of order $q = 2$, we choose $p = 2$ for the RTS-RK method. Moreover, we consider the AN-RK method with $h = 0.01$, expecting that the first integral will not be conserved. We observe in Fig. 7.6 that the method (7.2) conserves the angular momentum, while for AN-RK method the approximate conservation of the quadratic first integral shown in (7.22) is lost when integrating (7.42) over long time.

7.7.5 Conservation of Hamiltonians

Let us consider the pendulum problem, which is given by the Hamiltonian $Q\colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$Q(v, w) = \frac{v^2}{2} - \cos w,$$

where $y = (v, w)^\top \in \mathbb{R}^2$. We wish to study the validity of Theorem 7.33, i.e., show that the mean error on the Hamiltonian is of order $\mathcal{O}(h^q)$ for time spans of polynomial length and then it grows proportionally to the square root of time. We consider the initial condition $(v_0, w_0) = (1.5, -\pi)$ and integrate the equation employing RTS-RK based on the implicit midpoint method ($q = 2$) choosing $p = q$, which is the optimal scaling of the noise. We choose uniform time steps, vary their mean $h \in \{0.2, 0.1, 0.05, 0.025\}$, integrate the dynamical system up to the final time $T = 10^6$ and study the time evolution of the mean numerical error on the Hamiltonian Q. Results are shown in Fig. 7.7, where it is possible to notice that the error is bounded by $\mathcal{O}(h^q)$ (horizontal black lines) for long time spans. After this stationary phase, the error on the Hamiltonian appears to grow as the square root of time. The oscillations of the error which are shown in Fig. 7.7 are present even when integrating the pendulum system with a deterministic symplectic scheme. Moreover, considering $T = 10^3$, the time step $h \in \{0.2, 0.1\}$ and keeping all other parameters as above, we compute the mean Hamiltonian and represent it in Fig. 7.7 together with an approximate confidence interval. We arbitrarily fix a confidence interval at two standard deviations from the mean, and we employ it to show the path-wise variability of the value of the Hamiltonian. As expected, the variability decreases dramatically with respect to the time step h.

7.7.6 Bayesian inference

For the last numerical experiment we consider the Hénon–Heiles equation, a Hamiltonian system with energy $Q\colon \mathbb{R}^4 \to \mathbb{R}$ defined by

$$Q(v, w) = \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2 + w_1^2 w_2 - \frac{1}{3}w_2^3, \tag{7.44}$$

where $v, w \in \mathbb{R}^2$ are the velocity and position respectively and where we denote by $y = (v, w)^\top \in \mathbb{R}^4$ the solution. We consider an initial condition such that $Q(y_0) = 0.13$, for which the system exhibits a chaotic behaviour [66]. In the spirit of Section 7.6, we are interested in recovering the true value of the initial condition y_0 through a single observation y_{obs} of the solution (v, w) at a fixed time $t_{\text{obs}} = 10$. The exact forward operator $\mathcal{G}$ is therefore defined as $\mathcal{G}(y_0) = \varphi_{t_{\text{obs}}}(y_0)$. Noise is then set to be a Gaussian random variable $\beta \sim \mathcal{N}(0, \sigma^2 I)$, where $\sigma = 5 \cdot 10^{-4}$, and we fix a standard Gaussian prior on the initial condition, i.e., $\mu_0 = \mathcal{N}(0, I)$. We choose the observational

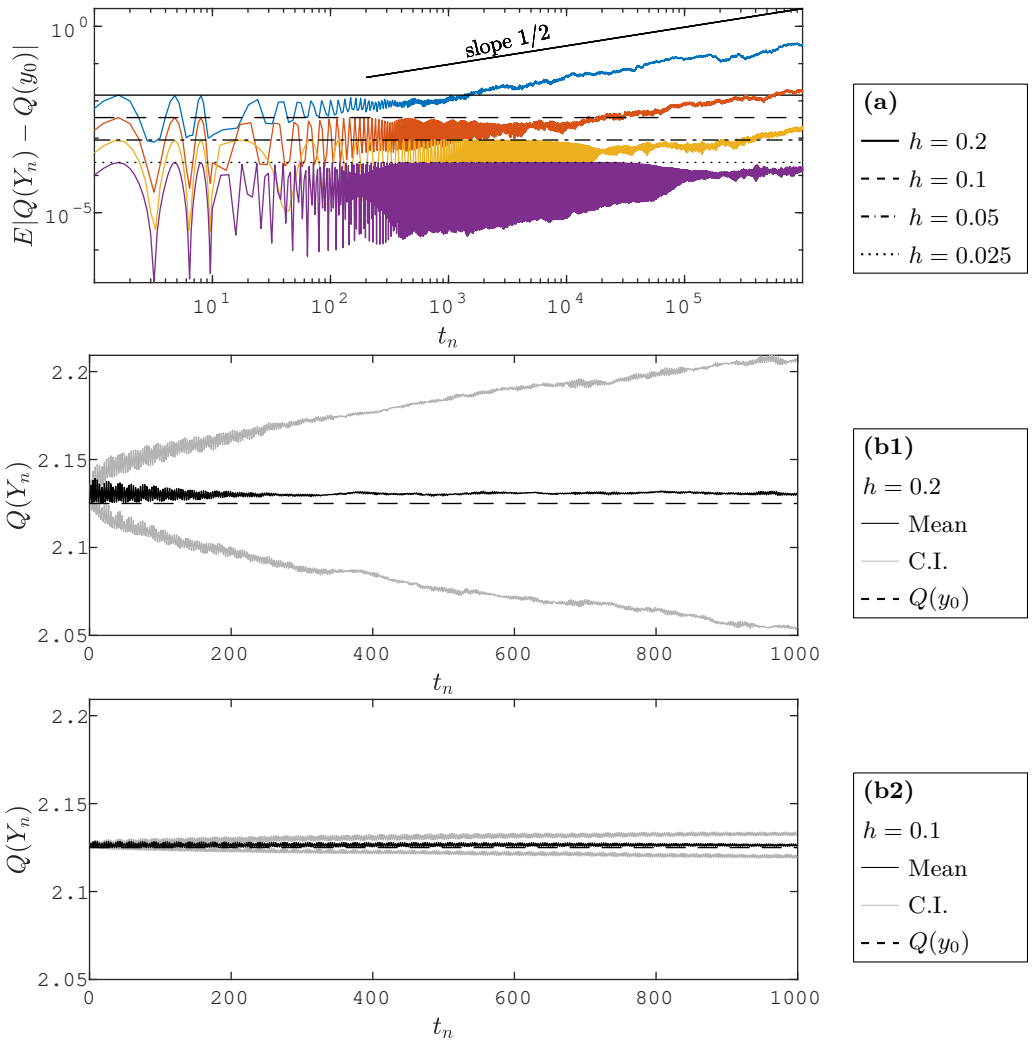


Figure 7.7 – **(a)**: Time evolution of the mean error for the pendulum problem and different values of the time step h. The black lines represent the theoretical estimate given by Theorem 7.33, while the colored lines represent the experimental results. The mean was computed by averaging 20 realisations of the numerical solution. **(b1)** and **(b2)**: Time evolution of the mean Hamiltonian for two different values of the time step. The mean Hamiltonian is depicted together with an approximate confidence interval, whose width is proportional to the standard deviation of the Hamiltonian over 200 trajectories.

noise to have a small variance (i.e., of order $\mathcal{O}(10^{-8})$) as in this case classical solvers present the misleading overconfident behaviour explained in Section 6.2 and Section 7.6.

Since the equation is Hamiltonian, we choose to employ a classical second-order ($q = 2$) symplectic method, the Störmer–Verlet scheme [60, 130, 142], for which one step is defined in the general

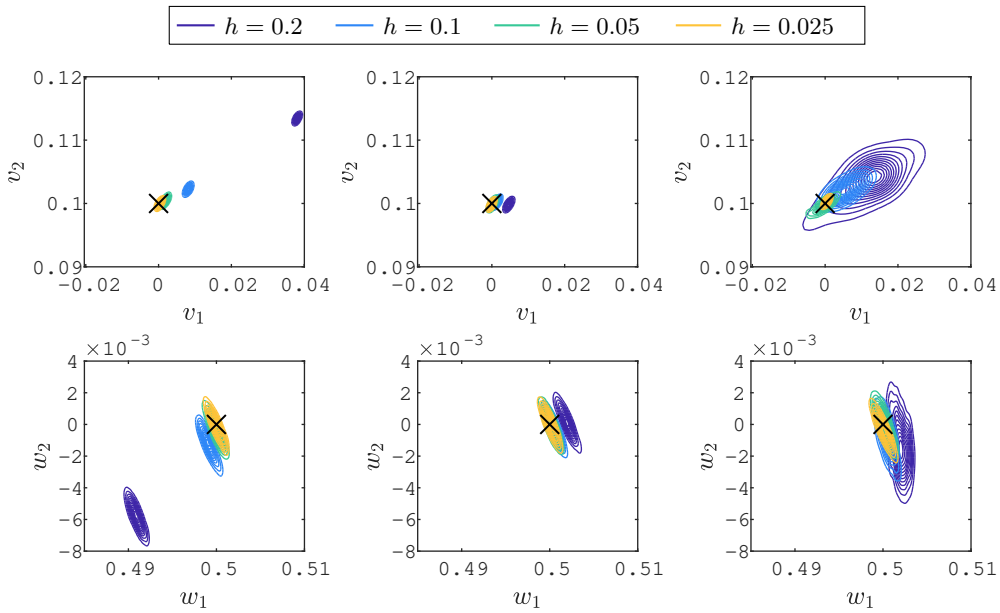


Figure 7.8 – Posterior distributions for the initial position and velocity of the Hénon-Heiles system with different values of $h = \{0.2, 0.1, 0.05, 0.025\}$. First row: initial velocity v_0. Second row: initial position w_0. First column: deterministic Heun's method. Second column: deterministic Störmer–Verlet scheme. Third column: RTS-RK Störmer–Verlet ($p = 2$).

case as

$$
\begin{aligned}
v_{n+1/2} &= v_n - \frac{h}{2}\nabla_w Q(v_n, w_n),\\
w_{n+1} &= w_n + \frac{h}{2}\left(\nabla_v Q(v_{n+1/2}, w_n) + \nabla_v Q(v_{n+1/2}, w_{n+1})\right),\\
v_{n+1} &= v_{n+1/2} - \frac{h}{2}\nabla_w Q(v_{n+1/2}, w_{n+1}).
\end{aligned}
$$

As the Hamiltonian Q given by (7.44) is separable, i.e., $Q(v, w) = Q_1(v) + Q_2(w)$, where $Q_1, Q_2 \colon \mathbb{R}^2 \to \mathbb{R}$, the Störmer–Verlet scheme simplifies to

$$
\begin{aligned}
v_{n+1/2} &= v_n - \frac{h}{2}\nabla_w Q_2(w_n),\\
w_{n+1} &= w_n + h\nabla_v Q_1(v_{n+1/2}),\\
v_{n+1} &= v_{n+1/2} - \frac{h}{2}\nabla_w Q_2(w_{n+1}).
\end{aligned}
$$

Hence, in the separable case the Störmer–Verlet scheme is explicit and the evaluation of the flow consists only of three evaluations of the derivatives of Q. We then employ this method both with a fixed time step h and as a basic integrator for the RTS-RK method (with uniformly distributed time steps and $p = 2$), thus computing the posterior distributions μ_h and $\widetilde{\mu}_{h,\mathrm{mar}}$ defined in Section 7.6, respectively. Moreover, we compute the posterior distribution given by a non-symplectic method, the Heun's scheme, which is a classical second order method. For the deterministic integrator, we generate samples from the distributions with the MH algorithm, and for the measure $\widetilde{\mu}_{h,\mathrm{mar}}$ we employ the PMMH. Finally, we vary the time step h for the three methods above in order to study whether the approximate posterior concentrates towards the true value of the initial condition.

We can observe in Fig. 7.8 that the posterior distributions given by Heun's method are concentrated away from the true value of the initial condition for the larger values of the time step. In fact, Heun's method is not symplectic, and a deviation on the energy Q is produced when integrating the dynamical system forward in time. Hence, initial conditions with a different energy level with respect to the observation are mapped by the approximate forward model to points which are close to the observations, and as a result the posterior distribution is concentrated far from the true value. This behaviour is corrected using the Störmer–Verlet method due to its symplecticity. However, we remark that the posterior distribution for $h = 0.2$ is still concentrated on a biased value of the initial condition, without any indication of this bias given by the posterior's variance. Applying the RTS-RK method together with PMMH instead gives nested posterior distributions whose variance quantifies the uncertainty of the numerical solver. This favourable behaviour is possible due to the numerical error quantification of probabilistic methods, which has been already shown in [37, 39], together with the good energy conservation properties of the RTS-RK method when a symplectic integrator is used as its deterministic component as proved in Theorem 7.33.

7.8 Proof of Technical Results

In this section, we prove technical result which were left unproved in the text in order to enhance readability.

7.8.1 A Modified Stochastic Differential Equation

In Remark 7.13, we claim the existence of a modified stochastic differential equation (SDE) whose solution is well approximated by the RTS-RK method. Let us denote by $\widetilde{f}$ the function defining the modified equation corresponding to the numerical flow ψ_h truncated after l terms, i.e.,

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \ldots + h^l f_{l+1}(y).$$

Details about the construction of such a function can be found in Section 7.5.2. In particular, analyticity of the function f is needed for a rigorous backward error analysis to hold. Therefore, we will refer in this section to Assumption 7.27 (see Section 7.5.2). For the additive noise method presented in [39], the authors consider the SDE

$$\mathrm{d}Y = \widetilde{f}(Y)\,\mathrm{d}t + \sqrt{Qh^{2p}}\,\mathrm{d}W, \tag{7.45}$$

where W is a d-dimensional standard Brownian motion. It is possible to show [39, Theorem 2.4] that the solution of (7.45) satisfies

$$|\mathbb{E}\left[\Psi(Y_N) - \Psi(Y(T) \mid Y_0 = y\right]| \leq Ch^{2p},$$

where $T = Nh$ and Y_N is the numerical solution given by the additive noise method after N steps. Here, we present a similar construction for the RTS-RK method. In particular, let us consider the modified SDE

$$\mathrm{d}\widetilde{Y} = \left(\widetilde{f}(\widetilde{Y}) + \frac{1}{2}Ch^{2p}\partial_{tt}\psi_h(\widetilde{Y})\right)\mathrm{d}t + \sqrt{Ch^{2p}\partial_t\psi_h(\widetilde{Y})\partial_t\psi_h(\widetilde{Y})^\top}\,\mathrm{d}W, \tag{7.46}$$

where C is given in Assumption 7.2.(iii). Let us denote by $\widetilde{\mathcal{L}}$ the generator of (7.46), which can be written explicitly as

$$\widetilde{\mathcal{L}} = \left(\widetilde{f} + \frac{1}{2}Ch^{2p}\partial_{tt}\psi_h\right)\cdot\nabla + \frac{1}{2}Ch^{2p}\partial_t\psi_h\partial_t\psi_h^\top : \nabla^2,$$

and, adopting the semi-group notation, it satisfies

$$\mathbb{E}\left[\Psi(\widetilde{Y}(h)) \mid \widetilde{Y}(0) = y\right] = e^{h\widetilde{\mathcal{L}}}\Psi(y).$$

In the following lemma, we consider the error over one step between the numerical solution given by the RTS-RK method and the solution of (7.46) in the weak sense. The proof is inspired by the calculations presented in [39, Section 2.4].

Lemma 7.38. *Under the assumptions of Lemma 7.7 and if Assumption 7.27 holds, then*

$$\left|\mathbb{E}\left[\Psi(Y_1) - \Psi(\widetilde{Y}(h)) \mid Y_0 = y\right]\right| \leq Ch^{2p+1},$$

where C is a positive constant independent of h and of y, $\widetilde{Y}$ is the solution of (7.46) and Y_1 is the numerical solution given by the RTS-RK method after one step.

Proof. Let us consider the modified ODE

$$\widehat{y}'(t) = \widetilde{f}(\widehat{y}), \tag{7.47}$$

and denote its flow as $\widehat{\varphi}_t$. The generator $\widehat{\mathcal{L}} = \widetilde{f} \cdot \nabla$ satisfies, adopting the semi-group notation,

$$\Psi(\widehat{\varphi}_h(y)) = e^{h\widehat{\mathcal{L}}}\Psi(y).$$

We can now compute the distance between the solution to (7.46) and (7.47) as

$$\begin{aligned}
e^{h\widetilde{\mathcal{L}}}\Psi(y) - e^{h\widehat{\mathcal{L}}}\Psi(y) &= e^{h\widetilde{f}\cdot\nabla}\left(e^{\frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h\cdot\nabla + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h\partial_t\psi_h^\top:\nabla^2} - I\right)\Psi(y) \\
&= (1+\mathcal{O}(h))\left(\frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h\cdot\nabla \right. \\
&\qquad \left. + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h\partial_t\psi_h^\top : \nabla^2 + \mathcal{O}\left(h^{4p+1}\right)\right)\Psi(y) \\
&= \frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h\cdot\nabla\Psi(y) + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h\partial_t\psi_h^\top : \nabla^2\Psi(y) + \mathcal{O}\left(h^{4p+1}\right).
\end{aligned}$$

Let us recall that equation (7.8) gives

$$\begin{aligned}
e^{h\mathcal{L}_h}\Psi(y) - \Psi(\psi_h(y)) &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\psi_h(y)\cdot\nabla\Psi(y) \\
&\quad + \frac{1}{2}Ch^{2p+1}\partial_t\psi_h(y)\partial_t\psi_h(y)^\top : \nabla^2\Psi(y) + \mathcal{O}(h^{2p+1}),
\end{aligned}$$

which implies that

$$e^{h\widetilde{\mathcal{L}}}\Psi(y) - e^{h\mathcal{L}_h}\Psi(y) = e^{h\widehat{\mathcal{L}}}\Psi(y) - \Psi(\psi_h(y)) + \mathcal{O}(h^{2p+1}).$$

Now, the theory of backward error analysis (see Section 7.5.2 or e.g. [60, Chapter IX]) guarantees that

$$e^{h\widehat{\mathcal{L}}}\Psi(y) - \Psi(\psi_h(y)) = \mathcal{O}(h^{q+l+2}).$$

Choosing $l = 2p - q - 1$, we have therefore

$$e^{h\widetilde{\mathcal{L}}}\Psi(y) - e^{h\mathcal{L}_h}\Psi(y) = \mathcal{O}(h^{2p+1}),$$

which is the desired result. □

The error can be then propagated to final time as in Theorem 7.12, as presented in the following theorem.

Theorem 7.39. *Under the assumptions of Lemma 7.38 and Theorem 7.12, and if there exists a constant $L > 0$ independent of h such that for all $\Psi \in \mathcal{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$*

$$\sup_{u\in\mathbb{R}^d} \left| e^{h\widetilde{\mathcal{L}}}\Psi(u) \right| \leq (1 + Lh) \sup_{u\in\mathbb{R}^d} |\Psi(u)| ,$$

then it holds

$$\left| \mathbb{E}\left[\Psi(Y_N) - \Psi(\widetilde{Y}(T)) \mid Y_0 = y \right] \right| \leq Ch^{2p},$$

where $T = Nh$ and C is a positive constant independent of h and of y, $\widetilde{Y}$ is the solution of (7.46) and Y_N is the numerical solution given by the RTS-RK method after N steps.

Proof. The proof follows by replacing $\mathcal{L}$ with $\widetilde{\mathcal{L}}$ and Lemma 7.7 with Lemma 7.38 in the proof of Theorem 7.12. □

7.8.2 Proof of Lemma 7.30

In the following, we denote by $[\![a, b]\!]$ the interval $[\![a, b]\!] = [a, b]$ if $a < b$ and $[\![a, b]\!] = [b, a]$ if $a \geq b$. Let us first consider $r \geq 2$ and the function $\gamma_r(x) = x^r e^{-r\kappa/x}$, whose first derivative is given by

$$\gamma_r'(x) = rx^{r-2}(x + \kappa)e^{-r\kappa/x}.$$

Under Assumption 7.29 we have that $H_j \leq Mh$ almost surely, and hence for any $t \in [\![h, H_j]\!]$

$$|\gamma_r'(t)| \leq r(Mh)^{r-2}(Mh + \kappa)e^{-r\kappa/(Mh)},$$

where we exploited that $e^{-r\kappa/x}$ is a growing function of x. The fundamental theorem of calculus gives

$$\begin{aligned} |\gamma_r(H_j)| &= \left| \gamma_r(h) + \int_h^{H_j} \gamma_r'(t)\,\mathrm{d}t \right| \\ &\leq \gamma_r(h) + r(Mh)^{r-2}(Mh + \kappa)e^{-r\kappa/(Mh)} \left| H_j - h \right| , \quad \text{almost surely.} \end{aligned}$$

Taking expectation on both sides and since by (7.30) it holds $|\eta_j|^r \leq C\gamma_r(H_j)$ we obtain

$$\mathbb{E}\left[|\eta_j|^r \right] \leq C \left(\gamma_r(h) + rM^{r-2}h^{p+r-3/2}(Mh + \kappa)e^{-r\kappa/(Mh)} \right),$$

which proves the desired inequality. This is because Assumption 7.29 and Assumption 7.2.(ii) imply that $M \geq 1$, and because Mh can be bounded by M. Let us now consider $r = 1$. In this case we have for $t \in [\![h, H_j]\!]$

$$|\gamma_1'(t)| \leq (mh)^{-1}(Mh + \kappa)e^{-\kappa/(Mh)}, \quad \text{almost surely.}$$

Hence, we apply the same reasoning as above and obtain almost surely

$$|\gamma_1(H_j)| \leq \gamma_1(h) + (mh)^{-1}(Mh + \kappa)e^{-\kappa/(Mh)} \left| H_j - h \right| ,$$

which implies the desired result by proceeding as above. □

7.8.3 Proof of Lemma 7.31

We first expand the square as

$$\left(\sum_{j=0}^{n-1}\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)\right)^2 = \sum_{j=0}^{n-1}\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)^2 + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)\left(\sum_{k=q}^{N-1} a_{ik}+b_i\right). \tag{7.48}$$

Then, we expand the square in the first sum and obtain

$$\begin{aligned}\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)^2 &= \left(\sum_{k=q}^{N-1} a_{jk}\right)^2 + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk}\\ &= \sum_{k=q}^{N-1} a_{jk}^2 + 2\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl} + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk}\\ &= a_{jq}^2 + \sum_{k=q+1}^{N-1} a_{jk}^2 + 2\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl} + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk}.\end{aligned} \tag{7.49}$$

We then rewrite the term appearing in the double sum in (7.48) as

$$\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)\left(\sum_{k=q}^{N-1} a_{ik}+b_i\right) = a_{jq}a_{iq} + \sum_{k=q}^{N-1}\sum_{\substack{l=q\\ l+k>2q}}^{N-1} a_{jk}a_{il} + b_j\sum_{k=q}^{N-1} a_{ik} + b_i\sum_{k=q}^{N-1} a_{jk} + b_ib_j \tag{7.50}$$

Substituting the expressions (7.49) and (7.50) in (7.48), we finally get

$$\left(\sum_{j=0}^{n-1}\left(\sum_{k=q}^{N-1} a_{jk}+b_j\right)\right)^2 = \sum_{j=0}^{n-1} a_{jq}^2 + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} a_{jq}a_{iq} + R(a) + S(a,b),$$

where the remainder $R(a)$ can be written as $R = R_1 + R_2 + R_3$ where

$$R_1(a) = \sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1} a_{jk}^2, \qquad R_2(a) = 2\sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl},$$
$$R_3(a) = 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\sum_{k=q}^{N-1}\sum_{\substack{l=q\\ l+k>2q}}^{N-1} a_{jk}a_{il},$$

and the remainder $S(a,b)$ can be written as $S = S_1 + S_2 + S_3 + S_4$ where

$$S_1(a,b) = \sum_{j=0}^{n-1} b_j^2, \qquad S_2(a,b) = 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} b_ib_j,$$
$$S_3(a,b) = 2\sum_{j=1}^{n-1}\sum_{k=q}^{N-1} b_ja_{jk}, \qquad S_4(a,b) = 2\sum_{j=1}^{n-1}\sum_{i=0}^{n-1}\left(b_j\sum_{k=q}^{N-1} a_{ik} + b_i\sum_{k=q}^{N-1} a_{jk}\right),$$

which proves the desired result. □

7.8.4 Proof of Lemma 7.32

In the following, all the constants are independent of h and n, but can depend on N and q. Moreover, since $h < 1$, we often apply $h^r \leq h^s$ for $r \geq s$. We first notice that, under Assumption 7.4 and Assumption 7.27, we get for all $j = 0, \dots, n-1$ and $k = q, \dots, N-1$

$$\begin{aligned} |\Delta_{j,k}| &= |Q_{k+1}(Y_j) - Q_{k+1}(Y_{j+1})| \\ &\leq C \left\| \psi_0(Y_j) - \psi_{H_j}(Y_j) \right\| \\ &\leq C_\Delta \, |H_j| \,, \end{aligned} \tag{7.51}$$

almost surely and where C_Δ is independent of h. Above, we exploited that Q_{k+1} is Lipschitz continuous for all $k = q, \dots, N+1$ due to Assumption 7.27. Let us now consider $R(\Delta)$. Due to (7.51) and to Assumption 7.29, we have

$$\begin{aligned} \mathbb{E}\left[(H_j^k - h^k)^2 \Delta_{j,k}^2\right] &\leq C_\Delta^2 \, \mathbb{E}\left[H_j^{k+1} - H_j h^k)^2\right] \\ &= C_\Delta^2 \Big(h^{2(k+1)} + C_{2(k+1)} h^{2p+2(k+1)-1} + h^{2(k+1)} + C_2 h^{2p+2k+1} \\ &\qquad -2h^{2k+2} - 2C_{k+2} h^{2p+2k+1} \Big) \\ &= C_\Delta^2 \left((C_{2(k+1)} + C_2 - 2C_{k+2}) h^{2p+2k+1} \right) \\ &\leq C h^{2p+2k+1}, \end{aligned} \tag{7.52}$$

where $C > 0$ is a positive constant. Now, since $k \geq q+1$, we get

$$\mathbb{E}\left[(H_j^k - h^k)^2 \Delta_{j,k}^2\right] \leq C h^{2(p+q+1)}.$$

Hence, for $R_1(\Delta)$ there exists a constant $\widetilde{C}_1$ such that

$$\mathbb{E}\left[R_1(\Delta)\right] \leq \widetilde{C}_1 n h^{2(p+q+1)}.$$

We now proceed to the second remainder $R_2(\Delta)$. Applying the Cauchy–Schwarz inequality and (7.52) we get

$$\begin{aligned} \mathbb{E}\left[(H_j^k - h^k)\Delta_{j,k}(H_j^l - h^l)\Delta_{j,l}\right] &\leq \mathbb{E}\left[(H_j^k - h^k)^2 \Delta_{j,k}^2\right]^{1/2} \mathbb{E}\left[(H_j^l - h^l)^2 \Delta_{j,l}^2\right]^{1/2} \\ &\leq C h^{2p+k+l+1}, \end{aligned}$$

where $C > 0$ is a positive constant. Now, since in the definition of $R_2(a)$ in (7.51) we have $k \geq q+1$ and $l \geq q$, we have here $k + l \geq 2q+1$. Therefore, there exists a constant $\widetilde{C}_2$ such that

$$\mathbb{E}\left[R_2(\Delta)\right] \leq \widetilde{C}_2 n h^{2(p+q+1)}.$$

We now consider the term $R_3(\Delta)$. Since H_i and H_j are independent for $i \neq j$, we have

$$\mathbb{E}\left[(H_j^k - h^k)\Delta_{j,k}(H_i^l - h^l)\Delta_{i,l}\right] = \mathbb{E}\left[(H_j^k - h^k)\Delta_{j,k}\right] \mathbb{E}\left[(H_i^l - h^l)\Delta_{i,l}\right].$$

Computing the two factors singularly, we have due to (7.51) and to Assumption 7.29

$$\begin{aligned} \mathbb{E}\left[(H_j^k - h^k)\Delta_{j,k}\right] &\leq C_\Delta \, \mathbb{E}\left[H_j^{k+1} - H_j h^k\right] \\ &= C_\Delta C_{k+1} h^{2p+k}, \end{aligned} \tag{7.53}$$

and analogously for $\mathbb{E}\left[(H_i^l - h^l)\Delta_{i,l}\right]$. Then, since $k + l \geq 2q+1$

$$\mathbb{E}\left[(H_j^k - h^k)\Delta_{j,k}(H_i^l - h^l)\Delta_{i,l}\right] \leq C_\Delta^2 C_{k+1} C_{l+1} h^{2(2p+q+1/2)}. \tag{7.54}$$

Hence, we have for a constant $\widetilde{C}_3 > 0$

$$\mathbb{E}\left[R_3(\Delta)\right] \leq \widetilde{C}_3 n^2 h^{2(2p+q+1/2)}.$$

Finally, replacing $t_n = nh$, we can write for a constant $C > 0$

$$\begin{aligned}\mathbb{E}\left[R(\Delta)\right] &\leq (\widetilde{C}_1 + \widetilde{C}_2) n h^{2(p+q+1)} + \widetilde{C}_3 n^2 h^{2(2p+q+1/2)} \\ &= (\widetilde{C}_1 + \widetilde{C}_2) t_n h^{2(p+q+1/2)} + \widetilde{C}_3 t_n^2 h^{2(2p+q-1/2)}.\end{aligned}$$

Let us now consider $S(\Delta, \eta)$. First, we notice that under the assumption $p \geq 3/2$ we have for any $r \geq 1$, $\min\{r, p + r - 3/2\} = r$, and therefore Lemma 7.30 simplifies to

$$\mathbb{E}\left[|\eta_j|^r\right] \leq C h^r e^{-r\kappa/(Mh)}.$$

We first consider $S_1(\Delta, \eta)$. Applying Lemma 7.30 with $r = 2$, we obtain for a constant $\widehat{C}_1 > 0$

$$\mathbb{E}\left[S_1(\Delta, \eta)\right] \leq \widehat{C}_1 n h^2 e^{-2\kappa/(Mh)}.$$

For the second term $S_2(\Delta, \eta)$, we have by (7.30) that $|\eta_i| \leq C H^i e^{-\kappa/H_i}$ and $\eta_j \leq C H^j e^{-\kappa/H_j}$ almost surely. These two bounds are independent for $i \neq j$ and therefore, applying Lemma 7.30 with $r = 1$, we have for a constant $\widehat{C}_2 > 0$

$$\mathbb{E}\left[S_2(\Delta, \eta)\right] \leq \widehat{C}_2 n^2 h^2 e^{-2\kappa/(Mh)}.$$

We now consider the third remainder $S_3(\Delta, \eta)$. Applying the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}\left[\eta_j (H_j^k - h^k)\Delta_{j,k}\right] \leq \mathbb{E}\left[\eta_j^2\right]^{1/2} \mathbb{E}\left[(H_j^k - h^k)^2 \Delta_{j,k}^2\right]^{1/2}.$$

Applying Lemma 7.30 with $r = 2$ to the first factor and (7.52) to the second we get

$$\begin{aligned}\mathbb{E}\left[\eta_j (H_j^k - h^k)\Delta_{j,k}\right] &\leq C h e^{-\kappa/(Mh)} h^{p+k+1/2} \\ &= C h^{p+k+3/2} e^{-\kappa/(Mh)}\end{aligned}$$

Now, since $k \geq q$, we have for a constant $\widehat{C}_3 > 0$

$$\mathbb{E}\left[S_3(\Delta, \eta)\right] \leq \widehat{C}_3 n h^{p+q+3/2} e^{-\kappa/(Mh)}.$$

Finally, we consider the last term $S_4(\Delta, \eta)$. Since by (7.30) it holds $|\eta_j| \leq C H_j e^{-\kappa/H_j}$ almost surely, and this bound is independent of H_i for $i \neq j$, applying (7.53) and Lemma 7.30 we have

$$\begin{aligned}\mathbb{E}\left[\eta_j (H_i^k - h^k)\Delta_{i,k}\right] &= \mathbb{E}\left[\eta_j\right] \mathbb{E}\left[(H_i^k - h^k)\Delta_{i,k}\right] \\ &\leq C h e^{-\kappa/(Mh)} h^{2p+k},\end{aligned}$$

which, since $k \geq q$, implies that there exists a constant $\widehat{C}_4 > 0$ such that

$$\mathbb{E}\left[S_4(\Delta, \eta)\right] \leq \widehat{C}_4 n^2 h^{2p+q+1} e^{-\kappa/(Mh)}.$$

Finally, replacing $t_n = nh$, we can write

$$\begin{aligned}\mathbb{E}\left[S(\Delta, \eta)\right] &\leq (\widehat{C}_1 n h^2 + \widehat{C}_2 n^2 h^2) e^{-2\kappa/(Mh)} + (\widehat{C}_3 n h^{p+q+3/2} + \widehat{C}_4 n^2 h^{2p+q+1}) e^{-\kappa/(Mh)} \\ &= (\widehat{C}_1 t_n h + \widehat{C}_2 t_n^2) e^{-2\kappa/(Mh)} + (\widehat{C}_3 t_n h^{p+q+1/2} + \widehat{C}_4 t_n^2 h^{2p+q-1}) e^{-\kappa/(Mh)},\end{aligned}$$

which completes the proof. □

8 Probabilistic Error Estimators with Random Mesh FEM

In this chapter we introduce the random mesh finite element method (RM-FEM) for elliptic PDEs. The RM-FEM is, to our knowledge, the second proposal in literature of a perturbation-based probabilistic method for PDEs in the sense of Definition 6.1, after the one presented in [39]. In particular, the RM-FEM is based on the finite element method (FEM) and proceeds to construct a probability measure over the solution by randomizing the discretization. In fact, we note here that the RM-FEM shares the basic ideas with the RTS-RK method presented in Chapter 7, where the randomization affects the space discretization instead of the time discretization. While Bayesian inverse problems remain a relevant application for the RM-FEM, in the framework of elliptic PDEs we are able to propose probabilistic a posteriori error estimator, which allow for adaptivity and goal-oriented implementations of the FEM. The content of this chapter is based on our article [7], and is one of the original contributions of this thesis.

The outline of this chapter is as follows. In Section 8.1 we state the problem of interest, introduce the RM-FEM and the main assumptions and notation required by our analysis. We then present the two main applications of the RM-FEM, i.e., a posteriori error estimators and Bayesian inverse problems, in Sections 8.2 and 8.3, respectively. For both applications, a series of numerical experiments in the one and two-dimensional cases illustrate the usefulness and efficiency of the RM-FEM. In Section 8.4 we present a rigorous a priori and a posteriori error analysis.

8.1 Random Mesh Finite Element Method

8.1.1 Notation

Let $d = 1, 2, 3$ and $D \subset \mathbb{R}^d$ be an open bounded domain with sufficiently smooth boundary ∂D. For $v \in \mathbb{R}^d$, we denote by $\|v\|_2$ the Euclidean norm on $\mathbb{R}^d$. We denote by $L^2(D)$ the space of square integrable functions, by $(\cdot, \cdot)$ the natural $L^2(D)$ inner product, and by $H^p(D)$ the Sobolev space of functions with p weak derivatives in $L^2(D)$. Moreover, we denote by $H_0^1(D)$ the space of functions in $H^1(D)$ vanishing on ∂D in the sense of traces, by $H^{-1}(D)$ the dual of $H_0^1(D)$ and by $\langle \cdot, \cdot \rangle$ the natural pairing between $H^{-1}(D)$ and $H_0^1(D)$. We equip the space $H_0^1(D)$ with the norm $\|v\|_{H_0^1(D)} = \|\nabla v\|_{L^2(D)}$, i.e. the $H^1(D)$ seminorm.

For an event space Ω, with a σ-algebra $\mathcal{A}$ and a probability measure P, we let the triple $(\Omega, \mathcal{A}, P)$ denote a probability space. For an event $A \in \mathcal{A}$, we say that A occurs almost surely (a.s.) if $P(A) = 1$. For $n \in \mathbb{N}$ we call random variables the measurable functions $X \colon \Omega \to \mathbb{R}^n$, and denote by $L^2(\Omega)$ the space of square integrable random variables, with associated inner product. Denoting by $\mathcal{B}(\mathbb{R}^n)$ the Borel σ-algebra on $\mathbb{R}^n$, we say that a probability measure μ_X on the

measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ satisfying $\mu_X(B) = P(X^{-1}(B))$ for all $B \in \mathcal{B}(\mathbb{R}^n)$ is the measure induced by X, or equivalently the distribution of X. For a set of random variables $\{X_i\}_{i=1}^n$ which are independent and identically distributed, we say they are i.i.d., and denoting by μ their common induced measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, we write $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu$.

8.1.2 Problem and Method Presentation

Let $\kappa \in L^\infty(D, \mathbb{R}^{d\times d})$, $f \in H^{-1}(D)$ and u be the weak solution of the partial differential equation (PDE)

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) &= f, \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{8.1}$$

i.e., the function $u \in V \equiv H_0^1(D)$ satisfying

$$a(u,v) = F(v), \qquad a(u,v) \coloneqq \int_D \kappa \nabla u \cdot \nabla v \,\mathrm{d}x, \quad F(v) \coloneqq \langle f, v\rangle, \tag{8.2}$$

for all functions $v \in V$. We assume there exist positive constants $\underline{\kappa}$ and $\bar{\kappa}$ such that for all $\xi \in \mathbb{R}^d$

$$\underline{\kappa} \left\|\xi\right\|_2^2 \leq \kappa\xi \cdot \xi \leq \bar{\kappa} \left\|\xi\right\|_2^2,$$

where $\left\|\cdot\right\|_2$ is the Euclidean norm on $\mathbb{R}^d$, so that there exist constants $m, M > 0$ such that for all $u, v \in V$ it holds

$$|a(u,v)| \leq M \left\|u\right\|_V \left\|v\right\|_V, \quad |a(u,u)| \geq m \left\|u\right\|_V^2.$$

The Lax–Milgram theorem then guarantees that the problem (8.2) is well-posed.

Let N be a positive integer and let $\mathcal{T}_h = \bigcup_{i=1}^N K_i$ be a partition of D, where for all $i = 1, \ldots, N$, the element $K_i \subset D$ is a segment, triangle or tetrahedron for $d = 1, 2, 3$ respectively. We denote by $h_i = \operatorname{diam}(K_i)$ the radius of the smallest ball containing K_i, and by $h = \max_i h_i$ the maximum radius, indexing the mesh $\mathcal{T}_h$. We denote by $\mathcal{V}_h$ the set of all vertices of the elements of $\mathcal{T}_h$, and in particular as $\mathcal{V}_h^I \subset \mathcal{V}_h$ the set of vertices which do not lie on the boundary of D, and by $\mathcal{V}_h^B = \mathcal{V}_h \setminus \mathcal{V}_h^I$. Moreover, we denote by N_I the number of internal vertices, i.e., $N_I = \left|\mathcal{V}_h^I\right|$. We assume the partition to be conforming, i.e., if two elements have non-empty intersection, than the latter consists of a point (for $d = 1$), of either a vertex or a side (for $d = 2$), and of either a vertex, a segment or a face (for $d = 3$). We then denote by $V_h \subset V$, $\dim(V_h) < \infty$ the space of continuous piecewise linear finite elements on $\mathcal{T}_h$, i.e.,

$$V_h \coloneqq \{v \in V \colon v\big|_K \in \mathbb{P}_1,\ \forall K \in \mathcal{T}_h\},$$

where $\mathbb{P}_1$ is the space of linear functions. Let us remark that imposing $u_h = 0$ on ∂D yields $\dim(V_h) = N_I$. The FEM proceeds by finding $u_h \in V_h$ such that

$$a(u_h, v_h) = F(v_h), \tag{8.3}$$

for all $v_h \in V_h$, which is equivalent to solving the linear system $A\mathbf{u} = \mathbf{f}$, where

$$\mathbf{u}_j = u_h(x_j), \quad x_j \in \mathcal{V}_h^I, \quad A_{ij} = a(\varphi_j, \varphi_i), \quad \mathbf{f}_j = F(\varphi_j), \quad i,j = 1, \ldots, N_I,$$

and where $\{\varphi_j\}_{j=1}^{N-1}$ are the Legendre basis functions defined on the internal vertices of $\mathcal{T}_h$. The assumptions on κ guarantee that A is symmetric positive definite, and in turn that $\mathbf{u}$ is uniquely defined and the problem (8.3) is well-posed.

We now introduce the random-mesh finite element method (RM-FEM), which is based on a random perturbation of the mesh $\mathcal{T}_h$ obtained by moving the internal vertices. First, we here

detail how we build perturbed meshes and which kind of random perturbations we consider to be admissible. Let $p \geq 1$, $\alpha \coloneqq \{\alpha_i \colon \Omega \to \mathbb{R}^d\}_{i=1}^{N_I}$ be a sequence of random variables and let us define the set of internal points $\widetilde{\mathcal{V}}_h^I = \{\widetilde{x}_i\}_{i=1}^{N_I}$ where

$$\widetilde{x}_i \coloneqq x_i + h^p \alpha_i. \tag{8.4}$$

We then define the set of perturbed vertices as $\widetilde{\mathcal{V}}_h = \widetilde{\mathcal{V}}_h^I \cup \mathcal{V}_h^B$, i.e., the vertices on the boundary are left unchanged. The perturbed mesh is then simply $\widetilde{\mathcal{T}}_h = \bigcup_{i=1}^N \widetilde{K}_i$, where each element $\widetilde{K}_i$ has the same vertices as its corresponding element K_i in the original mesh, modulo the random perturbation (8.4). In other words, we compute the internal points of the perturbed mesh following (8.4), and keep the connectivity structure of the original mesh $\mathcal{T}_h$. Clearly, the mesh so defined is not conforming for any sequence of random variable α, for which we therefore introduce an assumption.

Assumption 8.1. The sequence of random variables α is such that

(i) its components α_i admit densities F_{α_i} with respect to the Lebesgue measure on $\mathbb{R}^d$, which satisfy $\mathrm{supp}(F_{\alpha_i}) \subset B_{r_i}$, where $B_{r_i} \subset \mathbb{R}^d$ is the ball centered in the origin and of radius $r_i > 0$, and which are radial, i.e., $F_{\alpha_i}(x) = F_{\alpha_i}(\|x\|_2)$,

(ii) the perturbed mesh $\widetilde{\mathcal{T}}_h$ is conforming a.s.

Let us remark that the assumption (i) actually implies for all $p \geq 1$ the assumption (ii) a.s., provided the radii r_i are chosen small enough. We assume in (i) the densities F_{α_i} to be radial functions so that the random perturbations do not have a privileged direction.

Example 8.2. In the one-dimensional case, let $0 = x_0 < x_1 < \ldots < x_N = 1$ so that we have $N_I = N - 1$. Denoting $K_i = (x_i, x_{i-1})$ we call $\bar{h}_i$ the minimum element size for the two intervals sharing the point x_i as a vertex, i.e., $\bar{h}_i \coloneqq \min\{h_i, h_{i+1}\}$. Then, a choice of random variables satisfying Assumption 8.1 is given by

$$\alpha_i = \left(h^{-1}\bar{h}_i\right)^p \bar{\alpha}_i, \quad i = 1, \ldots, N-1, \quad \{\bar{\alpha}_i\}_{i=1}^{N-1} \overset{\text{i.i.d.}}{\sim} \mathcal{U}\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right),$$

where for a set $D \in \mathbb{R}^d$ we denote by $\mathcal{U}(D)$ the uniform distribution over D. With this choice, indeed, we have that $\widetilde{x}_i < \widetilde{x}_{i-1}$ a.s., and therefore the perturbed mesh is conforming. In the two-dimensional case, we introduce for $i = 1, \ldots, N_I$ the notation

$$\Delta_i = \{K \in \mathcal{T}_h \colon K \text{ has } x_i \text{ as a vertex}\}.$$

Analogously to the one-dimensional case, we write $\bar{h}_i \coloneqq \min_{j : K_j \in \Delta_i} h_j$. In this case, it is possible to verify that choosing for all $i = 1, \ldots, N_I$

$$\alpha_i = (h^{-1}\bar{h}_i)^p \bar{\alpha}_i, \quad i = 1, \ldots, N_I, \quad \{\bar{\alpha}_i\}_{i=1}^{N_I} \overset{\text{i.i.d.}}{\sim} \mathcal{U}\left(B_{1/2}\right),$$

then α satisfies Assumption 8.1. We verify this graphically in Fig. 8.1, where we show a realization of the perturbed mesh based on a generic Delaunay mesh and on a structured mesh on $D = (0,1)^2$ along with the sets where the perturbed points are constrained to belong a.s. We notice that for $p > 1$ the magnitude of the perturbations clearly tends to vanish. Finally, we remark that similar admissible perturbations can be introduced in higher dimensions.

Having defined the perturbed mesh, we now proceed with describing the RM-FEM. Let $\widetilde{V}_h$ be the space of continuous piecewise linear finite elements on $\widetilde{\mathcal{T}}_h$. Let moreover $\{\widetilde{\varphi}_i\}_{i=1}^{N_I}$ be the Legendre basis functions defined on the internal vertices of $\widetilde{\mathcal{T}}_h$ and $\widetilde{\mathcal{I}} \colon \mathcal{C}^0(D) \cap V \to \widetilde{V}_h$ be the Lagrange interpolation operator onto $\widetilde{V}_h$, i.e., for a function $v \in \mathcal{C}^0(D) \cap V$ and for $x \in D$ we define

$$\widetilde{\mathcal{I}} v(x) \coloneqq \sum_{i=1}^{N_I} v(x_i) \widetilde{\varphi}_i(x).$$

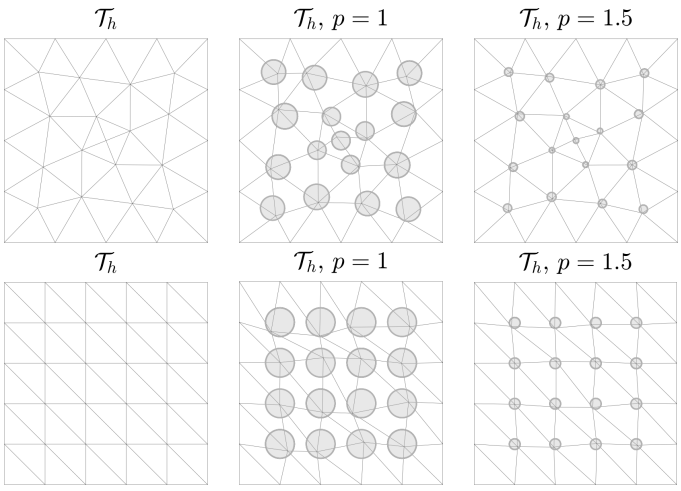


Figure 8.1 – A realization of $\widetilde{\mathcal{T}}_h$ for $p = \{1, 1.5\}$ based on two meshes $\mathcal{T}_h$ of $D = (0,1)^2$. On the first line, a Delaunay mesh. On the second line, a structured mesh. The regions where the perturbed points are included a.s. are depicted by light grey circles.

We are then interested in the two functions belonging to the finite element space $\widetilde{V}_h$ whose definition we give below.

Definition 8.3. Let $u_h \in V_h$ be defined in (8.3). We define the RM-FEM interpolant as the random function $\widetilde{\mathcal{I}}u_h \in \widetilde{V}_h$, where $\widetilde{\mathcal{I}}$ is the Lagrange interpolant onto $\widetilde{V}_h$.

Definition 8.4. Given the random finite element space $\widetilde{V}_h$, we define the RM-FEM solution as the unique random function $\widetilde{u}_h \in \widetilde{V}_h$ such that

$$a(\widetilde{u}_h, \widetilde{v}_h) = F(\widetilde{v}_h),$$

for all $\widetilde{v}_h \in \widetilde{V}_h$.

Remark 8.5. Clearly, either for any fixed $p \geq 1$ and $h \to 0$ or for any fixed $h < 1$ and $p \to \infty$, the functions u_h, $\widetilde{\mathcal{I}}u_h$ and $\widetilde{u}_h$ tend to coincide. We visualize this for u_h and $\widetilde{u}_h$ in Fig. 8.2, where we simply fix $\kappa = 1$ and the right-hand side f such that $u = \sin(2\pi x)$ in (8.1), and consider the effects of increasing p and decreasing h. For this simple problem, we notice that for $p = 2$ and $N = 20$ the FEM solution u_h and the RM-FEM solution $\widetilde{u}_h$ are almost indistinguishable.

Remark 8.6. All the quantities distinguished by a tilde (e.g., $\widetilde{\mathcal{T}}_h$, $\widetilde{V}_h$, $\widetilde{\mathcal{I}}$) are random variables with values in appropriate spaces. For example $\widetilde{u}_h$ is a random function $\widetilde{u}_h \colon \Omega \times D \to \mathbb{R}$, such that $\Omega \times D \ni (\omega, x) \mapsto \widetilde{u}_h(\omega, x)$. For economy of notation, in the following we drop the argument ω from all random variables.

Remark 8.7. The coefficient p in (8.4) has the same role as the coefficient identified by the same symbol in both [6, 39], i.e., it controls the variability of the probabilistic solutions by tuning the variability of the noise which is applied to the method.

Remark 8.8. Let us remark that the RM-FEM interpolant $\widetilde{\mathcal{I}}u_h$ is well-defined even allowing the vertices of $\mathcal{T}_h$ which lay on the boundary ∂D to be perturbed, as far as the perturbation moves them inside the domain D. The random RM-FEM interpolant $\widetilde{\mathcal{I}}u_h$ does not in this case belong to the space V in this case since it is not defined on the whole domain D and does not satisfy boundary conditions. For practical applications, one can nevertheless employ the RM-FEM interpolant defined on a smaller domain, which results from a perturbation of all vertices of $\mathcal{T}_h$, including those on the boundaries.

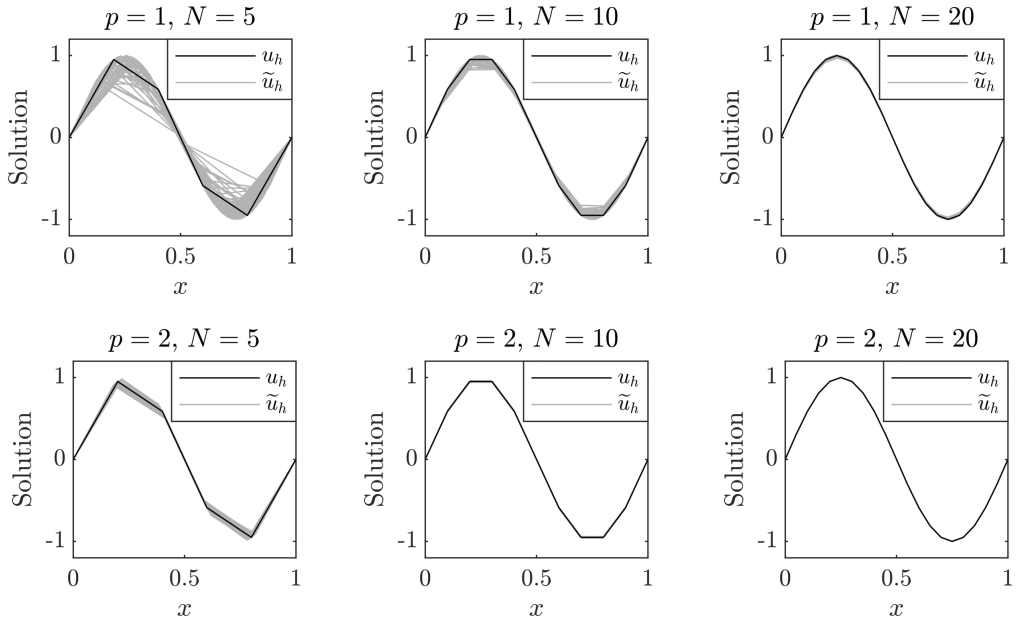


Figure 8.2 – Comparison between the RM-FEM and the FEM solutions. We display the solution u_h and 50 realizations of $\widetilde{u}_h$, by row respectively for $p = \{1, 2\}$ and by column for $N = \{5, 10, 20\}$.

Before proceeding with the two main applications of the RM-FEM, i.e., a posteriori error estimators and Bayesian inverse problems, we state an a priori error estimate, which suggests how to balance the sources of error due to numerical discretization and to the randomization of the method, respectively.

Theorem 8.9. *Let the solution u of (8.1) satisfy $u \in H^2(D)$. Then, it holds*

$$\|\widetilde{u}_h - u\|_V \leq Ch, \quad a.s.,$$

for a constant $C > 0$ independent of h. Moreover, if $p = 1$ in (8.4) the numerical and discretization errors are balanced with respect to h, i.e., it holds

$$\|u_h - \widetilde{u}_h\|_V = \mathcal{O}(h) = \mathcal{O}(\|u - u_h\|_V), \quad a.s.$$

This results indicates that one should fix $p = 1$ in (8.4) in order to obtain a family of probabilistic solutions whose statistical properties should reflect the true error. This is crucial when the RM-FEM is employed in a pipeline of computations such as Bayesian inverse problems, which will be presented in detail in Section 8.3. The proof of Theorem 8.9 is elementary and discussed in Section 8.4.1.

8.2 A Posteriori Error Estimators based on the RM-FEM

The first and foremost application of the RM-FEM is deriving a posteriori error estimators which are entirely based on the statistical information carried on by the mesh perturbation. We say that a quantity $\mathcal{E}_h$ is an a posteriori error estimator if it gives an error estimate on the numerical approximation and is computable only by knowledge of the numerical solution. Moreover, if there exist constants C_{up} and C_{low} independent of h and of u such that

$$C_{\text{low}}\mathcal{E}_h \leq \|u - u_h\|_V \leq C_{\text{up}}\mathcal{E}_h, \tag{8.5}$$

we say that the a posteriori error estimator is reliable and efficient, respectively. Indeed, the upper bound above guarantees that when the estimator is small, so is the numerical error. The lower bound, instead, gives an insurance on the quality of the estimator, as it shows that the estimation of the error is not exceedingly pessimistic. There exist in the literature a huge number of a posteriori error estimators, and we refer the reader to the surveys given e.g. in [12, 141]. Most a posteriori error estimators are expressed in the form

$$\mathcal{E}_h = \left(\sum_{K \in \mathcal{T}_h} \eta_K^2 \right)^{1/2},$$

where the η_K are local quantities depending on the solution and the data on the element K and its neighbors. For example, in the two-dimensional case a valid a posteriori error estimator is given by the expression of its local components

$$\eta_K^2 = h_K^2 \left\| f \right\|_{L^2(K)}^2 + h_K \left\| [\![\nabla u_h \cdot \nu_K]\!] \right\|_{L^2(\partial K)}^2,$$

where $[\![\cdot]\!]$ is the jump operator and ν_K denotes the unitary vector normal to the boundary of K (see e.g. [140, Section 3] or [12, Chapter 2]). Other a posteriori error estimators are based on recovered gradients, which are employed as surrogates of the gradient of the exact solution to estimate the error. A notable member of these methodologies is the Zienkiewicz–Zhu (ZZ) patch recovery technique [147, 148], which is proved to be superconvergent on special meshes, and which is in practice widely employed on any mesh.

It has been heuristically noted for ODEs in [25, 31, 127] that information on the variability of a probabilistic solution can be employed to estimate the error and thus adapt the numerical discretization. Indeed, building probabilistic solution to otherwise deterministic problems should pursue the goal of quantifying numerical errors through uncertainty. Guided by this observation, we now introduce two probabilistic error estimators for elliptic PDEs.

Definition 8.10. Let $\widetilde{\mathcal{I}} u_h$ be the RM-FEM interpolant defined in Definition 8.3 and for each $K \in \mathcal{T}_h$, let us denote by $\widetilde{K} \in \widetilde{\mathcal{T}}_h$ its corresponding element in $\widetilde{\mathcal{T}}_h$. We define the first RM-FEM a posteriori error estimator as

$$\widetilde{\mathcal{E}}_{h,1} \coloneqq \left(\sum_{K \in \mathcal{T}_h} \widetilde{\eta}_{K,1}^2 \right)^{1/2}, \quad \text{with} \quad \widetilde{\eta}_{K,1}^2 = h_K^{-(p-1)} \, \mathbb{E}\left[\left\| \nabla (u_h - \widetilde{\mathcal{I}} u_h) \right\|_{L^2(\widetilde{K})}^2 \right].$$

Moreover, we define the second RM-FEM a posteriori error estimator as

$$\widetilde{\mathcal{E}}_{h,2} \coloneqq \left(\sum_{K \in \mathcal{T}_h} \widetilde{\eta}_{K,2}^2 \right)^{1/2}, \quad \text{with} \quad \widetilde{\eta}_{K,2}^2 = h_K^{-(2p-2)} \, |K| \, \mathbb{E}\left[\left\| \nabla u_h \big|_K - \nabla \widetilde{\mathcal{I}} u_h \big|_{\widetilde{K}} \right\|^2 \right].$$

Remark 8.11. The scaling factors $h_K^{-(p-1)}$ and $h_K^{-(2p-2)}$ in the definition of $\widetilde{\eta}_{K,1}$ and $\widetilde{\eta}_{K,2}$ are necessary to obtain well-calibrated error estimators. This is made clearer in the one-dimensional case by the analysis presented in Section 8.4.2. For higher dimensions, they can be partially explained with the ansatz (8.13), especially for the first estimator $\widetilde{\mathcal{E}}_{h,1}$, and they appear in practice to be the correct scaling.

Remark 8.12. Computing the estimator $\widetilde{\mathcal{E}}_{h,1}$ is more involved than the estimator $\widetilde{\mathcal{E}}_{h,2}$. Indeed, for the latter it is sufficient to compute the interpolant $\widetilde{\mathcal{I}} u_h$ and the gradients over each element of u_h and of the interpolant. For $\widetilde{\mathcal{E}}_{h,1}$, instead, one has to compute on each element $\widetilde{K}$ the quantity

$$\left\| \nabla (u_h - \widetilde{\mathcal{I}} u_h) \right\|_{L^2(\widetilde{K})}.$$

By construction, each element $\widetilde{K}$ overlaps with the elements corresponding to its neighbors in the original mesh in a non-trivial manner, and if $d > 1$ one has to rely to the construction of a "super-mesh" (see e.g. [42, 43]) such that on each of its elements the quantity $\nabla(u_h - \widetilde{\mathcal{I}}u_h)$ is constant. A super-mesh has to be built for each realization of the perturbed mesh $\widetilde{\mathcal{T}}_h$, which could therefore be expensive.

In this article, we show in the one-dimensional case that the estimators given in Definition 8.10 are reliable and efficient in the sense of (8.5). In the statement of our theoretical result, which is given below, we make use of a quantity $\Lambda \in \mathbb{R}$ which is of higher order in most practical scenarios and which is defined as

$$\Lambda^2 \coloneqq h^\zeta \sum_{j=1}^{N} \int_{K_j} (f(x) + C_j)^2 \,\mathrm{d}x, \tag{8.6}$$

where for each K_j the real constant C_j will be specified in the analysis of Section 8.4.2 (see e.g. [19, Equation (8.7)]). Moreover, we consider one-dimensional meshes which are λ-quasi-uniform, i.e., we assume there exists a constant $\lambda \in (1, \infty)$ such that it holds

$$\frac{1}{\lambda} \leq \frac{h_j}{h_{j-1}} \leq \lambda, \quad j = 2, \dots, N,$$

uniformly in h. Finally, we consider perturbations satisfying

$$\alpha_i = \left(h^{-1}\bar{h}_i\right)^p \bar{\alpha}_i, \quad i = 1, \dots, N-1,$$

where $\bar{h}_i = \min\{h_i, h_{i+1}\}$ and for a i.i.d. sequence of random variables $\{\bar{\alpha}_i\}_{i=1}^{N-1}$ such that $|\bar{\alpha}_1| \leq 1/2$ a.s. These perturbations are indeed the same as the ones presented in Example 8.2, but without the assumption of $\{\bar{\alpha}_i\}_{i=1}^{N}$ to be uniformly distributed, which is not necessary in the following. In practice, a uniform distribution is nevertheless advisable, as it is still general enough and satisfies the radial assumption of Assumption 8.1(i). We moreover introduce the following technical assumption on the perturbation.

Assumption 8.13. Let the family of meshes $\mathcal{T}_h$ be λ-quasi-uniform, let p be the coefficient introduced in (8.4) and assume that for all h and p there exists $C > 0$ such that

$$4h^{p-1}\frac{\mathbb{E}\left|\bar{\alpha}_1\right|^2}{\mathbb{E}\left|\bar{\alpha}_1\right|} + C < 1 + \lambda^{-(p-1)}.$$

Remark 8.14. We note that Assumption 8.13 holds for $p > 1$ and h sufficiently small, and is therefore not restrictive in practice.

We can now state the main result involving a posteriori error estimators.

Theorem 8.15. *Let the dimension $d = 1$, let $p > 1$ in* (8.4) *and let Assumption 8.1 hold. Moreover, let $\widetilde{\mathcal{E}}_{h,1}$, $\widetilde{\mathcal{E}}_{h,2}$ and Λ be given in Definition 8.10 and* (8.6) *respectively and let the family of meshes $\mathcal{T}_h$ be λ-quasi-uniform. Then, there exists $C > 0$ independent of h and of the solution u such that it holds for $k \in \{1, 2\}$*

$$\|u - u_h\|_V \leq \widetilde{C}(\widetilde{\mathcal{E}}_{h,k}^2 + \Lambda^2)^{1/2},$$

up to higher order terms in h and under Assumption 8.13 for $k = 1$. If additionally $\kappa \in \mathcal{C}^2(D)$ and $f \in \mathcal{C}^1(D)$, then there exist constants $\widetilde{C}_{\mathrm{low}}$ and $\widetilde{C}_{\mathrm{up}}$ independent of h and of the solution u such that for $k \in \{1, 2\}$ it holds

$$\widetilde{C}_{\mathrm{low}}\widetilde{\mathcal{E}}_{h,k} \leq \|u - u_h\|_V \leq \widetilde{C}_{\mathrm{up}}\widetilde{\mathcal{E}}_{h,k},$$

up to higher order terms in h and under Assumption 8.13 for $k = 1$.

Let us notice that the estimators given in Definition 8.10 involve the computation of an expectation with respect to the random perturbations of the mesh, and therefore a Monte Carlo simulation is needed in practice. Let N_{MC} be a positive integer, $k \in \{1,2\}$ and $\{\widetilde{E}_{h,k}^{(i)}\}_{i=1}^{M}$ be i.i.d. realizations of the estimator $\widetilde{\mathcal{E}}_{h,k}$, obtained with independent perturbations of the mesh. Then, in practice we compute

$$\widetilde{E}_{h,k} \coloneqq \frac{1}{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} \widetilde{E}_{h,k}^{(i)}. \tag{8.7}$$

Remark 8.16. It could be suggested that the application of Monte Carlo techniques increases dramatically the simulation time. We argue that in practice the computational overhead is not relevant, mainly for three reasons. First, due to Theorem 6.7 the variance of Monte Carlo estimators drawn from probabilistic numerical methods decreases with respect to the discretization size h. Hence, the number of simulations M does not need to be large, nor increasing if $h \to 0$, to guarantee a good quality of the estimator. The same arguments hold for the RM-FEM, too. Second, the Monte Carlo estimation is completely parallelizable, thus reducing the cost by a factor equal to the number of available computing units. Finally, the computation of the RM-FEM interpolant $\widetilde{\mathcal{I}}u_h$ is not computationally involved, and neither is when repeated N_{MC} times.

8.2.1 Numerical Experiments

We now present numerical experiments on one and two-dimensional test cases to demonstrate the validity of our a posteriori error estimators. In particular, we are interested in determining whether the probabilistic error estimators introduced in Definition 8.10 are indeed reliable estimators for the numerical error in the FEM, and in employing these estimators for local refinements of the mesh. Setting a tolerance $\gamma > 0$, our goal is building a mesh $\mathcal{T}_h$ such that

$$\frac{\|u - u_h\|_V}{\|u_h\|_V} \leq \gamma. \tag{8.8}$$

Replacing the numerator with $\widetilde{\mathcal{E}}_{h,k}$, $k \in \{1,2\}$, we notice that the condition (8.8) is satisfied if it holds for all $K \in \mathcal{T}_h$

$$\widetilde{\eta}_{K,k} \leq \frac{\gamma \|u_h\|_V}{\widetilde{C}_{\mathrm{up}}\sqrt{N}} =: \gamma_{\mathrm{loc}}. \tag{8.9}$$

Indeed, in this case

$$\|u - u_h\|_V^2 \leq \widetilde{C}_{\mathrm{up}}^2 \widetilde{\mathcal{E}}_{h,k}^2 = \widetilde{C}_{\mathrm{up}}^2 \sum_{K \in \mathcal{T}_h} \widetilde{\eta}_{K,k}^2 \leq \gamma^2 \|u_h\|_V^2,$$

and thus (8.8) holds. Let us remark that $\widetilde{C}_{\mathrm{up}}^2$ is not known a priori in practice, and therefore we just decide to employ the condition (8.9) fixing $\widetilde{C}_{\mathrm{up}} = 1$ in our experiments. We therefore adapt the mesh by computing the local contributions and comparing them with γ_{loc}, thus locally refining the mesh if the condition (8.9) is not met, and coarsening if the local estimators are excessively small with respect to γ_{loc}.

In the following we employ for both the one and the two-dimensional cases the uniform distributions given in Example 8.2 for the random perturbations of the points. In light of Lemma 8.19 and Lemma 8.20, we decide to correct the estimators by normalizing them with respect to the random perturbations. In particular, in the following, the estimators are normalized as $\widetilde{\mathcal{E}}_{h,1} \leftarrow \widetilde{\mathcal{E}}_{h,1} / \mathbb{E}\,\|\bar{\alpha}_1\|$ and $\widetilde{\mathcal{E}}_{h,2} \leftarrow \widetilde{\mathcal{E}}_{h,2} / \mathbb{E}\,\|\bar{\alpha}_1\|^2$.

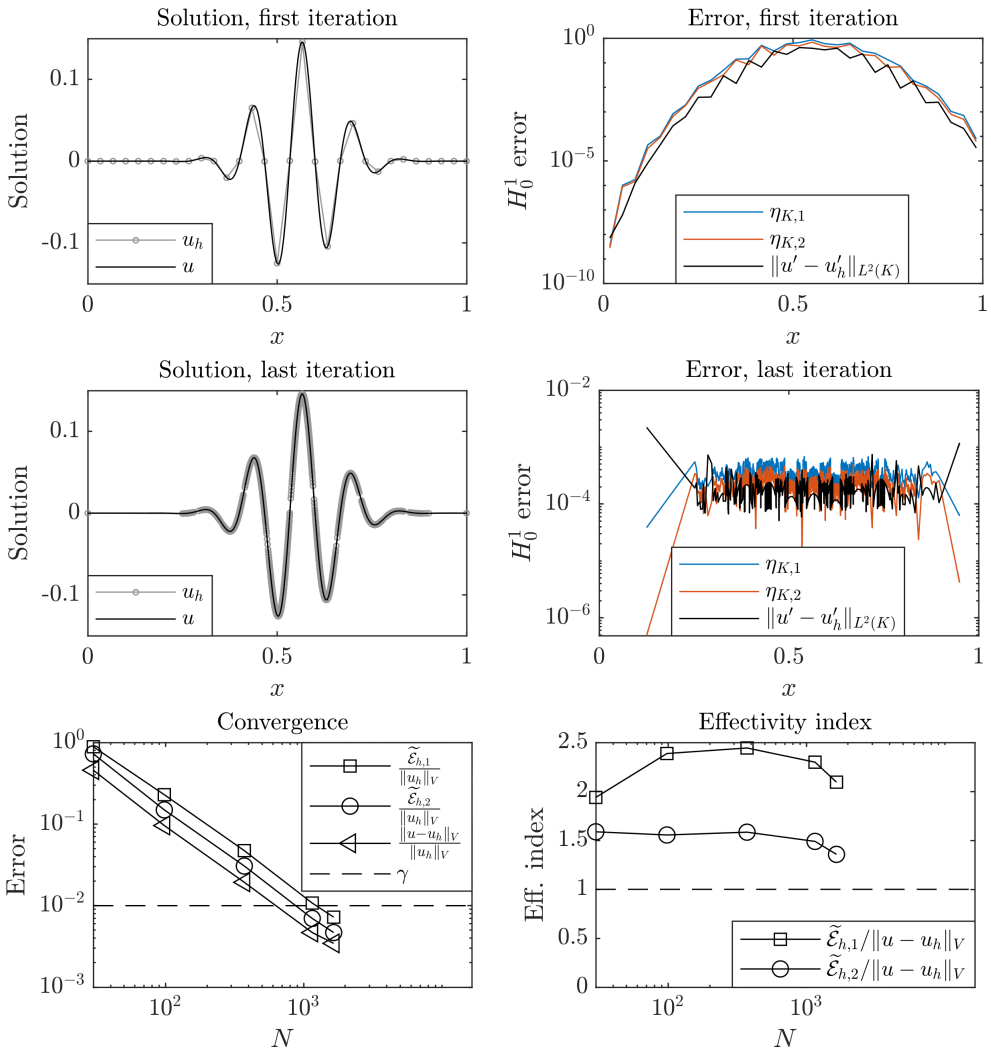


Figure 8.3 – Results for the one-dimensional experiment of Section 8.2.1. First and second rows: numerical and exact solutions u_h and u on the left, local contributions to the error estimators of Definition 8.10 compared with the true error $\|u - u_h\|_K$ on the right, at initialization and termination of the adaptivity procedure. Third row: on the left convergence of the global error $\|u - u_h\|_V$ and of the estimator $\mathcal{E}_h$ until the tolerance γ, on the right the effectivity index.

One-Dimensional Case

We first consider $d = 1$ and the two-point boundary value problem (8.15) with κ and the exact solution u given by

$$\kappa(x) = 1 + x^3, \quad u(x) = x^3 \sin(a\pi x) \exp(-b(x - 0.5)^2),$$

where we fix $a = 15$ and $b = 50$, and where we choose the right-hand side f so that u is indeed the solution. As a goal, we set the tolerance $\gamma = 10^{-2}$ in (8.8) and stop the algorithm when condition (8.9) is met by all elements of the mesh. We consider the RM-FEM implemented with uniform random variables as in Example 8.2 and fix $p = 3$ in (8.4). Moreover, we consider $N_{\mathrm{MC}} = 20$ realizations of the probabilistic mesh to approximate the error estimator as in (8.7). We then compute both the error estimators given in Definition 8.10 and employ $\widetilde{\mathcal{E}}_{h,1}$ for adapting the

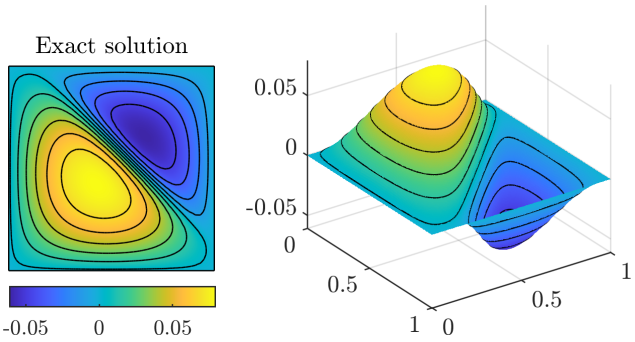


Figure 8.4 – Exact solution u_1 for the experiment of Section 8.2.1. Both the contour and the three-dimensional view highlight the steep gradient that features u_1.

mesh by refinement and coarsening, guided by the condition (8.9). The adaptivity algorithm is initialized with a mesh $\mathcal{T}_h$ built on $N = 30$ elements of equal size and proceeds by refinement and coarsening. Results, given in Fig. 8.3, confirm the validity of our probabilistic error estimators. In particular, we remark that the local error estimators succeed in identifying the regions where the mesh has to be refined, thus getting a solution with an approximately equal distribution of the error over the domain. Both probabilistic estimators, moreover, succeed in bounding the global error until the tolerance is reached, with the estimator $\widetilde{\mathcal{E}}_{h,2}$ which appears to be more efficient than $\widetilde{\mathcal{E}}_{h,1}$.

Two-Dimensional Case

We now present two numerical experiments conducted in the two-dimensional case. In particular, for both experiments we only focus on the computation of $\widetilde{\mathcal{E}}_{h,2}$ in Definition 8.10, since in view of Remark 8.12 this second estimator is computationally easier to implement than $\widetilde{\mathcal{E}}_{h,1}$ for $d > 1$. To account for errors on the boundary elements, we decide for these experiments to perturb all points, including those on the boundaries, following Remark 8.8. In order for $\widetilde{\mathcal{I}}u_h$, and thus $\widetilde{\mathcal{E}}_{h,2}$ to be well-defined, we reflect the perturbed boundary points symmetrically to the boundary ∂D in case they are outside the domain. For both experiments, we implement the RM-FEM with a uniform distribution for the random perturbations, as described in Example 8.2. Moreover, we fix $p = 3$ and compute the Monte Carlo approximation (8.7) on $N_{\mathrm{MC}} = 500$ realizations of the random mesh. For the adaptivity algorithm, we start from a coarse mesh and apply regular local refinements if the condition (8.9) is not met by the local error estimator $\widetilde{\eta}_{K,2}$. In the two-dimensional case we do not apply coarsening to the mesh.

We first consider $D = (0,1)^2$, the conductivity $\kappa = 1$, so that (8.1) reduces to $-\Delta u = f$ with homogeneous Dirichlet boundary conditions. Moreover, we choose the right-hand side f such that

$$u_1(x,y) = -x(1-x)y(1-y)\arctan\left(\beta\left(\frac{x+y}{\sqrt{2}} - \frac{4}{5}\right)\right),$$

where $\beta > 0$. The solution has a steep transition around the line $\{y = 4\sqrt{2}/5 - x\}$, whose steepness is proportional to the parameter β. In Fig. 8.4, we show the exact solution for $\beta = 20$, which we fix for this experiment. We initialize the adaptivity procedure with a mesh with maximum element size $h = 1/5$ and proceed with adaptation until a tolerance $\gamma = 0.1$. In Fig. 8.5 we show the convergence of $\widetilde{\mathcal{E}}_{h,2}$ with respect to the convergence of the true error, as well as the the effictivity index for this experiment. We can see that the estimator indeed captures the error globally. In Fig. 8.6, we show the behavior of the local contributions $\widetilde{\eta}_{K,2}$ with respect to the

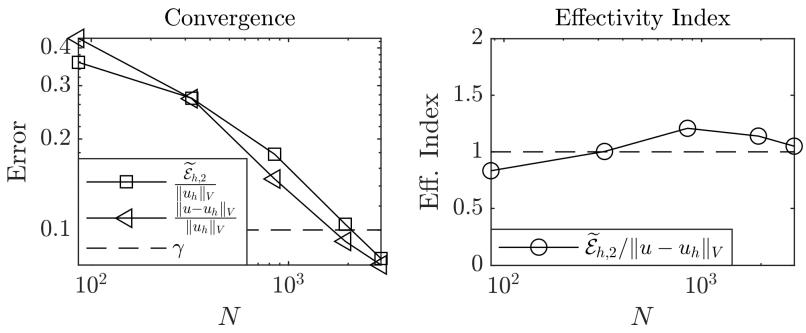


Figure 8.5 – Error convergence and effectivity index for the first experiment (function u_1) of Section 8.2.1

true error on each element, as well as the mesh adaptation. We can see that the error estimator succeeds in identifying the region where gradients are the steepest and proposes a mesh which appears adapted to this problem.

We then consider the L-shaped domain with the re-entrant corner on the origin, i.e. $D = (-1,1)^2 \setminus (-1,0)^2$. We set $\kappa = 1$, $f = 0$ and fix a inhomogeneous Dirichlet boundary conditions $u = g$ on ∂D, with g chosen such that the exact solution satisfies

$$u_2(r,\vartheta) = r^{2/3}\sin\left(\frac{2}{3}\left(\vartheta + \frac{\pi}{2}\right)\right),$$

where $(r,\vartheta) \in \mathbb{R}^+ \times (0, 2\pi]$ are the polar coordinates in $\mathbb{R}^2$. The exact solution of this problem is given in Fig. 8.7. Let us remark that the gradient of the exact solution is singular at the re-entrant corner, and we expect the mesh to be refined consequently at the singularity. For this experiment, we fix the tolerance $\gamma = 0.03$, and initialize the mesh to have a maximum element size of $h = 1/3$. Results, given in Fig. 8.8 and Fig. 8.9, show on the one hand that the estimator reproduces well the behavior of the global error during adaptation, and on the other hand that the mesh is progressively refined at the singularity as expected.

8.3 The RM-FEM for Bayesian Inverse Problems

In this section, we consider the application of the RM-FEM to Bayesian inverse problems. Let us recall that an introduction to Bayesian inverse problems is given in Chapter 1, and a general introduction of how to fit probabilistic methods in this framework is given in Section 6.2, where we also motivate the insertion of probabilistic forward maps into inverse problems. Here, we give only the details which are pertinent to the particular case of elliptic PDEs, and refer the reader to the aforementioned introductions for theoretical results and further details on implementation. In particular, we consider the framework of [49, Section 3.4] and introduce the parameterized PDE

$$\begin{aligned} -\nabla\cdot(\exp(\vartheta)\nabla u) &= f, \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{8.10}$$

where D is an open bounded set of $\mathbb{R}^d$ and $\vartheta\colon D \to \mathbb{R}$ is a scalar function. We let ϑ be such that problem (8.10) is well-posed, i.e., $\kappa = \exp(\vartheta) \in L^\infty(D)$ and $\kappa \geq \underline{\kappa} > 0$, and we denote by X the space of admissible values for ϑ. We introduce the solution operator $\mathcal{S}\colon X \to V$ such that $\mathcal{S}\colon \vartheta \mapsto u$, and the observation operator $\mathcal{O}\colon V \to \mathbb{R}^m$, which maps the solution of the PDE to point evaluations inside the domain on points $x^* = x_1^*, \dots, x_m^*$, i.e. $\mathcal{O}\colon u \mapsto y \coloneqq (u(x_1^*), \dots, u(x_m^*))^\top$. Moreover, we denote by $\mathcal{G} = \mathcal{O}\circ\mathcal{S}$, $\mathcal{G}\colon X \to \mathbb{R}^m$, the so-called forward operator, which maps the

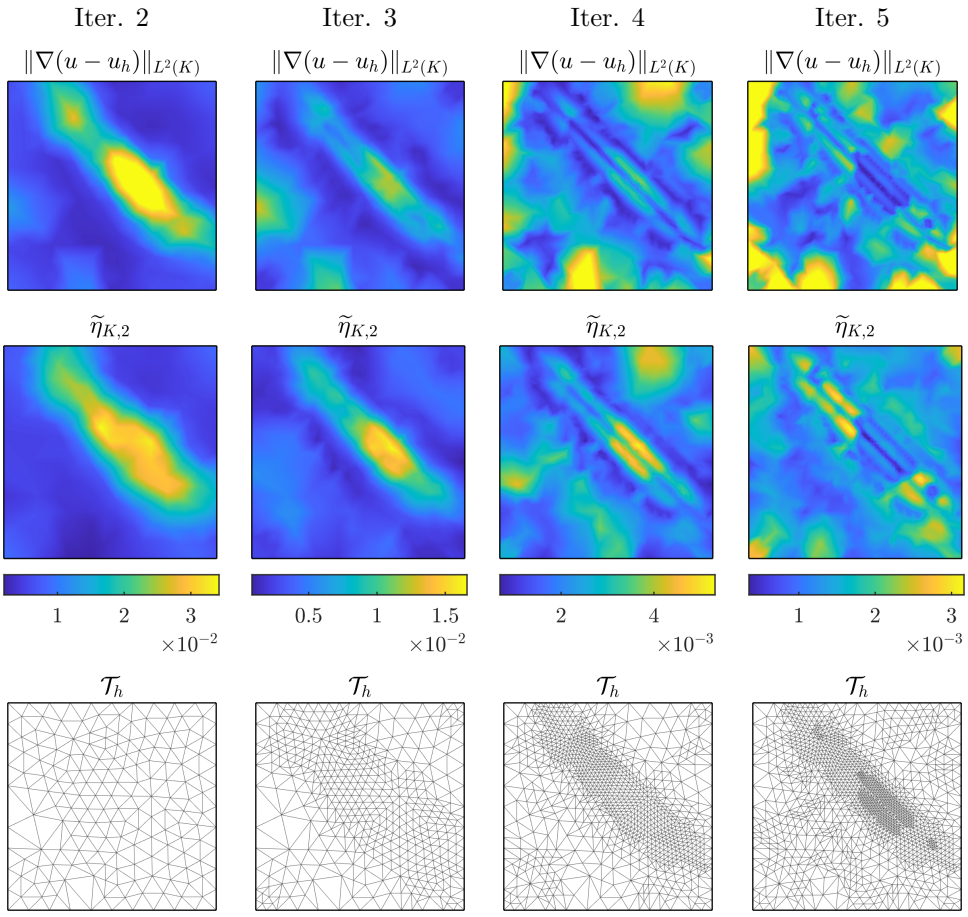


Figure 8.6 – Per row: True local error, local error estimator $\widetilde{\eta}_{K,2}$ and mesh $\mathcal{T}_h$ at each iteration of the adaptivity algorithm for the function u_1 of Section 8.2.1. The color bar is shared by the first and the second rows.

parameter to the observations. We then have the Gaussian observation model

$$y = \mathcal{G}(\vartheta) + \beta, \quad \beta \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma \in \mathbb{R}^{m\times m}$ is a non-singular covariance matrix on $\mathbb{R}^m$. Given an observation $y^* = (u(x_1^*), \dots, u(x_m^*))^\top$ associated to an unknown value $\vartheta^* \in X$ and corrupted by observational noise $\beta \in \mathbb{R}^m$ the inverse problem can then be stated as:

$$\text{find } \vartheta^* \in X \text{ given observations } y^* = \mathcal{G}(\vartheta^*) + \beta. \tag{8.11}$$

As discussed in Chapter 1, the randomness and the mismatch between the dimension of the unknown and of the observation make problem (8.11) ill-posed. We choose to restrict ourselves to the space $\mathcal{H} = \mathcal{C}^0(\overline{D}) \cap V$, which is a valid subspace of admissible values for ϑ, i.e., $\mathcal{H} \subset X$. We then consider a Gaussian prior measure $\mu_0 = \mathcal{N}(0, \Gamma_0)$, where $\Gamma_0 = -\Delta^{-\alpha}$ with $\alpha > d/2$ and where we equip the Laplacian with homogeneous boundary conditions. Fractional powers of the Laplacian should be understood as per [131, Section 2]. With this choice, we have that $\mu_0(\mathcal{H}) = 1$, and therefore μ_0 is a valid prior on the space $\mathcal{H}$ (see e.g. [131, Theorem 3.1]). Moreover, we remark that the forward operator $\mathcal{G}$ satisfies Assumption 1.1 due to the discussion of [131, Section

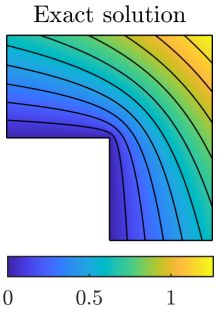


Figure 8.7 – Exact solution u_2 for the experiment of Section 8.2.1

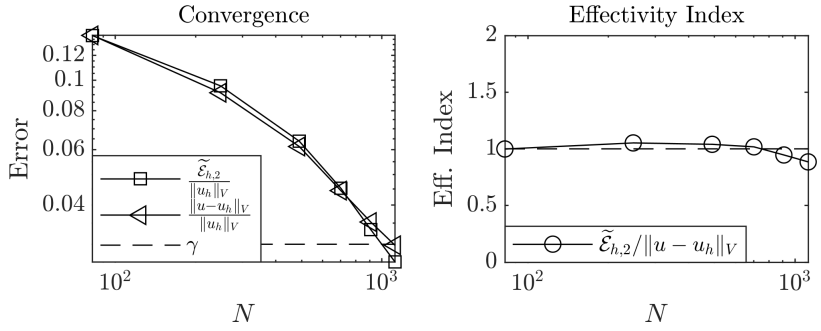


Figure 8.8 – Error convergence and effectivity index for the second experiment (function u_2) of Section 8.2.1

3.3]. Hence, the Bayesian inverse problem is well-posed and the posterior is given by

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{Z}\exp(-\Phi(\vartheta; y)),$$

where Z is the normalization constant

$$Z = \int_{\mathcal{H}} \exp(-\Phi(\vartheta; y))\,\mathrm{d}\mu_0(\vartheta),$$

and where for any $y \in \mathbb{R}^m$ the potential $\Phi(\cdot; y)\colon X \to \mathbb{R}$ is given due to the Gaussian assumption on the noise $\beta \sim \mathcal{N}(0, \Sigma)$ by

$$\Phi(\vartheta; y) = \frac{1}{2}\left\|\Sigma^{-1/2}\left(\mathcal{G}(\vartheta) - y\right)\right\|_2^2.$$

We consider the approximated posterior μ_h, obtained applying the forward map $\mathcal{G}_h = \mathcal{O}\circ\mathcal{S}_h$, where $\mathcal{O}$ is the observation operator defined above, and where $\mathcal{S}_h\colon X \to V_h$ maps a log-conductivity in the FEM solution $u_h \in V_h$. Moreover, we define following Section 6.2 the random posterior $\widetilde{\mu}_h$, which is obtained by discretizing the forward map with the RM-FEM as $\widetilde{\mathcal{G}}_h = \mathcal{O} \circ \widetilde{\mathcal{S}}_h$, where the solution operator $\widetilde{\mathcal{S}}_h\colon X \to \widetilde{V}_h$ maps a log-conductivity $\vartheta \in X$ to the RM-FEM solution $\widetilde{u}_h \in \widetilde{V}_h$ of Definition 8.4.

An approximate solution of the inverse problem is computed in practice employing the truncated Karhunen–Loève (KL) expansion based on the eigenfunctions of the prior covariance Γ_0, as described in Section 1.2, and by approximating the posterior $\widetilde{\mu}_h$ with its Monte Carlo surrogate $\widetilde{\mu}_{h,\mathrm{MC}}$ defined in Section 6.2. We note that the marginal approximation $\widetilde{\mu}_{h,\mathrm{mar}}$ could be employed, too, but due to the high dimension of the problem such a choice would entail a significant additional computational cost, as per Remark 6.10.

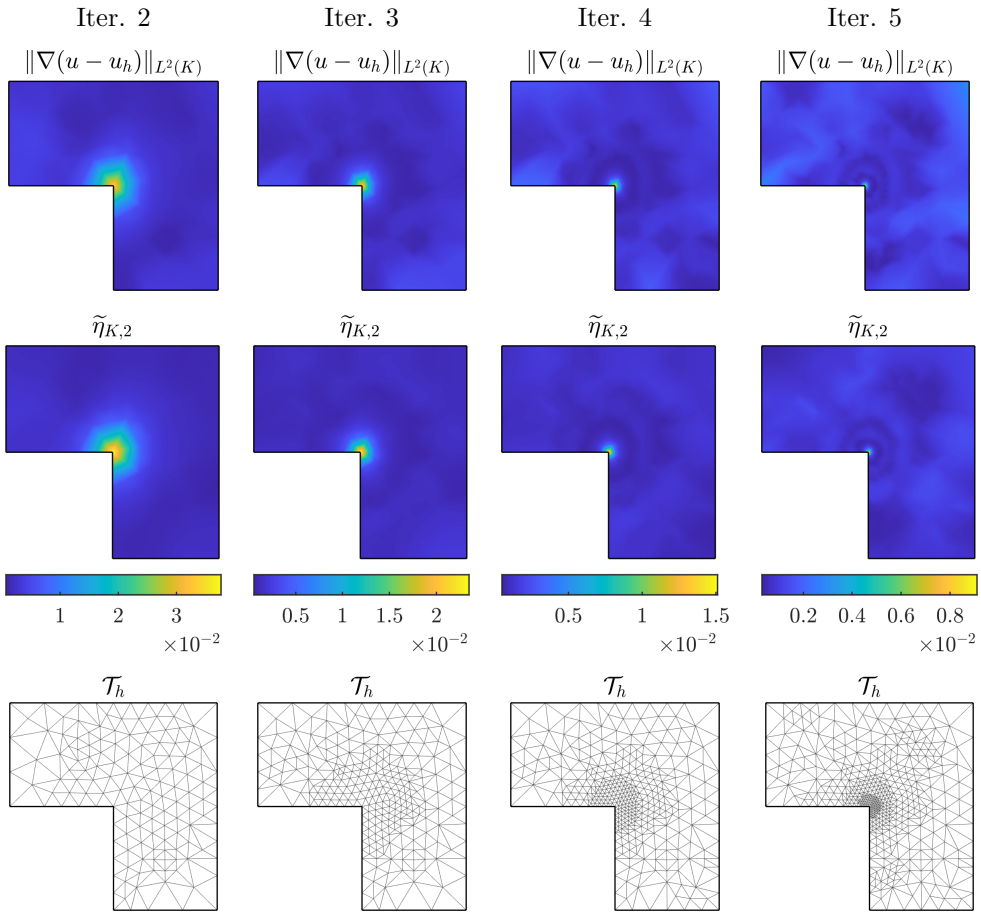


Figure 8.9 – Per row: True local error, local error estimator $\widetilde{\eta}_{K,2}$ and mesh $\mathcal{T}_h$ at each iteration of the adaptivity algorithm for the function u_2 of Section 8.2.1. The color bar is shared by the first and the second rows.

8.3.1 Numerical Experiments

In this section we present numerical experiments highlighting the beneficial effects of adopting the probabilistic framework of the RM-FEM in the context of Bayesian inverse problems.

One-Dimensional Case

We first consider $D = (0,1)$ and solve the inverse problem presented above for two different true diffusion fields κ^*. In both cases, we consider the prior on $\mathcal{H}$ to be given by $\mathcal{N}(0, \Gamma_0)$, with $\Gamma_0^{-1} = -\mathrm{d}^2/\mathrm{d}x^2$ with homogeneous boundary conditions, so that the Bayesian inverse problem is well-posed. First, we consider $\kappa_1^* = \exp(\vartheta_1^*)$, where the log-conductivity $\vartheta_1^* \in \mathcal{H}$ is given by

$$\vartheta_1^*(x) = \sum_{j=1}^{4} \xi_j \sqrt{\lambda_j} \varphi_j(x),$$

with $\xi_1 = \xi_2 = 1$, $\xi_3 = \xi_4 = 1/4$, and where $\{(\lambda_i, \varphi_i)\}_{i=1}^4$ are the first four ordered eigenpairs of Γ_0. Second, we consider $\vartheta_2^* \in X \cap \mathcal{H}^C$, so that the true conductivity does not belong to the domain in which we solve the inverse problem, but it is still admissible for (8.10) to be well-posed.

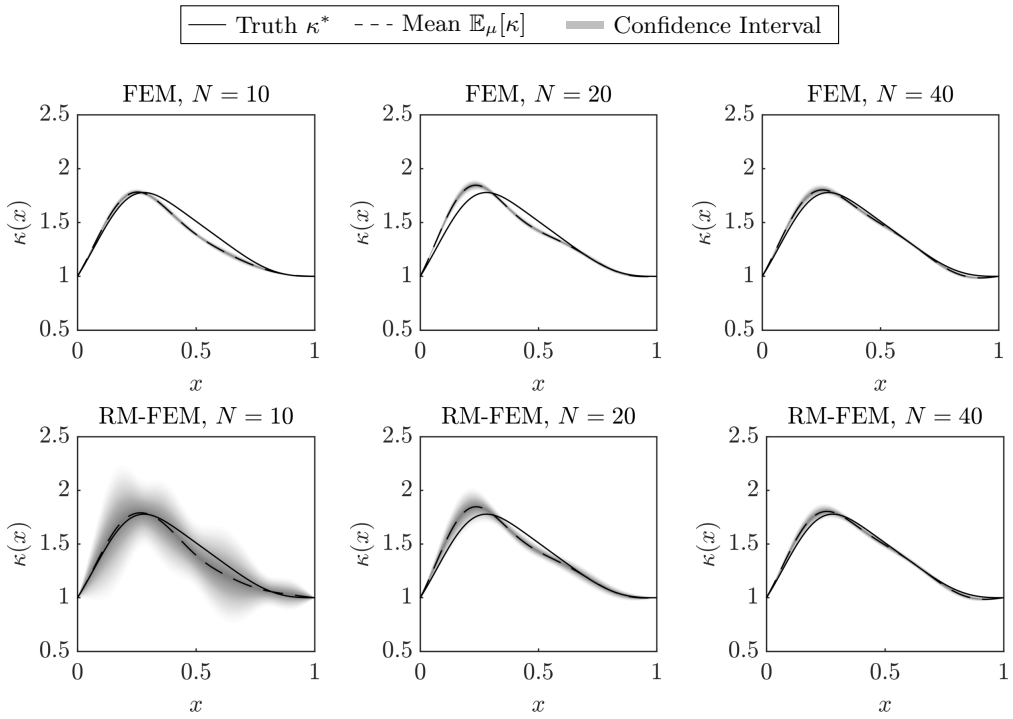


Figure 8.10 – Numerical results for κ_1^* in Section 8.3.1. In all plots, the solid line represents the true conductivity, the dashed line is the posterior mean, and the shaded grey area is a confidence interval. In the first row, results are obtained by approximating the forward map with FEM, and in the second with the RM-FEM.

In particular, we consider the discontinuous conductivity

$$\kappa_2^*(x) = \begin{cases} 1.5, & 0.2 < x < 0.6, \\ 0.5, & 0.6 < x < 0.8, \\ 1, & \text{otherwise}, \end{cases}$$

and infer $\vartheta_2^* = \log(\kappa_2^*)$. For both problems, we choose the right-hand side in (8.10) as $f(x) = \sin(2\pi x)$. Synthetic observations are obtained as point evaluations of a reference solution on points $x_i^* = i/10$, for $i = 1, \dots, 9$, corrupted by Gaussian noise $\mathcal{N}(0, 10^{-8} I)$. The forward map is approximated with FEM and RM-FEM. The mesh $\mathcal{T}_h$ for the FEM is equally spaced, and we vary the number of elements $N = \{10, 20, 40\}$. For the RM-FEM, we consider $p = 1$ in (8.4) as per Theorem 8.9 and implement the random perturbations with an uniform distribution as in Example 8.2.

We sample with the MH algorithm from the posterior distributions μ_h and $\widetilde{\mu}_h$, with $M_{\mathrm{MC}} = 2 \cdot 10^5$ for μ_h and with $M_{\mathrm{MC}} = 50$ and $M_{\mathrm{chain}} = 2 \cdot 10^5$ for $\widetilde{\mu}_h$. Knowing for the first conductivity κ_1^* that the true conductivity is fully determined by four coefficients, we fix the truncation index $N_{\mathrm{KL}} = 4$ in the Karhunen–Loève expansion. For the second conductivity κ_2^*, we fix $N_{\mathrm{KL}} = 9$. We then approximate the mean and pointwise standard deviation for the deterministic and probabilistic posteriors, respectively. Moreover, we arbitrarily fix a pointwise confidence interval at twice the standard deviation away from the mean. Numerical results are given in Fig. 8.10 and Fig. 8.11. Results highlight that for a coarse approximation, specifically for $N = 10$, the posterior distribution μ_h is overconfident on the result. Indeed, the posterior mean fails to capture precisely the true conductivity in both the continuous and discontinuous case, and the confidence interval

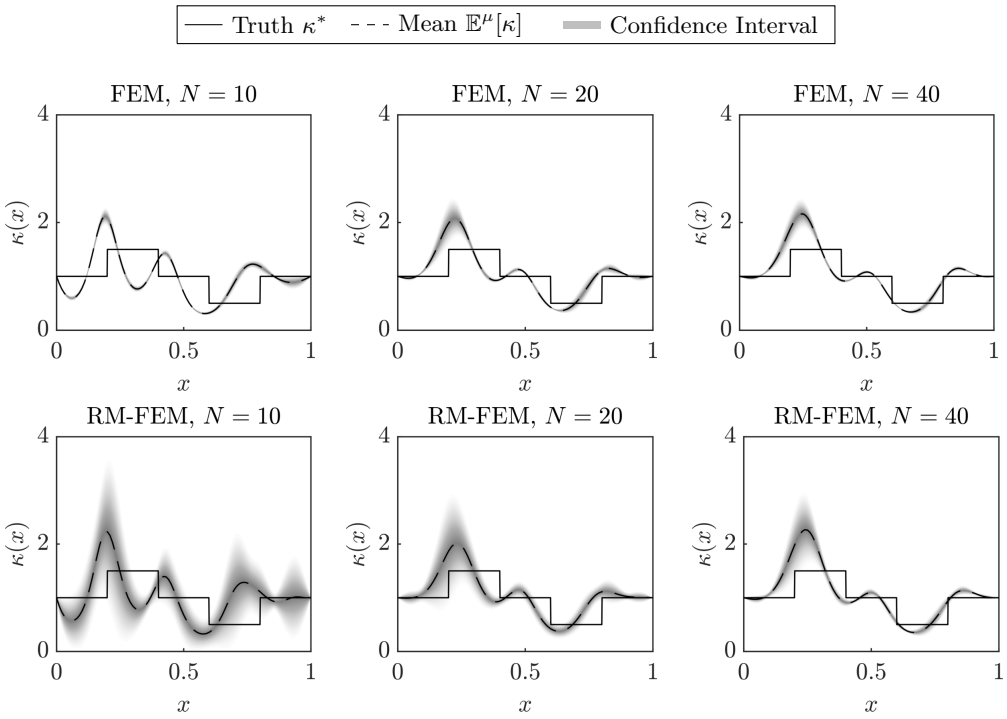


Figure 8.11 – Numerical results for κ_2^* in Section 8.3.1 In all plots, the solid line represents the true conductivity, the dashed line is the posterior mean, and the shaded grey area is a confidence interval. In the first row, results are obtained by approximating the forward map with FEM, and in the second with the RM-FEM.

is extremely sharply concentrated around the mean. Conversely, the distribution $\widetilde{\mu}_h$ based on the probabilistic forward model accounts better for the uncertainty due to numerical discretization. Increasing the number of elements N, the mean computed under μ_h and $\widetilde{\mu}_h$ tends to approximate better the true conductivity field. In particular, for $N = 40$ the posteriors μ_h and $\widetilde{\mu}_h$ are already practically undistinguishable and are close to the true field. Moreover, let us remark that while the width of the confidence interval seems independent of N for μ_h, it shrinks coherently to the discretization for $\widetilde{\mu}_h$. Finally, we note that for κ_2^* even for larger values of N the posterior $\widetilde{\mu}_h$ seems to capture with its uncertainty local errors in the solution of the inverse problem. Indeed, the posterior mean is particularly off the true field on the left side of the domain, where the confidence interval is wider with respect to areas where the solution is more accurate.

Two-Dimensional Case

We consider now a two dimensional example on the domain $D = (0,1)^2$. We fix a Gaussian prior μ_0 on $\mathcal{H}$ for the log-conductivity ϑ chosen as $\mu_0 = \mathcal{N}(0, \Gamma_0)$, where $\Gamma_0 = -\Delta^{-1.3}$ with homogeneous boundary conditions, so that the inverse problem is well-posed. We fix $N_{\mathrm{KL}} = 6$ and let the true conductivity $\kappa^* = \exp(\vartheta^*)$ in (8.10) be given by

$$\vartheta^* = \sum_{i=1}^{6} \sqrt{\lambda_i}\varphi_i \xi_i^*,$$

where $\{(\lambda_i, \varphi_i)\}_{i=1}^6$ are the first six ordered eigenpairs of Γ_0, and where $\xi_i^* = (-1)^{i+1} \cdot 10$ for $i = 1, 2, \ldots, 6$. Let us remark that $\vartheta^* \in \mathcal{H}$. The right-hand side in (8.10) is chosen as

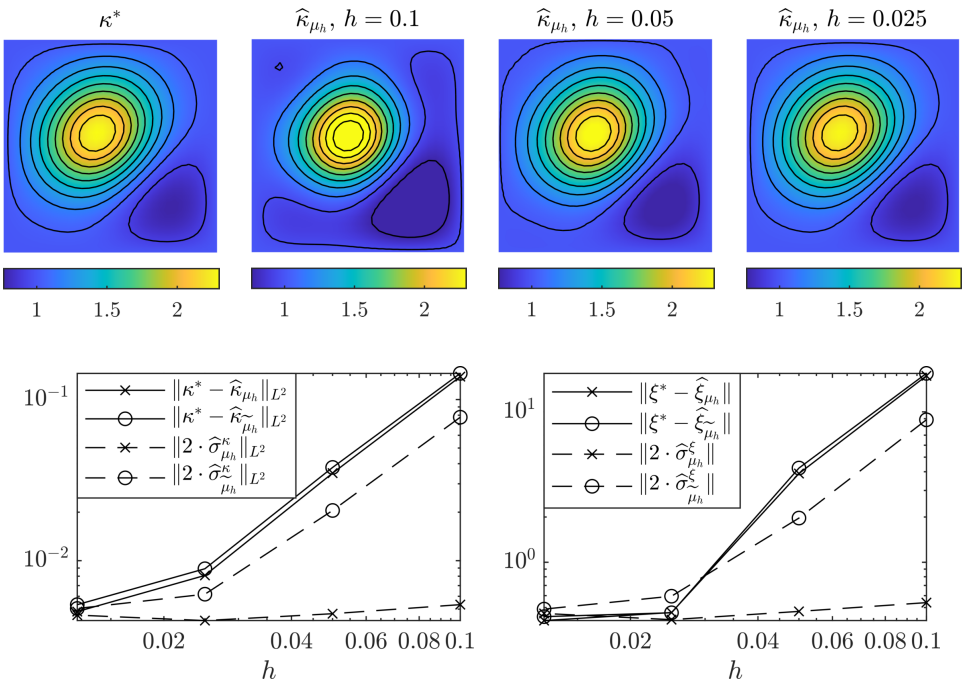


Figure 8.12 – Numerical results for Section 8.3.1. First row: True conductivity field κ^* and posterior mean field $\widehat{\kappa}_{\mu_h}$ estimated with MCMC and different values of h. Second row: Mean error vs standard deviation under μ_h and $\widetilde{\mu}_h$. On the left, $L^2(D)$ error on the mean field vs $L^2(D)$ norm of the punctual standard deviation under μ_h and $\widetilde{\mu}_h$. On the right, error with respect to exact KL coefficients, and standard deviations under μ_h and $\widetilde{\mu}_h$.

$f(x, y) = 8\pi^2 \sin(2\pi x) \sin(2\pi y)$. Synthetic observations are obtained by evaluating a reference solution on $m = 50$ random locations sampled from $\mathcal{U}(D)$ and then corrupted by an observational noise distributed as $\mathcal{N}(0, 10^{-6} I)$. We then approximate the forward map in the inverse problem with the FEM and the RM-FEM. We choose a structured mesh $\mathcal{T}_h$ as the one in Example 8.2 (or the second row of Fig. 8.1). In particular, in this case we let h denote the constant length of the short side of the triangular elements, i.e., the inverse of the number of subdivisions of each side of D. In particular, we consider $h = 0.1 \cdot 2^{-i}$, $i = 0, 1, \ldots, 3$. The RM-FEM is implemented with $p = 1$ in (8.4) as per Theorem 8.9, and with an uniform choice for the random perturbations as the one described in Example 8.2.

Employing the notation introduced in Section 6.2, we then sample from the posterior distributions μ_h and $\widetilde{\mu}_h$ employing the RAM method of [143] considering only the first $N_{\mathrm{KL}} = 6$ coefficients in the KL expansion. In particular, we consider $N_{\mathrm{MC}} = 10^5$ samples for the deterministic case, and for the probabilistic case we generate $N_{\mathrm{chain}} = 10^5$ samples for $N_{\mathrm{MC}} = 24$ parallel chains, each corresponding to a realization of the random mesh in the RM-FEM. We then compute for each value of h the mean and standard deviation of the field computed under μ_h (resp. $\widetilde{\mu}_h$) and denote their Monte Carlo approximations as $\widehat{\kappa}_{\mu_h}$ and $\widehat{\sigma}^{\kappa}_{\mu_h}$ (resp. $\widehat{\kappa}_{\widetilde{\mu}_h}$, $\widehat{\sigma}^{\kappa}_{\widetilde{\mu}_h}$). Moreover, we consider the statistics of the 6-dimensional coefficient σ of the KL expansion, and denote by $\widehat{\xi}_{\mu_h}$ and $\widehat{\sigma}^{\xi}_{\mu_h}$ the Monte Carlo approximation of mean and standard deviation computed under μ_h (resp. $\widehat{\xi}_{\widetilde{\mu}_h}$, $\widehat{\sigma}^{\xi}_{\widetilde{\mu}_h}$). We show in Fig. 8.12 the posterior mean $\widehat{\kappa}_{\mu_h}$ for three values of h, compared to the truth κ^*, and remark that the mean approximation is sensibly better for smaller values of h. The mean value under the probabilistic posterior $\widetilde{\mu}_h$ is not shown, as it is essentially equal to the deterministic case. The beneficial effect of employing the RM-FEM-based posterior distribution

$\widetilde{\mu}_h$, with respect to the FEM-based posterior μ_h, consists of the approximate equalities

$$\left\|\widehat{\sigma}^{\kappa}_{\widetilde{\mu}_h}\right\|_{L^2(D)} = \mathcal{O}\left(\left\|\kappa^* - \widehat{\kappa}_{\widetilde{\mu}_h}\right\|_{L^2(D)}\right), \qquad \left\|\widehat{\sigma}^{\xi}_{\widetilde{\mu}_h}\right\| = \mathcal{O}\left(\left\|\xi^* - \widehat{\xi}_{\widetilde{\mu}_h}\right\|\right),$$

which indicate that the error on the conductivity field, or on the coefficients of its KL expansion, are well represented by the uncertainty in the posterior distribution. This is shown in Fig. 8.12, where we notice that under μ_h the standard deviation is practically independent of h and small with respect to the error on the solution of the inverse problem. Conversely, under $\widetilde{\mu}_h$ we have that the posterior standard deviation converges accordingly to the error, both for the L^2-norm of the error on the mean and for the coefficients of the KL expansion.

8.4 Error Analysis for the RM-FEM

In this section, we present our a priori and our a posteriori error analysis for the RM-FEM. Let us remark that while the a priori error analysis is carried on for a general space dimension d and the adaptive algorithm has been shown to be efficient in higher dimensions (see Section 8.2.1), we present a rigorous a posteriori error analysis only in case $d = 1$. Conversely, in the a priori analysis we fix the coefficient $p = 1$ in (8.4), whereas in the a posteriori analysis we consider general perturbations, i.e., general coefficients $p \geq 1$ in the same equality.

8.4.1 A Priori Error Estimates

We first prove the a priori error estimate given in Theorem 8.9. The convergence properties of the FEM for the elliptic problem (8.1) are well-established. In particular, without any additional assumptions on the exact solution, i.e., when $u \in V$, it holds $\|u - u_h\|_V \to 0$ for $h \to 0$. Under the more restrictive assumption $u \in H^2(D) \cap V$, we have a linear convergence rate, i.e.

$$\|u - u_h\|_V \leq Ch\,|u|_{H^2(D)}\,, \tag{8.12}$$

for a constant $C > 0$, which is independent of h and u [26, 33, 117]. It is desirable that the RM-FEM is endowed with the same property. Moreover, we wish the error due to randomization to be balanced with the error due to the FEM discretization, which is shown in the proof of Theorem 8.9 below.

Proof of Theorem 8.9. Since (8.12) holds independently of the mesh, we have

$$\|u - \widetilde{u}_h\|_V \leq \widetilde{C}h\,|u|_{H^2(D)}\,, \quad \text{a.s.},$$

for a constant $\widetilde{C}$ independent of h and u and of the coefficient p in (8.4). Hence, by the triangle inequality we have for $p = 1$

$$\|u_h - \widetilde{u}_h\|_V \leq \|u - u_h\|_V + \|u - \widetilde{u}_h\|_V \leq (C + \widetilde{C})h\,|u|_{H^2(D)}\,, \quad \text{a.s.},$$

i.e., we have $\mathcal{O}(\|u_h - \widetilde{u}_h\|_V) = \mathcal{O}(\|u - u_h\|_V)$, which shows the desired result. □

Let us remark that we have shown above that the probabilistic solution converges with the same rate with respect to h in case $p = 1$, but we have not considered the case $p > 1$, for which the probabilistic term may be of higher order. Indeed, a preliminary theoretical and numerical investigation leads us to conjecture that

$$\left(\mathbb{E}\,\|u_h - \widetilde{u}_h\|_V^2\right)^{1/2} \leq Ch^{(p+1)/2}, \tag{8.13}$$

so that, at least in the mean-square sense, the error due to randomization should converge faster than the error due to discretization if $p > 1$.

8.4.2 A Posteriori Error Analysis in the One-Dimensional Case

In this section we prove our main result for the a posteriori error estimator of the RM-FEM given in Definition 8.10, namely Theorem 8.15. Our goal is to prove in the one-dimensional case that the probabilistic a posteriori error estimators are reliable and efficient, i.e., that there exist positive constants $\widetilde{C}_{\text{low}}$ and $\widetilde{C}_{\text{up}}$ independent of h and u such that

$$\widetilde{C}_{\text{low}}\widetilde{\mathcal{E}}_{h,k} \leq \|u - u_h\|_V \leq \widetilde{C}_{\text{low}}\widetilde{\mathcal{E}}_{h,k}. \tag{8.14}$$

for $k = \{1, 2\}$. Consider the elliptic two-point boundary value problem

$$\begin{aligned} -(\kappa u')' &= f, \quad \text{in } D, \\ u(0) &= u(1) = 0, \end{aligned} \tag{8.15}$$

where $\kappa \in L^\infty(D)$ satisfies $\kappa(x) \geq \underline{\kappa}$ almost everywhere in D, and where we assume $f \in L^1(D)$. We recall that the notation for one-dimensional problems has been introduced and discussed in Example 8.2 and at the end of Section 8.2. Additionally, we introduce here for a function w which is piecewise constant on each $K_i \in \mathcal{T}_h$ the jump operator

$$[\![w]\!]_{x_i} \coloneqq w\big|_{K_i} - w\big|_{K_{i+1}}, \qquad i = 1, \dots, N-1, \quad [\![w]\!]_{x_0} = [\![w]\!]_{x_N} = 0.$$

Our strategy for proving that the error estimator introduced in Definition 8.10 satisfies (8.14) relies on showing it is equivalent to known valid estimators. In particular, we consider the following estimator, defined in [19, Definition 6.3].

Definition 8.17. Let κ be the diffusion coefficient of (8.15) satisfy $\kappa \in \mathcal{C}^0(D)$ and $\kappa \geq \underline{\kappa} > 0$. We define the error estimator

$$\mathcal{E}_h^2 \coloneqq \sum_{j=1}^{N} \eta_j^2, \quad \eta_j \coloneqq \left\|\kappa^{-1}\ell_j\right\|_{L^2(K_j)},$$

with $\ell_j \colon K_j \to \mathbb{R}$ the linear function defined by $\ell_j(x_{j-1}) = \tau_{j,1}$, $\ell_j(x_j) = -\tau_{j,0}$ where

$$\tau_{j,k} = \frac{h_j}{h_{j-k+1} + h_{j-k}} [\![u_h']\!]_{x_{j-k}} \kappa(x_{j-k}).$$

Clearly, the quantity $\mathcal{E}_h$ is computable up to quadrature error due to the approximation of the local estimators η_j. Let us finally introduce more precisely the higher-order quantity Λ appearing in (8.6), i.e.,

$$\Lambda^2 \coloneqq h^\zeta \sum_{j=1}^{N} \int_{K_j} (f(x) + \ell_j'(x))^2 \,\mathrm{d}x, \tag{8.16}$$

where $\zeta \in (0, 1)$ is arbitrary and ℓ_j are the linear functions employed in Definition 8.17. We can now state the main result concerning the estimator $\mathcal{E}_h$, which summarizes [19, Theorems 8.1 and 8.2].

Theorem 8.18. *Let $\mathcal{E}_h$ and Λ be defined in Definition 8.17 and* (8.16)*, respectively. Then, it holds up to higher order terms in h*

$$\|u - u_h\|_V \leq C\left(\mathcal{E}_h^2 + \Lambda^2\right)^{1/2},$$

for a constant C independent of h and of the solution u. If moreover the family of meshes $\mathcal{T}_h$ is λ-quasi-uniform and if $\kappa \in \mathcal{C}^2(D)$ and $f \in \mathcal{C}^1(D)$ then, up to higher order terms, it holds

$$C_{\text{low}}\mathcal{E}_h \leq \|u - u_h\|_V \leq C_{\text{up}}\mathcal{E}_h,$$

for constants C_{low}, C_{up} independent of h and u.

We recall that in the one dimensional case the probabliistic error estimators for the RM-FEM are given by

$$\widetilde{\mathcal{E}}_{h,1} \coloneqq \left(\sum_{i=1}^{N} \widetilde{\eta}^2_{K_i,1}\right)^{1/2}, \quad \text{with} \quad \widetilde{\eta}^2_{K_i,1} = h_i^{-(p-1)}\,\mathbb{E}\left[\left\|u_h' - (\widetilde{\mathcal{I}}u_h)'\right\|^2_{L^2(\widetilde{K}_i)}\right].$$
$$\widetilde{\mathcal{E}}_{h,2} \coloneqq \left(\sum_{i=1}^{N} \widetilde{\eta}^2_{K_i,2}\right)^{1/2}, \quad \text{with} \quad \widetilde{\eta}^2_{K_i,2} = h_i^{-(2p-3)}\,\mathbb{E}\left[\left\|u_h'\big|_K - (\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}}\right\|^2\right].$$

Our strategy to prove Theorem 8.15 relies on showing that the deterministic estimator $\mathcal{E}_h$ of Definition 8.17, as well as its probabilistic counterparts $\widetilde{\mathcal{E}}_{h,1}$ and $\widetilde{\mathcal{E}}_{h,2}$ above are all equivalent to the quantity

$$\mathcal{J}(u_h) \coloneqq \sum_{i=1}^{N-1} \bar{h}_i \left[\!\left[u_h'\right]\!\right]^2,$$

i.e., the sum of all squared jumps of the derivatives on the internal nodes. We first prove the equivalence for $\widetilde{\mathcal{E}}_{h,1}$.

Lemma 8.19. *Let Assumption 8.1 hold. Then, if the mesh is λ-quasi-uniform it holds*

$$\left(\frac{\mathbb{E}\,|\bar{\alpha}_1|\left(1+\lambda^{-(p-1)}\right)}{2} - 2h^{p-1}\,\mathbb{E}\,|\bar{\alpha}_1|^2\right)\mathcal{J}^2(u_h) \leq \widetilde{\mathcal{E}}^2_{h,1} \leq \frac{\mathbb{E}\,|\bar{\alpha}_1|\left(1+\lambda^{p-1}\right)}{2}\mathcal{J}^2(u_h),$$

where $\widetilde{\mathcal{E}}_{h,1}$ is given in Definition 8.10.

Proof. Let $\widetilde{K}_i$, $i = 1, \ldots, N$, be a generic element of the perturbed mesh and let us compute the derivative of the interpolant on $\widetilde{K}_i$, which is given by

$$(\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i} = \frac{u_h(\widetilde{x}_i) - u_h(\widetilde{x}_{i-1})}{\widetilde{x}_i - \widetilde{x}_{i-1}},$$

where an exact Taylor expansion allows to compute

$$u_h(\widetilde{x}_{i-1}) = u_h(x_{i-1}) + h^p \alpha_{i-1} u_h'(\widetilde{x}_{i-1}).$$

Hence, it holds

$$(\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i} = \frac{x_i - x_{i-1}}{\widetilde{x}_i - \widetilde{x}_{i-1}} u_h'\big|_{K_i} + h^p \frac{\alpha_i u_h'(\widetilde{x}_i) - \alpha_{i-1} u_h'(\widetilde{x}_{i-1})}{\widetilde{x}_i - \widetilde{x}_{i-1}},$$

which we can rewrite rearranging terms as

$$(\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i} - u_h'\big|_{K_i} = h^p \frac{\alpha_i\left(u_h'(\widetilde{x}_i) - u_h'\big|_{K_i}\right) + \alpha_{i-1}\left(u_h'\big|_{K_i} - u_h'(\widetilde{x}_{i-1})\right)}{\widetilde{x}_i - \widetilde{x}_{i-1}}. \tag{8.17}$$

It is clear then that the expression above depends on the signs of the variables α_{i-1} and α_i. For simplicity of notation, we therefore introduce the events $A_{i,j}^{(s_i,s_j)} \in \mathcal{A}$, where $s_i, s_j \in \{+,-\}$, defined as

$$A_{i,j}^{(s_i,s_j)} \coloneqq \{\omega \in \Omega \colon \alpha_i(\omega) \in \mathbb{R}^{s_i}, \alpha_j(\omega) \in \mathbb{R}^{s_j}\}.$$

We now define $e_h \coloneqq u_h - \widetilde{\mathcal{I}} u_h$ and write for any $i = 1, \dots, N$

$$\mathbb{E}\, \|e_h'\|_{L^2(\widetilde{K}_i)}^2 = \mathbb{E}\left(I_{i-1,i} + I_i + I_{i+1,i}\right),$$

with

$$I_{i,j} \coloneqq \int_{K_i \cap \widetilde{K}_j} (e_h')^2 \,\mathrm{d}x,$$

and where we write $I_i \coloneqq I_{i,i}$ and adopt the convention $I_{0,1} = I_{N+1,N} = 0$. In what follows we study $\mathbb{E}\, I_{i,j}$. We first consider I_i, which we express by the law of total expectation as

$$\mathbb{E}\, I_i = \sum_{s_{i-1},s_i \in \{-,+\}} \mathbb{E}\left[I_i \mid A_{i-1,i}^{(s_{i-1},s_i)}\right] P(A_{i-1,i}^{(s_{i-1},s_i)}). \tag{8.18}$$

In the trivial case $\alpha_{i-1} > 0$ and $\alpha_i < 0$, i.e., if $A_{i-1,i}^{(+,-)}$ occurs, we have $\widetilde{K}_i \cap K_i = \widetilde{K}_i$ and therefore $\mathbb{E}[I_i \mid A_{i-1,i}^{(+,-)}] = 0$. If $A_{i-1,i}^{(-,-)}$ occurs, the equality (8.17) simplifies to

$$(\widetilde{\mathcal{I}} u_h)'\big|_{\widetilde{K}_i} - u_h'\big|_{K_i} = -\frac{h^p \alpha_{i-1}}{\widetilde{x}_i - \widetilde{x}_{i-1}} \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}.$$

Since in this case $\left|K_i \cap \widetilde{K}_i\right| = \widetilde{x}_i - x_{i-1}$, integrating yields

$$I_i = \frac{h^{2p}(\widetilde{x}_i - x_{i-1})}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \alpha_{i-1}^2 \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2. \tag{8.19}$$

Similar calculations allow to show that if $A_{i-1,i}^{(+,+)}$ occurs, it holds

$$I_i = \frac{h^{2p}(x_i - \widetilde{x}_{i-1})}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \alpha_i^2 \left[\!\left[u_h'\right]\!\right]_{x_i}^2, \tag{8.20}$$

Finally, if $A_{i-1,i}^{(-,+)}$ occurs, we get

$$I_i = \frac{h^{2p}(x_i - x_{i-1})}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \xi_i^2. \tag{8.21}$$

where we denote

$$\xi_i \coloneqq \alpha_{i-1} \left[\!\left[u_h'\right]\!\right]_{x_{i-1}} + \alpha_i \left[\!\left[u_h'\right]\!\right]_{x_i}. \tag{8.22}$$

We thus have an expression for $\mathbb{E}\, I_i$ due to (8.18). We now turn to $I_{i-1,i}$. Since $\widetilde{K}_i \cap K_{i-1} = \emptyset$ if $\alpha_{i-1} > 0$, we have by the law of total expectation

$$\mathbb{E}\, I_{i-1,i} = \mathbb{E}\left[I_{i-1,i} \mid A_{i-1,i}^{(-,-)}\right] P(A_{i-1,i}^{(-,-)}) + \mathbb{E}\left[I_{i-1,i} \mid A_{i-1,i}^{(-,+)}\right] P(A_{i-1,i}^{(-,+)}). \tag{8.23}$$

Let us remark that adding and subtracting $u_h'\big|_{K_i}$ yields

$$(\widetilde{\mathcal{I}} u_h)'\big|_{\widetilde{K}_i} - u_h'\big|_{K_{i-1}} = (\widetilde{\mathcal{I}} u_h)'\big|_{\widetilde{K}_i} - u_h'\big|_{K_i} - \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}.$$

The same computations employed for I_i allow to conclude that

$$I_{i-1,i} = \begin{cases} -h^p \alpha_{i-1} \left(\dfrac{h^p \alpha_{i-1}}{\widetilde{x}_i - \widetilde{x}_{i-1}} + 1\right)^2 \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2, & \text{if } A_{i-1,i}^{(-,-)} \text{ occurs,} \\ -h^p \alpha_{i-1} \left(\dfrac{h^p}{\widetilde{x}_i - \widetilde{x}_{i-1}} \xi_i + \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}\right)^2, & \text{if } A_{i-1,i}^{(-,+)} \text{ occurs.} \end{cases}$$

which, replaced into (8.23) gives the final expression for $\mathbb{E}\, I_{i-1,i}$. Similarly, for $I_{i+1,i}$ we have

$$\mathbb{E}\, I_{i+1,i} = \mathbb{E}\left[I_{i+1,i} \mid A_{i-1,i}^{(+,+)}\right] P(A_{i-1,i}^{(+,+)}) + \mathbb{E}\left[I_{i+1,i} \mid A_{i-1,i}^{(-,+)}\right] P(A_{i-1,i}^{(-,+)}),$$

where

$$I_{i+1,i} = \begin{cases} h^p \alpha_i \left(\dfrac{h^p \alpha_i}{\widetilde{x}_i - \widetilde{x}_{i-1}} - 1\right)^2 [\![u_h']\!]_{x_i}^2, & \text{if } A_{i-1,i}^{(+,+)} \text{ occurs,} \\ h^p \alpha_i \left(\dfrac{h^p}{\widetilde{x}_i - \widetilde{x}_{i-1}} \xi_i - [\![u_h']\!]_{x_i}\right)^2, & \text{if } A_{i-1,i}^{(-,+)} \text{ occurs.} \end{cases}$$

We now reassemble the quantity $I_i + I_{i-1,i} + I_{i+1,i}$ by grouping terms with regards to their conditioning on the sign of (α_{i-1}, α_i). In particular, some algebraic simplifications yield

$$I_i + I_{i-1,i} + I_{i+1,i} = \begin{cases} h^p \alpha_i [\![u_h']\!]_{x_i}^2 - \dfrac{h^{2p} \alpha_i^2}{\widetilde{x}_i - \widetilde{x}_{i-1}} [\![u_h']\!]_{x_i}^2, & \text{if } A_{i-1,i}^{(+,+)} \text{ occurs,} \\ -h^p \alpha_{i-1} [\![u_h']\!]_{x_{i-1}}^2 - \dfrac{h^{2p} \alpha_{i-1}^2}{\widetilde{x}_i - \widetilde{x}_{i-1}} [\![u_h']\!]_{x_{i-1}}^2, & \text{if } A_{i-1,i}^{(-,-)} \text{ occurs,} \\ h^p \alpha_i [\![u_h']\!]_{x_i}^2 - h^p \alpha_{i-1} [\![u_h']\!]_{x_{i-1}}^2 - \dfrac{h^{2p}}{\widetilde{x}_i - \widetilde{x}_{i-1}} \xi_i^2, & \text{if } A_{i-1,i}^{(-,+)} \text{ occurs.} \end{cases}$$

We now can compute the estimator $\widetilde{\mathcal{E}}_{h,1}$ by summing its local contributions, as in

$$\begin{aligned} \widetilde{\mathcal{E}}_{h,1}^2 &= \sum_{i=1}^N \eta_{K,1}^2 = \sum_{i=1}^N h_i^{-(p-1)} \|e_h'\|_{L^2(\widetilde{K}_i)}^2 \\ &= \sum_{i=1}^N h_i^{-(p-1)} \mathbb{E}(I_i + I_{i-1,i} + I_{i+1,i}) =: J_1 + J_2, \end{aligned}$$

where J_1 and J_2 are given by

$$J_1 := \frac{h^p}{2} \sum_{i=1}^N h_i^{-(p-1)} \left([\![u_h']\!]_{x_i}^2 \mathbb{E}\left[\alpha_i \mid \alpha_i > 0\right] - [\![u_h']\!]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1} \mid \alpha_i < 0\right]\right),$$

$$\begin{aligned} J_2 := -\frac{h^{2p}}{4} \sum_{i=1}^N h_i^{-(p-1)} \Bigg(&[\![u_h']\!]_{x_{i-1}}^2 \mathbb{E}\left[\frac{\alpha_{i-1}^2}{\widetilde{x}_i - \widetilde{x}_{i-1}} \mid A_{i-1,i}^{(-,-)}\right] + [\![u_h']\!]_{x_i}^2 \mathbb{E}\left[\frac{\alpha_i^2}{\widetilde{x}_i - \widetilde{x}_{i-1}} \mid A_{i-1,i}^{(+,+)}\right] \\ &+ \mathbb{E}\left[\frac{\xi_i^2}{\widetilde{x}_i - \widetilde{x}_{i-1}} \mid A_{i-1,i}^{(-,+)}\right]\Bigg). \end{aligned}$$

Let us consider J_1 and J_2 separately. Rearranging the sum, noticing that under Assumption 8.1(i) it holds $\mathbb{E}\left[\alpha_i \mid \alpha_i > 0\right] = -\mathbb{E}\left[\alpha_i \mid \alpha_i < 0\right] = \mathbb{E}\,|\alpha_i|$ and recalling that $\alpha_i = (\bar{h}_i h^{-1})^p \bar{\alpha}_i$, we obtain

$$\begin{aligned} J_1 &= \frac{h^p}{2} \sum_{i=1}^{N-1} \left(h_i^{-(p-1)} + h_{i+1}^{-(p-1)}\right) [\![u_h']\!]_{x_i}^2 \mathbb{E}\left[\alpha_i \mid \alpha_i > 0\right] \\ &= \frac{\mathbb{E}\,|\bar{\alpha}_1|}{2} \sum_{i=1}^{N-1} \left(h_i^{-(p-1)} + h_{i+1}^{-(p-1)}\right) \bar{h}_i^p [\![u_h']\!]_{x_i}^2. \end{aligned}$$

Now, let us remark that if the mesh is λ-quasi-uniform, it holds

$$\left(1 + \lambda^{-(p-1)}\right) \bar{h}_i \le \left(h_i^{-(p-1)} + h_{i+1}^{-(p-1)}\right) \bar{h}_i^p \le \left(1 + \lambda^{p-1}\right) \bar{h}_i,$$

which implies

$$\frac{\mathbb{E}\,|\bar{\alpha}_1| \left(1 + \lambda^{-(p-1)}\right)}{2} \mathcal{J}^2(u_h) \le J_1 \le \frac{\mathbb{E}\,|\bar{\alpha}_1| \left(1 + \lambda^{p-1}\right)}{2} \mathcal{J}^2(u_h). \tag{8.24}$$

We now turn to J_2. Clearly, we have $J_2 \leq 0$, which implies the desired upper bound together with (8.24). For the lower bound, we remark that in both cases $A_{i-1,i}^{(+,+)}$ and $A_{i-1,i}^{(-,-)}$ occur, we have that $\widetilde{x}_i - \widetilde{x}_{i-1} \geq h_i/2$, and if $A_{i-1,i}^{(-,+)}$ occurs, we have $\widetilde{x}_i - \widetilde{x}_{i-1} \geq h_i$. Hence, simplifying the conditioning in the first and second terms, we obtain

$$J_2 \geq -\frac{h^{2p}}{4} \sum_{i=1}^{N} h_i^{-p} \Big(2 \left[\!\left[u_h' \right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0 \right] + 2 \left[\!\left[u_h' \right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0 \right] + \mathbb{E}\left[\xi_i^2 \mid A_{i-1,i}^{(-,+)} \right] \Big).$$

We now consider ξ_i given in (8.22) and use $(a+b)^2 \leq 2(a^2+b^2)$ for $a = \alpha_{i-1} \left[\!\left[u_h' \right]\!\right]_{x_{i-1}}$ and $b = \alpha_i \left[\!\left[u_h' \right]\!\right]_{x_i}$ to obtain

$$\mathbb{E}\left[\xi_i^2 \mid A_{i-1,i}^{(-,+)} \right] \leq 2 \left[\!\left[u_h' \right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0 \right] + 2 \left[\!\left[u_h' \right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0 \right].$$

Therefore

$$J_2 \geq -h^{2p} \sum_{i=1}^{N} h_i^{-p} \Big(\left[\!\left[u_h' \right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0 \right] + \left[\!\left[u_h' \right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0 \right] \Big).$$

Rewriting the sum and replacing the definition of α_i yields

$$J_2 \geq -\sum_{i=1}^{N-1} \bar{h}_i^{2p} \left[\!\left[u_h' \right]\!\right]_{x_i}^2 \left(h_i^{-p} \mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i > 0 \right] + h_{i+1}^{-p} \mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i < 0 \right] \right).$$

Now $\bar{h}_i = \min\{h_i, h_{i+1}\}$ implies $h_i^{-p} \leq \bar{h}_i^{-p}$ and $h_{i+1}^{-p} \leq \bar{h}_i^{-p}$, which gives

$$\begin{aligned} J_2 &\geq -2 \sum_{i=1}^{N-1} \bar{h}_i^{p} \left[\!\left[u_h' \right]\!\right]_{x_i}^2 \left(\frac{1}{2} \mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i > 0 \right] + \frac{1}{2} \mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i < 0 \right] \right) \\ &\geq -2 \, \mathbb{E}\left| \bar{\alpha}_1 \right|^2 \sum_{i=1}^{N-1} \bar{h}_i^{p} \left[\!\left[u_h' \right]\!\right]_{x_i}^2, \end{aligned}$$

where we applied the law of total expectation on the second line. Finally, we have $\bar{h}_i \leq h$ and $p \geq 1$, which yield

$$J_2 \geq -2 \, \mathbb{E}\left| \bar{\alpha}_1 \right|^2 h^{p-1} \mathcal{J}^2(u_h).$$

Combining this with (8.24) then yields the desired lower bound and thus concludes the proof. □

Let us remark that the coefficient appearing in the lower bound of Lemma 8.19 is positive if Assumption 8.13 holds. We now prove the equivalence of the estimator $\widetilde{\mathcal{E}}_{h,2}$ given in Definition 8.10 with $\mathcal{J}(u_h)$.

Lemma 8.20. *Let Assumption 8.1 hold and let the mesh $\mathcal{T}_h$ be λ-quasi uniform. Then, it holds*

$$\frac{\mathbb{E}\left| \bar{\alpha}_1 \right|^2}{2(1+\lambda)^2 \lambda^{2p-1}} \mathcal{J}^2(u_h) \leq \widetilde{\mathcal{E}}_{h,2}^2 \leq 3 \, \mathbb{E}\left| \bar{\alpha}_1 \right|^2 \mathcal{J}^2(u_h),$$

where $\widetilde{\mathcal{E}}_{h,2}$ is given in Definition 8.10.

Proof. As $|K_i| = h_i$, we have

$$\widetilde{\mathcal{E}}_{h,2} = \sum_{i=1}^{N} h_i^{-(2p-3)} \, \mathbb{E}\left[\left| u_h'\big|_{K_i} - (\widetilde{\mathcal{I}} u_h)'\big|_{\widetilde{K}_i} \right|^2 \right].$$

Proceeding similarly to (8.19), (8.20) and (8.21) and applying the law of total expectation, we obtain

$$\begin{aligned}\mathbb{E}\left[\left|u_h'\big|_{K_i} - (\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i}\right|^2\right] &= \frac{h^{2p}}{4}\left[\!\left[u_h'\right]\!\right]_{x_i}^2 \mathbb{E}\left[\frac{\alpha_i^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(+,+)}\right] \\ &\quad + \frac{h^{2p}}{4}\left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\frac{\alpha_{i-1}^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,-)}\right] \\ &\quad + \frac{h^{2p}}{4}\mathbb{E}\left[\frac{\xi_i^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,+)}\right],\end{aligned} \tag{8.25}$$

where we recall the notation ξ_i introduced in (8.22). Let us first consider the lower bound. Since $\xi_i^2 \geq 0$ a.s., and $\widetilde{x}_i - \widetilde{x}_{i-1} \leq (1+\lambda)h_i$ a.s. under the assumption that the mesh is λ-quasi-uniform, we have

$$\begin{aligned}\mathbb{E}\left[\left|u_h'\big|_{K_i} - (\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i}\right|^2\right] &\geq \frac{h^{2p}h_i^{-2}}{4(1+\lambda)^2}\Big(\left[\!\left[u_h'\right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0\right] \\ &\quad + \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0\right]\Big).\end{aligned}$$

Assembling the sum, rearranging terms and recalling that $\alpha_i = (h^{-1}\bar{h}_i)^p\bar{\alpha}_i$ with $\bar{h}_i = \min\{h_i, h_{i+1}\}$, we then obtain

$$\begin{aligned}\widetilde{\mathcal{E}}_{h,2}^2 &\geq \frac{1}{2(1+\lambda)^2}\sum_{i=1}^{N-1}\bar{h}_i^{2p}h_i^{1-2p}\left[\!\left[u_h'\right]\!\right]_{x_i}^2\left(\frac{1}{2}\mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i > 0\right] + \frac{1}{2}\mathbb{E}\left[\bar{\alpha}_i^2 \mid \bar{\alpha}_i < 0\right]\right) \\ &\geq \frac{\mathbb{E}\left|\bar{\alpha}_1\right|^2}{2(1+\lambda)^2\lambda^{2p-1}}\sum_{i=1}^{N-1}\bar{h}_i\left[\!\left[u_h'\right]\!\right]_{x_i}^2 = \frac{\mathbb{E}\left|\bar{\alpha}_1\right|^2}{2(1+\lambda)^2\lambda^{2p-1}}\mathcal{J}^2(u_h),\end{aligned}$$

where we employed the law of total expectation and the inequality $h_i^{1-2p} \leq \lambda^{1-2p}\bar{h}_i^{1-2p}$ on the second line. Hence, we proved the lower bound. For the upper bound, using again the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we obtain

$$\xi_i^2 \leq 2\alpha_i^2\left[\!\left[u_h'\right]\!\right]_{x_i}^2 + 2\alpha_{i-1}^2\left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2, \quad \text{a.s,}$$

so that

$$\begin{aligned}\mathbb{E}\left[\frac{\xi_i^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,+)}\right] &\leq 2\left[\!\left[u_h'\right]\!\right]_{x_i}^2 \mathbb{E}\left[\frac{\alpha_i^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,+)}\right] \\ &\quad + 2\left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\frac{\alpha_{i-1}^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,+)}\right].\end{aligned}$$

Under $A_{i-1,i}^{(-,+)}$, we have $\widetilde{x}_i - \widetilde{x}_{i-1} \geq h_i$, which implies

$$\mathbb{E}\left[\frac{\xi_i^2}{(\widetilde{x}_i - \widetilde{x}_{i-1})^2} \mid A_{i-1,i}^{(-,+)}\right] \leq 2h_i^{-2}\left(\left[\!\left[u_h'\right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0\right] + \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0\right]\right).$$

Then, considering that under $A_{i-1,i}^{(+,+)}$ or $A_{i-1,i}^{(-,-)}$ it holds $\widetilde{x}_i - \widetilde{x}_{i-1} \geq h_i/2$ and plugging into (8.25) we have

$$\mathbb{E}\left[\left|u_h'\big|_{K_i} - (\widetilde{\mathcal{I}}u_h)'\big|_{\widetilde{K}_i}\right|^2\right] \leq \frac{3}{2}h_i^{-2}h^{2p}\left(\left[\!\left[u_h'\right]\!\right]_{x_i}^2 \mathbb{E}\left[\alpha_i^2 \mid \alpha_i > 0\right] + \left[\!\left[u_h'\right]\!\right]_{x_{i-1}}^2 \mathbb{E}\left[\alpha_{i-1}^2 \mid \alpha_{i-1} < 0\right]\right).$$

We can therefore reassemble and rearrange the sum following the same procedure as for the lower bound, which, together with $h_i^{1-2p} \leq \bar{h}_i^{1-2p}$, yields

$$\widetilde{\mathcal{E}}_{h,2}^2 \leq 3\,\mathbb{E}\left|\bar{\alpha}_1\right|^2\mathcal{J}^2(u_h),$$

which proves the desired result. □

We finally prove the equivalence of the deterministic error estimator $\mathcal{E}_h$ given in Definition 8.17 with the quantity $\mathcal{J}(u_h)$.

Lemma 8.21. *Let the mesh $\mathcal{T}_h$ be λ-quasi-uniform. Then, it holds*

$$\frac{\lambda m^2}{6(1+\lambda)^3 M^2}\mathcal{J}^2(u_h) \leq \mathcal{E}_h^2 \leq \frac{2\lambda^2 M^2}{3(1+\lambda)m^2}\mathcal{J}^2(u_h),$$

where $\mathcal{E}_h$ is given in Definition 8.17 and where $m = \underline{\kappa}$ and $M = \|\kappa\|_{L^\infty(D)}$.

Proof. Simple algebraic computations yield

$$\|\ell_j\|^2_{L^2(K_j)} = \frac{h_j}{3}\left(\tau_{j,0}^2 - \tau_{j,0}\tau_{j,1} + \tau_{j,1}^2\right),$$

where ℓ_j are the linear functions employed in Definition 8.17. Applying the inequalities $(a^2 + b^2)/2 \leq a^2 - ab + b^2 \leq 2(a^2 + b^2)$ we obtain

$$\frac{h_j}{6M^2}\left(\tau_{j,0}^2 + \tau_{j,1}^2\right) \leq \eta_j^2 \leq \frac{2h_j}{3m^2}\left(\tau_{j,0}^2 + \tau_{j,1}^2\right).$$

We now remark that if the mesh $\mathcal{T}_h$ is λ-quasi-uniform and under the assumptions on κ it holds for $k \in \{0, 1\}$

$$\frac{m^2}{(1+\lambda)^2}\left[\!\left[u_h'\right]\!\right]^2_{x_{j-k}} \leq \tau_{j,k}^2 \leq \frac{\lambda^2 M^2}{(1+\lambda)^2}\left[\!\left[u_h'\right]\!\right]^2_{x_{j-k}},$$

which, in turn, implies

$$\frac{m^2 h_j}{6(1+\lambda)^2 M^2}\left(\left[\!\left[u_h'\right]\!\right]^2_{x_{j-1}} + \left[\!\left[u_h'\right]\!\right]^2_{x_j}\right) \leq \eta_j^2 \leq \frac{2\lambda^2 M^2 h_j}{3(1+\lambda)^2 m^2}\left(\left[\!\left[u_h'\right]\!\right]^2_{x_{j-1}} + \left[\!\left[u_h'\right]\!\right]^2_{x_j}\right).$$

We now focus on the upper bound. Reassembling the global error estimator $\mathcal{E}_h$, we have

$$\begin{aligned}
\mathcal{E}_h^2 &\leq \frac{2\lambda^2 M^2}{3(1+\lambda)^2 m^2}\sum_{j=1}^{N} h_j\left(\left[\!\left[u_h'\right]\!\right]^2_{x_{j-1}} + \left[\!\left[u_h'\right]\!\right]^2_{x_j}\right) \\
&= \frac{2\lambda^2 M^2}{3(1+\lambda)^2 m^2}\sum_{j=1}^{N-1}(h_j + h_{j+1})\left[\!\left[u_h'\right]\!\right]^2_{x_j} \\
&\leq \frac{2\lambda^2 M^2}{3(1+\lambda)m^2}\mathcal{J}^2(u_h),
\end{aligned}$$

where we recall $\bar{h}_j = \min\{h_j, h_{j+1}\}$, so that $h_j + h_{j+1} \leq (1+\lambda)\bar{h}_j$. We conclude the proof proceeding similarly for the lower bound as in Lemma 8.20. □

We can finally prove Theorem 8.15 and conclude the error analysis.

Proof of Theorem 8.15. Let us first consider $\widetilde{\mathcal{E}}_{h,1}$. Under Assumption 8.13, we have for the lower bound of Lemma 8.19

$$\left(\frac{\mathbb{E}\,|\bar{\alpha}_1|\left(1+\lambda^{-(p-1)}\right)}{2} - 2h^{p-1}\,\mathbb{E}\,|\bar{\alpha}_1|^2\right)\mathcal{J}^2(u_h) \geq C\,\mathbb{E}\,|\bar{\alpha}_1|\,\mathcal{J}^2(u_h)$$

for a constant $C > 0$. Hence, due to Lemma 8.21 we have that there exists a constant $\widehat{C}$ such that

$$\widetilde{\mathcal{E}}_{h,1} \geq \widehat{C}\mathcal{E}_h,$$

and therefore, Theorem 8.18 implies

$$\|u - u_h\|_V \leq C_{\mathrm{up}} \mathcal{E}_h \leq C_{\mathrm{up}} \widehat{C} \widetilde{\mathcal{E}}_{h,1},$$

which yields the desired upper bound with $\widetilde{C}_{\mathrm{up}} = \widehat{C} C_{\mathrm{up}}$. The lower bound follows equivalently under the additional regularity required by Theorem 8.18. Similarly, the results for $\widetilde{\mathcal{E}}_{h,2}$ follows from Lemmas 8.19 and 8.21, together with Theorem 8.18. □

9 Conclusion of Part II

In this second part of the thesis we introduced two probabilistic numerical methods for differential equations, namely the random time step Runge–Kutta method (RTS-RK) for ODEs and the random mesh finite element method (RM-FEM) for elliptic PDEs.

In Chapter 6 we introduced a general framework for the field of probabilistic numerics (PN). In particular, we divided the field of PN in two macro-areas: the perturbation-based and the measure-valued numerical methods. Moreover, we defined two notions of convergence for probabilistic methods and we showed a contraction result for Monte Carlo estimators drawn from perturbation-based schemes. We furthermore detailed how PN and the Bayesian approach can be combined to enhance the quality of the solution of inverse problems in terms of uncertainty quantification. Finally, we demonstrated by an example involving ODEs the differences between perturbation-based and measure-valued probabilistic methods, and argued what are the advantages and disadvantages of each of these two classes.

In Chapter 7 we presented the RTS-RK, a novel perturbation-based probabilistic scheme for ODEs built on Runge–Kutta integrators and on a random selection of the time steps. After analyzing the weak and mean-square convergence properties of the scheme, we focused on its geometric properties. In this regard, we have shown that the RTS-RK preserves the geometric properties of the Runge–Kutta method it is built on, and in particular the conservation of first integrals and the approximation of Hamiltonians over long time intervals are guaranteed for the RTS-RK. We remark that the predecessor of the RTS-RK, the additive noise method of [39, 86], fails instead to represent the uncertainty of the underlying Runge–Kutta integrator on a large class of geometric ODEs. Finally, we applied the RTS-RK to Bayesian inference problems and we showed heuristically its advantageous properties in this context. The validity of our analysis is corroborated by an extensive series of numerical examples, which show the potential of the RTS-RK and its relevance within the field of PN.

In Chapter 8 we introduced a novel probabilistic methods for PDEs based on the FEM and random meshes, the RM-FEM. We demonstrated how our methodology can be successful when employed in pipelines of computations, such as Bayesian inverse problems. We also show a rigorous use of probabilistic methods for a posteriori error estimators, often speculated in the field. A series of numerical experiments in the one and two-dimensional case illustrate the potential of the RM-FEM for both inverse problems and mesh adaptation.

The area of PN has been enriched in the last decade with a relatively large amount of contributions from researches belonging to different areas: numerical analysis, optimization, statistics and eventually machine learning. Nevertheless, PN as it is meant nowadays is a relatively recent

field of research, and several interesting questions are still open. In particular, some future work that is associated to the content of this second part of the thesis and which we believe could be relevant is:

(i) Extending the result on Monte Carlo estimators of Theorem 6.7 to the deterministic approximations of the random measures arising when a PN method is applied to Bayesian inverse problems;

(ii) Introducing and analyzing probabilistic a posteriori error estimators for time-dependent problems, thus designing a PN-based routine for time step adaptation;

(iii) Generalizing the analysis of the RM-FEM to higher-dimensional elliptic PDEs;

(iv) Applying the mesh refinement strategy presented in Chapter 8 and based on the RM-FEM to nonlinear PDEs, for which a posteriori a error estimators can be proved to be reliable and efficient only in special cases;

(v) Combining the RM-FEM and the RTS-RK (or other perturbation-based probabilistic methods) in a random-space/random-time adaptive scheme for parabolic and/or hyperbolic PDEs.

A Probability Theory

We briefly cover here some topics of probability theory and stochastic calculus. Our goal is twofold. On the one hand, we set here our notation on probability theory, which is adopted throughout the thesis. On the other hand, we introduce for completeness a series of standard results which are repeatedly employed in different chapters of this thesis. Let us remark that for the sake of simplicity we oftentimes restrict the setting for results which could be presented and applied to a broader scope. The frameworks we consider are nevertheless sufficient for providing a theoretical basis to this thesis. In order to mitigate the lack of generalization, we provide throughout the chapter references to broader and deeper discussions about the different topics.

A.1 Weak Convergence

We introduce in this section the notion of weak convergence of probability measures and of random variables on metric spaces. For a wider discussion about the topic and its implications, we refer the reader to [23], [69, Chapter 18], [121, Appendix A], and [113, Chapter 3.5]

Definition A.1. Let $(\mathcal{H}, \lambda)$ be a metric space with the Borel σ-algebra $\mathcal{B}(\mathcal{H})$. Let $\{\mu_n\}_{n\geq 0}$ be a sequence of probability measures on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, and let μ be another measure on the same space. We say that μ_n converges weakly to μ, in symbols $\mu_n \Rightarrow \mu$, if for all continuous and bounded functions f defined on $\mathcal{H}$ it holds

$$\lim_{n\to\infty} \int_{\mathcal{H}} f(x)\mu_n(\mathrm{d}x) = \int_{\mathcal{H}} f(x)\mu(\mathrm{d}x).$$

Remark A.2. Weak convergence could be defined differently to Definition A.1 by employing a series of equivalent conditions which have to be satisfied by μ_n and μ. The equivalence between these criteria often goes under the name of Portmanteau theorem [23, Theorem 2.1].

Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X\colon \Omega \to \mathcal{H}$, we denote by μ_X the probability measure induced by X on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, i.e., for any $B \in \mathcal{B}(\mathcal{H})$

$$\mu_X(B) = P(\{\omega\colon X(\omega) \in B\}) = P(X^{-1}(B)).$$

We now give the definition of weak convergence of random variables.

Definition A.3. Let $\{(\Omega_n, \mathcal{F}_n, P_n)\}_{n\geq 0}$ be a sequence of probability spaces and let $\{X_n\colon \Omega_n \to \mathcal{H}\}_{n\geq 0}$ be a sequence of random variables. Moreover, let $(\Omega, \mathcal{F}, P)$ and $X\colon \Omega \to \mathcal{H}$ be another probability space and another random variable. We say that X_n converges weakly to X, in symbols $X_n \Rightarrow X$, if $\mu_{X_n} \Rightarrow \mu_X$.

Remark A.4. Let $X_n\colon \Omega \to \mathbb{R}$ and $X\colon \Omega \to \mathbb{R}$ be real-valued random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$. Moreover, denote by $F_{X_n}\colon \mathbb{R} \to \mathbb{R}$ (resp. F_X) the probability function

$$F_{X_n}(x) = P\left(X_n^{-1}(-\infty, x)\right),$$

and equivalently for X and F_X. Oftentimes, in this setting it is said that X_n converges weakly to X if $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$ where F_X is continuous. It is possible to show that this notion of weak convergence and the one of Definition A.3 are in fact equivalent for real-valued random variables.

The following result guarantees that weak convergence of random variables is preserved by continuous mappings.

Theorem A.5. *Let $\{(\Omega_n, \mathcal{F}_n, P_n)\}_{n\geq 0}$ be a sequence of probability spaces and let $\{X_n\colon \Omega_n \to \mathcal{H}_1\}_{n\geq 0}$ be a sequence of random variables with values on a metric space $(\mathcal{H}_1, \lambda_1)$. Moreover, let $(\Omega, \mathcal{F}, P)$ and $X\colon \Omega \to \mathcal{H}_1$ be another probability space and another random variable, such that $X_n \Rightarrow X$. Let $f\colon \mathcal{H}_1 \to \mathcal{H}_2$, where $(\mathcal{H}_2, \lambda_2)$ is a metric space, be a continuous function. Then, it holds $f(X_n) \Rightarrow f(X)$.*

We conclude by recalling that weak convergence is indeed the weakest convergence for random variables. In particular, convergence in probability for a sequence of random variables implies its weak convergence. The converse is not true, unless the limit of the sequence is a constant c. Only in this case, we have that $X_n \Rightarrow c$ implies $X_n \to c$ in probability.

A.2 The Radon–Nykodim Theorem

In this section we briefly present the Radon–Nykodim theorem, a change of measure formula which is a fundamental tool of probability theory. Complete discussions on this topic can be found in [69, Chapter 28], [88, Chapter 6] or in [75, Chapter 9]

Definition A.6. Let $(\Omega, \mathcal{A})$ be a measurable space and let P and Q be two probability measures defined on $(\Omega, \mathcal{A})$. We say that P is absolutely continuous with respect to Q, in symbols $P \ll Q$, if $P(A) = 0$ for all $A \in \mathcal{A}$ such that $Q(A) = 0$. If furthermore $Q \ll P$ we say that the probability measures P and Q are equivalent, and write $P \sim Q$.

We now introduce the Radon–Nykodim theorem.

Theorem A.7. *Let $(\Omega, \mathcal{A})$ be a measurable space and let P and Q be two probability measures on $(\Omega, \mathcal{A})$, such that $P \ll Q$. Then, there exists a $\mathcal{A}$-measurable non-negative random variable X such that*

$$P(A) = \int_A X \, \mathrm{d}Q,$$

for all $A \in \mathcal{A}$. Furthermore, X is unique Q-a.s.

We call X the Radon–Nykodim derivative of P with respect to Q and write

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) = X(\omega),$$

omitting the argument ω for economy of notation when it does not compromise clarity. The following result gives the conditions on the Radon–Nykodim derivative so that it is invertible.

Theorem A.8. *With the notation of Theorem A.7, if $X > 0$ Q-a.s., then $P \sim Q$ and*

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) = X^{-1}(\omega),$$

for all $\omega \in \Omega$.

A.3 Connections between SDEs and PDEs

In this section, we consider the connections between SDEs and some specific parabolic PDEs. In particular, we discuss the backward Kolmogorov equation (BKE) and the Fokker–Planck equation (FPE). For a broader discussion on the topic we refer the reader, e.g., to [110]. Let $F\colon \mathbb{R}^d \to \mathbb{R}^d$ and $G\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$ and let $X = (X_t, 0 \le t \le T)$ be the $\mathbb{R}^d$-valued stochastic process solution to the autonomous Itô SDE

$$\mathrm{d}X_t = F(X_t)\,\mathrm{d}t + G(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \tag{A.1}$$

where W_t is a m-dimensional Brownian motion. We let F and G satisfy the usual assumptions under which (A.1) admits a unique strong solution. We call generator of the SDE (A.1) the differential operator $\mathcal{L}$ which acts on any function $u \in \mathcal{C}_b^0(\mathbb{R}^d)$, the space of continuous and bounded functions, as

$$\mathcal{L}u = F \cdot \nabla u + \frac{1}{2} G G^\top : \nabla^2 u.$$

where we denote by $\nabla^2 u$ the Hessian of u and by : the Frobenius inner product for matrices, i.e., for two square matrices $A, B \in \mathbb{R}^{d\times d}$ we have $A : B = \mathrm{tr}(A^\top B)$. Heuristically, the generator quantifies the infinitesimal average rate of change of the function u computed on the solution of (A.1). In particular, denoting by $\mathcal{P}_t$ the family of operators on $\mathcal{C}_b^0(\mathbb{R}^d)$, indexed by $t \ge 0$ and such that

$$(\mathcal{P}_t u)(x) = \mathbb{E}[u(X_t) \mid X_0 = x],$$

we have that $\mathcal{L}$ is defined with the limit

$$\mathcal{L}u \coloneqq \lim_{t\to 0} \frac{\mathcal{P}_t u - u}{t} \tag{A.2}$$

provided it exists. Let us remark that the operator $\mathcal{P}_t$ forms a semigroup with respect to t, which is often called the Markov semigroup associated to (A.1).

The first parabolic PDE which we introduce is the BKE. Let $T > 0$ and u be the solution to the final-value PDE

$$\begin{aligned} -\partial_t u(t,x) &= \mathcal{L}u(t,x), \quad x \in \mathbb{R}^d, 0 \le t < T, \\ u(T,x) &= \varphi(x), \qquad\quad x \in \mathbb{R}^d. \end{aligned} \tag{A.3}$$

for any function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ which is sufficiently smooth for (A.3) to be well-posed. Then the following result, which is referred to in literature as the Kolmogorov representation formula, holds.

Theorem A.9. *Let X be the solution of* (A.1) *and let $\mathcal{L}$ be the generator* (A.2)*. Moreover, let φ, F and G be smooth enough so that equation* (A.3) *is well-posed. Then, the solution u of the BKE* (A.3) *satisfies*

$$\mathbb{E}\left[\varphi(X_t) \mid X_T = x\right] = u(t,x),$$

where $\mathbb{E}$ denotes expectation with respect to the Brownian motion W.

The proof of Theorem A.9 is given in [110, Chapter 2]. In words, solving the BKE allows to compute the average of functionals of the solution of the SDE without simulating the SDE itself. Since we assumed the equation (A.1) to be autonomous, i.e., the functions F and G do not have explicit dependence on the time variable, the BKE can be rewritten after a time inversion as an initial-value PDE which reads

$$\begin{aligned} \partial_t u(t,x) &= \mathcal{L}u(t,x), \quad x \in \mathbb{R}^d, 0 \leq t < T, \\ u(0,x) &= \varphi(x), \qquad x \in \mathbb{R}^d. \end{aligned}$$

Similarly to Theorem A.9, the Kolmogorov representation for this form of the BKE reads

$$\mathbb{E}\left[\varphi(X_t) \mid X_0 = x\right] = u(t,x).$$

We now present the FPE, which describes the time evolution of the probability density of the solution of (A.1). Let us denote by $\mathcal{L}^*$ the L^2-adjoint of $\mathcal{L}$, which acts on a sufficiently function $\rho\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ as

$$\mathcal{L}^*\rho = -\nabla \cdot (F\rho) + \frac{1}{2} GG^\top : \nabla^2 \rho. \tag{A.4}$$

The FPE is the initial-value parabolic PDE

$$\begin{aligned} \partial_t \rho(t,x) &= \mathcal{L}^*\rho(t,x), \quad x \in \mathbb{R}^d, t > 0, \\ \rho(0,x) &= \rho_0(x), \qquad x \in \mathbb{R}^d, \end{aligned} \tag{A.5}$$

for an initial condition ρ_0 such that $\rho_0(x) \geq 0$ for all $x \in \mathbb{R}^d$ and such that in the weak sense

$$\int_{\mathbb{R}^d} \rho_0(x)\,\mathrm{d}x = 1,$$

i.e., for ρ_0 a probability density function on $\mathbb{R}^d$.

The following result, due to Kolmogorov, provides the connection between the density of the solution of its associated SDE and the FPE.

Theorem A.10. *Let the initial condition x_0 of* (A.1) *admit a density ρ_0 with respect to the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Moreover, let ρ_0, F and G be smooth enough so that equation* (A.5) *is well-posed. Let μ_t denote the measure on $\mathbb{R}^d$ induced by the solution X of* (A.1) *at time t, and assume that μ_t admits a density $\rho_X(t,y)$ with respect to the Lebesgue measure. Then, the function ρ_X is the unique solution of the FPE* (A.5).

The proof of Theorem A.10 can be found e.g. in [110, Chapter 4], where boundary conditions are treated and several properties of the FPE are analyzed in details.

A.4 Ergodic Processes

In this section we consider a class of ergodic stochastic processes. For the sake of simplicity, we restrict ourselves to diffusion processes, i.e., stochastic processes which can be written as the solutions of SDEs of the form (A.1). For a complete discussion on ergodic processes, we refer the reader to [22, 51, 110, 112].

Heuristically, we say that a stochastic process $X = (X_t, t \geq 0)$ is ergodic if its induced measure μ_t on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ tends to an invariant measure for $t \to \infty$, which we here denote by μ_∞. The FPE defined in (A.5), which describes the time evolution of the measure μ_t of X_t, plays therefore

a crucial role in the definition and in the study of ergodic processes. If the measure μ_t tends to an invariant measure μ_∞, then the density ρ_∞ should satisfy $\partial_t \rho_\infty = 0$, which plugging into (A.5) yields

$$\mathcal{L}^* \rho_\infty = 0,$$

which we call the stationary FPE. Indeed, this characterization of the invariant density can be employed to define ergodic diffusion processes, as in the definition below.

Definition A.11. Let $\mathcal{L}^*$ be defined in (A.4). We say that $X = (X_t, t \geq 0)$ solution of (A.1) is ergodic if and only if there exists a unique ρ_∞ such that

$$\int_{\mathbb{R}^d} \rho_\infty(x)\,\mathrm{d}x = 1, \quad \mathrm{Ker}(\mathcal{L}^*) = \mathrm{span}\{\rho^\infty\}.$$

We call the measure μ_∞ such that $\mu_\infty(\mathrm{d}x) = \rho_\infty(x)\,\mathrm{d}x$ the invariant measure, and ρ_∞ the invariant density.

The ergodicity of X has an implication for the kernel of the generator $\mathcal{L}$. In particular, we have that since $\dim(\mathrm{Ker}(\mathcal{L}^*)) = 1$, then $\dim(\mathrm{Ker}(\mathcal{L})) = 1$, and therefore that

$$\mathrm{Ker}(\mathcal{L}) = \mathrm{span}\{1\}, \tag{A.6}$$

where 1 denotes constant functions. It is trivial to notice that $\mathcal{L}1 = 0$ by the definition of $\mathcal{L}$. The fact that the dimension of the kernel of $\mathcal{L}$ does not exceed one is implied by Fredholm's alternative, which we state here and which is given e.g. in [52, Appendix D] or [113, Theorem 2.42].

Theorem A.12. *Let $\mathcal{L}$ be a compact operator on a Hilbert space. Then exactly one of the two following alternatives holds*

(i) *$\mathcal{L}u = f$ and $\mathcal{L}^*\rho = g$ have a unique solution for all f and g,*

(ii) $\dim(\mathrm{Ker}(\mathcal{L})) = \dim(\mathrm{Ker}(\mathcal{L}^*))$.

The proof of the Fredholm's alternative can be found e.g. in [52]. Clearly, in case X is ergodic alternative (ii) and thus (A.6) hold. Let us further remark that in this case (or more in general, if alternative (ii) holds), the non-homogeneous equations $\mathcal{L}u = f$ and $\mathcal{L}^*\rho = g$ admit a solution if and only if a centering condition is verified. Indeed, multiplying the first equation by ρ_∞ and integrating we obtain

$$(\mathcal{L}u, \rho_\infty) = (f, \rho_\infty).$$

Now, on the right hand side it clearly holds $(\mathcal{L}u, \rho_\infty) = (u, \mathcal{L}^*\rho_\infty) = 0$, so that for $\mathcal{L}u = f$ to be solvable we need $(f, \rho_\infty) = 0$. In other words, the equation $\mathcal{L}u = f$ is solvable if and only if $f \in \mathrm{Ker}^\perp(\mathcal{L}^*)$. In the same way, we obtain that the equation $\mathcal{L}^*\rho = g$ is solvable if and only if $(g, 1) = 0$, i.e., if g is a function with zero average.

Example A.13. Let $d = 1$ and α and σ be positive real numbers. We consider the Ornstein–Uhlenbeck process X, i.e. the solution of

$$\mathrm{d}X_t = -\alpha X_t\,\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t,$$

with a given initial condition X_0. Direct calculations with the Itô formula allow to compute the exact solution of the SDE above, which reads

$$X_t = X_0 + \sqrt{2\sigma}\int_0^t e^{-\alpha(t-s)}\,\mathrm{d}W_s, \tag{A.7}$$

from which one can easily infer that

$$X_t \sim \mathcal{N}\left(e^{-\alpha t}, \left(1 - e^{-2\alpha t}\right) \frac{\sigma}{\alpha}\right).$$

Taking informally the limit for $t \to \infty$ shows that the invariant distribution is the Gaussian $\mu_\infty = \mathcal{N}(0, \sigma/\alpha)$. Let us verify this employing the theoretical tools introduced above. The adjoint of the generator is in this case given by

$$\mathcal{L}^* \rho = \alpha x \rho' + \alpha \rho + \sigma \rho''.$$

Replacing the density ρ_∞, which is given by

$$\rho_\infty = \sqrt{\frac{\alpha}{2\pi\sigma}} \exp\left(-\frac{\alpha x^2}{2\sigma}\right),$$

we obtain $\mathcal{L}^* \rho_\infty = 0$, which shows that μ_∞ is indeed the invariant measure of (A.7).

Ergodic processes are characterized by the fundamental property that time averages tend asymptotically to space averages mediated by their invariant measure. We now formalize this result, known as the ergodic theorem.

Theorem A.14. *Let $X = (X_t, t \geq 0)$ be an ergodic process with values in $\mathbb{R}^d$, and let μ_∞ denote its invariant measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then, for all continuous functions φ it holds*

$$\lim_{T \to \infty} \frac{1}{T} \int_0^t \varphi(X_t) = \mathbb{E}^{\mu_\infty}[\varphi(X)], \tag{A.8}$$

almost surely.

The rate of convergence in (A.8) can be quantified in specific cases. In particular, if the convergence rate is exponential in time, we say that the process is geometrically ergodic, which we formalize in the definition below.

Definition A.15. The process $X = (X_t, t \geq 0)$ solution of (A.1) is geometrically ergodic if there exist constants $C, \lambda > 0$ such that for all measurable $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfying for some integer $q > 0$

$$\varphi(x) \leq 1 + \|x\|_2^q,$$

it holds

$$|\mathbb{E}\, \varphi(X_t) - \mathbb{E}^{\mu_\infty}[\varphi(X)]| \leq C \left(1 + \|X_0\|_2^q\right) e^{-\lambda t},$$

where $\mathbb{E}$ denotes expectation with respect to the Brownian motion.

A.5 Martingales

In this section, we set our notation and give a brief introduction on the theory of martingales, as well as a short series of convergence results. The topics introduced here are treated in a broader and deeper manner in the classic books [90, 110, 120].

We start by recalling the definition of a martingale.

Definition A.16. Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $\mathcal{F}_t$, and let $M = (M_t, t \geq 0)$ be a $\mathcal{F}_t$-adapted process with values in $\mathbb{R}^d$. If

(i) $\mathbb{E}\,[|M_t|] < \infty$ for all $t \geq 0$,

(ii) $\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s$, for all $0 \leq s \leq t$,

then we say that M is a $\mathcal{F}_t$-martingale. If moreover $\mathbb{E}[|M_t|^2] < \infty$ for all $t \geq 0$, we say that M is a square-integrable martingale.

Let us remark that since M_t is $\mathcal{F}_t$-adapted, the property (i) can be rewritten as

$$\mathbb{E}[M_t - M_s \mid \mathcal{F}_s] = 0,$$

which can be a convenient rewriting in some applications. A notable example of a martingale is given by the Itô integral. In particular, let W be a standard Brownian motion and let Z_t be a real-valued stochastic process which is $\mathcal{F}_t$-adapted, where $\mathcal{F}_t$ is the natural filtration generated by W, and such that

$$\int_0^t \mathbb{E}\left[|Z_s|^2\right] \mathrm{d}s < \infty \tag{A.9}$$

Then, the stochastic integral

$$I_t = \int_0^t Z_s \, \mathrm{d}W_s, \tag{A.10}$$

is a square-integrable $\mathcal{F}_t$-martingale.

An important quantity associated to any martingale is its quadratic variation. Let $X = (X_t, t \geq 0)$ be a real-valued stochastic process. The quadratic variation of X, denoted $\langle X \rangle := (\langle X \rangle_t, t \geq 0)$, is the process defined by

$$\langle X \rangle_t := \lim_{n_\Delta \to \infty} \sum_{i=1}^{n_\Delta} (X_{t_i} - X_{t_{i-1}})^2,$$

where for $\Delta > 0$, the partition $0 = t_0 < t_1 < \cdots < t_{n_\Delta} = t$ of the interval $[0, t]$ has characteristic size Δ, and where the limit is taken in probability. If X is a martingale, its quadratic variation can be alternatively defined as follows (see e.g. [90, Chapter 1§8], [113, Chapter 3]).

Definition A.17. Let $M = (M_t, t \geq 0)$ be a $\mathcal{F}_t$-martingale with values in $\mathbb{R}^d$. Then, let Q be a non-decreasing $\mathbb{R}^{d \times d}$-valued process such that the process $M \otimes M - Q$ is an $\mathcal{F}_t$-martingale. We call Q the quadratic variation of M and write $Q = \langle M \rangle$.

Under the condition that the martingale M is square-integrable, its quadratic variation is a well-defined stochastic process. This is given in by the following result (see e.g. [120, Chapter IV]).

Theorem A.18. *Let $M = (M_t, t \geq 0)$ be a square-integrable $\mathcal{F}_t$-martingale with values in $\mathbb{R}^d$. Then, its quadratic variation $\langle M \rangle$ exists and is unique.*

For the Itô integral I given in (A.10) the quadratic variation can be computed explicitly, as shown in the following result. While existence and uniqueness of the quadratic variation is guaranteed because I is square-integrable under (A.9), proving that the explicit expression is indeed given by (A.11) relies on the Itô isometry.

Corollary A.19. *Let Z_t be a $\mathcal{F}_t$-adapted stochastic process with values in $\mathbb{R}^d$ and let*

$$I_t = \int_0^t Z_s \, \mathrm{d}W_s.$$

Then, if (A.9) *holds, the quadratic variation of I is given by*

$$\langle I \rangle_t = \int_0^t Z_s Z_s^\top \, \mathrm{d}s. \tag{A.11}$$

We conclude by presenting convergence results for continuous-time martingales. We restrict the setting to the one-dimensional case, and remark that generalizations and broader discussions on the topic can be found in [70, 90]. The first convergence theorem we state is the strong law of large numbers for martingales.

Theorem A.20. *Let $M = (M_t, t \geq 0)$ be a real-valued and square-integrable martingale such that $\langle M \rangle_\infty = \infty$ a.s. Then,*

$$\lim_{t\to\infty} \frac{M_t}{\langle M \rangle_t} = 0, \quad a.s.$$

Intuitively, Theorem A.20 is a generalization of

$$\lim_{t\to\infty} \frac{W_t}{t} = 0, \tag{A.12}$$

a.s. and for a Brownian motion W, which can be shown to hold by considering that $\widetilde{W} = (\widetilde{W}_t = tW(1/t), t \geq 0)$, and $\widetilde{W}_0 = 0$ a.s., is a Brownian motion. Indeed, it holds $\langle W \rangle_t = t$, so that (A.12) is a particular instance of Theorem A.20. We now introduce the central limit theorem (CLT) for martingales, which describe the long-time behavior of sufficiently well-behaved martingales after being appropriately rescaled.

Theorem A.21. *Let $M = (M_t, t \geq 0)$ be a real-valued and square-integrable martingale such that*

$$\lim_{t\to\infty} \mathbb{E}\left[\left|\frac{\langle M \rangle_t}{t} - C\right|\right] = 0,$$

for some $C > 0$. Then, it holds for $t \to \infty$

$$\frac{M_t}{\sqrt{t}} \Rightarrow Z, \qquad Z \sim \mathcal{N}(0, C).$$

The last convergence result we present is the functional CLT for martingales. This result guarantees weak convergence of a sequence of multivariate continuous martingales to a rescaled multi-dimensional Brownian motion. Convergence is granted under the condition that the quadratic variation of the elements of the sequence converges weakly to a linear function of time. We refer to [144], [51, Chapter 7] or [70, Section VIII.3b] for a proof.

Theorem A.22. *Let $\{M_n\}_{n=1,2,\dots}$ be a sequence of continuous square-integrable $\mathcal{F}_t$-martingales with values in $\mathbb{R}^d$ and let $C \in \mathbb{R}^{d\times d}$ be a symmetric positive semi-definite matrix. If for $n \to \infty$*

$$\langle M_n \rangle_t \Rightarrow Ct,$$

then, for $n \to \infty$

$$M_n \Rightarrow \sqrt{C}W, \quad with \quad \sqrt{C}\sqrt{C}^\top = C$$

where W is a d-dimensional standard Brownian motion.

Bibliography

[1] A. Abdulle, *Fourth order Chebyshev methods with recurrence relation*, SIAM J. Sci. Comput., 23 (2002), pp. 2041–2054.

[2] A. Abdulle, *A priori and a posteriori error analysis for numerical homogenization: a unified framework*, Ser. Contemp. Appl. Math. CAM, 16 (2011), pp. 280–305.

[3] A. Abdulle and A. Di Blasio, *Numerical homogenization and model order reduction for multiscale inverse problems*, Multiscale Model. Simul., 17 (2019), pp. 399–433.

[4] A. Abdulle and A. Di Blasio, *A Bayesian Numerical Homogenization Method for Elliptic Multiscale Inverse Problems*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 414–450.

[5] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden, *The heterogeneous multiscale method*, Acta Numer., 21 (2012), pp. 1–87.

[6] A. Abdulle and G. Garegnani, *Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration*, Stat. Comput., 30 (2020), pp. 907–932.

[7] A. Abdulle and G. Garegnani, *A probabilistic finite element method based on random meshes: A posteriori error estimators and Bayesian inverse problems*, Comput. Methods Appl. Mech. Engrg., 384 (2021), p. 113961.

[8] A. Abdulle, G. Garegnani, G. A. Pavliotis, A. M. Stuart, and A. Zanoni, *Drift estimation of multiscale diffusions based on filtered data*, to appear in Found. Comput. Math., (2021).

[9] A. Abdulle, G. Garegnani, and A. Zanoni, *Ensemble Kalman Filter for Multiscale Inverse Problems*, Multiscale Model. Simul., 18 (2020), pp. 1565–1594.

[10] A. Abdulle and A. A. Medovikov, *Second order Chebyshev methods based on orthogonal polynomials*, Numer. Math., 90 (2001), pp. 1–18.

[11] A. Abdulle, G. A. Pavliotis, and A. Zanoni, *Eigenfunction martingale estimating functions and filtered data for drift estimation of discretely observed multiscale diffusions*. arXiv preprint arXiv:2104.10587, 2021.

[12] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000.

[13] Y. Aït-Sahalia and J. Jacod, *High-frequency financial econometrics*, Princeton University Press, 2014.

[14] Y. Aït-Sahalia, P. A. Mykland, and L. Zhang, *How often to sample a continuous-time process in the presence of market microstructure noise*, Rev. Financ. Stud, 18 (2005), pp. 351–416.

[15] C. Andrieu, A. Doucet, and R. Holenstein, *Particle Markov chain Monte Carlo methods*, J. R. Stat. Soc. Ser. B. Stat. Methodol., (2010), pp. 269 – 342.

[16] C. Andrieu and G. O. Roberts, *The pseudo-marginal approach for efficient Monte Carlo computations*, Ann. Statist., 37 (2009), pp. 697–725.

[17] R. Azencott, A. Beri, A. Jain, and I. Timofeyev, *Sub-sampling and parametric estimation for multiscale dynamics*, Commun. Math. Sci., 11 (2013), pp. 939–970.

[18] R. Azencott, A. Beri, and I. Timofeyev, *Adaptive sub-sampling for parametric estimation of Gaussian diffusions*, J. Stat. Phys., 139 (2010), pp. 1066–1089.

[19] I. Babuška and W. C. Rheinboldt, *A posteriori error analysis of finite element solutions for one-dimensional problems*, SIAM J. Numer. Anal., 18 (1981), pp. 565–589.

[20] I. V. Basawa and B. L. S. Prakasa Rao, *Statistical inference for stochastic processes*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London-New York, 1980. Probability and Mathematical Statistics.

[21] G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.

[22] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic analysis for periodic structures*, North-Holland Publishing Co., Amsterdam, 1978.

[23] P. Billingsley, *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, second ed., 1999. A Wiley-Interscience Publication.

[24] J. P. N. Bishwal, *Parameter estimation in stochastic differential equations*, vol. 1923 of Lecture Notes in Mathematics, Springer, Berlin, 2008.

[25] N. Bosch, P. Hennig, and F. Tronarp, *Calibrated adaptive probabilistic ODE solvers*. arXiv preprint arXiv:2012.08202, 2020.

[26] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.

[27] A.-P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980), Soc. Brasil. Mat., Rio de Janeiro, 1980, pp. 65–73.

[28] D. Calvetti, M. Dunlop, E. Somersalo, and A. M. Stuart, *Iterative updating of model error for Bayesian inversion*, Inverse Problems, 34 (2018), pp. 025008, 38.

[29] D. Calvetti, O. Ernst, and E. Somersalo, *Dynamic updating of numerical model discrepancy using sequential sampling*, Inverse Problems, 30 (2014), pp. 114019, 19.

[30] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[31] O. A. Chkrebtii and D. A. Campbell, *Adaptive step-size selection for state-space probabilistic differential equation solvers*, Stat. Comput., 29 (2019), pp. 1285–1295.

[32] O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami, *Bayesian solution uncertainty quantification for differential equations*, Bayesian Anal., 11 (2016), pp. 1239–1267.

[33] P. G. Ciarlet, *The finite element method for elliptic problems.*, vol. 40 of Classics Appl. Math., SIAM, Philadelphia, 2002.

[34] D. Cioranescu and P. Donato, *An introduction to homogenization*, vol. 17 of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, New York, 1999.

[35] E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, and A. M. Stuart, *Calibrate, emulate, sample*, J. Comput. Phys., 424 (2021), pp. 109716, 20.

[36] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami, *Probabilistic numerical methods for partial differential equations and Bayesian inverse problems.* arXiv preprint arXiv:1605.07811, 2017.

[37] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami, *Probabilistic numerical methods for PDE-constrained Bayesian inverse problems*, AIP Conference Proceedings, 1853 (2017), p. 060001.

[38] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami, *Bayesian probabilistic numerical methods*, SIAM Rev., 61 (2019), pp. 756–789.

[39] P. R. Conrad, M. Girolami, S. Särkkä, A. M. Stuart, and K. Zygalakis, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput., 27 (2017), pp. 1065–1082.

[40] C. J. Cotter and G. A. Pavliotis, *Estimating eddy diffusivities from noisy Lagrangian observations*, Commun. Math. Sci., 7 (2009), pp. 805–838.

[41] 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 (2013), pp. 424–446.

[42] M. Croci and P. E. Farrell, *Complexity bounds on supermesh construction for quasi-uniform meshes*, J. Comput. Phys., 414 (2020), pp. 109459, 7.

[43] M. Croci, M. B. Giles, M. E. Rognes, and P. E. Farrell, *Efficient white noise sampling and coupling for multilevel Monte Carlo with nonnested meshes*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1630–1655.

[44] D. Crommelin and E. Vanden-Eijnden, *Fitting timeseries by continuous-time Markov chains: a quadratic programming approach*, J. Comput. Phys., 217 (2006), pp. 782–805.

[45] D. Crommelin and E. Vanden-Eijnden, *Reconstruction of diffusions using spectral data from timeseries*, Commun. Math. Sci., 4 (2006), pp. 651–668.

[46] D. Crommelin and E. Vanden-Eijnden, *Data-based inference of generators for Markov jump processes using convex optimization*, Multiscale Model. Simul., 7 (2009), pp. 1751–1778.

[47] D. Crommelin and E. Vanden-Eijnden, *Diffusion estimation from multiscale data by operator eigenpairs*, Multiscale Model. Simul., 9 (2011), pp. 1588–1623.

[48] M. Dashti and A. M. Stuart, *Uncertainty quantification and weak approximation of an elliptic inverse problem*, SIAM J. Numer. Anal., 49 (2011), pp. 2524–2542.

[49] M. Dashti and A. M. Stuart, *The Bayesian Approach to Inverse Problems*, in Handbook of Uncertainty Quantification, Springer, 2016, pp. 1–118.

[50] W. E, D. Liu, and E. Vanden-Eijnden, *Analysis of multiscale methods for stochastic differential equations*, Comm. Pure Appl. Math., 58 (2005), pp. 1544–1585.

[51] S. N. Ethier and T. G. Kurtz, *Markov processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1986. Characterization and convergence.

[52] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[53] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res., (1994), pp. 10143–10162.

[54] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dyn., 53 (2003), pp. 343–367.

[55] G. Evensen, *Data assimilation: the ensemble Kalman filter*, Springer Science & Business Media, 2009.

[56] S. Gailus and K. Spiliopoulos, *Statistical inference for perturbed multiscale dynamical systems*, Stochastic Process. Appl., 127 (2017), pp. 419–448.

[57] S. Gailus and K. Spiliopoulos, *Discrete-time statistical inference for multiscale diffusions*, Multiscale Model. Simul., 16 (2018), pp. 1824–1858.

[58] M. Girolami, E. Febrianto, G. Yin, and F. Cirak, *The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions*, Comput. Methods Appl. Mech. Engrg., 375 (2021), pp. 113533, 32.

[59] E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math., 25 (1997), pp. 219–227.

[60] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics 31, Springer-Verlag, Berlin, second ed., 2006.

[61] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2008.

[62] E. Hairer and G. Wanner, *Solving ordinary differential equations II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2002.

[63] M. Hairer, A. M. Stuart, and S. J. Vollmer, *Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions*, Ann. Appl. Probab., 24 (2014), pp. 2455–2490.

[64] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[65] P. Hennig, M. A. Osborne, and M. Girolami, *Probabilistic numerics and uncertainty in computations*, Proc. A., 471 (2015), pp. 20150142, 17.

[66] M. Hénon and C. Heiles, *The applicability of the third integral of motion: Some numerical experiments*, Astronom. J., 69 (1964), pp. 73–79.

[67] D. Z. Huang, T. Schneider, and A. M. Stuart, *Unscented Kalman inversion*. arXiv preprint arXiv:2102.01580, 2021.

[68] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), pp. 045001, 20.

[69] J. Jacod and P. Protter, *Probability essentials*, Universitext, Springer-Verlag, Berlin, second ed., 2003.

[70] J. Jacod and A. N. Shiryaev, *Limit theorems for stochastic processes*, vol. 288 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, second ed., 2003.

[71] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[72] S. Kalliadasis, S. Krumscheid, and G. A. Pavliotis, *A new framework for extracting coarse-grained models from time series with multiscale structure*, J. Comput. Phys., 296 (2015), pp. 314–328.

[73] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 35–45.

[74] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, vol. 113 of Graduate Texts in Mathematics, Springer-Verlag, New York, second ed., 1991.

[75] A. F. Karr, *Probability*, Springer Texts in Statistics, Springer-Verlag, New York, 1993.

[76] H. Kersting and P. Hennig, *Active uncertainty calibration in Bayesian ODE solvers*, in Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), AUAI Press, 2016, pp. 309–318.

[77] H. Kersting, T. J. Sullivan, and P. Hennig, *Convergence rates of Gaussian ODE filters*, Stat. Comput., 30 (2020), pp. 1791–1816.

[78] M. Kessler and M. Sørensen, *Estimating equations based on eigenfunctions for a discretely observed diffusion process*, Bernoulli, 5 (1999), pp. 299–314.

[79] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, no. 23 in Applications of Mathematics: Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, Berlin and New York, 1992.

[80] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of Monte Carlo methods*, vol. 706, John Wiley & Sons, 2013.

[81] S. Krumscheid, G. A. Pavliotis, and S. Kalliadasis, *Semiparametric drift and diffusion estimation for multiscale diffusions*, Multiscale Model. Simul., 11 (2013), pp. 442–473.

[82] S. Krumscheid, M. Pradas, G. A. Pavliotis, and S. Kalliadasis, *Data-driven coarse graining in action: Modeling and prediction of complex systems*, Physical Review E, 92 (2015), p. 042139.

[83] Y. A. Kutoyants, *Statistical inference for ergodic diffusion processes*, Springer Series in Statistics, Springer-Verlag London, Ltd., London, 2004.

[84] T. Lelièvre and G. Stoltz, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), pp. 681–880.

[85] H. C. Lie, M. Stahn, and T. J. Sullivan, *Randomised one-step time integration methods for deterministic operator differential equations*. arXiv preprint arXiv:2103.16506, 2021.

[86] H. C. Lie, A. M. Stuart, and T. J. Sullivan, *Strong convergence rates of probabilistic integrators for ordinary differential equations*, Stat. Comput., 29 (2019), pp. 1265–1283.

[87] 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 (2018), pp. 1600–1629.

[88] R. S. Liptser and A. N. Shiryaev, *Statistics of random processes. I*, vol. 5 of Applications of Mathematics (New York), Springer-Verlag, Berlin, expanded ed., 2001. General theory, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability.

[89] R. S. Liptser and A. N. Shiryaev, *Statistics of random processes. II*, vol. 6 of Applications of Mathematics (New York), Springer-Verlag, Berlin, expanded ed., 2001. Applications, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability.

[90] R. S. Liptser and A. N. Shiryayev, *Theory of martingales*, vol. 49 of Mathematics and its Applications (Soviet Series), Kluwer Academic Publishers Group, Dordrecht, 1989. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze].

[91] E. N. Lorenz, *Deterministic nonperiodic flow*, J. Atmos. Sci., 20 (1963), pp. 130–141.

[92] T. Matsuda and Y. Miyatake, *Estimation of Ordinary Differential Equation Models with Discretization Error Quantification*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 302–331.

[93] J. C. Mattingly, N. S. Pillai, and A. M. Stuart, *Diffusion limits of the random walk Metropolis algorithm in high dimensions*, Ann. Appl. Probab., 22 (2012), pp. 881–930.

[94] J. C. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.

[95] G. N. Milstein, *Numerical integration of stochastic differential equations*, vol. 313 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1995. Translated and revised from the 1988 Russian original.

[96] G. N. Milstein and M. V. Tretyakov, *Stochastic numerics for mathematical physics*, Scientific Computing, Springer-Verlag, Berlin and New York, 2004.

[97] G. N. Milstein and M. V. Tretyakov, *Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients*, SIAM J. Numer. Anal., 43 (2005), pp. 1139–1154.

[98] S. Moskow and M. Vogelius, *First-order corrections to the homogenised eigenvalues of a periodic composite medium. a convergence proof*, Proc. Roy. Soc. Edinburgh, 127A (1997), pp. 1263–1299.

[99] J. Nolen and G. Papanicolaou, *Fine scale uncertainty in parameter estimation for elliptic equations*, Inverse Problems, 25 (2009), pp. 115021, 22.

[100] J. Nolen, G. A. Pavliotis, and A. M. Stuart, *Multiscale modeling and inverse problems*, in Numerical analysis of multiscale problems, vol. 83 of Lect. Notes Comput. Sci. Eng., Springer, Heidelberg, 2012, pp. 1–34.

[101] C. J. Oates, J. Cockayne, R. G. Aykroyd, and M. Girolami, *Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment*, J. Amer. Statist. Assoc., 114 (2019), pp. 1518–1531.

[102] C. J. Oates and T. J. Sullivan, *A modern retrospective on probabilistic numerics*, Stat. Comput., 29 (2019), pp. 1335–1351.

[103] J. Oesterle, N. Krämer, P. Hennig, and P. Berens, *Numerical uncertainty can critically affect simulations of mechanistic models in neuroscience.* biorXiv preprint bioRxiv:2021.04.27.441605, 2021.

[104] S. C. Olhede, A. M. Sykulski, and G. A. Pavliotis, *Frequency domain estimation of integrated volatility for Itô processes in the presence of market-microstructure noise*, Multiscale Model. Simul., 8 (2010), pp. 393–427.

[105] L. F. Olsen, *An enzyme reaction with a strange attractor*, Phys. Lett. A, 94 (1983), pp. 454 – 457.

[106] H. Owhadi, *Bayesian numerical homogenization*, Multiscale Model. Simul., 13 (2015), pp. 812–828.

[107] H. Owhadi, *Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games*, SIAM Rev., 59 (2017), pp. 99–149.

[108] H. Owhadi and L. Zhang, *Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients*, J. Comput. Phys., 347 (2017), pp. 99–128.

[109] A. Papavasiliou, G. A. Pavliotis, and A. M. Stuart, *Maximum likelihood drift estimation for multiscale diffusions*, Stochastic Process. Appl., 119 (2009), pp. 3173–3210.

[110] G. A. Pavliotis, *Stochastic processes and applications*, vol. 60 of Texts in Applied Mathematics, Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[111] G. A. Pavliotis, Y. Pokern, and A. M. Stuart, *Parameter estimation for multiscale diffusions: an overview*, in Statistical methods for stochastic differential equations, vol. 124 of Monogr. Statist. Appl. Probab., CRC Press, Boca Raton, FL, 2012, pp. 429–472.

[112] G. A. Pavliotis and A. M. Stuart, *Parameter estimation for multiscale diffusions*, J. Stat. Phys., 127 (2007), pp. 741–781.

[113] G. A. Pavliotis and A. M. Stuart, *Multiscale methods: averaging and homogenization*, vol. 53 of Texts in Applied Mathematics, Springer, New York, 2008.

[114] G. A. Pavliotis, A. M. Stuart, and U. Vaes, *Derivative-free bayesian inversion using multiscale dynamics.* arXiv preprint arXiv:2102.00540, 2021.

[115] Y. Pokern, A. M. Stuart, and J. H. van Zanten, *Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs*, Stochastic Process. Appl., 123 (2013), pp. 603–628.

[116] Y. Pokern, A. M. Stuart, and E. Vanden-Eijnden, *Remarks on drift estimation for diffusion processes*, Multiscale Model. Simul., 8 (2009), pp. 69–95.

[117] A. Quarteroni, *Numerical Models for Differential Problems*, vol. 2 of Modeling, Simulation & Applications, Springer, 2009.

[118] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Inferring solutions of differential equations using noisy multi-fidelity data*, J. Comput. Phys., 335 (2017), pp. 736–746.

[119] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Machine learning of linear differential equations using Gaussian processes*, J. Comput. Phys., 348 (2017), pp. 683–693.

[120] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*, vol. 293 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, third ed., 1999.

[121] H. Rieder, *Robust asymptotic statistics*, Springer Series in Statistics, Springer-Verlag, New York, 1994.

[122] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Texts in Statistics, Springer-Verlag, New York, second ed., 2004.

[123] S. Salsa, *Partial differential equations in action*, vol. 99 of Unitext, Springer, [Cham], third ed., 2016. From modelling to theory, La Matematica per il 3+2.

[124] F. Santambrogio, *Optimal transport for applied mathematicians*, vol. 87 of Progress in Nonlinear Differential Equations and their Applications, Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[125] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290.

[126] M. Schober, D. Duvenaud, and P. Hennig, *Probabilistic ODE solvers with Runge–Kutta means*, in Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 739–747.

[127] M. Schober, S. Särkkä, and P. Hennig, *A probabilistic model for the numerical solution of initial value problems*, Stat. Comput., 29 (2019), pp. 99–122.

[128] R. D. Skeel and C. W. Gear, *Does variable step size ruin a symplectic integrator?*, Physica, 60 (1992), pp. 311–313.

[129] J. Skilling, *Bayesian solution of ordinary differential equations*, in Maximum entropy and Bayesian methods, Springer, 1992, pp. 23–37.

[130] C. Störmer, *Sur les trajectoires des corpuscules électrisés*, Arch. sci. phys. nat. Genève, 24 (1907), pp. 5–18, 113–158, 221–247.

[131] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[132] T. J. Sullivan, *Introduction to uncertainty quantification*, vol. 63 of Texts in Applied Mathematics, Springer, Cham, 2015.

[133] T. J. Sullivan, *Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors*, Inverse Probl. Imaging, 11 (2017), pp. 857–874.

[134] O. Teymur, H. C. Lie, T. J. Sullivan, and B. Calderhead, *Implicit probabilistic integrators for ODEs*, in Advances in Neural Information Processing Systems, 2018, pp. 7244–7253.

[135] O. Teymur, K. Zygalakis, and B. Calderhead, *Probabilistic linear multistep methods*, in Advances in Neural Information Processing Systems, 2016, pp. 4321–4328.

[136] V. Thomée, *Galerkin finite element methods for parabolic problems*, vol. 25 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[137] F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig, *Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective*, Stat. Comput., 29 (2019), pp. 1297–1315.

[138] P. J. van der Houwen and B. P. Sommeijer, *On the internal stability of explicit, m-stage Runge–Kutta methods for large m-values*, Z. Angew. Math. Mech., 60 (1980), pp. 479–485.

[139] H. van Zanten, *A multivariate central limit theorem for continuous local martingales*, Statist. Probab. Lett., 50 (2000), pp. 229–235.

[140] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.

[141] R. Verfürth, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[142] L. Verlet, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Physical Review, 159 (1967), pp. 98–103.

[143] M. Vihola, *Robust adaptive Metropolis algorithm with coerced acceptance rate*, Stat. Comput., 22 (2012), pp. 997–1008.

[144] W. Whitt, *Proofs of the martingale FCLT*, Probab. Surv., 4 (2007), pp. 268–302.

[145] Y. Ying, J. Maddison, and J. Vanneste, *Bayesian inference of ocean diffusivity from Lagrangian trajectory data*, Ocean Model., 140 (2019).

[146] L. Zhang, P. A. Mykland, and Y. Aït-Sahalia, *A tale of two time scales: determining integrated volatility with noisy high-frequency data*, J. Amer. Statist. Assoc., 100 (2005), pp. 1394–1411.

[147] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

[148] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates. II. Error estimates and adaptivity*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

Curriculum Vitae

Personal data

Name	Giacomo Garegnani
Date of birth	October 20, 1992
Nationality	Italian and Swiss

Education

2017 - 2021 **PhD in Mathematics**
École Polytechnique Fédérale de Lausanne, Switzerland.
Thesis advisor: Professor A. Abdulle.

2014 - 2017 **MSc in Computational Science and Engineering**
École Polytechnique Fédérale de Lausanne, Switzerland.
Thesis advisor: Professor A. Abdulle.

2011 - 2014 **BSc in Mathematical Engineering**
Politecnico di Milano, Italy.
Thesis advisor: Professor F. Tomarelli.

Work Experience

2016 **Software Engineering Intern**
MindMaze – Lausanne, Switzerland.

2015 - 2016 **R&D Intern**
STMicroelectronics – Crolles, France.

PhD Publications

[1] A. Abdulle and G. Garegnani, *A probabilistic finite element method based on random meshes: A posteriori error estimators and Bayesian inverse problems.* Comput. Methods Appl. Mech. Engrg., 384 (2021), p. 113961.

[2] A. Abdulle, G. Garegnani, G. A. Pavliotis, A. M. Stuart, and A. Zanoni, *Drift estimation of multiscale diffusions based on filtered data.* to appear in Found. Comput. Math., (2021).

[3] A. Abdulle, G. Garegnani, and A. Zanoni, *Ensemble Kalman Filter for Multiscale Inverse Problems*, Multiscale Model. Simul., 18 (2020), pp. 1565–1594.

[4] A. Abdulle and G. Garegnani, *Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration*, Stat. Comput., 30 (2020), pp. 907–932.

Other Publications

[5] G. Garegnani, V. Fiori, G. Gouget, F. Monsieur, C. Tavernier, *Wafer level measurements and numerical analysis of self-heating phenomena in nano-scale SOI MOSFETs*, Microelectron. Reliab., 63 (2016), pp. 90 – 96.

[6] G. Garegnani, V. Fiori, S. Gallois-Garreignot, R. Gonella, *Numerical analysis of thermal effects in SOI MOSFET flip-chip packages: multi-scale studies on isolated transistors and global simulations*, Electronics System-Integration Technology Conference, Grenoble (2016).

Presentations

- UQ Hybrid Seminar – RWTH Aachen University (Aachen, Germany, January 2021);
  Seminar: *Filtering the data: An alternative to subsampling for drift estimation of multiscale diffusions.*
- SIAM Conference on Uncertainty Quantification (Garching, Germany, March 2020) – Canceled due to Covid-19 pandemic;
  Talk: *Model misspecification and uncertainty quantification for drift estimation in multiscale diffusion processes.*
- Imperial College London (London, UK, February 2020);
  Seminar: *A pre-processing technique for asymptotically correct drift estimation in multiscale diffusion processes.*
- California Institute of Technology (Pasadena, US, August 2019);
  Seminar: *Bayesian inference of multiscale differential equations.*
- MATHICSE Retreat (Champéry, Switzerland, June 2019);
  Talk: *Bayesian inference of multiscale diffusion processes.*
- FoMICS-DADSi Summer School on Data Assimilation (Lugano, Switzerland, September 2018);
  Talk: *Probabilistic Runge–Kutta methods for uncertainty quantification of numerical errors in geometric integration.*
- AIMS Conference on Dynamical Systems, Differential Equations and Applications (Taipei, Taiwan, July 2018);
  Talk: *Uncertainty quantification of numerical errors in geometric integration via random time steps.*
- MATHICSE Retreat (Sainte-Croix, Switzerland, June 2018);
  Talk: *Probabilistic geometric integration of ordinary differential equations.*
- Swiss Numerics Colloquium (Zürich, Switzerland, April 2018);
  Talk: *Random time steps geometric integrators of ordinary differential equations for uncertainty quantification of numerical errors*
- Max Planck Institute for Intelligent Systems (Tübingen, Germany, March 2018);
  Seminar: *Uncertainty quantification of numerical errors in geometric integration via random time steps.*

- MATHICSE Retreat (Leysin, Switzerland, June 2017);
  Talk: *Probabilistic Runge–Kutta methods for ODEs: Chaotic problems and geometric properties.*

Academic visits

- Imperial College (London, UK, February 2020)
  One-week visit to Professor Grigorios A. Pavliotis' group
- California Institute of Technology (Pasadena, USA, August – September 2019)
  Five-weeks visit to Professor Andrew M. Stuart's group
- Max Planck Institute for Intelligent Systems (Tübingen, Germany, March 2018)
  One-week visit to Professor Philip Hennig's group

Distinctions

- SIAM travel award for the SIAM Conference on Uncertainty Quantification 2020
- President of the EPFL chapter of SIAM (2017 - 2018)

Teaching

Co-Supervised MSc Master Projects

- A. Zanoni, *Ensemble Kalman Filter for Multiscale Inverse Problems* (2019)
- A. Stankovic, *Probabilistic methods for differential equations: Adaptivity and Bayesian inverse problems* (2018)

Co-Supervised MSc Semester Projects

- D. Hamm, *Numerical study of an iterative filtering method for drift estimation of multiscale diffusions* (2021)
- A. Van De Velde, *Parameter estimation in multiscale Langevin dynamics with particle filters and Monte Carlo methods* (2020)
- W. Reise *Probabilistic Solvers for Ordinary Differential Equations* (2019)

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.