FIRST LINE OF TITLE
SECOND LINE OF TITLE

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the registrar's office.

Thèse n. 1234 2020
présentée le 28 février 2020
à la Faculté des sciences de base
laboratoire ANMC
programme doctoral en Mathématiques
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Giacomo Garegnani

acceptée sur proposition du jury :

Prof Name Surname, président du jury
Prof Assyr Abdulle, directeur de thèse
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur

Lausanne, EPFL, 2020

To my parents...

Acknowledgements

La science progresse mieux
quand on rigole

Assyr Abdulle

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Lausanne, February 28, 2020 D. K.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Résumé

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Contents

1 Introduction

2 Probabilistic methods for ODEs

2.1 Introduction

A variety of methods for integrating ordinary differential equations (ODEs) has been studied in the last decades, [31, 32, 30], with an emphasis on building accurate and stable deterministic approximations of the exact solution. In general, these methods are based on a time discretization on which the solution of the ODE is approximated via an iterative deterministic algorithm. Given a time step h, which indicates the refinement of the discretization, all these methods provide a point value for the approximation of the solution and guarantee that in the asymptotic limit of $h \to 0$ the numerical approximation will coincide with the exact solution. However, for some problems such as chaotic systems or inference problems having a distributional solution can help to quantify the uncertainty introduced by the numerical discretization without invoking the asymptotic limit $h \to 0$.

In recent years, probabilistic numerical methods for differential equations have been proposed [19, 15, 58] in order to quantify the uncertainty introduced by the time discretization in a statistical manner. A review summarizing the recent advancements in the field of probabilistic numerical can be found in [47, 18]. In general, these methods proceed iteratively to establish a probability measure over the numerical solution, thus providing a richer information than a single point value. In particular, probabilistic solvers offer a quantitative characterisation of late time errors by tuning the noise introduced by the method according to the accuracy of the solver. In this way, it is possible to obtain a reliable approach for capturing the sensitivity of the solution to numerical error, while transferring the convergence properties of classical deterministic integrators to the introduced probability measure in a consistent manner.

In the following, we will first show two examples motivating the probabilistic approach, and then present the main contributions of this work.

2.1.1 Motivating examples

Probabilistic integrators for ODEs do not provide more accurate solutions than classical deterministic methods nor are they computationally cheaper. Nevertheless, they can be useful in a variety of different problems, among which we identified the integration of chaotic dynamical systems and the solution of Bayesian inverse problems, which are briefly presented here.

Chaotic differential equations

Let us consider the Lorenz system [42], which is defined by the following ODE

$$\begin{aligned} y_1' &= \eta(y_2 - y_1), & y_1(0) &= -10, \\ y_2' &= y_1(\rho - y_3) - y_2, & y_2(0) &= -1, \\ y_3' &= y_1 y_2 - \beta y_3, & y_3(0) &= 40. \end{aligned} \tag{2.1}$$

It is well-known that for $\rho = 28$, $\eta = 10$, $\beta = 8/3$, this equation has a chaotic behaviour, i.e., a small perturbation forces the trajectories to deviate from the true solution. Integrating numerically (2.1) the error which is introduced at each time step is indeed a perturbation, thus any numerical solution cannot be considered reliable. In order to explore the state space of this chaotic dynamical system, we introduce a 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. 2.1 we show $M = 20$ numerical trajectories given by a second-order Runge–Kutta method for three different scales of the noise. It is possible to remark that in all the three cases, the numerical solutions almost coincide up to some time $\bar{t}$, thus diverging and showing 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. Nevertheless, it is unclear how to choose σ^2 so that the amount of noise that is introduced is balanced with the numerical error. Probabilistic methods for differential equations such as the one presented in this work and the one introduced by Conrad et al. [19] provide a rigorous analysis that suggests how to introduce a source of artificial noise in a consistent manner.

Bayesian inference

Problems of Bayesian inference are most often used to justify the usefulness of probabilistic methods for differential equations. The impact of a probabilistic component in the numerical approximation of inverse problems involving ODEs has already been presented in several works (e.g., [19, 15, 17]). In particular, the common underlying idea of these works is that if a deterministic integrator with a fixed finite time step is employed to approximate the solution of the ODE appearing in an inverse problem, the numerical error introduced by deterministic solvers can lead to inappropriate and non-predictive posterior concentrations. In the limit of

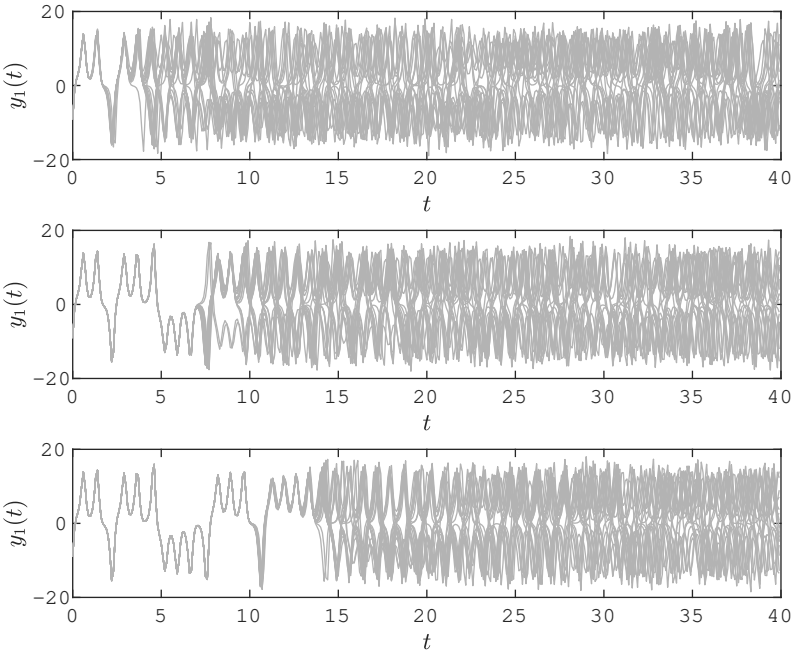

Figure 2.1 – First component $y_1(t)$ of the solution of (2.1) with decreasing Gaussian perturbations on the initial condition from top to bottom ($\sigma = 10^{-1}, 10^{-3}, 10^{-5}$, respectively).

an infinitely refined time discretization the posterior distributions obtained with a classical numerical method will indeed tend to the true distribution, but for a fixed time step (i.e., for a fixed computational budget) numerical error can lead to posterior concentrations away from the true value of the parameter of interest. These inappropriate solutions to inverse problems can be corrected by employing a probabilistic integrator to solve the ODE, thus obtaining posterior distributions that reflect the uncertainty given by the numerical solver (see fig. 2.3 at page 30 for an example).

2.1.2 Contributions

The method we analyse in this paper is inspired from the work of Conrad et al. [19], where a probabilistic method for ODEs is presented. This method consists of perturbing a deterministic numerical solution (e.g. arising from a Runge–Kutta discretization) with an additive source of noise at each time step. By appropriately scaling the random term, they manage to obtain a probabilistic solution without altering the convergence of the underlying deterministic scheme.

An additive noise contribution could nonetheless produce disruptive effects on favourable geometric features of deterministic schemes. A direct example of this non-robust behaviour 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. Chemical reactions with small population size for one species at some time of the evolution are typical physical examples. In particular, an additive random term could force the solution on the negative plane with a non-zero probability, and this probability could become non-negligibly big in case the magnitude of one component of the solution is small. Other geometric properties of an underlying ODE are also destroyed when perturbing the flow by a noisy forcing term.

Motivated by these issues, we present in this work a new probabilistic method for ODEs based on a random selection of the time steps. Hence, the randomness of the scheme becomes intrinsic in contrast to the additive noise method. For this new robust probabilistic integrator, we are able to prove strong and weak convergence towards the exact solution of the underlying ODE. Precisely, setting the variance of the random time steps to be proportional to some power of a deterministic time step allows to retrieve the rates of the underlying Runge–Kutta integrator.

It has been pointed out by Kersting and Hennig [36] that probabilistic methods based on sampling should be equipped with a criterion to choose the number of samples, so that computational effort is not wasted or, conversely, the sample size is not insufficient to describe the dynamics in a probabilistic fashion. In order to address this issue, in this work we show that Monte Carlo estimators drawn from our probabilistic solver converge with respect to the time step in the mean square sense independently of the sample size. We are able to prove a similar property for the scheme proposed in [19].

A large variety of dynamical systems is characterised by geometrical properties of their flow map [30]. Most notably, Hamiltonian systems, which are employed for modelling a variety of physical phenomena, are endowed with the property of symplecticity. It is possible to obtain good approximations of the solutions of Hamiltonian systems via mimicking numerically the geometric properties of the exact flow, i.e., by employing symplectic integrators. In particular, for symplectic integrators the energy function conserved by the exact flow is approximately conserved by numerical trajectories over long time spans, which in turn guarantees high-quality numerical solutions at the price of a rather low computational effort. While geometric properties of Runge–Kutta schemes have been analysed extensively in the deterministic case, they have not been considered yet for probabilistic numerical methods. The method we present in this work, being only an intrinsic modification of a Runge–Kutta integrator, is endowed with the geometric properties of its deterministic counterpart. In particular, we first show that our probabilistic scheme inherits the property of exact conservation of first integrals of the considered dynamics. Then, we show that in Hamiltonian systems the good approximation of the energy function given by symplectic schemes is preserved by our randomisation procedure over polynomially long times.

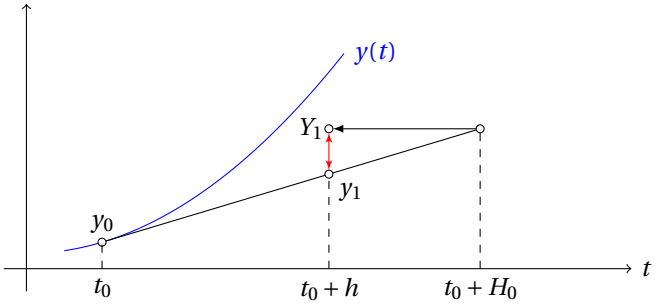

Figure 2.2 – Graphical representation of one step of the RTS-RK method with $\Psi_h(y) = y + hf(y)$. The red arrow is the stochastic contribution due to random time-stepping.

2.1.3 Outline

The paper is organised as follows. In section 2.2 we introduce the setting for probabilistic numerics and present our novel numerical scheme. We then show in section 2.3 and section 2.4 the properties of weak and mean square convergence of the numerical solution towards the exact solution of the ODE. In section 2.5 we analyse the accuracy of Monte Carlo estimators drawn from the numerical solution. The geometric properties of the numerical scheme are presented in section 2.6 and section 2.7, while in section 2.8 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 variety of numerical experiments confirming our theoretical results in section 2.9.

2.2 Random time step Runge–Kutta method

Let us consider a Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ and the ODE

$$y' = f(y), \quad y(0) = y_0 \in \mathbb{R}^d. \tag{2.2}$$

In the following, we will write for simplicity the solution $y(t)$ of (2.2) in terms of the flow of the ODE. In particular, we consider the family $\{\varphi_t\}_{t\geq 0}$ of functions $\varphi_t\colon \mathbb{R}^d \to \mathbb{R}^d$ such that

$$y(t) = \varphi_t(y_0). \tag{2.3}$$

Given a time step h, let us consider a Runge–Kutta method which deterministically approximates the solution $\varphi_t(y_0)$ of (2.2). In particular, we can write the numerical solution y_k approximating $\varphi_{t_k}(y_0)$, with $t_k = kh$ in terms of the numerical flow $\{\Psi_t\}_{t\geq 0}$, with $\Psi_t\colon \mathbb{R}^d \to \mathbb{R}^d$, which is uniquely determined by the coefficients of the method, as

$$y_{k+1} = \Psi_h(y_k), \quad k = 0, 1, \dots. \tag{2.4}$$

In order to provide a probabilistic interpretation of the numerical solution rather than a series of point values, Conrad et al. propose the scheme defined by

$$Y_{k+1} = \Psi_h(Y_k) + \xi_k(h), \quad k = 0, 1, \ldots, \tag{2.5}$$

where Y_k is a random variable approximating $y(t_k)$ with $Y_0 = y_0$, and $\xi_k(h)$ are appropriately scaled independent and identically distributed (i.i.d.) random variables with values in $\mathbb{R}^d$. Maintaining the same notation as in (2.5), in this work we propose a random time-stepping Runge–Kutta method (RTS-RK), i.e., the scheme defined defined by the recurrence relation

$$Y_{k+1} = \Psi_{H_k}(Y_k), \quad k = 0, 1, \ldots, \tag{2.6}$$

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. 2.2. Let us finally remark that the sequence Y_k, $k = 0, 1, \ldots$, form a homogeneous Markov chain, as the transition probability is independent of the index k.

Remark 2.1. We note that in terms of computational cost simulating the two methods (2.5) and (2.6) is equivalent.

2.2.1 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 2.2. The i.i.d. random variables H_k satisfy for all $k = 0, 1, \ldots$

1. $H_k > 0$ a.s.,
2. there exists $h > 0$ such that $\mathbb{E}\, H_k = h$,
3. there exist $p \geq 1/2$ and $C > 0$ independent of k such that the scaled random variables $Z_k := H_k - h$ satisfy
$$\mathbb{E}\, Z_k^2 = C h^{2p+1}. \tag{2.7}$$

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 2.3. Let us consider the random variables $\{H_k\}_{k\geq 0}$ such that

$$H_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(h - h^{p+1/2}, h + h^{p+1/2}), \quad 0 < h < 1, \quad p \geq 1/2. \tag{2.8}$$

We easily verify that the assumptions 1 and 2 are verified as $h < 1$, and that 3 is verified with

$C = 1/3$. Another choice of random variables could simply be

$$H_k \overset{\text{i.i.d.}}{\sim} \log\mathcal{N}\big(\log h - \log\sqrt{1+h^{2p}}, \log(1+h^{2p})\big), \tag{2.9}$$

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 2.4. The Runge–Kutta method defined by the numerical flow $\{\Psi_t\}_{t\geq 0}$ is of order q, i.e., for h small enough, there exists a constant $C > 0$ such that

$$\|\Psi_h(y) - \varphi_h(y)\| \leq Ch^{q+1}, \quad \forall y \in \mathbb{R}^d. \tag{2.10}$$

Remark 2.5. 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,

1. 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 (2.9), may force the solution outside D,
2. 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 2.3, which allow the time steps to be small enough almost surely. For the first issue, more sophisticated techniques of path rejection could be employed [45], but the mean-square convergence properties which will be examined in section 2.4 would not hold.

In order to tackle the second issue presented in the Remark above, we introduce a further assumption.

Assumption 2.6. 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 to allow 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.

2.3 Weak convergence analysis

The first property of the RTS-RK method we wish to analyze is its weak convergence, which gives an indication about the behavior of the numerical solution (2.6) in the mean sense. In the following, we denote by $\mathscr{C}^l_b(\mathbb{R}^d,\mathbb{R})$ the functions in $\mathscr{C}^l(\mathbb{R}^d,\mathbb{R})$ with all derivatives up to order l bounded uniformly in $\mathbb{R}^d$. Moreover, we consider the integration of (2.2) over the finite length domain $[0,T]$, where $T > 0$ is the final time. Let us define the weak order of convergence.

Definition 2.7. The numerical method (2.6) has weak order r for (2.2) if for any sufficiently smooth function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ there exists a constant $C > 0$ independent of h such that

$$|\mathbb{E}\,\Phi(Y_k) - \Phi(y(kh))| \le Ch^r, \tag{2.11}$$

for all $k = 1, 2, \dots, N$ and $T = Nh$.

Let us introduce the Lie derivative of the flow $\mathscr{L} = f \cdot \nabla$, which allows us to adopt the semi-group notation for the exact solution of (2.2) (see e.g. [30, Section III.5.1] or [54, Section 4.3]) and write for any smooth function Φ

$$\Phi(\varphi_h(y)) = e^{h\mathscr{L}}\Phi(y). \tag{2.12}$$

Moreover, let us recall that the probabilistic numerical solution $\{Y_k\}_{k\ge 0}$ forms a homogeneous Markov chain. Therefore, given $h > 0$ there exists an operator $\mathscr{P}_h$, the generator [51, Section 2.3], such that

$$\mathbb{E}\big(\Phi(Y_{k+1}) \mid Y_k = y\big) = (\mathscr{P}_h\Phi)(y). \tag{2.13}$$

In order to have an analogy with the notation (2.12), we adopt the exponential form of the infinitesimal generator and denote in the following $\mathscr{P}_h = e^{h\mathscr{L}_h}$, 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

$$\mathbb{E}\big(\Phi(Y_{k+1}) \mid Y_0 = y\big) = e^{h\mathscr{L}_h}\,\mathbb{E}\big(\Phi(Y_k) \mid Y_0 = y\big). \tag{2.14}$$

We can now state a result of local weak convergence of the probabilistic numerical solution.

Lemma 2.8 (Weak local order)**.** *Let assumption 2.2, assumption 2.4 and assumption 2.6 hold and let f in* (2.2) *be sufficiently smooth. If $\mathbb{E}|H_0^4| < \infty$, there exists a constant $C > 0$ independent of h and y such that for any function $\Phi \in \mathscr{C}_b^l(\mathbb{R}^d, \mathbb{R})$, with $l = \max\{q, 3\}$*

$$|\mathbb{E}(\Phi(Y_1) \mid Y_0 = y) - \Phi(\varphi_h(y))| \le Ch^{\min\{2p+1,\,q+1\}}. \tag{2.15}$$

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 Φ computed on the numerical solution as

$$\begin{aligned}
\Phi(Y_1) &= \Phi(\Psi_{H_0}(Y_0)) \\
&= \Phi\Big(\Psi_h(Y_0) + (H_0 - h)\partial_t\Psi_h(Y_0) + \frac{1}{2}(H_0 - h)^2\partial_{tt}\Psi_h(Y_0) + \mathscr{O}(|H_0 - h|^3)\Big) \\
&= \Phi(\Psi_h(Y_0)) + \Big((H_0 - h)\partial_t\Psi_h(Y_0) + \frac{1}{2}(H_0 - h)^2\partial_{tt}\Psi_h(Y_0)\Big)\cdot\nabla\Phi(\Psi_h(Y_0)) \\
&\quad + \frac{1}{2}(H_0 - h)^2\partial_t\Psi_h(Y_0)\partial_t\Psi_h(Y_0)^\top : \nabla^2\Phi(\Psi_h(Y_0)) + \mathscr{O}(|H_0 - h|^3),
\end{aligned} \tag{2.16}$$

where we denote by $\nabla^2\Phi$ the Hessian matrix of Φ, 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 2.2 we get

$$\begin{aligned} e^{h\mathscr{L}_h}\Phi(y)-\Phi(\Psi_h(y)) &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h(y)\cdot\nabla\Phi(\Psi_h(y)) \\ &\quad+\frac{1}{2}Ch^{2p+1}\partial_t\Psi_h(y)\partial_t\Psi_h(y)^\top:\nabla^2\Phi(\Psi_h(y))+\mathscr{O}(h^{3p+3/2}), \end{aligned} \tag{2.17}$$

where we exploited Hölder's inequality for the last term. Moreover, expanding Φ in y we get

$$\begin{aligned} \Phi(\Psi_h(y)) &= \Phi\left(\Psi_0(y)+h\partial_t\Psi_0(y)+\mathscr{O}(h^2)\right) \\ &= \Phi(y)+\mathscr{O}(h), \end{aligned} \tag{2.18}$$

which implies

$$\begin{aligned} e^{h\mathscr{L}_h}\Phi(y)-\Phi(\Psi_h(y)) &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h(y)\cdot\nabla\Phi(y) \\ &\quad+\frac{1}{2}Ch^{2p+1}\partial_t\Psi_h(y)\partial_t\Psi_h(y)^\top:\nabla^2\Phi(y)+\mathscr{O}(h^{2p+1}). \end{aligned} \tag{2.19}$$

Let us remark that due to the smoothness of the flow we have

$$e^{h\mathscr{L}}\Phi(y)-\Phi(\Psi_h(y))=\mathscr{O}(h^{q+1}). \tag{2.20}$$

Combining (2.20) and (2.19) we have the one-step weak error of the probabilistic method on the original ODE, i.e.,

$$e^{h\mathscr{L}}\Phi(y)-e^{h\mathscr{L}_h}\Phi(y)=\mathscr{O}(h^{\min\{2p+1,q+1\}}), \tag{2.21}$$

which proves the desired result. □

Remark 2.9. Let us remark that rigorously if $\partial_{tt}\Psi_h(y)$ is bounded independently of y then the equality (2.16) holds. In fact, as it can be noticed in (2.17), a 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 [19].

Assumption 2.10. The function f and the distribution of the random time steps H_k, $k=0,1,\ldots$, are such that the operator $e^{h\mathscr{L}_h}$ satisfies for all functions $\psi\in\mathscr{C}_b^q(\mathbb{R}^d,\mathbb{R})$ and a positive constant L,

$$\sup_{u\in\mathbb{R}^d}|e^{h\mathscr{L}_h}\psi(u)|\le(1+Lh)\sup_{u\in\mathbb{R}^d}|\psi(u)|, \tag{2.22}$$

where L may depend on f and on the distribution of the random time steps, but not on ψ or h.

Remark 2.11. Let us remark that in order for Ψ_h to satisfy assumption 2.4, i.e., for Ψ_h to be of order q, the right hand side f must be of class $\mathscr{C}_b^q(\mathbb{R}^d,\mathbb{R}^d)$ (see, e.g. [31, Theorem II.3.1]). Therefore, in order to apply the bound (2.22) to composite functions $\Phi\circ\varphi_h:\mathbb{R}^d\to\mathbb{R}$ where $\Phi\in\mathscr{C}_b^\infty(\mathbb{R}^d,\mathbb{R})$, by the chain rule we need assumption 2.10 to hold for functions in $C_b^q(\mathbb{R}^d,\mathbb{R})$. This fact will be exploited in the proof of theorem 2.13 below.

We now give a lemma useful for bounding discrete sequences, which is taken from [44, Lemma 1.6].

Lemma 2.12. *Suppose that for arbitrary N and $k = 0,\dots,N$ we have*

$$e_k \le (1 + Ah)e_{k-1} + Bh^r, \tag{2.23}$$

where $h = T/N$, $A > 0$, $B \ge 0$, $r \ge 1$ and $e_k \ge 0$, $k = 0,\dots,N$. Then

$$e_k \le e^{AT}e_0 + \frac{B}{A}(e^{AT} - 1)h^{r-1}. \tag{2.24}$$

The proof of lemma 2.12 follows from the discrete Grönwall inequality. We can now state the main result on weak convergence.

Theorem 2.13 (Weak order)**.** *Let the assumptions of lemma 2.8 and assumption 2.10 hold. Then, there exists a constant $C > 0$ independent of h and of the initial condition such that for all functions $\Phi \in \mathscr{C}_b^l(\mathbb{R}^d, \mathbb{R})$, with $l = \max\{q, 3\}$*

$$|\mathbb{E}\Phi(Y_k) - \Phi(y(kh))| \le Ch^{\min\{2p,q\}}, \tag{2.25}$$

for all $k = 1, 2, \dots, N$ and $T = Nh$.

Proof. Let us introduce the following notation

$$\begin{aligned} w_k(u) &= \Phi(\varphi_{t_k}(u)), \\ W_k(u) &= \mathbb{E}(\Phi(Y_k) \mid Y_0 = u). \end{aligned} \tag{2.26}$$

By the triangle inequality and the Markov property (2.14), we have

$$\begin{aligned} \sup_{u\in\mathbb{R}^d} |W_k(u) - w_k(u)| \le &\sup_{u\in\mathbb{R}^d} |e^{h\mathscr{L}} w_{k-1}(u) - e^{h\mathscr{L}_h} w_{k-1}(u)| \\ &+ \sup_{u\in\mathbb{R}^d} |e^{h\mathscr{L}_h} w_{k-1}(u) - e^{h\mathscr{L}_h} W_{k-1}(u)|. \end{aligned} \tag{2.27}$$

We then apply lemma 2.8 to the first term and assumption 2.10 to the second and denote $e_k := \sup_{u\in\mathbb{R}^d} |W_k(u) - w_k(u)|$, thus obtaining

$$e_k \le Ch^{\min\{2p+1,q+1\}} + (1 + Lh)e_{k-1}. \tag{2.28}$$

We can therefore apply lemma 2.12 with $A = L$ and $r = \min\{2p + 1, q + 1\}$, and therefore get for a constant $C > 0$

$$\sup_{u\in\mathbb{R}^d} |w_k(u) - W_k(u)| \le Ch^{\min\{2p,q\}}, \tag{2.29}$$

which is the desired result. □

Remark 2.14. In [19], 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 the Appendix.

Remark 2.15. 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 (2.25) is legitimate and does not induce time misalignment between true and numerical solutions. This basic property applies to all results in the following.

2.4 Mean square convergence analysis

The second property of (2.6) we analyze is its mean square order of convergence, which gives an indication on the path-wise distance between each realisation of the numerical solution and the exact solution of (2.2). Let us define the mean square order of convergence.

Definition 2.16. The numerical method (2.6) has mean square order of convergence r for (2.2) if there exists a constant $C > 0$ independent of h and of the initial condition y_0 such that

$$\left(\mathbb{E}\|Y_k - y(kh)\|^2\right)^{1/2} \leq Ch^r \tag{2.30}$$

for all $k = 1, 2, \ldots, N$ and $T = Nh$.

Remark 2.17. Let us remark that the mean square convergence is stronger than the traditional strong convergence, since, by Jensen's inequality

$$\mathbb{E}\,\|Y_k - y(kh)\| \leq \left(\mathbb{E}\|Y_k - y(kh)\|^2\right)^{1/2} \leq Ch^r. \tag{2.31}$$

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 2.18 (Local mean square convergence)**.** *Under assumption 2.2, assumption 2.4 and assumption 2.6 the numerical solution Y_1 given by one step of the RTS-RK method* (2.6) *satisfies*

$$\left(\mathbb{E}\|Y_1 - y(h)\|^2\right)^{1/2} \leq Ch^{\min\{p+1/2, q+1\}}, \tag{2.32}$$

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 triangular and Young's inequalities we have for all $y \in \mathbb{R}^d$

$$\mathbb{E}\|\Psi_{H_0}(y) - \varphi_h(y)\|^2 \leq 2\,\mathbb{E}\|\Psi_{H_0}(y) - \Psi_h(y)\|^2 + 2\|\Psi_h(y) - \varphi_h(y)\|^2. \tag{2.33}$$

We now consider assumption 2.4 and assumption 2.2, thus getting

$$\begin{aligned}\mathbb{E}\|\Psi_{H_0}(y)-\varphi_h(y)\|^2 &\le 2L_\Psi^2\mathbb{E}|H_0-h|^2+2C_1h^{2(q+1)}\\ &= 2L_\Psi^2C_2h^{2p+1}+2C_1h^{2(q+1)}\\ &\le C^2h^{2\min\{p+1/2,q+1\}},\end{aligned} \tag{2.34}$$

where C_1 and C_2 are the constants given in assumption 2.4 and assumption 2.2 respectively. This is the desired result with $C=\max\{2L_\Psi^2C_2,2C_1\}^{1/2}$. □

As a consequence of the one-step convergence, we can prove a result of global mean square convergence.

Theorem 2.19 (Global mean square convergence)**.** *Let f be globally Lipschitz and $t_k=kh$ for $k=1,2,\dots,N$, where $Nh=T$. Then, under the assumptions of lemma 2.18 the numerical solution given by* (2.6) *satisfies*

$$\sup_{k=1,2,\dots,N}\left(\mathbb{E}\|Y_k-y(t_k)\|^2\right)^{1/2}\le Ch^{\min\{p,q\}}, \tag{2.35}$$

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 lemmas.

Lemma 2.20. *Given the ODE* (2.2) *with f globally Lipschitz, then for any y and w in $\mathbb{R}^d$ and $0<h<1$ we have*

$$\|\varphi_h(y)-\varphi_h(w)\|\le(1+Ch)\|y-w\|, \tag{2.36}$$

$$\|\varphi_h(y)-\varphi_h(w)-(y-w)\|\le Ch\|y-w\|, \tag{2.37}$$

where C is a positive constant independent of h and of the initial condition y_0.

The proof of lemma 2.20 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 2.19. 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 $e_k^2:=\mathbb{E}\|Y_k-y(t_k)\|^2$. Adding and subtracting the exact flow applied to the numerical solution, we obtain

$$\begin{aligned}e_{k+1}^2=&\,\mathbb{E}\|\Psi_{H_k}(Y_k)-\varphi_h(Y_k)\|^2+\mathbb{E}\|\varphi_h(Y_k)-\varphi_h(y(t_k))\|^2\\ &+2\mathbb{E}\Big(\big(\varphi_h(Y_k)-\varphi_h(y(t_k))\big)^\top\big(\Psi_{H_k}(Y_k)-\varphi_h(Y_k)\big)\Big).\end{aligned} \tag{2.38}$$

Let us consider the three terms in (2.38) separately. For the first term, we have by lemma 2.18

$$\mathbb{E}\|\Psi_{H_k}(Y_k)-\varphi_h(Y_k)\|^2\le Ch^{\min\{2p+1,2(q+1)\}}. \tag{2.39}$$

For the second term, due to (2.36), we have

$$\mathbb{E}\|\varphi_h(Y_k) - \varphi_h(y(t_k))\|^2 \leq (1+Ch)^2 e_k^2. \tag{2.40}$$

Let us now define $Z = \varphi_h(Y_k) - \varphi_h(y(t_k)) - (Y_k - y(t_k))$. Then we can rewrite the inner product as

$$\begin{aligned}\mathbb{E}\Big(\big(\varphi_h(Y_k) - \varphi_h(y(t_k))\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) &= \mathbb{E}\Big(\big(Y_k - y(t_k)\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) \\ &\quad + \mathbb{E}\Big(Z^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big).\end{aligned} \tag{2.41}$$

We bound the two terms in (2.41) separately. For the first term, by the law of total expectation, we have

$$\begin{aligned}\mathbb{E}\Big(\big(Y_k - y(t_k)\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) &= \mathbb{E}\mathbb{E}\Big(\big(Y_k - y(t_k)\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big) \mid Y_k\Big) \\ &= \mathbb{E}\Big(\big(Y_k - y(t_k)\big)^\top \mathbb{E}\big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k) \mid Y_k\big)\Big).\end{aligned} \tag{2.42}$$

Applying Cauchy–Schwarz inequality to the outer expectation we get

$$\begin{aligned}\mathbb{E}\Big(\big(Y_k - y(t_k)\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) &\leq \Big(\mathbb{E}\|\mathbb{E}\big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k) \mid Y_k\big)\|^2\Big)^{1/2} e_k \\ &\leq Ch^{\min\{2p+1,q+1\}} e_k,\end{aligned} \tag{2.43}$$

where we applied lemma 2.8. We now consider the second term in (2.41). By the Cauchy–Schwarz inequality we have

$$\mathbb{E}\Big(Z^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) \leq \big(\mathbb{E}\|Z\|^2\big)^{1/2} \big(\mathbb{E}\|\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\|^2\big)^{1/2}. \tag{2.44}$$

We now apply (2.37) and lemma 2.18 to obtain

$$\mathbb{E}\Big(Z^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) \leq Ch^{\min\{p+3/2,q+2\}} e_k. \tag{2.45}$$

We can hence bound the scalar product in (2.41) with Young's inequality and assuming $h < 1$ as

$$\begin{aligned}\mathbb{E}\Big(\big(\varphi_h(Y_k) - \varphi_h(y(t_k))\big)^\top \big(\Psi_{H_k}(Y_k) - \varphi_h(Y_k)\big)\Big) &\leq Ch^{\min\{p+3/2,q+1\}} e_k \\ &\leq \frac{he_k^2}{2} + C\frac{h^{\min\{2p+2,2q+1\}}}{2}.\end{aligned} \tag{2.46}$$

Combining (2.39), (2.40) and (2.46), we have

$$e_{k+1}^2 \leq Ch^{\min\{2p+1,2q+1\}} + (1+Ch)e_k^2, \tag{2.47}$$

which implies the desired result by lemma 2.12 and since $e_0 = 0$. □

Remark 2.21. Let us remark that the difference between global and local orders of convergence, i.e., between (2.32) and (2.35), 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 2.22. As for the additive noise method proposed in [19], 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.

2.5 Mean square convergence of Monte Carlo estimators

The third property we analyze is the mean-square convergence of Monte Carlo estimators drawn from the random time-stepping Runge–Kutta method. Let us consider a function $\Phi \in \mathscr{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$ with Lipschitz constant L_Φ and a final time $T > 0$. Moreover, let us introduce the notation $Z = \Phi(y(T))$ and $Z_N = \mathbb{E}\,\Phi(Y_N)$, where N is such that $T = Nh$. In general, the quantity Z_N is not accessible, and we have to replace it by its Monte Carlo estimator

$$\widehat{Z}_{N,M} = M^{-1} \sum_{i=1}^{M} \Phi(Y_N^{(i)}). \tag{2.48}$$

where M is the number of realisations of the numerical solution and we denote by $\{Y_N^{(i)}\}_{i=1}^M$ a set of i.i.d. realisations of the numerical solution. Hence, we are interested in studying the mean square error of the Monte Carlo estimator, which is defined as

$$\mathrm{MSE}(\widehat{Z}_{N,M}) = \mathbb{E}(Z - \widehat{Z}_{N,M})^2. \tag{2.49}$$

In the following result, we prove that this quantity converges to zero independently of the number of trajectories M, in the limit $h \to 0$.

Theorem 2.23. *Under assumption 2.2, assumption 2.6 and assumption 2.4, the Monte Carlo estimator $\widehat{Z}_{N,M}$ satisfies*

$$\mathrm{MSE}(\widehat{Z}_{N,M}) \leq C\Big(h^{2\min\{2p,q\}} + \frac{h^{2\min\{p,q\}}}{M}\Big), \tag{2.50}$$

where C is a positive constant independent of h and M.

Proof. Thanks to the classic decomposition of the MSE, we have

$$\mathrm{MSE}(\widehat{Z}_{N,M}) = \mathrm{Var}\,\widehat{Z}_{N,M} + \big(\mathbb{E}(\widehat{Z}_{N,M} - Z)\big)^2. \tag{2.51}$$

Due to the unbiasedness of the Monte Carlo estimator $\widehat{Z}_{N,M}$ and applying theorem 2.13 to the second term, we have

$$\mathrm{MSE}(\widehat{Z}_{N,M}) \leq \mathrm{Var}\,\widehat{Z}_{N,M} + Ch^{2\min\{2p,q\}}. \tag{2.52}$$

The variance of the estimator can be trivially bounded by exploiting the Lipschitz continuity

of Φ and the independence of the samples as

$$\begin{aligned} \operatorname{Var} \widehat{Z}_{N,M} &= M^{-1} \operatorname{Var}\big(\Phi(Y_N)\big) \\ &\leq M^{-1} \mathbb{E}\big(\Phi(Y_N) - \Phi(y(T))\big)^2 \\ &\leq M^{-1} L_\Phi^2 \mathbb{E}\|Y_N - y(T)\|^2. \end{aligned} \tag{2.53}$$

Applying theorem 2.19 we get

$$\operatorname{Var} \widehat{Z}_{N,M} \leq M^{-1} L_\Phi^2 C h^{2\min\{p,q\}}, \tag{2.54}$$

which proves the desired result. □

Let us remark that with the choice $p = q$, which is the minimum p for which the order of convergence of the underlying deterministic method is not affected by the probabilistic setting, we have $\mathrm{MSE}(\widehat{Z}_{N,M}) \leq C h^{2q}$ with $M = 1$. Hence, the Monte Carlo estimators drawn from (2.6) converge in the mean square sense independently of the number of samples M in (2.48). In the sub-optimal case $p < q$, one should carefully select the number of trajectories M so that the two terms in (2.50) are balanced. In particular, this would lead to

$$M = \begin{cases} \mathcal{O}(1), & \text{if } p \geq q, \\ \mathcal{O}(h^{2(p-q)}), & \text{if } p < q \leq 2p, \\ \mathcal{O}(h^{-2p}), & \text{if } 2p < q, \end{cases} \tag{2.55}$$

where the notation $M = \mathcal{O}(h^r)$ for a real number r means that there exist constants C_1 and C_2 such that $C_1 h^r \leq M \leq C_2 h^r$.

Remark 2.24. Let us remark that in order to have uncertainty quantification for a fixed value $h > 0$ it is necessary to draw a sample with $M > 1$, since otherwise the probability distribution over the numerical solution would be a Dirac delta. theorem 2.23 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.

2.6 Conservation of first integrals

Numerical methods for ODEs are often studied in terms of their geometric properties [30]. In particular, we investigate here whether the random choice of time steps in (2.6) spoils the properties of the underlying deterministic Runge–Kutta method. Let us recall the definition of first integral for an ODE.

Definition 2.25. Given a function $I\colon \mathbb{R}^d \to \mathbb{R}$, then $I(y)$ is a first integral of (2.2) 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 2.26. 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 $H_0 > 0$

$$I(Y_1) = v^\top y_0 + H_0 \sum_{i=1}^s b_i v^\top f(y_0 + H_0 \sum_{j=1}^s a_{ij} K_j), \tag{2.56}$$

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 $n \geq 3$ [30, 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 2.27. *Let $I(y)$ be a first integral for* (2.2) *and Ψ_h be the numerical flow of a Runge–Kutta scheme for* (2.2)*. If the scheme defined by Ψ_h conserves $I(y)$ for any $h > 0$, then the numerical method* (2.6) *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 [30, 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 [30, 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 2.28. *If the Runge–Kutta scheme defined by Ψ_h conserves quadratic first integrals then the numerical method* (2.6) *conserves quadratic first integrals almost surely.*

Proof. This result is a direct consequence of theorem 2.27. □

The properties above for the RTS-RK method are not satisfied by the additive noise method presented in [19]. 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 (2.6) conserves linear first integrals in the strong sense. For the additive noise numerical method (2.5), we have

$$\begin{aligned} I(Y_1) &= v^\top y_0 + h\textstyle\sum_{i=1}^s b_i v^\top f(y_0 + h\sum_{j=1}^s a_{ij} K_j) + v^\top \xi_0(h), \\ &= v^\top (y_0 + \xi_0(h)). \end{aligned} \tag{2.57}$$

If the random variable ξ_0 is zero-mean, then $\mathbb{E}\, I(Y_1) = I(y_0)$ and iteratively along the solution $\mathbb{E}\, I(Y_k) = I(y_0)$. 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

$$\begin{aligned} I(Y_1) &= (\Psi_h(y_0) + \xi_0(h))^\top S(\Psi_h(y_0) + \xi_0(h)) \\ &= I(y_0) + 2\xi_0(h)^\top S\Psi_h(y_0) + \xi_0(h)^\top S\xi_0(h). \end{aligned} \tag{2.58}$$

If the random variables are zero-mean and if there exists a matrix Q such that $\mathbb{E}\,\xi_0(h)\xi_0(h)^\top = Qh^{2p+1}$ for some $p \geq 1$ (Assumption 1 in [19]) we then have

$$\mathbb{E}\, I(Y_1) = I(y_0) + Q : Sh^{2p+1}. \tag{2.59}$$

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 2.27 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.

2.7 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

$$y' = J^{-1}\nabla Q(y), \quad y(0) = y_0 \in \mathbb{R}^{2d}, \tag{2.60}$$

where the matrix $J \in \mathbb{R}^{2d\times 2d}$ is defined as

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{2.61}$$

and where I is the identity matrix in $\mathbb{R}^{d\times d}$. The Hamiltonian Q is a first integral for (2.60), 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 2.29 (Definition VI.2.2 in [30])**.** 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

$$(g')^\top J g' = J. \tag{2.62}$$

It is well-known that the flow $\varphi_t\colon \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ of any system of the form (2.60) 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 [30, 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 (2.60).

2.7.1 Symplecticity of the RTS-RK method

It has been pointed out [30, Section VIII.1] that applying an adaptive step size technique to a symplectic method can destroy its symplecticity. Therefore, Skeel and Gear [59] 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

$$V^\top J V = J, \quad V = \partial_y \Psi_{\tau(y,h)}(y) + \partial_t \Psi_{\tau(y,h)}(y) \partial_y \tau(y,h)^\top. \tag{2.63}$$

Let us now consider the RTS-RK method based on a symplectic deterministic integrator. We have the following lemma.

Lemma 2.30. *If the flow Ψ_h of the deterministic integrator is symplectic, then the flow of the random time-stepping probabilistic method* (2.6) *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 2.2. Hence, τ is independent of y, i.e., $\partial_y \tau(y,h) = 0$, and with the notation introduced above

$$V = \partial_y \Psi_{\tau(h)}(y). \tag{2.64}$$

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.

2.7.2 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 $\mathbb{E}|Q(Y_n) - Q(y_0)|$ 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 [30, 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. [30, Section IX.7].

Assumption 2.31. 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

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \dots, \tag{2.65}$$

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.,

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \dots + h^{N-1} f_N(y). \tag{2.66}$$

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. [30, Theorem IX.7.6])

$$\|\widetilde{\varphi}_h(y) - \Psi_h(y)\| \leq Che^{-\kappa/h}, \tag{2.67}$$

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. [30, Section IX.8]) that for a Hamiltonian system (2.60) and a symplectic integrator the modified equation is still a Hamiltonian system, i.e., there exists a modified Hamiltonian $\widetilde{Q}$ defined as

$$\widetilde{Q}(y) = Q(y) + h^q Q_{q+1}(y) + \dots + h^{N-1} Q_N(y), \tag{2.68}$$

such that $\widetilde{f} = J^{-1}\nabla\widetilde{Q}$. The estimate (2.67) implies that the modified Hamiltonian is almost conserved by the symplectic integrator. In particular, if Q is Lipschitz, we have

$$|\widetilde{Q}(\Psi_h(y)) - \widetilde{Q}(y)| \leq Che^{-\kappa/h}. \tag{2.69}$$

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. [30, Theorem IX.8.1.] or [11].

Theorem 2.32. *Under assumption 2.31 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*

$$\begin{aligned} \widetilde{Q}(y_n) &= \widetilde{Q}(y_0) + \mathcal{O}(e^{-\kappa/2h}), \\ Q(y_n) &= Q(y_0) + \mathcal{O}(h^q). \end{aligned} \tag{2.70}$$

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 2.30, 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

$$\widehat{Q}_j(y) = Q(y) + H_j^q Q_{q+1}(y) + \ldots + H_j^{N-1} Q_N(y). \tag{2.71}$$

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

$$\eta_j := \widehat{Q}_j(\Psi_{H_j}(y)) - \widehat{Q}_j(y), \tag{2.72}$$

which, in light of (2.69), satisfy

$$|\eta_j| \leq CH_j e^{-\kappa/H_j}, \tag{2.73}$$

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 2.33. There exists $\bar{r} > 1$ such that for any $1 < r < \bar{r}$, the random time steps $\{H_j\}_{j\geq 0}$ satisfy

$$\mathbb{E}\, H_j^r = h^r + C_r h^{2p+r-1}, \tag{2.74}$$

where p is defined in assumption 2.2 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

$$\begin{aligned}\mathbb{E}(H_j^r - h^r) &= C_r h^{2p+r-1},\\ \mathbb{E}(H_j^r - h^r)^2 &= (C_{2r} - 2C_r) h^{2p+2r-1}.\end{aligned} \tag{2.75}$$

Then, for any $r, s > 1$ such that $r + s < R$, it holds

$$\begin{aligned}\mathbb{E}(H_j^{r+s} - h^{r+s}) &= \widehat{C}_{r,s} h^s \mathbb{E}(H_j^r - h^r),\\ \mathbb{E}(H_j^{r+s} - h^{r+s})^2 &= \widetilde{C}_{r,s} h^{2s} \mathbb{E}(H_j^r - h^r)^2,\end{aligned} \tag{2.76}$$

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 2.33 is satisfied for the uniform random time steps $H_j \overset{\text{i.i.d.}}{\sim} \mathscr{U}(h - h^{p+1/2}, h + h^{p+1/2})$ introduced in example 2.3. Let us now prove a bound on the random variables η_j defined in (2.72).

Lemma 2.34. *Suppose that assumption 2.2, assumption 2.6 and assumption 2.33 hold true, and suppose that* $0 < h \leq 1$. *Then the random variables* η_j *satisfy*

$$\mathbb{E}|\eta_j|^r \leq C h^{\min\{r, p+r-3/2\}} e^{-r\kappa/(Mh)}, \tag{2.77}$$

where $C > 0$ *is independent of* h *and for all* $r \in \mathbb{N}$ *with* $r \geq 1$.

Proof. The proof is given in the Appendix. □

Let us furthermore introduce two lemmas, which will be employed for proving long-time conservation of Hamiltonians. Let us remark that in lemma 2.35 the values n, q, N indicate generic positive integers.

Lemma 2.35. *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*

$$\Big(\sum_{j=0}^{n-1}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)\Big)^2 = \sum_{j=0}^{n-1} a_{jq}^2 + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} a_{jq}a_{iq} + R(a) + S(a,b), \tag{2.78}$$

where the remainder $R(a)$ *can be written as* $R = R_1 + R_2 + R_3$, *with*

$$\begin{aligned}R_1(a) &= \sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1} a_{jk}^2, \qquad & R_2(a) &= 2\sum_{j=0}^{n-1}\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl},\\ R_3(a) &= 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\sum_{k=q}^{N-1}\sum_{\substack{l=q\\ l+k>2q}}^{N-1} a_{jk}a_{il}, & &\end{aligned} \tag{2.79}$$

and the remainder $S(a,b)$ can be written as $S = S_1 + S_2 + S_3 + S_4$, with

$$\begin{aligned} S_1(a,b) &= \sum_{j=0}^{n-1} b_j^2, & S_2(a,b) &= 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} b_i b_j, \\ S_3(a,b) &= 2\sum_{j=1}^{n-1}\sum_{k=q}^{N-1} b_j a_{jk}, & S_4(a,b) &= 2\sum_{j=1}^{n-1}\sum_{i=0}^{n-1}\Big(b_j\sum_{k=q}^{N-1} a_{ik} + b_i\sum_{k=q}^{N-1} a_{jk}\Big). \end{aligned} \tag{2.80}$$

Proof. The proof is given in the Appendix. □

Lemma 2.36. *Let assumption 2.2 hold with $p \geq 3/2$ and $h < 1$, and let assumption 2.4, assumption 2.31 and assumption 2.33 hold. Moreover, let q be specified in assumption 2.4 and N be the truncation index of the modified right hand side* (2.66)*. Let us consider the sets of real-valued random variables $\Delta := \{\Delta_{j,k}(H_j^k - h^k), j = 0,\dots,n-1,\ k = q,\dots,N-1\}$, where $\Delta_{j,k} := Q_{k+1}(Y_j) - Q_{k+1}(Y_{j+1})$ and $\eta := \{\eta_j, j = 0,\dots,n-1\}$. Then, with the notation of lemma 2.35, there exist positive constants C_1, C_2 independent of h and n, but possibly dependent on q and N, such that*

$$\begin{aligned} \mathbb{E}\,R(\Delta) &\leq C_1\big(t_n h^{2(p+q+1/2)} + t_n^2 h^{2(2p+q-1/2)}\big), \\ \mathbb{E}\,S(\Delta,\eta) &\leq C_2\big((t_n h + t_n^2)e^{-2\kappa/(Mh)} + (t_n h^{p+q+1/2} + t_n^2 h^{2p+q-1})e^{-\kappa/(Mh)}\big), \end{aligned} \tag{2.81}$$

where $t_n = nh$.

Proof. The proof is given in the Appendix. □

It is now possible to prove a result of long conservation of the Hamiltonian for symplectic RTS-RK methods.

Theorem 2.37. *Let $0 < h \leq 1$. Suppose that assumption 2.2 holds for $p \geq 3/2$, that assumption 2.6 and assumption 2.31 hold, and that assumption 2.33 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*

$$\mathbb{E}|Q(Y_n) - Q(y_0)| \leq Ch^q, \tag{2.82}$$

for time intervals of length

$$t_n = \mathcal{O}\big(\min\{h^{1-2p}, e^{\kappa/(4Mh)}h^{-(2p+2q-1)/4}, e^{\kappa/(2Mh)}\}\big) \tag{2.83}$$

where p is given in assumption 2.2 and M in assumption 2.33.

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

$$\widetilde{Q}(Y_n) - \widetilde{Q}(y_0) = \sum_{j=0}^{n-1} \big(\widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j)\big). \tag{2.84}$$

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

$$\begin{aligned} \widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j) &= \widetilde{Q}(Y_{j+1}) - \widehat{Q}_j(Y_{j+1}) + \widehat{Q}_j(Y_{j+1}) - \widetilde{Q}(Y_j) \\ &= \widetilde{Q}(Y_{j+1}) - \widehat{Q}_j(Y_{j+1}) + \widehat{Q}_j(Y_j) - \widetilde{Q}(Y_j) + \eta_j. \end{aligned} \tag{2.85}$$

Hence, by applying the definition (2.68) of $\widetilde{Q}$ and (2.71) of $\widehat{Q}_j$, we get

$$\widetilde{Q}(Y_{j+1}) - \widetilde{Q}(Y_j) = \sum_{k=q}^{N-1} (H_j^k - h^k)\Delta_{j,k} + \eta_j, \tag{2.86}$$

where $\Delta_{j,k}$ is defined in lemma 2.36. Going back to (2.84), applying Jensen's inequality and lemma 2.35 we obtain

$$\begin{aligned} \big(\mathbb{E}|\widetilde{Q}(Y_n) - \widetilde{Q}(y_0)|\big)^2 &\le \mathbb{E}\Big(\sum_{j=0}^{n-1}\Big(\sum_{k=q}^{N-1}(H_j^k - h^k)\Delta_{j,k} + \eta_j\Big)\Big)^2 \\ &= \sum_{j=0}^{n-1} \mathbb{E}\big((H_j^q - h^q)^2\Delta_{j,q}^2\big) \\ &\quad + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} \mathbb{E}\big((H_j^q - h^q)\Delta_{j,q}(H_i^q - h^q)\Delta_{i,q}\big) + \mathbb{E}\,R(\Delta) + \mathbb{E}\,S(\Delta,\eta). \end{aligned} \tag{2.87}$$

The first term in (2.87) satisfies

$$\Big(\sum_{j=0}^{n-1} \mathbb{E}\big((H_j^q - h^q)^2\Delta_{j,q}^2\big)\Big)^{1/2} \le C\sqrt{t_n}h^{p+q}, \tag{2.88}$$

due to (2.158). Now, considering (2.165), we obtain that the second term in (2.87) satisfies

$$\Big(2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} \mathbb{E}\big((H_j^q - h^q)\Delta_{j,q}(H_i^q - h^q)\Delta_{i,q}\big)\Big)^{1/2} \le C t_n h^{2p+q-1}. \tag{2.89}$$

For the remainder term $\mathbb{E}\,R(\Delta)$, due to lemma 2.36 we get

$$(\mathbb{E}\,R(\Delta))^{1/2} \le C\big(\sqrt{t_n}h^{p+q+1/2} + t_n h^{2p+q-1/2}\big). \tag{2.90}$$

For the remainder term $\mathbb{E}\,S(\Delta,\eta)$, due to lemma 2.36 and since $h \le 1$ and $p \ge 3/2$ by assumption,

we get

$$\begin{aligned}(\mathbb{E}\,S(\Delta,\eta))^{1/2} &\le C\Big(t_n^2\big(e^{-2\kappa/(Mh)}+h^{p+q+1/2}e^{-\kappa/(Mh)}\big)\Big)^{1/2}\\ &\le C\,t_n\big(e^{-\kappa/(Mh)}+h^{(2p+2q+1)/4}e^{-\kappa/(2Mh)}\big).\end{aligned} \tag{2.91}$$

Finally, taking the square root of both sides of (2.87), replacing the expressions we obtained above and since $h \le 1$, we get that the modified Hamiltonian satisfies

$$\mathbb{E}|\widetilde{Q}(Y_n)-\widetilde{Q}(y_0)| \le C\Big(\sqrt{t_n}h^{p+q}+t_n h^{2p+q-1}+t_n\big(e^{-\kappa/(Mh)}+h^{(2p+2q+1)/4}e^{-\kappa/(2Mh)}\big)\Big). \tag{2.92}$$

Hence, imposing for a constant $C>0$

$$t_n \le C\min\{h^{1-2p}, e^{\kappa/(4Mh)}h^{-(2p+2q-1)/4}, e^{\kappa/(2Mh)}\}, \tag{2.93}$$

and since exponential terms are dominated by polynomial terms (see e.g. [30, Theorem IX.8.1]), we obtain

$$\mathbb{E}|\widetilde{Q}(Y_n)-\widetilde{Q}(y_0)| \le Ch^q. \tag{2.94}$$

Finally, applying the triangle inequality, since for all $y\in\mathbb{R}^d$ it holds $|Q(y)-\widetilde{Q}(y)|\le Ch^q$ by definition of the modified Hamiltonian $\widetilde{Q}$ and due to (2.94) we get

$$\begin{aligned}\mathbb{E}|Q(Y_n)-Q(y_0)| &\le \mathbb{E}|Q(Y_n)-\widetilde{Q}(Y_n)|+\mathbb{E}|Q(y_0)-\widetilde{Q}(y_0)|+\mathbb{E}|\widetilde{Q}(Y_n)-\widetilde{Q}(y_0)|\\ &\le Ch^q,\end{aligned} \tag{2.95}$$

which is the desired result. □

Remark 2.38. The result of theorem 2.37 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 2.33 arbitrarily close to 1 and we have

$$\mathbb{E}|Q(Y_n)-Q(y_0)| = \mathcal{O}(h^q), \tag{2.96}$$

for exponentially long time spans $t_n = \mathcal{O}\big(e^{\kappa/(2h)}\big)$, which is consistent with the theory of deterministic symplectic integrators summarised by theorem 2.32.

Remark 2.39. It has been observed (see for example [29, 30]) 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 2.37 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 2.40. As it can be noticed in the proof of lemma 2.36, we introduce the assumption $p\ge 3/2$ in order to simplify the terms composing the remainder $S(\Delta,\eta)$. In case $1\le p<3/2$, 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 2.41. Let us remark that in order for (2.91) 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 (2.88) to (2.91) as $t_n \leq 1$, and the desired result would still hold.

2.8 Bayesian inference

It has been recently shown [19, 15, 41] that probabilistic methods for ordinary and partial differential equations guarantee robust results (with respect to the numerical discretization error) in the context of Bayesian inverse problems. In this section, we briefly introduce a Bayesian inverse problem in the ODE setting and illustrate how the RTS-RK method can be employed in this framework.

Let us consider a function $f_\vartheta \colon \mathbb{R}^d \to \mathbb{R}^d$ which depends on a real parameter $\vartheta \in \Theta$, where Θ is an open subset of $\mathbb{R}^n$ and the ODE

$$y'_\vartheta = f_\vartheta(y), \quad y_\vartheta(0) = y_0 \in \mathbb{R}^d. \tag{2.97}$$

In order to simplify the notation, we consider y_0 to be a fixed initial condition. In general, y_0 could depend itself on ϑ. In the classical setting of numerical analysis, the main problem of interest is to determine the solution y_ϑ given the parameter ϑ. The inverse problem we consider is instead to determine ϑ through observations of the solution y_ϑ (or quantities derived from it). In the Bayesian setting, the inverse problem is recast in terms of probability distributions, and the goal is to establish a probability measure on ϑ, known as the posterior measure, given observed data and a probability measure, known as the prior, which captures all knowledge on the parameter available beforehand.

Let us denote by $z \in \mathbb{R}^m$ the observable and by $\mathcal{G} \colon \Theta \to \mathbb{R}^m$ the forward operator, which can be written as $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$, where $\mathcal{S}$ is the solution operator and $\mathcal{O}$ is the observation operator. In this case, $\mathcal{S} \colon \mathbb{R}^n \to \mathcal{C}([0,T])$ is the operator mapping ϑ into the solution y_ϑ, and $\mathcal{O} \colon \mathcal{C}([0,T]) \to \mathbb{R}^m$ maps the solution into the observable. Observations are then given by evaluations of the forward model corrupted by noise. In particular, we model noise as a Gaussian random variable $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ independent of ϑ, so that observations read

$$z = \mathcal{G}(\vartheta) + \varepsilon. \tag{2.98}$$

Under these assumptions, the likelihood of the observations can be written as

$$\pi(z \mid \vartheta) = e^{-V_z(\vartheta)}, \tag{2.99}$$

where the function $V_z \colon \Theta \to \mathbb{R}$, called the potential or negative log-likelihood, is given by

$$V_z(\vartheta) = \frac{1}{2}\big(\mathcal{G}(\vartheta) - z\big)^\top \Sigma_\varepsilon^{-1} \big(\mathcal{G}(\vartheta) - z\big). \tag{2.100}$$

The second building block of Bayesian inverse problems is the prior distribution, which we denote by $\pi_0(\vartheta)$. The prior encodes all the knowledge on the parameter that is known before observations are provided. In the following, we adopt a common abuse of notation, confounding measures and their probability density function.

Once the likelihood model and the prior distribution are established, it is possible to compute the posterior distribution $\pi(\vartheta \mid z)$ via Bayes' theorem, i.e.,

$$\pi(\vartheta \mid z) = \frac{\pi(z \mid \vartheta)\pi_0(\vartheta)}{\mathcal{Z}(z)}, \tag{2.101}$$

where $\mathcal{Z}(z)$ is the normalising constant given by

$$\mathcal{Z}(z) = \int_\Theta \pi(z \mid \vartheta)\pi_0(\vartheta)\, \mathrm{d}\vartheta. \tag{2.102}$$

Let us denote by $\mathcal{G}^h(\vartheta)$ the forward model where the solution operator is approximated by a Runge–Kutta method with time step h, and consequently with $V_z^h(\vartheta)$ and $\pi^h(z \mid \vartheta)$ the potential and the likelihood function obtained replacing $\mathcal{G}(\vartheta)$ with $\mathcal{G}^h(\vartheta)$. We can then define analogously the approximated posterior distribution $\pi^h(\vartheta \mid z)$ via Bayes' formula. In the following, we assume that the posteriors $\pi(\vartheta \mid z)$ and $\pi^h(\vartheta \mid z)$ are absolutely continuous with respect to the Lebesgue density. In [61, Theorem 4.6], Stuart proves that the posterior distribution $\pi^h(\vartheta \mid z)$ converges to $\pi(\vartheta \mid z)$ with respect to h with the same rate as $V_z^h(\vartheta)$ converges to $V_z(\vartheta)$. There, convergence is shown with respect to the Hellinger distance for a Gaussian prior, which is defined for probability density functions which are absolutely continuous with respect to the Lebesgue density as

$$d_{\mathrm{Hell}}\big(\pi^h(\vartheta \mid z), \pi(\vartheta \mid z)\big)^2 = \frac{1}{2}\int_\Theta \left(\sqrt{\pi^h(\vartheta \mid z)} - \sqrt{\pi(\vartheta \mid z)}\right)^2 \mathrm{d}\vartheta. \tag{2.103}$$

Hence, when there is no restriction in computational resources and it is possible to choose h small, the approximated posterior distribution can be made arbitrarily close to the true posterior. The result is proved in [61] under the hypothesis of a Gaussian prior, but can be extended to a wider class of thin-tailed priors as done in [24] and to heavy-tailed priors as done in [62].

In this work we consider the case when h is fixed, and in particular we are interested in the case where the numerical error dominates the noise contribution. It has been shown via examples in [19, 17] that in this small noise limit the approximated posterior distributions can be overly confident on the value of the parameter. In particular, the expectation of ϑ computed under the posterior distribution exhibits a bias with respect to the true value, which is not highlighted by the dispersion of the posterior itself. This undesirable phenomenon can be corrected by means of a probabilistic method, as the one presented by Conrad et al. in [19] or the RTS-RK method, to approximate the potential $V_z(\vartheta)$. Let us denote by $\xi \in \mathcal{X}$ the auxiliary random variable introduced by the probabilistic method. In the case of RTS-RK, we have $\xi = (H_0, H_1, \dots, H_{N-1})^\top$ and $\mathcal{X} \subset \mathbb{R}_+^N$. The likelihood function, denoted as $\pi_{\mathrm{pr}}^h(z \mid \vartheta)$ is

then defined by

$$\pi^h_{\text{prob}}(z \mid \vartheta) = \mathbb{E}^{\xi}\, e^{-V_z^{h,\xi}(\vartheta)}. \tag{2.104}$$

where $V_z^{h,\xi}$ is the approximation of the potential function given by the probabilistic method. The corresponding posterior distribution π^h_{pr} is then defined by

$$\pi^h_{\text{prob}}(\vartheta \mid z) = \frac{\pi^h_{\text{prob}}(z \mid \vartheta)\pi_0(\vartheta)}{\mathbb{E}^{\xi}\, \mathcal{Z}^{h,\xi}(z)}, \tag{2.105}$$

where the normalising constant is given by $\mathbb{E}^{\xi}\, \mathcal{Z}^{h,\xi}(z)$, where

$$\mathcal{Z}^{h,\xi}(z) = \int_{\Theta} e_z^{-V^{h,\xi}}(\vartheta)\pi_0(\vartheta)\,\mathrm{d}\vartheta. \tag{2.106}$$

Modifying the posterior in this manner allows to obtain qualitatively better results, which account for the uncertainty introduced by the numerical solver. Moreover, this posterior distribution still converges to the true posterior for $h \to 0$ as proved in [41], where (2.105) is called the marginal posterior.

In order to sample from the posteriors defined above we employ Markov chain Monte Carlo (MCMC) algorithms. In particular, due to the manner in which the probabilistic posterior (2.105) is defined, the pseudo-marginal Metropolis–Hastings (PMMH) algorithm [9] is a suitable choice for sampling. We note that in case of a deterministic approximation of the forward model, the standard random walk Metropolis–Hastings can be employed.

2.8.1 Analytical posteriors in a linear problem

If the forward operator $\mathcal{G}$ is linear, the prior on the unknown parameter is Gaussian, and the negative log-likelihood is given by (4.2), then there is an explicit formula for the corresponding posterior distribution. Let us hence consider the following one dimensional ODE

$$y'(t) = -y(t), \quad y(0) = \vartheta. \tag{2.107}$$

Given $h > 0$, we consider the inferential problem of determining the true initial condition ϑ^* from a single observation $z = \varphi_h(\vartheta^*) + \varepsilon$, where $\varphi_h(\vartheta^*) = \vartheta^* e^{-h}$ is the true solution at time $t = h$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a source of noise. In this case, the parameter space is $\Theta = \mathbb{R}$ and the forward operator $\mathcal{G}$ is defined by $\mathcal{G}\colon \mathbb{R} \to \mathbb{R}$, $\mathcal{G}\colon \vartheta \mapsto \vartheta e^{-h}$. In the following, we verify heuristically the convergence of the posterior distributions obtained with deterministic and probabilistic integrators with respect to a vanishing noise scale. If a Gaussian prior $\pi_0 = \mathcal{N}(0,1)$ is given for ϑ, the true posterior distribution is computable analytically and is given by

$$\pi(\vartheta \mid z) = \mathcal{N}\Big(\vartheta; \frac{ze^{-h}}{\sigma^2 + e^{-2h}}, \frac{\sigma^2}{\sigma^2 + e^{-2h}}\Big), \tag{2.108}$$

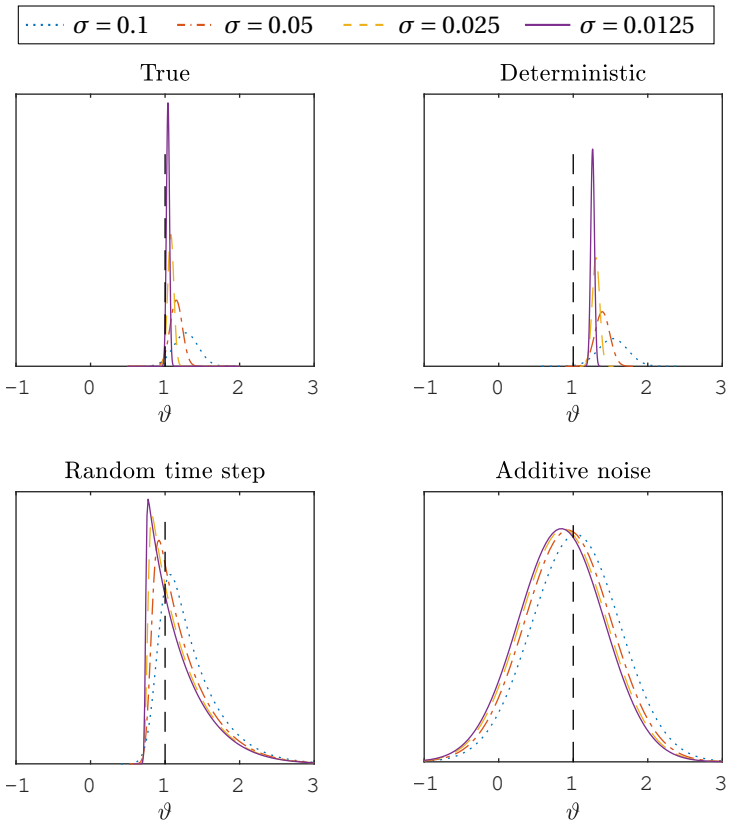

Figure 2.3 – Analytical posterior distributions in the linear case of section 2.8.1 for the true solution and its approximations with the deterministic explicit Euler method and the two probabilistic versions with additive noise (2.5) and with random time steps (2.6). In this case, $h = 0.5$ and the 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.

where $\mathcal{N}(x; \mu, \alpha^2)$ is the density of a Gaussian random variable of mean μ and variance α^2 evaluated in x. Consistently, if $\sigma^2 \to 0$, we have that $z \to \vartheta^* e^{-h}$ and therefore $\pi(\vartheta \mid z) \to \delta_{\vartheta^*}$.

If we approximate $\varphi_h(\vartheta)$ for a given initial condition ϑ with a single step of the explicit Euler method (i.e., with step size h), we get $\Psi_h(\vartheta) = (1-h)\vartheta$. Computing the posterior distribution obtained with this approximation leads to

$$\pi^h(\vartheta \mid z) = \mathcal{N}\left(\vartheta; \frac{(1-h)z}{\sigma^2 + (1-h)^2}, \frac{\sigma^2}{\sigma^2 + (1-h)^2}\right). \tag{2.109}$$

In the limit of $\sigma^2 \to 0$, we get in this case that the posterior distribution tends to $\pi^h(\vartheta \mid z) \to \delta_{\bar{\vartheta}}$, where $\bar{\vartheta} = e^{-h}\vartheta^*/(1-h)$. The posterior distribution is hence tending to a biased Dirac delta with respect to the true value.

Let us consider the additive noise method (2.5) applied to the explicit Euler method, i.e., the

random approximation $y(h) \approx Y_1$, where $Y_1 = (1-h)\vartheta + \xi$ and where $\xi \sim \mathcal{N}(0, h^3)$, so that the method converges consistently with the deterministic method. In this case, the posterior distribution that we denote by $\pi^h_{\text{prob,AN}}$ is given by

$$\pi^h_{\text{prob,AN}}(\vartheta \mid z) = \mathcal{N}\Big(\vartheta; \frac{(1-h)z}{\widetilde{\sigma}^2 + (1-h)^2}, \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}^2 + (1-h)^2}\Big). \tag{2.110}$$

where $\widetilde{\sigma}^2 = \sigma^2 + h^3$. Hence, taking the limit $\sigma^2 \to 0$ gives

$$\pi^h_{\text{prob,AN}}(\vartheta \mid z) \to \mathcal{N}\Big(\vartheta; \frac{(1-h)e^{-h}\vartheta^*}{h^3 + (1-h)^2}, \frac{h^3}{h^3 + (1-h)^2}\Big), \tag{2.111}$$

which shows that while the asymptotic mean is still biased with respect to the true value, the uncertainty in the forward model is reflected by a positive variance. Let us now consider the random time step explicit Euler with step size distribution $H \sim \mathcal{U}(h - h^{p+1/2}, h + h^{p+1/2})$. In this case, the forward model is given by

$$Y_1 = \vartheta - H\vartheta = (1-h)\vartheta + U\vartheta, \quad U \sim \mathcal{U}(-h^{p+1/2}, h^{p+1/2}). \tag{2.112}$$

Hence, disregarding all multiplicative constants that are independent of ϑ and setting $p = q = 1$, we get the posterior

$$\pi^h_{\text{prob,RTS}}(\vartheta \mid z) \propto \exp\Big(-\frac{\vartheta^2}{2}\Big)\frac{1}{\vartheta}\Big(\Phi\Big(\frac{((1-h)+h^{3/2})\vartheta - z}{\sigma}\Big) - \Phi\Big(\frac{((1-h)-h^{3/2})\vartheta - z}{\sigma}\Big)\Big), \tag{2.113}$$

where Φ denotes the cumulative distribution function of a standard Gaussian random variable. Since we require in assumption 2.2.1 that $H > 0$ almost surely, the time step H cannot be Gaussian and the closed-form expression of the posterior is not as neatly defined as in the additive noise case. In the limit for $\sigma \to 0$, we get the limiting distribution

$$\pi^h_{\text{prob,RTS}}(\vartheta \mid z) \propto \exp\Big(-\frac{\vartheta^2}{2}\Big)\frac{1}{\vartheta}\chi_{\{y_{\min} \le \vartheta \le y_{\max}\}}, \tag{2.114}$$

where $y_{\min}$ and $y_{\max}$ are given by

$$y_{\min} = \frac{e^{-h}\vartheta^*}{((1-h)+h^{3/2})}, \quad y_{\max} = \frac{e^{-h}\vartheta^*}{((1-h)-h^{3/2})}. \tag{2.115}$$

It is hence possible to remark that for the RTS-RK method the variance of the posterior distribution is not collapsing to zero for $\sigma \to 0$ as in the deterministic case.

We fix $h = 0.5$ 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. 2.3 we show the posteriors (2.108), (2.109), (2.111) and (2.113), which confirm our claim, i.e., that probabilistic methods take into account the variability in the forward model caused by the numerical approximation and transfer it to the posterior belief.

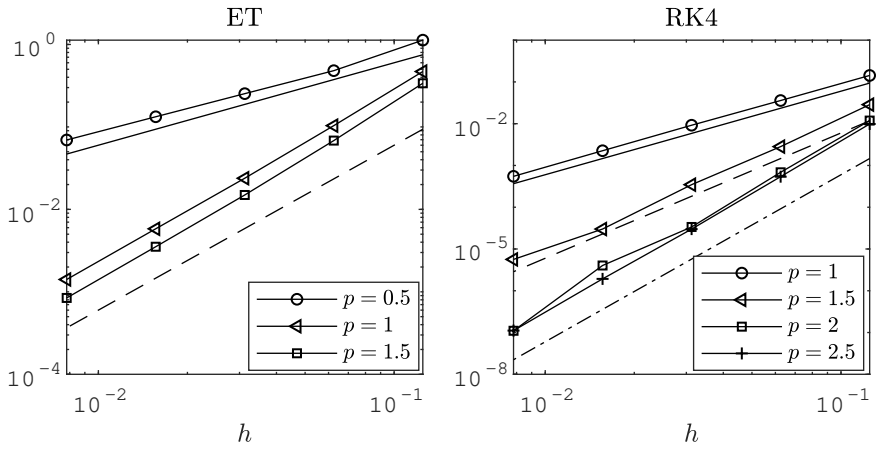

Figure 2.4 – Weak order of convergence for the random time-stepping explicit trapezoidal (ET) and fourth-order Runge–Kutta (RK4) as a function of the value of p of assumption 2.2. In the left figure, reference slopes 1 and 2 are displayed (solid and dashed lines), while in the right figure reference slopes 2, 3 and 4 are displayed (solid, dashed and dash-dotted lines).

2.9 Numerical experiments

In this section, we present a series of numerical experiments that illustrate the versatility and usefulness of our new random time stepping method. These experiments also corroborate the theoretical results presented in the previous sections.

2.9.1 Weak order of convergence

In order to verify the result predicted in theorem 2.19, we consider the FitzHugh–Nagumo equation, which is defined as

$$\begin{aligned} y_1' &= c\big(y_1 - \frac{y_1^3}{3} + y_2\big), \qquad y_1(0) = -1, \\ y_2' &= -\frac{1}{c}(y_1 - a + b y_2), \quad y_2(0) = 1, \end{aligned} \tag{2.116}$$

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 and the classic fourth-order Runge–Kutta method. The random steps are uniform as in example 2.3. 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 2.2 in order to verify the theoretical result of theorem 2.13. In particular, we consider $p \in \{0.5, 1, 1.5\}$ for the explicit trapezoidal rule and $p \in \{1, 1.5, 2, 2.5\}$ for the classic fourth order Runge–Kutta method. The function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ of the solution we consider is defined as $\Phi(x) \coloneqq x^\top x$. Finally, we consider 10^6 trajectories of the numerical solution in order to approximate the expectation with a Monte Carlo sum.

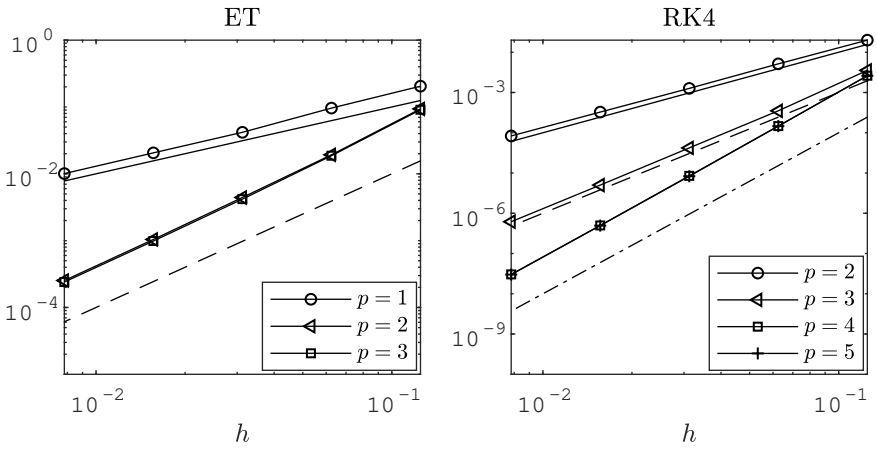

Figure 2.5 – Mean square order of convergence for the random time-stepping explicit trapezoidal (ET) and fourth-order Runge–Kutta (RK4) as a function of the value of p of assumption 2.2. In the left figure, reference slopes 1 and 2 are displayed (solid and dashed lines), while in the right figure reference slopes 2, 3 and 4 are displayed (solid, dashed and dash-dotted lines).

Results (fig. 2.4) show that the order of convergence predicted theoretically is confirmed by numerical experiments.

2.9.2 Mean square order of convergence

We now verify the weak order of convergence predicted in theorem 2.13. For this experiment we consider the ODE (2.116) as well, with the same time scale T and parameters as in section 2.9.1. The reference solution at final time is generated in this case as well with a high-order method on a fine time scale. We consider as deterministic solvers the explicit trapezoidal rule and the classic fourth order Runge–Kutta method, which verify assumption 2.4 with $q = 2$ and $q = 4$ respectively. Moreover, we consider uniform random time steps as in example 2.3, where we vary the value of p in assumption 2.2 in order to verify the order of convergence predicted in theorem 2.19. In particular, we consider $p \in \{1,2,3\}$ for the explicit trapezoidal rule and $p \in \{2,3,4,5\}$ for the classic fourth order Runge–Kutta method. We vary the mean time step h taken by the random time steps H_n in the range $h_i = 0.125 \cdot 2^{-i}$, with $i = 0,1,\ldots,4$. Then, we simulate 10^3 realizations of the numerical solution Y_{N_i}, with $N_i = T/h_i$ for $i = 0,1,\ldots,4$, and compute the approximate mean square order of convergence for each value of h with a Monte Carlo mean. Results (fig. 2.5) show that the orders predicted theoretically by theorem 2.19 are confirmed numerically.

2.9.3 Mean-square convergence of Monte Carlo estimators

We shall now verify numerically the validity of theorem 2.23. We consider the ODE (2.116), with final time $T = 1$ and the same parameters as above. In this case as well, we consider the explicit

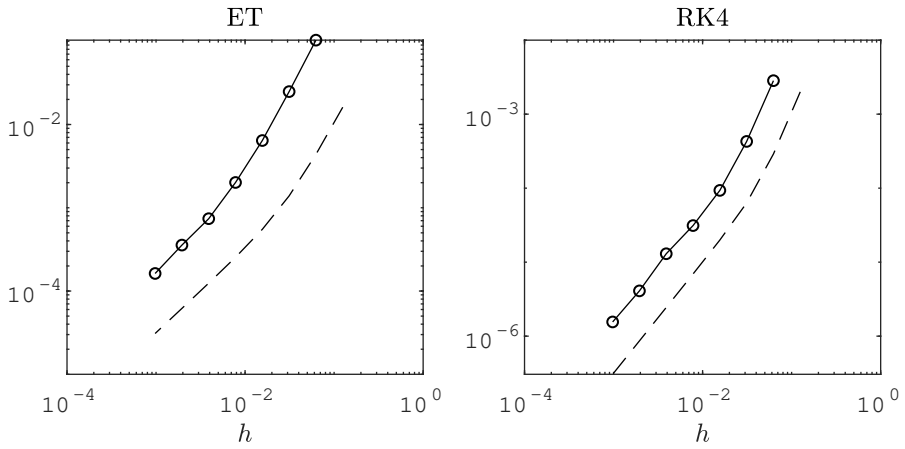

Figure 2.6 – Convergence of the square root of the MSE of the Monte Carlo estimator for the random time-stepping explicit trapezoidal (ET) (left figure) and fourth-order Runge–Kutta (RK4) (right figure) with respect to the time step h. The dashed line corresponds to the order predicted in theorem 2.23 with $M = 10^3$ for ET and $M = 10^4$ for RK4.

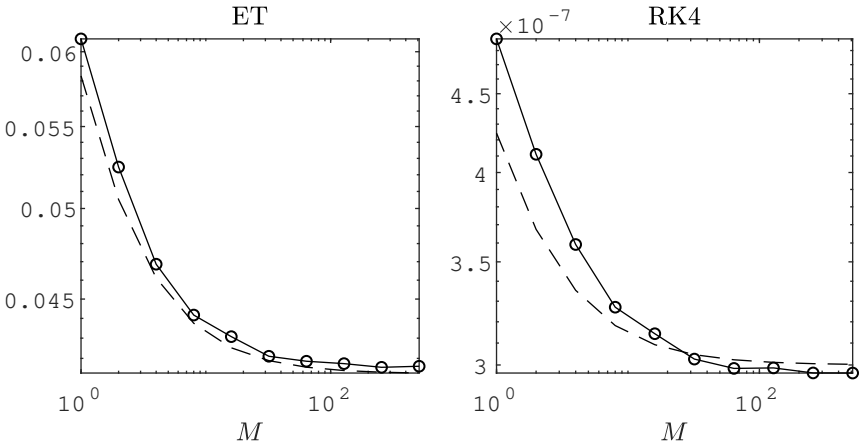

Figure 2.7 – Convergence of the square root of the MSE of the Monte Carlo estimator for the random time-stepping explicit trapezoidal (ET) (left figure) and fourth-order Runge–Kutta (RK4) (right figure) with respect to the number of trajectories M. The dashed line corresponds to the order predicted in theorem 2.23 with $h = 0.05$ for ET and $h = 0.01$ for RK4.

trapezoidal rule and the fourth-order explicit Runge–Kutta method with uniform random time steps having mean $h_i = 0.125 \cdot 2^{-i}$ with $i = 0, 1, \ldots, 7$. 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 presented in theorem 2.23 dominates, while in the regime of small h, the higher order of the first term makes the second term larger in magnitude. This behaviour results in the change of slope in the convergence plot which can be observed in fig. 2.6, both in the theoretical estimate and in the numerical results. We perform the same experiment using the fourth-order explicit Runge–Kutta method, 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 (2.50) reduces to

$$\mathrm{MSE}(\widehat{Z}_{N,M}) \leq Ch^{2q}\Big(1 + \frac{1}{M}\Big). \tag{2.117}$$

In fig. 2.7 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.

2.9.4 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

$$\mathrm{O_2} + 2\mathrm{NADH} + 2\mathrm{H^+} \rightarrow 2\mathrm{H_2O} + 2\mathrm{NAD^+}, \tag{2.118}$$

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 [49] to describe the time-evolution of the chemical system is the following

$$\begin{aligned} &\mathrm{B} + \mathrm{X} \xrightarrow{k_1} 2\mathrm{X}, \quad 2\mathrm{X} \xrightarrow{k_2} 2\mathrm{Y}, \quad \mathrm{A} + \mathrm{B} + \mathrm{Y} \xrightarrow{k_3} 3\mathrm{X}, \\ &\mathrm{X} \xrightarrow{k_4} \mathrm{P}, \quad \mathrm{Y} \xrightarrow{k_5} \mathrm{Q}, \quad \mathrm{X_0} \xrightarrow{k_6} \mathrm{X}, \\ &\mathrm{A_0} \xleftrightarrow{k_7} \mathrm{A}, \quad \mathrm{B_0} \xrightarrow{k_8} \mathrm{B}. \end{aligned} \tag{2.119}$$

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

$$\begin{aligned} \mathrm{A}' &= k_7(\mathrm{A_0} - \mathrm{A}) - k_3\mathrm{ABY}, & \mathrm{A}(0) &= 6, \\ \mathrm{B}' &= k_8\mathrm{B_0} - k_1\mathrm{BX} - k_3\mathrm{ABY}, & \mathrm{B}(0) &= 58, \\ \mathrm{X}' &= k_1\mathrm{BX} - 2k_2\mathrm{X}^2 + 3k_3\mathrm{ABY} - k_4\mathrm{X} + k_6\mathrm{X_0}, & \mathrm{X}(0) &= 0, \\ \mathrm{Y}' &= 2k_2\mathrm{X}^2 - k_5\mathrm{Y} - k_3\mathrm{ABY}, & Y(0) &= 0, \end{aligned} \tag{2.120}$$

where $\mathrm{A_0} = 8$, $\mathrm{B_0} = 1$, $\mathrm{X_0} = 1$ and the real parameters k_i, $i = 1, \dots, 8$ representing the reaction rates take values

$$\begin{aligned} &k_1 = 0.35, \quad k_2 = 250, \quad k_3 = 0.035, \quad k_4 = 20, \\ &k_5 = 5.35, \quad k_6 = 10^{-5}, \quad k_7 = 0.1, \quad k_8 = 0.825. \end{aligned} \tag{2.121}$$

It has been shown [49] 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, numerically we observe 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 2.3. 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 additive noise method (2.5) and the random time-stepping scheme (2.6) to equation (2.120). We choose $h = 0.05$ as the mean of uniformly distributed time steps for (2.6) and as the time step for (2.5), while we employ the Runge–Kutta–Chebyshev method (RKC) [63] as deterministic integrator. Since 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 [6, 1] could also be used as deterministic solvers for the RTS-RK method. It can be seen in fig. 2.8 that the RTS-RK method conserves the positivity of the numerical solution while capturing 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. 2.8.

2.9.5 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

$$\begin{aligned} w_1' &= v_1, \quad v_1' = -\frac{w_1}{\|q\|^3} - \frac{\delta w_1}{\|q\|^5}, \\ w_2' &= v_2, \quad v_2' = -\frac{w_2}{\|q\|^3} - \frac{\delta w_2}{\|q\|^5}, \end{aligned} \tag{2.122}$$

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

$$w_1(0) = 1 - e, \quad w_2(0) = 0, \quad v_1(0) = 0, \quad v_2(0) = \sqrt{(1+e)/(1-e)}, \tag{2.123}$$

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

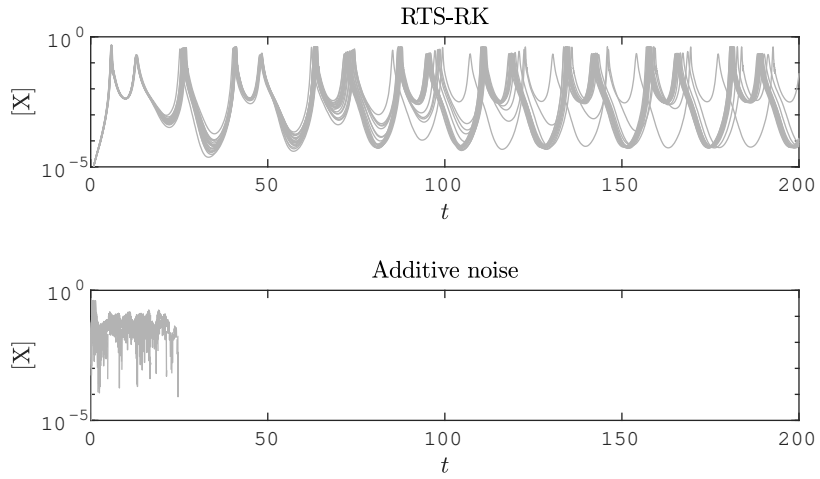

Figure 2.8 – Fifty trajectories of the numerical value of the concentration of the X species for the random time-stepping and additive noise methods (above and below respectively).

momentum, which reads

$$I(v, w) = w_1 v_2 - w_2 v_1. \tag{2.124}$$

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 2.27, we expect therefore that the random time-stepping method (2.6) implemented with Ψ_h given by the implicit midpoint rule also conserves quadratic first integrals. We integrate (2.122) 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 additive noise method (2.5) with $h = 0.01$, expecting that the first integral will not be conserved. We observe in fig. 2.9 that the method (2.6) conserves the angular momentum, while for the method (2.5) the approximate conservation of the quadratic first integral shown in (2.59) is lost when integrating (2.122) over long time.

2.9.6 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

$$Q(v, w) = \frac{v^2}{2} - \cos w, \tag{2.125}$$

where $y = (v, w)^\top \in \mathbb{R}^2$. We wish to study the validity of theorem 2.37, 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

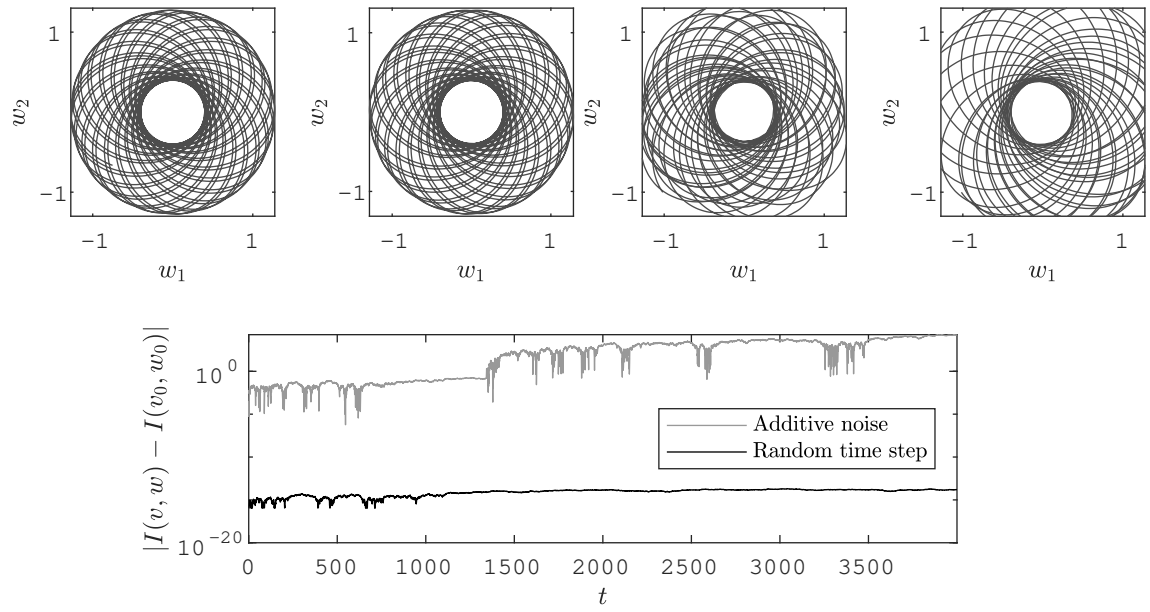

Figure 2.9 – Trajectories of (2.122) given by the RTS-RK method (2.6) for $0 \leq t \leq 200$ and $3800 \leq t \leq 4000$ (first and second figures), and by the additive noise method (2.5) for $0 \leq t \leq 200$ and $200 \leq t \leq 400$ (third and fourth figures). Error on the angular momentum I defined in (2.124) for $0 \leq t \leq 4000$ given by the two methods.

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. 2.10, 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. 2.10 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. 2.10 together with an approximate confidence interval. We arbitrarily compute the confidence interval as $(\mathbb{E}\,Q(Y_n) - 2\mathrm{Var}Q(Y_n)^{1/2}, \mathbb{E}\,Q(Y_n) + 2\mathrm{Var}Q(Y_n)^{1/2})$, 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.

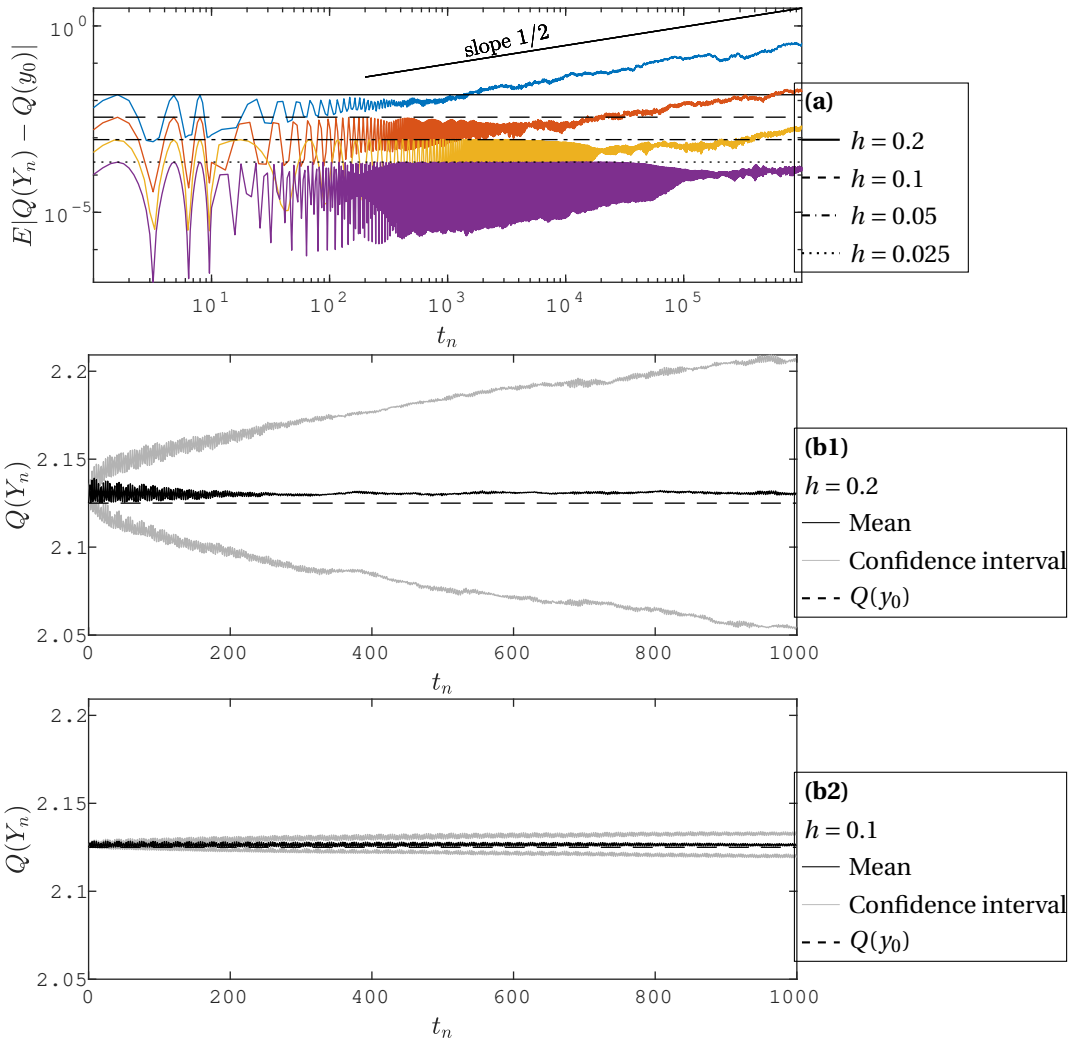

Figure 2.10 – **(a)**: Time evolution of the mean error for the pendulum problem and different values of the time step h. The black lines represent the theoretical estimate given by theorem 2.37, while the colored lines represent the experimental results. The mean was computed by averaging 20 realisations 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.

2.9.7 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

$$Q(v,w) = \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2 + w_1^2 w_2 - \frac{1}{3}w_2^3, \tag{2.126}$$

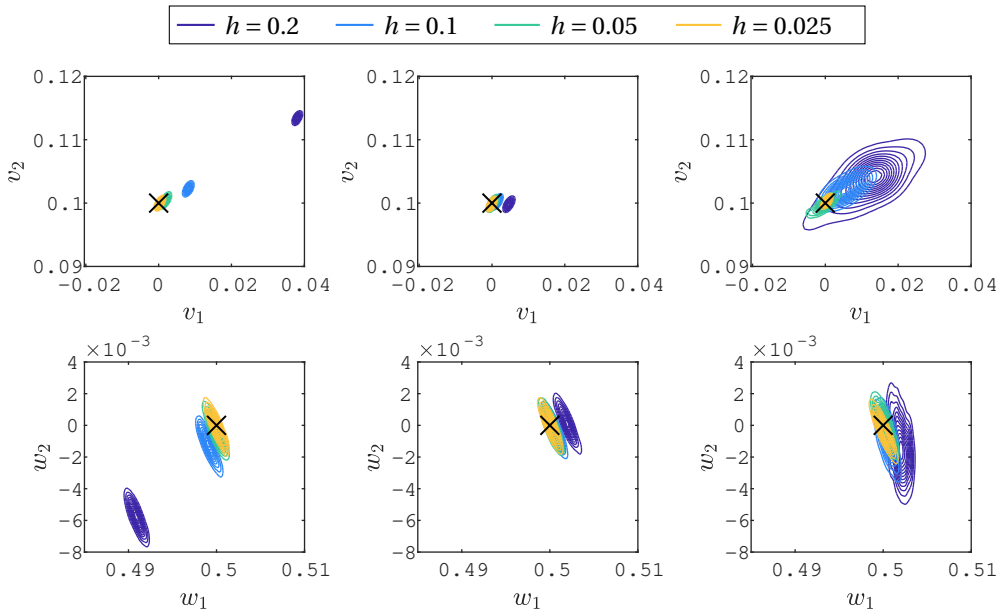

Figure 2.11 – Posterior distributions for the initial position and velocity of the Hénon-Heiles system with different values of $h = \{0.2, 0.1, 0.05, 0.025\}$. First row: initial velocity v_0. Second row: initial position w_0. First column: deterministic Heun's method. Second column: deterministic Störmer–Verlet scheme. Third column: RTS-RK Störmer–Verlet ($p = 2$).

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 [34]. In the spirit of section 2.8, 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 $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$, where $\sigma_\varepsilon = 5 \cdot 10^{-4}$, and we fix a standard Gaussian prior on the initial condition, i.e., $\pi_0 = \mathcal{N}(0, I)$, so that the likelihood is given by (4.7). We choose the observational 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 2.8.

Since the equation is Hamiltonian, we choose to employ a classical second-order ($q = 2$) symplectic method, the Störmer–Verlet scheme [60, 65, 30], for which one step is defined in the general case as

$$\begin{aligned} v_{n+1/2} &= v_n - \frac{h}{2}\nabla_w Q(v_n, w_n), \\ w_{n+1} &= w_n + \frac{h}{2}\big(\nabla_v Q(v_{n+1/2}, w_n) + \nabla_v Q(v_{n+1/2}, w_{n+1})\big), \\ v_{n+1} &= v_{n+1/2} - \frac{h}{2}\nabla_w Q(v_{n+1/2}, w_{n+1}). \end{aligned} \tag{2.127}$$

As the Hamiltonian Q given by (2.126) 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

$$\begin{aligned} v_{n+1/2} &= v_n - \frac{h}{2}\nabla_w Q_2(w_n), \\ w_{n+1} &= w_n + h\nabla_v Q_1(v_{n+1/2}), \\ v_{n+1} &= v_{n+1/2} - \frac{h}{2}\nabla_w Q_2(w_{n+1}). \end{aligned} \tag{2.128}$$

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 $\pi^h(y_0 \mid y_{\text{obs}})$ and $\pi^h_{\text{prob}}(y_0 \mid y_{\text{obs}})$ defined in section 2.8, respectively. Moreover, we compute the posterior distribution given by a non-symplectic method, the Heun's scheme, which is a classical second order method. The time step h is varied for the three methods above in order to study whether the approximate posterior concentrates towards the true posterior distribution $\pi(y_0 \mid y_{\text{obs}})$.

We can observe in fig. 4.3 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 [19, 17], 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 2.37.

2.10 Conclusion

In this work we introduced the RTS-RK method, a novel probabilistic integrators for ODEs built on Runge–Kutta numerical integrators with random time steps. In particular, we analysed its weak and mean-square convergence properties, as well as the quality of Monte Carlo estimators drawn from the probabilistic solution. Geometric properties such as the conservation of first integrals and the approximation of Hamiltonians over long time intervals have been extensively treated theoretically. Finally, we showed heuristically the advantageous properties of the probabilistic approach in Bayesian inference problems with respect to the classic deter-

ministic approach when the discretization is not in the asymptotic regime $h \to 0$. The validity of our theoretical contributions is strengthened by an extensive series of numerical examples.

The RTS-RK method is partially inspired by the probabilistic method based on additive noise perturbations presented in [19] and further analysed in [40], with which it shares convergence properties and the advantageous behaviour in inverse problems. Nonetheless, our method fills the void of geometry-aware probabilistic integrators for ODEs, and thus it represents a step forward in the field of probabilistic numerics for differential equations.

Appendix

A modified stochastic differential equation

In remark 2.14, 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.,

$$\widetilde{f}(y) = f(y) + h^q f_{q+1}(y) + h^{q+1} f_{q+2}(y) + \ldots + h^l f_{l+1}(y). \tag{2.129}$$

Details about the construction of such a function can be found in section 2.7.2. 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 2.31 (see section 2.7.2). For the additive noise method presented in [19], the authors consider the SDE

$$\mathrm{d}Y = \widetilde{f}(Y)\,\mathrm{d}t + \sqrt{Qh^{2p}}\,\mathrm{d}W, \tag{2.130}$$

where W is a d-dimensional standard Brownian motion. It is possible to show [19, Theorem 2.4] that the solution of (2.130) satisfies

$$|\mathbb{E}\big(\Phi(Y_N) - \Phi(Y(T)) \mid Y_0 = y\big)| \leq Ch^{2p}, \tag{2.131}$$

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

$$\mathrm{d}\widetilde{Y} = \Big(\widetilde{f}(\widetilde{Y}) + \frac{1}{2}Ch^{2p}\partial_{tt}\Psi_h(\widetilde{Y})\Big)\,\mathrm{d}t + \sqrt{Ch^{2p}\partial_t\Psi_h(\widetilde{Y})\partial_t\Psi_h(\widetilde{Y})^\top}\,\mathrm{d}W, \tag{2.132}$$

where C is given in assumption 2.2.3. Let us denote by $\widetilde{\mathscr{L}}$ the generator of (2.132), which can be written explicitly as

$$\widetilde{\mathscr{L}} = \Big(\widetilde{f} + \frac{1}{2}Ch^{2p}\partial_{tt}\Psi_h\Big)\cdot\nabla + \frac{1}{2}Ch^{2p}\partial_t\Psi_h\partial_t\Psi_h^\top : \nabla^2, \tag{2.133}$$

and, adopting the semi-group notation, it satisfies

$$\mathbb{E}\big(\Phi(\widetilde{Y}(h)) \mid \widetilde{Y}(0) = y\big) = e^{h\widetilde{\mathscr{L}}}\Phi(y). \tag{2.134}$$

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 (2.132) in the weak sense. The proof is inspired by the calculations presented in [19, Section 2.4].

Lemma 2.42. *Under the assumptions of lemma 2.8 and if assumption 2.31 holds, then*

$$|\mathbb{E}\big(\Phi(Y_1) - \Phi(\widetilde{Y}(h)) \mid Y_0 = y\big)| \leq Ch^{2p+1}, \tag{2.135}$$

where C is a positive constant independent of h and of y, $\widetilde{Y}$ is the solution of (2.132) *and Y_1 is the numerical solution given by the RTS-RK method after one step.*

Proof. Let us consider the modified ODE

$$\widehat{y}'(t) = \widetilde{f}(\widehat{y}), \tag{2.136}$$

and denote its flow as $\widehat{\varphi}_t$. The generator $\widehat{\mathscr{L}} = \widetilde{f}\cdot\nabla$ satisfies, adopting the semi-group notation,

$$\Phi(\widehat{\varphi}_h(y)) = e^{h\widehat{\mathscr{L}}}\Phi(y). \tag{2.137}$$

We can now compute the distance between the solution to (2.132) and (2.136) as

$$\begin{aligned} e^{h\widetilde{\mathscr{L}}}\Phi(y) - e^{h\widehat{\mathscr{L}}}\Phi(y) &= e^{h\widetilde{f}\cdot\nabla}\Big(e^{\frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h\cdot\nabla + \frac{1}{2}Ch^{2p+1}\partial_t\Psi_h\partial_t\Psi_h^\top:\nabla^2} - I\Big)\Phi(y) \\ &= \big(1+\mathscr{O}(h)\big)\Big(\frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h\cdot\nabla + \frac{1}{2}Ch^{2p+1}\partial_t\Psi_h\partial_t\Psi_h^\top:\nabla^2 + \mathscr{O}\big(h^{4p+1}\big)\Big)\Phi(y) \\ &= \frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h\cdot\nabla\Phi(y) + \frac{1}{2}Ch^{2p+1}\partial_t\Psi_h\partial_t\Psi_h^\top:\nabla^2\Phi(y) + \mathscr{O}\big(h^{4p+1}\big). \end{aligned} \tag{2.138}$$

Let us recall that equation (2.19) gives

$$\begin{aligned} e^{h\mathscr{L}_h}\Phi(y) - \Phi(\Psi_h(y)) = &\frac{1}{2}Ch^{2p+1}\partial_{tt}\Psi_h(y)\cdot\nabla\Phi(y) \\ &+ \frac{1}{2}Ch^{2p+1}\partial_t\Psi_h(y)\partial_t\Psi_h(y)^\top:\nabla^2\Phi(y) + \mathscr{O}(h^{2p+1}), \end{aligned} \tag{2.139}$$

which implies that

$$e^{h\widetilde{\mathscr{L}}}\Phi(y) - e^{h\mathscr{L}_h}\Phi(y) = e^{h\widehat{\mathscr{L}}}\Phi(y) - \Phi(\Psi_h(y)) + \mathscr{O}(h^{2p+1}). \tag{2.140}$$

Now, the theory of backward error analysis (see section 2.7.2 or e.g. [30, Chapter IX]) guarantees that

$$e^{h\widehat{\mathscr{L}}}\Phi(y) - \Phi(\Psi_h(y)) = \mathscr{O}(h^{q+l+2}). \tag{2.141}$$

Choosing $l = 2p - q - 1$, we have therefore

$$e^{h\widetilde{\mathscr{L}}}\Phi(y) - e^{h\mathscr{L}_h}\Phi(y) = \mathscr{O}(h^{2p+1}), \tag{2.142}$$

which is the desired result. □

The error can be then propagated to final time as in theorem 2.13, as presented in the following theorem.

Theorem 2.43. *Under the assumptions of lemma 2.42 and theorem 2.13, and if there exists a constant $L > 0$ independent of h such that for all $\Phi \in \mathscr{C}_b^\infty(\mathbb{R}^d, \mathbb{R})$*

$$\sup_{u\in\mathbb{R}^d} |e^{h\widetilde{\mathscr{L}}}\Phi(u)| \leq (1 + Lh) \sup_{u\in\mathbb{R}^d} |\Phi(u)|, \tag{2.143}$$

then it holds

$$|\mathbb{E}\big(\Phi(Y_N) - \Phi(\widetilde{Y}(T)) \mid Y_0 = y\big)| \leq Ch^{2p}, \tag{2.144}$$

where $T = Nh$ and C is a positive constant independent of h and of y, $\widetilde{Y}$ is the solution of (2.132) *and Y_N is the numerical solution given by the RTS-RK method after N steps.*

Proof. The proof follows by replacing $\mathscr{L}$ with $\widetilde{\mathscr{L}}$ and lemma 2.8 with lemma 2.42 in the proof of theorem 2.13. □

Proof of lemma 2.34

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

$$\gamma_r'(x) = r x^{r-2}(x + \kappa)e^{-r\kappa/x}. \tag{2.145}$$

Under assumption 2.33 we have that $H_j \leq Mh$ almost surely, and hence for any $t \in [\![h, H_j]\!]$

$$|\gamma_r'(t)| \leq r(Mh)^{r-2}(Mh + \kappa)e^{-r\kappa/(Mh)}, \tag{2.146}$$

where we exploited that $e^{-r\kappa/x}$ is a growing function of x. The fundamental theorem of calculus gives

$$\begin{aligned} |\gamma_r(H_j)| &= \left|\gamma_r(h) + \int_h^{H_j} \gamma_r'(t)\,\mathrm{d}t\right| \\ &\leq \gamma_r(h) + r(Mh)^{r-2}(Mh + \kappa)e^{-r\kappa/(Mh)}|H_j - h|, \quad \text{almost surely.} \end{aligned} \tag{2.147}$$

Taking expectation on both sides and since by (2.73) it holds $|\eta_j|^r \leq C\gamma_r(H_j)$ we obtain

$$\mathbb{E}|\eta_j|^r \leq C\big(\gamma_r(h) + rM^{r-2}h^{p+r-3/2}(Mh + \kappa)e^{-r\kappa/(Mh)}\big), \tag{2.148}$$

which proves the desired inequality. This is because assumption 2.33 and assumption 2.2.2 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 $t \in [\![h, H_j]\!]$

$$|\gamma_1'(t)| \leq (mh)^{-1}(Mh+\kappa)e^{-\kappa/(Mh)}, \quad \text{almost surely.} \tag{2.149}$$

Hence, we apply the same reasoning as above and obtain almost surely

$$|\gamma_1(H_j)| \leq \gamma_1(h) + (mh)^{-1}(Mh+\kappa)e^{-\kappa/(Mh)}|H_j - h|, \tag{2.150}$$

which implies the desired result by proceeding as above. □

Proof of lemma 2.35

We first expand the square as

$$\begin{aligned}\Big(\sum_{j=0}^{n-1}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)\Big)^2 = \sum_{j=0}^{n-1}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)^2 \\ + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)\Big(\sum_{k=q}^{N-1} a_{ik} + b_i\Big).\end{aligned} \tag{2.151}$$

Then, we expand the square in the first sum and obtain

$$\begin{aligned}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)^2 &= \Big(\sum_{k=q}^{N-1} a_{jk}\Big)^2 + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk} \\ &= \sum_{k=q}^{N-1} a_{jk}^2 + 2\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl} + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk} \\ &= a_{jq}^2 + \sum_{k=q+1}^{N-1} a_{jk}^2 + 2\sum_{k=q+1}^{N-1}\sum_{l=q}^{k-1} a_{jk}a_{jl} + b_j^2 + 2b_j\sum_{k=q}^{N-1} a_{jk}.\end{aligned} \tag{2.152}$$

We then rewrite the term appearing in the double sum in (2.151) as

$$\begin{aligned}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)\Big(\sum_{k=q}^{N-1} a_{ik} + b_i\Big) = a_{jq}a_{iq} + \sum_{k=q}^{N-1}\sum_{\substack{l=q \\ l+k>2q}}^{N-1} a_{jk}a_{il} \\ + b_j\sum_{k=q}^{N-1} a_{ik} + b_i\sum_{k=q}^{N-1} a_{jk} + b_ib_j\end{aligned} \tag{2.153}$$

Substituting the expressions (2.152) and (2.153) in (2.151), we finally get

$$\Big(\sum_{j=0}^{n-1}\Big(\sum_{k=q}^{N-1} a_{jk} + b_j\Big)\Big)^2 = \sum_{j=0}^{n-1} a_{jq}^2 + 2\sum_{j=1}^{n-1}\sum_{i=0}^{j-1} a_{jq}a_{iq} + R(a) + S(a,b), \tag{2.154}$$

where the remainder $R(a)$ can be written as $R = R_1 + R_2 + R_3$ where

$$\begin{aligned} R_1(a) &= \sum_{j=0}^{n-1} \sum_{k=q+1}^{N-1} a_{jk}^2, \qquad & R_2(a) &= 2 \sum_{j=0}^{n-1} \sum_{k=q+1}^{N-1} \sum_{l=q}^{k-1} a_{jk} a_{jl}, \\ R_3(a) &= 2 \sum_{j=1}^{n-1} \sum_{i=0}^{j-1} \sum_{k=q}^{N-1} \sum_{\substack{l=q \\ l+k>2q}}^{N-1} a_{jk} a_{il}, \end{aligned} \tag{2.155}$$

and the remainder $S(a,b)$ can be written as $S = S_1 + S_2 + S_3 + S_4$ where

$$\begin{aligned} S_1(a,b) &= \sum_{j=0}^{n-1} b_j^2, \qquad & S_2(a,b) &= 2 \sum_{j=1}^{n-1} \sum_{i=0}^{j-1} b_i b_j, \\ S_3(a,b) &= 2 \sum_{j=1}^{n-1} \sum_{k=q}^{N-1} b_j a_{jk}, \qquad & S_4(a,b) &= 2 \sum_{j=1}^{n-1} \sum_{i=0}^{n-1} \Big(b_j \sum_{k=q}^{N-1} a_{ik} + b_i \sum_{k=q}^{N-1} a_{jk} \Big). \end{aligned} \tag{2.156}$$

which proves the desired result. □

Proof of lemma 2.36

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 2.4 and assumption 2.31, we get for all $j = 0, \dots, n-1$ and $k = q, \dots, N-1$

$$\begin{aligned} |\Delta_{j,k}| &= |Q_{k+1}(Y_j) - Q_{k+1}(Y_{j+1})| \\ &\leq C \|\Psi_0(Y_j) - \Psi_{H_j}(Y_j)\| \\ &\leq C_\Delta |H_j|, \end{aligned} \tag{2.157}$$

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 2.31. Let us now consider $R(\Delta)$. Due to (2.157) and to assumption 2.33, we have

$$\begin{aligned} \mathbb{E}(H_j^k - h^k)^2 \Delta_{j,k}^2 &\leq C_\Delta^2 \, \mathbb{E}(H_j^{k+1} - H_j h^k)^2 \\ &= C_\Delta^2 \big(h^{2(k+1)} + C_{2(k+1)} h^{2p+2(k+1)-1} + h^{2(k+1)} + C_2 h^{2p+2k+1} \\ &\quad - 2h^{2k+2} - 2C_{k+2} h^{2p+2k+1} \big) \\ &= C_\Delta^2 \big((C_{2(k+1)} + C_2 - 2C_{k+2}) h^{2p+2k+1} \big) \\ &\leq C h^{2p+2k+1}, \end{aligned} \tag{2.158}$$

where $C > 0$ is a positive constant. Now, since $k \geq q+1$, we get

$$\mathbb{E}(H_j^k - h^k)^2 \Delta_{j,k}^2 \leq C h^{2(p+q+1)}. \tag{2.159}$$

Hence, for $R_1(\Delta)$ there exists a constant $\widetilde{C}_1$ such that

$$\mathbb{E}\, R_1(\Delta) \le \widetilde{C}_1 n h^{2(p+q+1)}. \tag{2.160}$$

We now proceed to the second remainder $R_2(\Delta)$. Applying the Cauchy–Schwarz inequality and (2.158) we get

$$\begin{aligned}\mathbb{E}\left((H_j^k - h^k)\Delta_{j,k}(H_j^l - h^l)\Delta_{j,l}\right) &\le \left(\mathbb{E}\left((H_j^k - h^k)^2\Delta_{j,k}^2\right)^{1/2}\left(\mathbb{E}\left((H_j^l - h^l)^2\Delta_{j,l}^2\right)^{1/2}\right.\right.\\ &\le C h^{2p+k+l+1},\end{aligned} \tag{2.161}$$

where $C > 0$ is a positive constant. Now, since in the definition of $R_2(a)$ in (2.157) we have $k \ge q+1$ and $l \ge q$, we have here $k + l \ge 2q+1$. Therefore, there exists a constant $\widetilde{C}_2$ such that

$$\mathbb{E}\, R_2(\Delta) \le \widetilde{C}_2 n h^{2(p+q+1)}. \tag{2.162}$$

We now consider the term $R_3(\Delta)$. Since H_i and H_j are independent for $i \ne j$, we have

$$\mathbb{E}\left((H_j^k - h^k)\Delta_{j,k}(H_i^l - h^l)\Delta_{i,l}\right) = \mathbb{E}(H_j^k - h^k)\Delta_{j,k}\,\mathbb{E}(H_i^l - h^l)\Delta_{i,l}. \tag{2.163}$$

Computing the two factors singularly, we have due to (2.157) and to assumption 2.33

$$\begin{aligned}\mathbb{E}(H_j^k - h^k)\Delta_{j,k} &\le C_\Delta\, \mathbb{E}(H_j^{k+1} - H_j h^k)\\ &= C_\Delta C_{k+1} h^{2p+k},\end{aligned} \tag{2.164}$$

and analogously for $\mathbb{E}(H_i^l - h^l)\Delta_{i,l}$. Then, since $k + l \ge 2q+1$

$$\mathbb{E}\left((H_j^k - h^k)\Delta_{j,k}(H_i^l - h^l)\Delta_{i,l}\right) \le C_\Delta^2 C_{k+1} C_{l+1} h^{2(2p+q+1/2)}. \tag{2.165}$$

Hence, we have for a constant $\widetilde{C}_3 > 0$

$$\mathbb{E}\, R_3(\Delta) \le \widetilde{C}_3 n^2 h^{2(2p+q+1/2)}. \tag{2.166}$$

Finally, replacing $t_n = nh$, we can write for a constant $C > 0$

$$\begin{aligned}\mathbb{E}\, R(\Delta) &\le (\widetilde{C}_1 + \widetilde{C}_2) n h^{2(p+q+1)} + \widetilde{C}_3 n^2 h^{2(2p+q+1/2)}\\ &= (\widetilde{C}_1 + \widetilde{C}_2) t_n h^{2(p+q+1/2)} + \widetilde{C}_3 t_n^2 h^{2(2p+q-1/2)}.\end{aligned} \tag{2.167}$$

Let us now consider $S(\Delta, \eta)$. First, we notice that under the assumption $p \ge 3/2$ we have for any $r \ge 1$, $\min\{r, p + r - 3/2\} = r$, and therefore lemma 2.34 simplifies to

$$\mathbb{E}|\eta_j|^r \le C h^r e^{-r\kappa/(Mh)}. \tag{2.168}$$

We first consider $S_1(\Delta, \eta)$. Applying lemma 2.34 with $r = 2$, we obtain for a constant $\widehat{C}_1 > 0$

$$\mathbb{E}\, S_1(\Delta, \eta) \le \widehat{C}_1 n h^2 e^{-2\kappa/(Mh)}. \tag{2.169}$$

For the second term $S_2(\Delta,\eta)$, we have by (2.73) that $|\eta_i| \le CH^i e^{-\kappa/H_i}$ and $\eta_j \le CH^j e^{-\kappa/H_j}$ almost surely. These two bounds are independent for $i \neq j$ and therefore, applying lemma 2.34 with $r = 1$, we have for a constant $\widehat{C}_2 > 0$

$$\mathbb{E}\, S_2(\Delta,\eta) \le \widehat{C}_2 n^2 h^2 e^{-2\kappa/(Mh)}. \tag{2.170}$$

We now consider the third remainder $S_3(\Delta,\eta)$. Applying the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}\,\eta_j (H_j^k - h^k)\Delta_{j,k} \le (\mathbb{E}\,\eta_j^2)^{1/2} (\mathbb{E}(H_j^k - h^k)^2 \Delta_{j,k}^2)^{1/2}. \tag{2.171}$$

Applying lemma 2.34 with $r = 2$ to the first factor and (2.158) to the second we get

$$\begin{aligned} \mathbb{E}\,\eta_j (H_j^k - h^k)\Delta_{j,k} &\le Che^{-\kappa/(Mh)} h^{p+k+1/2} \\ &= Ch^{p+k+3/2} e^{-\kappa/(Mh)} \end{aligned} \tag{2.172}$$

Now, since $k \ge q$, we have for a constant $\widehat{C}_3 > 0$

$$\mathbb{E}\, S_3(\Delta,\eta) \le \widehat{C}_3 nh^{p+q+3/2} e^{-\kappa/(Mh)}. \tag{2.173}$$

Finally, we consider the last term $S_4(\Delta,\eta)$. Since by (2.73) it holds $|\eta_j| \le CH_j e^{-\kappa/H_j}$ almost surely, and this bound is independent of H_i for $i \neq j$, applying (2.164) and lemma 2.34 we have

$$\begin{aligned} \mathbb{E}\,\eta_j (H_i^k - h^k)\Delta_{i,k} &= \mathbb{E}\,\eta_j\, \mathbb{E}(H_i^k - h^k)\Delta_{i,k} \\ &\le Che^{-\kappa/(Mh)} h^{2p+k}, \end{aligned} \tag{2.174}$$

which, since $k \ge q$, implies that there exists a constant $\widehat{C}_4 > 0$ such that

$$\mathbb{E}\, S_4(\Delta,\eta) \le \widehat{C}_4 n^2 h^{2p+q+1} e^{-\kappa/(Mh)}. \tag{2.175}$$

Finally, replacing $t_n = nh$, we can write

$$\begin{aligned} \mathbb{E}\, S(\Delta,\eta) &\le (\widehat{C}_1 nh^2 + \widehat{C}_2 n^2 h^2) e^{-2\kappa/(Mh)} + (\widehat{C}_3 nh^{p+q+3/2} + \widehat{C}_4 n^2 h^{2p+q+1}) e^{-\kappa/(Mh)} \\ &= (\widehat{C}_1 t_n h + \widehat{C}_2 t_n^2) e^{-2\kappa/(Mh)} + (\widehat{C}_3 t_n h^{p+q+1/2} + \widehat{C}_4 t_n^2 h^{2p+q-1}) e^{-\kappa/(Mh)}, \end{aligned} \tag{2.176}$$

which completes the proof. □

3 Probabilistic methods for elliptic PDEs

3.1 Introduction

TO DO[4, 36, 37, 15, 19]

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) &= f, \quad \text{in } D, \\ u &= g, \quad \text{on } \partial D. \end{aligned} \tag{3.1}$$

important: review of probabilistic methods for PDEs and ODEs. Have PDEs really been treated already? How? Inverse problems: what is the current state of things? Has anyone gone infinite dimensional?

3.2 Random mesh probabilistic Finite Elements

Weak formulation: bilinear form $a\colon V \times V \to \mathbb{R}$ and a linear functional $F\colon V \to \mathbb{R}$ satisfying the usual continuity and coercivity constraints, look for $u \in V$ satisfying

$$a(u, v) = F(v), \tag{3.2}$$

for all functions $v \in V$. Galerkin formulation: for $V_h \subset V$ such that $\dim V_h < \infty$, find $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h, \tag{3.3}$$

for all $v_h \in V_h$. Given a triangulation $\mathcal{T}_h$ of the domain D, we choose V_h to be the space of linear finite elements, i.e., $V_h = X_h^1 \cap V$, where

$$X_h^1 = \{v_h \in C^0(\overline{D})\colon v_h|_K \in \mathcal{P}_1, \text{ for all } K \in \mathcal{T}_h\}, \tag{3.4}$$

and where $\mathcal{P}_1$ is the space of polynomials of degree at most one. The finite element space can be written then as $V_h = \text{span}\{\varphi_i\}_{i=1}^N$, where the basis $\{\varphi_i\}_{i=1}^N$ are the Lagrange basis functions.

Hence, each $v_h \in V_h$ can be written as $v_h = \sum_{i=1}^{N} v_i \varphi_i$, where v_i are the coefficients of v_h on the basis $\{\varphi_i\}_{i=1}^{N}$. Our probabilistic method is based on a randomly perturbed mesh $\widetilde{\mathcal{T}_h}$ which is defined as follows.

Definition 3.1. Let us consider a domain $D \subset \mathbb{R}^d$ and a triangulation $\mathcal{T}_h$ characterised by the maximum diameter $h > 0$ of its elements and by the set of vertices $\mathcal{N}_h = \{x_i\}_{i=1}^{N}$ such that $\mathcal{N}_h = \mathcal{N}_h^I \cup \mathcal{N}_h^B$, where $\mathcal{N}_h^I$ and $\mathcal{N}_h^B$ are the vertices in the interior of D and on ∂D respectively, and where we denote $N_I = |\mathcal{N}_h^I|$ and $N_B = |\mathcal{N}_h^B|$. Given a probability space (Ω, Σ, μ), the random mesh $\widetilde{\mathcal{T}_h}$ is defined by a sequence of random variables $\{\alpha_i\}_{i=1}^{N_I}$, $\alpha_i \colon \Omega \to \mathbb{R}^d$, which are used to perturb the internal nodes as

$$\widetilde{x}_i = x_i + \bar{h}_i^p \alpha_i, \; x_i \in \mathcal{N}_h^I \tag{3.5}$$

where $p \geq 1$ and $\bar{h}_i$ is defined as the minimum diameter of the elements K having x_i as a vertex, i.e.

$$\bar{h}_i = \min_{K \in \Delta(x_i)} h_K, \tag{3.6}$$

where $\Delta(x_i)$ is such set of elements. The vertices laying on ∂D in $\mathcal{T}_h$ are unperturbed in $\widetilde{\mathcal{T}_h}$.

Once the perturbed mesh $\widetilde{\mathcal{T}_h}$ is obtained, let us denote by $\widetilde{V}_h$ the piecewise linear finite element space defined on $\widetilde{\mathcal{T}_h}$. Let us remark that the space $\widetilde{V}_h = \widetilde{V}_h(\omega)$ is random itself, i.e., for each realisation of the random variables $\{\alpha_i\}_{i=1}^{N_I}$ we obtain a different perturbed finite element space.

Definition 3.2. With the notation above, the probabilistic solution $\widetilde{u}_h \colon \Omega \times D \to \mathbb{R}$ is a random field satisfying for all $\omega \in \Omega$

$$\widetilde{u}_h(\omega, \cdot) \in \widetilde{V}_h(\omega), \text{ s.t. } a(\widetilde{u}_h(\omega, \cdot), \widetilde{v}_h) = F(\widetilde{v}_h), \text{ for all } \widetilde{v}_h \in \widetilde{V}_h(\omega). \tag{3.7}$$

Let us finally introduce the following assumption on the random variables defining the mesh perturbation.

Assumption 3.3. The random variables α_i are chosen such that the perturbed mesh $\widetilde{\mathcal{T}_h}$ has the same topology of the mesh $\mathcal{T}_h$ (e.g., no exchange of vertices in one-dimension and no crossing edges in two-dimensions) almost surely.

3.2.1 Notation

We now introduce some basic notation which will be employed throughout this paper. Most of the notation is classic, but we report it here for completeness. The symbol $D \subset \mathbb{R}^2$ is employed for a bounded domain with smooth boundary, or for a convex polygon. The following symbols are employed for function spaces

- $L^p(D) = \left\{ v \colon D \to \mathbb{R}, \int_D v^p \, \mathrm{d}x < \infty \right\}$,

- $W^{q,p}(D) = \{v \in L^p(D), \sum_{|\alpha|\leq q} |D^\alpha v| \in L^p(D)\}$,
- $H^q(D) \equiv W^{q,2}(D)$,
- $H_0^q(D) = \{v \in H^q(D), v\big|_{\partial D} = 0\}$,
- $\mathscr{C}_0^l(D) = \{v \in \mathscr{C}^l(D), v\big|_{\partial D} = 0\}$.

For a function $v \in \mathscr{X}$ where $\mathscr{X}$ is any of the spaces above, we denote by $\|v\|_{\mathscr{X}}$ and $|v|_{\mathscr{X}}$ the usual norms and seminorms. Furthermore, for L^p and $W^{q,p}$, the usual meaning is given for $p = \infty$. For a vector $x \in \mathbb{R}^2$ we denote simply by $|x|$ its Euclidean norm. Moreover, we will employ the following symbols

- $\mathscr{T}_h$ is a triangulation of D satisfying assumption, and V_h is the space of linear finite elements with zero boundary conditions defined on $\mathscr{T}_h$,
- $\widetilde{\mathscr{T}_h}$ is a perturbation of $\mathscr{T}_h$ as for definition 3.1 such that assumption 3.3 holds, and $\widetilde{V}_h$ is the space of linear finite elements with zero boundary conditions defined on $\widetilde{\mathscr{T}_h}$.

Finally, if a function $v\colon D \to \mathbb{R}$ or $v\colon D \to \mathbb{R}^2$ is constant over a set $K \subset D$, we denote by $v\big|_K$ its constant value.

3.3 A priori error analysis

In this section, we analyse the convergence a priori of our method. In particular, we wish the family of probabilistic solutions to be close to the solution obtained with the original mesh, i.e., we will prove that

$$\|u_h - \widetilde{u}_h\|_{\mathscr{X}} \leq C\eta(h), \quad \text{a.s.}, \tag{3.8}$$

where $\eta\colon \mathbb{R} \to \mathbb{R}$ is such that $\eta(h) \to 0$ for $h \to 0$ and where $\mathscr{X} = \{H^1(D), L^\infty(D)\}$. Similarly to standard error analysis, we first introduce an interpolation result and then prove convergence in the above sense.

3.3.1 Interpolation analysis

In this section we consider the Legendre piecewise linear interpolants and their properties when they are employed to pass from the space V_h to the space $\widetilde{V}_h$. Let us first recall the definition of the Legendre interpolant.

Definition 3.4. Let $V = \mathscr{C}_0^0(D)$. We denote by $\Pi_h\colon V \to V_h$ and $\widetilde{\Pi}_h\colon V \to \widetilde{V}_h$ the Legendre piecewise linear interpolation operators on V_h and $\widetilde{V}_h$ respectively, i.e., for $v \in V$

$$\Pi_h v(x) = \sum x_j \in \mathscr{N}_h^I v(x_j)\varphi_j(x), \quad \widetilde{\Pi}_h v(x) = \sum \widetilde{x}_j \in \widetilde{\mathscr{N}_h^I} v(x_j)\widetilde{\varphi}_j(x), \tag{3.9}$$

where $\{\varphi_i\}_{i=1}^{N^I}$ and $\{\widetilde{\varphi}_i\}_{i=1}^{N^I}$ are the basis functions of V_h and $\widetilde{V}_h$ respectively.

In the following lemma we characterise the value that the Legendre interpolant $\widetilde{\Pi}_h$ assumes on the nodes of the original mesh $\mathcal{T}_h$. Let us remark that $V_h \subset C_0^0(D)$, thus the interpolant above can be employed on V_h.

Lemma 3.5. *With the notation of definition 3.4, it holds for all $v_h \in V_h$ and all $x_i \in \mathcal{N}_h^I$*

$$\begin{aligned} v_h(\widetilde{x}_i) &= v_h(x_i) + \bar{h}_i^p \alpha_i^\top \nabla v_h(\widetilde{x}_i), \\ \widetilde{\Pi}_h v_h(x_i) - v_h(x_i) &= \bar{h}_i^p \alpha_i^\top \Big(\nabla v_h(\widetilde{x}_i) - \nabla \widetilde{\Pi}_h v_h(x_i) \Big). \end{aligned} \tag{3.10}$$

Proof. We can now expand the function v_h, which is linear on the segment connecting x_i and $\widetilde{x}_i$, as

$$v_h(\widetilde{x}_i) = v_h(x_i) + \bar{h}_i^p \alpha_i^\top \nabla v_h(\widetilde{x}_i), \tag{3.11}$$

which is the first equality. Let us now denote $e_h = \widetilde{\Pi}_h v_h - v_h$. An exact Taylor expansion of the linear basis function $\widetilde{\varphi}_j$ gives

$$\begin{aligned} e_h(x_i) &= \sum_j v_h(\widetilde{x}_j)\varphi_j(x_i) - v_h(x_i) \\ &= \sum_j v_h(\widetilde{x}_j)\Big(\widetilde{\varphi}_j(\widetilde{x}_i) - \bar{h}_i^p \alpha_i^\top \nabla \widetilde{\varphi}_j(x_i)\Big) - v_h(x_i) \\ &= v_h(\widetilde{x}_i) - v_h(x_i) - \sum_j \bar{h}_i^p \alpha_i^\top v_h(\widetilde{x}_j) \nabla \widetilde{\varphi}_j(x_i). \end{aligned} \tag{3.12}$$

This, together with (3.11), yields

$$\begin{aligned} e_h(x_i) &= \bar{h}_i^p \alpha_i^\top \nabla v_h(\widetilde{x}_i) - \bar{h}_i^p \alpha_i^\top \sum_j v_h(\widetilde{x}_j) \nabla \widetilde{\varphi}_j(x_i) \\ &= \bar{h}_i^p \alpha_i^\top \Big(\nabla v_h(\widetilde{x}_i) - \nabla \widetilde{\Pi}_h v_h(x_i) \Big), \end{aligned} \tag{3.13}$$

which is the second desired equality and which therefore concludes the proof. □

We are not interested in all possible functions in V_h, but only in those which are close enough to a smooth function. The definition below sets the function space we consider in the following.

Definition 3.6. We denote by $V_h^{2,\infty} \subset V_h$ the space such that $v_h \in V_h^{2,\infty}$ if there exists $v \in W^{2,\infty}(D)$ satisfying

$$\| v - v_h \|_{W^{1,\infty}(D)} \leq C h |\log h| |v|_{W^{2,\infty}(D)}, \tag{3.14}$$

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

In the following Lemma, we provide a property of functions in $V_h^{2,\infty}$ which is quite consequent from the definition of the space. Since we repeatedly employ this result in the following, let us highlight it here.

Lemma 3.7. *Let $v_h \in V_h^{2,\infty}$. Then, if two triangles $K, K' \in \mathcal{T}_h$ share a vertex, it holds*

$$|\nabla v_h\big|_K - \nabla v_h\big|_{K'}| \leq Ch|\log h||v|_{W^{2,\infty}(D)}, \tag{3.15}$$

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

Proof. The proof follows from the triangle inequality. In particular, let $x \in K$ and $x' \in K'$. Then, there exists $v \in W^{2,\infty}$ such that

$$\begin{aligned} |\nabla v_h\big|_K - \nabla v_h\big|_{K'}| &\leq |\nabla v_h\big|_K - \nabla v(x)| + |\nabla v_h\big|_{K'} - \nabla v(x')| + |\nabla v(x) - \nabla v(x')| \\ &\leq 2|v_h - v|_{W^{1,\infty}(D)} + |v|_{W^{2,\infty}(D)}|x - x'|. \end{aligned} \tag{3.16}$$

The desired result follows from the definition of $V_h^{2,\infty}(D)$ and from the fact that since K and K' share a vertex, it is possible to bound $|x - x'| \leq Ch$. □

We now proceed to bound the difference between the gradient of v_h and of its interpolant $\widetilde{\Pi}_h v_h$ on a single element. In the proof, the notation for the reference triangle and for affine maps triangle is borrowed from [57, Chapter 4].

Lemma 3.8. *With the notation of definition 3.1 and definition 3.4, let $K \in \mathcal{T}_h$ be an element of the original mesh and let $\widetilde{K} \in \mathcal{T}_h$ be the corresponding element in the perturbed mesh. Then, it holds for all $v_h \in V_h^{2,\infty}$*

$$|\nabla v_h\big|_K - \nabla \widetilde{\Pi}_h v_h\big|_{\widetilde{K}}| \leq Ch^p|\log h||v|_{W^{2,\infty}(D)}, \tag{3.17}$$

where p is given in definition 3.1.

Proof. Let us denote by x_1, x_2, x_3 the vertices of K and by $\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3$ the corresponding vertices of $\widetilde{K}$. Let $\widehat{K}$ be the triangle with vertices $\widehat{x}_1 = (0,0)^\top$, $\widehat{x}_2 = (1,0)^\top$, $\widehat{x}_3 = (0,1)^\top$. We consider the affine maps $F_K \colon \widehat{K} \to K$ and $\widetilde{F}_K \colon \widehat{K} \to \widetilde{K}$ defined for all $\widehat{x} \in \widehat{K}$ as

$$F_K(\widehat{x}) = B_K\widehat{x} + b_K, \quad \widetilde{F}_K(\widehat{x}) = \widetilde{B}_K\widehat{x} + \widetilde{b}_K, \tag{3.18}$$

where $b_K = x_1$, $\widetilde{b}_K = \widetilde{x}_1$ and the matrices $B_K, \widetilde{B}_K \in \mathbb{R}^{2\times 2}$ are defined as

$$B_K = \Big(x_2 - x_1 \mid x_3 - x_1\Big), \quad B_{\widetilde{K}} = \Big(\widetilde{x}_2 - \widetilde{x}_1 \mid \widetilde{x}_3 - \widetilde{x}_1\Big), \tag{3.19}$$

so that $F_K(\widehat{x}_i) = x_i$ and $\widetilde{F}_K(\widehat{x}_i) = \widetilde{x}_i$ for $i = 1,2,3$. Furthermore, we define $\widehat{v}_h \colon \widehat{K} \to \mathbb{R}$ as $\widehat{v}_h := v_h \circ F_K$ and $\widetilde{\Pi}_h\widehat{v}_h \colon \widehat{K} \to \mathbb{R}$ as $\widetilde{\Pi}_h\widehat{v}_h := \widetilde{\Pi}_h v_h \circ \widetilde{F}_K$. Then, the chain rule yields

$$\nabla v_h\big|_K - \nabla\widetilde{\Pi}_h v_h\big|_{\widetilde{K}} = B_K^{-\top}\widehat{\nabla}\widehat{v}_h - B_{\widetilde{K}}^{-\top}\widehat{\nabla}\widetilde{\Pi}_h\widehat{v}_h, \tag{3.20}$$

where $\widehat{\nabla}$ is the gradient with respect to $\widehat{x}$. Let us write $\widetilde{B}_K = B_K + \Lambda$, where $\Lambda \in \mathbb{R}^{2\times 2}$ is the

random matrix defined as

$$\Lambda = \left(\bar{h}_2^p \alpha_2 - \bar{h}_1^p \alpha_1 \mid \bar{h}_3^p \alpha_3 - \bar{h}_1^p \alpha_1 \right), \tag{3.21}$$

and remark that we can rewrite $B_{\widetilde{K}}^{-\top}$ with algebraic operations as

$$B_{\widetilde{K}}^{-\top} = (B_K + \Lambda)^{-\top} = B_K^{-\top} (I + B_K^{-1} \Lambda)^{-\top} = B_K^{-\top} (I - \Gamma), \tag{3.22}$$

where the random matrix $\Gamma \in \mathbb{R}^{2\times 2}$ is given by the series expansion $\Gamma = \sum_{j=0}^{\infty} \Gamma_j$, with

$$\Gamma_j = (-1)^j (\Lambda^\top B_K^{-\top})^{j+1}. \tag{3.23}$$

Let us remark that since $|B_K^{-1}| \leq Ch^{-1}$ (see e.g. [16, Theorem 3.1.3] or [57, Lemma 4.3]) and due to assumption 3.3, the single addends Γ_j satisfy

$$|\Gamma_j| \leq |B_K^{-1}|^{j+1} |\Lambda|^{j+1} \leq Ch^{-j-1} h^{jp} \leq Ch^{j(p-1)-1}, \tag{3.24}$$

almost surely. Moreover, we have that $\Gamma_j = -\Gamma_{j-1}\Gamma_0$. Let us remark that

$$\widehat{\nabla} \widetilde{\Pi}_h \widehat{v}_h = \begin{pmatrix} \widetilde{\Pi}_h \widehat{v}_h(\widehat{x}_2) - \widetilde{\Pi}_h \widehat{v}_h(\widehat{x}_1) \\ \widetilde{\Pi}_h \widehat{v}_h(\widehat{x}_3) - \widetilde{\Pi}_h \widehat{v}_h(\widehat{x}_1) \end{pmatrix} = \begin{pmatrix} \widetilde{\Pi}_h v_h(\widetilde{x}_2) - \widetilde{\Pi}_h v_h(\widetilde{x}_1) \\ \widetilde{\Pi}_h v_h(\widetilde{x}_3) - \widetilde{\Pi}_h v_h(\widetilde{x}_1) \end{pmatrix}. \tag{3.25}$$

Since the interpolation is exact on the nodes of the mesh $\widetilde{\mathcal{T}_h}$ and due to lemma 3.5, this yields

$$\widehat{\nabla} \widetilde{\Pi}_h \widehat{v}_h = \begin{pmatrix} v_h(\widetilde{x}_2) - v_h(\widetilde{x}_1) \\ v_h(\widetilde{x}_3) - v_h(\widetilde{x}_1) \end{pmatrix} = \begin{pmatrix} v_h(x_2) - v_h(x_1) \\ v_h(x_3) - v_h(x_1) \end{pmatrix} + \gamma = \widehat{\nabla} \widehat{v}_h + \gamma, \tag{3.26}$$

where

$$\gamma = \begin{pmatrix} \bar{h}_2^p \alpha_2^\top \nabla v_h(\widetilde{x}_2) - \bar{h}_1^p \alpha_1^\top \nabla v_h(\widetilde{x}_1) \\ \bar{h}_3^p \alpha_3^\top \nabla v_h(\widetilde{x}_3) - \bar{h}_1^p \alpha_1^\top \nabla v_h(\widetilde{x}_1) \end{pmatrix}. \tag{3.27}$$

Employing the properties of the sequence of matrices Γ_j, we can now rewrite (3.20) as

$$\begin{aligned} \nabla v_h\big|_K - \nabla \widetilde{\Pi}_h v_h\big|_{\widetilde{K}} &= B_K^{-\top} (-\gamma + \Gamma \widehat{\nabla} \widehat{v}_h + \Gamma\gamma) \\ &= B_K^{-\top} \left(-\gamma + \Gamma_0 \widehat{\nabla} \widehat{v}_h + \sum_{j=1}^{\infty} (\Gamma_j \widehat{\nabla} \widehat{v}_h + \Gamma_{j-1}\gamma) \right) \\ &= B_K^{-\top} \sum_{j=0}^{\infty} \Gamma_{j-1} (\gamma - \Gamma_0 \widehat{\nabla} \widehat{v}_h), \end{aligned} \tag{3.28}$$

where $\Gamma_{-1} = -I$. We can now compute explicitly the difference $\gamma - \Gamma_0 \widehat{\nabla} \widehat{v}_h$ as

$$\begin{aligned}\gamma - \Gamma_0 \widehat{\nabla} \widehat{v}_h &= \gamma - \Lambda^\top B_K^{-\top} \widehat{\nabla} \widehat{v}_h = \gamma - \Lambda^\top \nabla v_h\big|_K \\ &= \begin{pmatrix} \bar{h}_2^p \alpha_2^\top \nabla v_h(\widetilde{x}_2) - \bar{h}_1^p \alpha_1^\top \nabla v_h(\widetilde{x}_1) \\ \bar{h}_3^p \alpha_3^\top \nabla v_h(\widetilde{x}_3) - \bar{h}_1^p \alpha_1^\top \nabla v_h(\widetilde{x}_1) \end{pmatrix} - \begin{pmatrix} (\bar{h}_2^p \alpha_2^\top - \bar{h}_1^p \alpha_1^\top) \nabla v_h\big|_K \\ (\bar{h}_3^p \alpha_3^\top - \bar{h}_1^p \alpha_1^\top) \nabla v_h\big|_K \end{pmatrix} \\ &= \begin{pmatrix} \bar{h}_2^p \alpha_2^\top \left(\nabla v_h(\widetilde{x}_2) - \nabla v_h\big|_K\right) + \bar{h}_1^p \alpha_1^\top \left(\nabla v_h\big|_K - \nabla v_h(\widetilde{x}_1)\right) \\ \bar{h}_3^p \alpha_3^\top \left(\nabla v_h(\widetilde{x}_3) - \nabla v_h\big|_K\right) + \bar{h}_1^p \alpha_1^\top \left(\nabla v_h\big|_K - \nabla v_h(\widetilde{x}_1)\right) \end{pmatrix}.\end{aligned} \tag{3.29}$$

Due to lemma 3.7 and to assumptions on the mesh, we have therefore

$$|\gamma - \Gamma_0 \widehat{\nabla} \widehat{v}_h| \leq Ch|\log h||v|_{W^{2,\infty}(D)}, \tag{3.30}$$

almost surely, which, replaced in (3.28), implies

$$|\nabla v_h\big|_K - \nabla \widetilde{\Pi}_h v_h\big|_{\widetilde{K}}| \leq h^{p+1}|\log h||v|_{W^{2,\infty}(D)}|B_K^{-\top}| \sum_{j=0}^{\infty} |\Gamma_{j-1}|. \tag{3.31}$$

From the definition of Γ_j, $j = -1, 0, \ldots, \infty$, we have

$$\sum_{j=0}^{\infty} |\Gamma_{j-1}| \leq \sum_{j=0}^{\infty} |\Lambda|^j |B_K^{-1}|^j \leq C \sum_{j=0}^{\infty} h^{(p-1)j}, \tag{3.32}$$

and since $h < 1$ and $p \geq 1$, this is bounded independently of h. Finally,

$$|\nabla v_h\big|_K - \nabla \widetilde{\Pi}_h v_h\big|_{\widetilde{K}}| \leq Ch^p|\log h||v|_{W^{2,\infty}(D)}, \tag{3.33}$$

which is the desired result. □

We can now prove an interpolation result in $L^\infty(D)$

Lemma 3.9. *With the notation of definition 3.6, let $v_h \in V_h^{2,\infty}$. Then, with the notation of definition 3.4, it holds*

$$\|v_h - \widetilde{\Pi}_h v_h\|_{L^\infty(D)} \leq Ch^{p+1}|\log h|, \tag{3.34}$$

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

Proof. Let us denote $e_h = \widetilde{\Pi}_h v_h - v_h$ and let us consider $x_i \in \mathcal{N}^I$. By definition $e_h(\widetilde{x}_i) = 0$ for all $i = 0, \ldots, N$ and due to lemma 3.5

$$e_h(x_i) = h^p \alpha_i \left(\nabla v_h(\widetilde{x}_i) - \nabla \widetilde{\Pi}_h v_h(x_i)\right) =: h^p \alpha_i \varepsilon_i. \tag{3.35}$$

Let us denote by K the element of $\mathcal{T}_h$ such that the corresponding element $\widetilde{K} \in \widetilde{\mathcal{T}_h}$ contains x_i. Furthermore, let us denote by K' the element in the original mesh containing $\widetilde{x}_i$. We refer

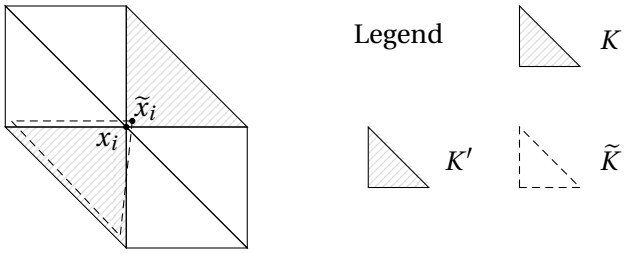

Figure 3.1 – Scheme for the proof of Lemma 3.9. The triangle with solid grey lines on the background is K, the triangle with dashed grey lines on the background is K' and the triangle with dashed borders is $\widetilde{K}$.

to fig. 3.1 for a schematic representation of these elements. With this notation, we have

$$\nabla v_h(\widetilde{x}_i) = \nabla v_h\big|_{K'}, \quad \nabla\widetilde{\Pi}_h v_h(x_i) = \nabla\widetilde{\Pi}_h v_h\big|_{\widetilde{K}}, \tag{3.36}$$

and we can then decompose ε_i as $\varepsilon_i = \varepsilon_{i,1} + \varepsilon_{i,2}$ with

$$\varepsilon_{i,1} = \nabla v_h\big|_{K'} - \nabla v_h\big|_K, \quad \varepsilon_{i,2} = \nabla v_h\big|_K - \nabla\widetilde{\Pi}_h v_h\big|_{\widetilde{K}}. \tag{3.37}$$

Due to lemma 3.7, we have

$$|\varepsilon_{i,1}| \leq Ch|\log h||v|_{W^{2,\infty}(D)}, \tag{3.38}$$

Moreover, due to lemma 3.8, we have

$$|\varepsilon_{i,2}| \leq Ch^p|\log h||v|_{W^{2,\infty}(D)} \tag{3.39}$$

Since $p \geq 1$, the triangular inequality yields $\varepsilon_i \leq Ch|\log h||v|_{W^{2,\infty}(D)}$ for each node. Replacing this bound in (3.35), we get for each $x_i \in \mathcal{N}_h^I$

$$|e_h(x_i)| \leq Ch^{p+1}\alpha_i|\log h||v|_{W^{2,\infty}(D)}. \tag{3.40}$$

Let us now remark that since by definition $e_h(\widetilde{x}_i) = 0$ for all modified nodes, and since e_h is linear on D, the maximum of e_h has to be realised on one of the nodes of the original mesh. Hence

$$\|e_h\|_{L^\infty(D)} = \max_{x_i \in \mathcal{N}_h^I} |e_h(x_i)| \leq Ch^{p+1}|\log h||v|_{W^{2,\infty}(D)}, \tag{3.41}$$

which implies the desired result. □

We now consider the interpolation error in $H^1(D)$.

Lemma 3.10. *With the notation of definition 3.6, let $v_h \in V_h^{2,\infty}$. Then, with the notation of definition 3.4, it holds*

$$\|v_h - \widetilde{\Pi}_h v_h\|_{H^1(D)} \leq Ch^{(p+1)/2}|\log h|, \tag{3.42}$$

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

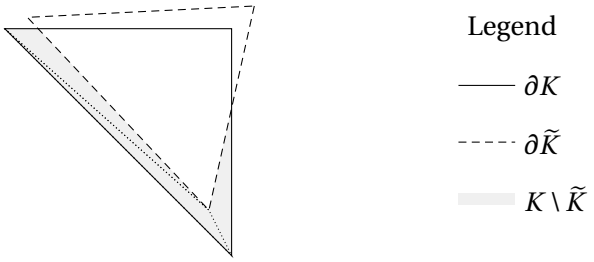

Figure 3.2 – Scheme for the proof of lemma 3.10. The triangle with solid border is $K \in \mathcal{T}_h$, the one with dashed border is $\widetilde{K} \in \widetilde{T}_h$. The area filled in grey is $K \setminus \widetilde{K}$, and the dotted lines give one of the possible subdivision in triangles of $K \setminus \widetilde{K}$.

Proof. First, let us recall that for any triangle of sides of length a, b, c and of area A it holds [22]

$$A \leq \frac{4\sqrt{3}}{9} \frac{abc}{a+b+c}. \tag{3.43}$$

Let us now consider an element $K \in \mathcal{T}_h$ and the corresponding element $\widetilde{K} \in \widetilde{\mathcal{T}_h}$. It is clear (see e.g. fig. 3.2) that it is possible to subdivide $K \setminus \widetilde{K}$ into a bounded number of triangles for which the length one side is bounded by Ch^p and the length of the two other side is bounded by Ch. Therefore, due to (3.43) we have

$$|K \setminus \widetilde{K}| \leq C \frac{h^{p+2}}{h+h+h^p} \leq Ch^{p+1}. \tag{3.44}$$

Moreover, we remark that

$$|K \cap \widetilde{K}| \leq |K| \leq Ch^2. \tag{3.45}$$

Let us now denote by N_K the number of triangles in which the set $K \setminus \widetilde{K}$ is divided and by $K_{\text{diff}}^{(i)}$, $i = 1, \dots, N_K$ these triangles. We have

$$\int_K |\nabla e_h|^2 \, \mathrm{d}x = \int_{K \cap \widetilde{K}} |\nabla e_h|^2 \, \mathrm{d}x + \sum_{i=1}^{N_K} \int_{K_{\text{diff}}^{(i)}} |\nabla e_h|^2 \, \mathrm{d}x. \tag{3.46}$$

Now, lemma 3.8 and (3.45) yield

$$\int_{K \cap \widetilde{K}} |\nabla e_h|^2 \, \mathrm{d}x = \int_{K \cap \widetilde{K}} |\nabla v_h\big|_K - \nabla \widetilde{\Pi}_h v_h\big|_{\widetilde{K}}|^2 \, \mathrm{d}x \leq Ch^{2p+2} |\log h|^2 |v|^2_{W^{2,\infty}(D)}. \tag{3.47}$$

Let us now consider the second term. Each triangle $K_{\text{diff}}^{(i)}$ intersects a finite number $N_K^{(i)}$ of triangles in the mesh $\widetilde{\mathcal{T}_h}$. We denote by $\widetilde{K}^{(i,j)}$ for $j = 1, \dots, N_K^{(i)}$ these triangles and by $\widetilde{K}_{\text{diff}}^{(i,j)} = \widetilde{K}^{(i,j)} \cap \widetilde{K}_{\text{diff}}^{(i)}$, for which we remark that it holds

$$|\widetilde{K}_{\text{diff}}^{(i,j)}| \leq |K_{\text{diff}}^{(i)}| \leq |K \setminus \widetilde{K}| \leq Ch^{p+1}. \tag{3.48}$$

Finally, for each i, j we denote by $K^{(i,j)}$ the element of $\mathcal{T}_h$ corresponding to $\widetilde{K}^{(i,j)} \in \mathcal{T}_h$ and

remark that it is a neighbour of K. Therefore

$$\int_{K_{\text{diff}}^{(i)}} |e_h|^2 \, \mathrm{d}x = \sum_{j=1}^{N_K^{(i)}} \int_{\widetilde{K}_{\text{diff}}^{(i,j)}} |e_h|^2 \, \mathrm{d}x, \tag{3.49}$$

where due to Young's inequality we have

$$\begin{aligned} \int_{\widetilde{K}_{\text{diff}}^{(i,j)}} |\nabla e_h|^2 \, \mathrm{d}x &= \int_{\widetilde{K}_{\text{diff}}^{(i,j)}} |\nabla v_h|_K - \nabla \widetilde{\Pi}_h v_h|_{\widetilde{K}^{(i,j)}}|^2 \, \mathrm{d}x \\ &\leq 2 \left(|\nabla v_h|_K - \nabla v_h|_{K^{(i,j)}}|^2 + |\nabla \widetilde{\Pi}_h v_h|_{\widetilde{K}^{(i,j)}} - \nabla v_h|_{K^{(i,j)}}|^2 \right) |\widetilde{K}_{\text{diff}}^{(i,j)}|. \end{aligned} \tag{3.50}$$

Due to lemma 3.7, we have first

$$|\nabla v_h|_K - \nabla v_h|_{K^{(i,j)}}|^2 \leq C h^2 |\log h|^2 |v|^2_{W^{2,\infty}(D)}, \tag{3.51}$$

and due to lemma 3.8, we obtain

$$|\nabla \widetilde{\Pi}_h v_h|_{\widetilde{K}^{(i,j)}} - \nabla v_h|_{K^{(i,j)}}|^2 \leq C h^{2p} |\log h|^2 |v|^2_{W^{2,\infty}(D)}. \tag{3.52}$$

Therefore, replacing these two inequalities and (3.44) in (3.50) and since $p \geq 1$ and $h < 1$, we obtain for a constant $C > 0$

$$\int_{\widetilde{K}_{\text{diff}}^{(i,j)}} |\nabla e_h|^2 \, \mathrm{d}x \leq C h^{p+3} |\log h|^2 |v|^2_{W^{2,\infty}(D)}. \tag{3.53}$$

Hence, we get

$$\int_{K \setminus \widetilde{K}} |\nabla e_h|^2 \, \mathrm{d}x \leq C N_K \left(\sum_{i=1}^{N_K} N_K^{(i)} \right) h^{p+3} \left(|\log h| + 1 \right)^2 |v|^2_{W^{2,\infty}(D)}. \tag{3.54}$$

Combining (3.47) and (3.54) and since $2p+2 \geq p+3$ for $p \geq 1$, we conclude that there exists a constant $C > 0$ independent of h but dependent on v such that

$$\int_K |\nabla e_h|^2 \, \mathrm{d}x \leq C h^{p+3} |\log h|^2. \tag{3.55}$$

Finally, due to assumption on the mesh, we have that $N \leq C h^2$ and therefore

$$|e_h|^2_{H^1(D)} = \sum_{K \in \mathcal{T}_h} \int_K |\nabla e_h|^2 \, \mathrm{d}x \leq C h^{p+1} |\log h|^2, \tag{3.56}$$

which implies the desired result. □

3.3.2 The sum space

In order to prove convergence of the probabilistic solution, and moreover the closeness of u_h to $\widetilde{u}_h$ in the sense of (3.8), we first need to define a convenient function space which is finite

dimensional but which contains both V_h and $\widetilde{V}_h$.

Definition 3.11. Let us denote by $V_h^+ \subset V$ the space of functions that can be written as the sum of a function in V_h and a function in $\widetilde{V}_h$, i.e., for any function $v_h^+ \in V_h^+$ there exists functions $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$ such that $v_h^+ = v_h + \widetilde{v}_h$.

Remark 3.12. Let us remark that in our setting of homogeneous boundary conditions $V_h \cap \widetilde{V}_h = \{0\}$ almost surely. Therefore, the space V_h^+ is given by the direct sum $V_h^+ = V_h \oplus \widetilde{V}_h$ and the decomposition of $v_h^+ \in V_h^+$ is unique. Moreover, since $\dim(V_h) = \dim(\widetilde{V}_h) = N_I$, we have $\dim(V_h^+) = 2N_I$. Moreover, let us remark that we are not building a so-called supermesh as in [26, 25, 21]

The following result characterizes the distance of the finite elements solutions on the spaces V_h and $\widetilde{V}_h$.

Lemma 3.13. *Let u_h and $\widetilde{u}_h$ be the solutions of*

$$a(u_h, v_h) = F(v_h), \quad a(\widetilde{u}_h, \widetilde{v}_h) = F(\widetilde{v}_h), \tag{3.57}$$

for all $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$. Then, it holds for all $v_h, w_h \in V_h$ and for all $\widetilde{v}_h, \widetilde{w}_h \in \widetilde{V}_h$

$$\|u_h - \widetilde{u}_h\|_V^2 \leq C\left(\|u_h^+ - w_h\|_V \|\widetilde{u}_h - v_h\|_V + \|u_h^+ - \widetilde{w}_h\|_V \|u_h - \widetilde{v}_h\|_V\right), \tag{3.58}$$

where $C > 0$ is a constant independent of h and where $u_h^+ \in V_h^+$ is the solution of

$$a(u_h^+, v_h^+) = F(v_h^+), \tag{3.59}$$

for all $v_h^+ \in V_h^+$.

Proof. Since V_h and $\widetilde{V}_h$ are both subspaces of V_h^+, we have due to Galerkin's orthogonality

$$\begin{aligned} a(u_h^+ - u_h, v_h) &= 0, \quad \forall v_h \in V_h, \\ a(u_h^+ - \widetilde{u}_h, \widetilde{v}_h) &= 0, \quad \forall \widetilde{v}_h \in \widetilde{V}_h, \end{aligned} \tag{3.60}$$

which means that u_h and $\widetilde{u}_h$ are the elliptic projection of u_h^+ onto V_h and $\widetilde{V}_h$ respectively. Hence, due to Cea's lemma

$$\|u_h^+ - u_h\|_V \leq C\|u_h^+ - w_h\|_V, \quad \|u_h^+ - \widetilde{u}_h\|_V \leq C\|u_h^+ - \widetilde{w}_h\|_V, \tag{3.61}$$

for all $w_h \in V_h$ and $\widetilde{w}_h \in \widetilde{V}_h$, where $C = M/\alpha$. Using the coercivity on V of $a(\cdot,\cdot)$, adding and subtracting $a(u_h^+, u_h - \widetilde{u}_h)$ and due to (3.60) we have for all $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$

$$\begin{aligned} \alpha\|u_h - \widetilde{u}_h\|_V^2 &\leq a(u_h - \widetilde{u}_h, u_h - \widetilde{u}_h) \\ &= a(u_h - u_h^+, u_h - \widetilde{u}_h) + a(u_h^+ - \widetilde{u}_h, u_h - \widetilde{u}_h) \\ &= a(u_h - u_h^+, v_h - \widetilde{u}_h) + a(u_h^+ - \widetilde{u}_h, u_h - \widetilde{v}_h). \end{aligned} \tag{3.62}$$

Due to the continuity of the bilinear form we then have for all $w_h \in V_h$ and $\widetilde{w}_h \in \widetilde{V}_h$

$$\begin{aligned}\alpha\|u_h - \widetilde{u}_h\|_V^2 &\leq M\Big(\|u_h^+ - u_h\|_V\|\widetilde{u}_h - v_h\|_V + \|u_h^+ - \widetilde{u}_h\|_V\|u_h - \widetilde{v}_h\|_V\Big) \\ &\leq \frac{M^2}{\alpha}\Big(\|u_h^+ - w_h\|_V\|\widetilde{u}_h - v_h\|_V + \|u_h^+ - \widetilde{w}_h\|_V\|u_h - \widetilde{v}_h\|_V\Big),\end{aligned} \tag{3.63}$$

which is the desired result. □

Let us remark that the Lemma above holds true for any choice of V_h and $\widetilde{V}_h$, not necessarily disjoint, and for any space V_h^+ such that $V_h \cup \widetilde{V}_h \subseteq V_h^+$. For the next result, we instead consider the setting in which V_h and $\widetilde{V}_h$ are the fixed and randomly perturbed finite element spaces of definition 3.1.

Lemma 3.14. *Let $u_h^+ \in V_h^+$ be the solution of* (3.59)*, and let us denote by z_h and $\widetilde{z}_h$ its unique components in V_h and $\widetilde{V}_h$, respectively, i.e., $u_h^+ = z_h + \widetilde{z}_h$. Then, it holds*

$$\|z_h - \widetilde{z}_h\|_V \leq Ch^r. \tag{3.64}$$

Proof. □

Corollary 3.15. *With the notation of lemma 3.14 and of definition 3.6, we have $z_h \in V_h^{2,\infty}$ and $\widetilde{z}_h \in \widetilde{V}_h^{2,\infty}$.*

Proof. Let us consider without loss of generality z_h. Due to (3.61), we have

$$\|u - u_h^+\|_V \leq \|u - u_h\|_V, \tag{3.65}$$

□

We now introduce a result of interpolation with the Legendre interpolants defined in definition 3.4.

Lemma 3.16. *Let Π_h and $\widetilde{\Pi}_h$ be defined in definition 3.4. Then, for all $v_h^+ \in V_h^+$ it holds*

$$\Pi_h v_h^+ - v_h^+ = \Pi_h \widetilde{v}_h - \widetilde{v}_h, \quad \widetilde{\Pi}_h v_h^+ - v_h^+ = \Pi_h v_h - v_h, \tag{3.66}$$

where $v_h \in V_h$, $\widetilde{v}_h \in \widetilde{V}_h$ and $v_h^+ = v_h + \widetilde{v}_h$.

Proof. The result is implied by the linearity of Π_h and $\widetilde{\Pi}_h$ and since the restriction of Π_h on V_h is the identity function (respectively, $\widetilde{\Pi}_h$ on $\widetilde{V}_h$). □

3.3.3 Convergence result

We now present here a classic convergence result for the finite elements method [16, Theorem 3.3.7], which allows to control the supremum of the error under smoothness assumptions on the solution.

Theorem 3.17. *Let u be the solution of* (3.2) *and $u_h \in V_h$ be the solution of* (3.3)*. Then, if $u \in W^{2,\infty}(D)$, it holds*

$$\begin{aligned} \|u - u_h\|_{L^\infty(D)} &\leq Ch^2 |\log h|^{3/2} |u|_{W^{2,\infty}(D)}, \\ |u - u_h|_{W^{1,\infty}(D)} &\leq Ch |\log h| |u|_{W^{2,\infty}(D)}, \end{aligned} \tag{3.67}$$

where $C > 0$ is a constant independent of h. In particular, with the notation of definition 3.6, we have that $u_h \in V_h^{2,\infty}$.

We can now prove the main result of a priori convergence for the probabilistic solution.

Theorem 3.18. *Let u be the solution of* (3.2) *and let u_h and $\widetilde{u}_h$ be the solutions of*

$$a(u_h, v_h) = F(v_h), \quad a(\widetilde{u}_h, \widetilde{v}_h) = F(\widetilde{v}_h), \tag{3.68}$$

for all $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$. Then, if $u \in W^{2,\infty}(D)$, it holds for $V = H_0^1(D)$

$$\|u_h - \widetilde{u}_h\|_V \leq \quad a.s., \tag{3.69}$$

and moreover, it holds

$$\|\widetilde{u}_h - u\| \leq \quad a.s. \tag{3.70}$$

Proof. Considering lemma 3.13 with $v_h = \Pi_h \widetilde{u}_h$, $w_h = \Pi_h u_h^+$, $\widetilde{v}_h = \widetilde{\Pi}_h u_h$ and $\widetilde{w}_h = \widetilde{\Pi}_h u_h^+$, we get

$$\|u_h - \widetilde{u}_h\|_V^2 \leq C \left(\|u_h^+ - w_h\|_V \|\widetilde{u}_h - v_h\|_V + \|u_h^+ - \widetilde{w}_h\|_V \|u_h - \widetilde{v}_h\|_V \right), \tag{3.71}$$

□

3.4 A posteriori error analysis

Several techniques exist for obtaining a posteriori error estimators in the framework of the FEM (see [64] for an overview), with the twofold goal of controlling the quality of numerical solutions and hence improve the meshing procedure to maximise efficiency. The main purpose of probabilistic numerical methods is to quantify the uncertainty introduced by approximate computations [33]. For the reasons above, we believe that deriving an error estimator from a family of numerical solutions fits perfectly in the probabilistic framework. In this section we present such a procedure for a probabilistic error estimation.

Assumption 3.19. Let $u_h^+ \in V_h^+$ be defined in (3.59). Then we assume there exists $0 \leq \beta < 1$ such that

$$\|u - u_h^+\| \leq \beta \|u - u_h\|, \tag{3.72}$$

where $\|u\|^2 = a(u,u)$. Moreover, there exists a constant $\gamma > 0$ such that

$$\|u_h - u_h^+\| \leq \gamma \|u_h - \widetilde{u}_h\|, \tag{3.73}$$

almost surely, where $\widetilde{u}_h$ is the probabilistic solution.

Let us remark that since $V_h \subset V_h^+$, we have $\beta \leq 1$ for the best approximation property of the Galerkin method and that assumption 3.19 is often denoted in literature as the saturation assumption.

Lemma 3.20. *Let us denote by $z_h \in V_h$ the function $z_h = w_h - u_h/2$. Then*

$$\|z_h - \widetilde{\Pi}_h z_h\| \leq \dots \tag{3.74}$$

Proof.

$$\|z_h\| \leq \frac{1}{2}\|w_h - \widetilde{w}_h\|. \tag{3.75}$$

□

Lemma 3.21. *Under ..., there exists $\gamma > 0$ independent of h and p such that*

$$\|u_h - u_h^+\| \leq \gamma \|u_h - \widetilde{u}_h\|, \tag{3.76}$$

almost surely in Ω.

Proof. Let us write $u_h^+ = w_h + \widetilde{w}_h$, where w_h and $\widetilde{w}_h$ are the two components of u_h^+ in V_h and $\widetilde{V}_h$ respectively. For any $v_h^+ \in V_h^+$, $v_h^+ = v_h + \widetilde{v}_h$, with $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$, by Galerkin orthogonality

$$\begin{aligned} a(u_h^+ - u_h, v_h^+) &= a(u_h^+ - u_h, \widetilde{v}_h) - a(u_h^+ - \widetilde{u}_h, \widetilde{v}_h) \\ &= a(\widetilde{u}_h - u_h, \widetilde{v}_h). \end{aligned} \tag{3.77}$$

Choosing $v_h^+ = u_h^+ - u_h$, we have $\widetilde{v}_h = \widetilde{w}_h$ and

$$\|u_h^+ - u_h\|^2 = a(\widetilde{u}_h - u_h, \widetilde{w}_h). \tag{3.78}$$

The same procedure applied to $u_h^+ - \widetilde{u}_h$ yields

$$\|u_h^+ - \widetilde{u}_h\|^2 = a(u_h - \widetilde{u}_h, w_h). \tag{3.79}$$

Hence

$$\|u_h^+ - u_h\|^2 + \|u_h^+ - \widetilde{u}_h\|^2 = a(u_h - \widetilde{u}_h, w_h - \widetilde{w}_h). \tag{3.80}$$

Let us introduce the functions $z_h = w_h - u_h/2 \in V_h$ and $\widetilde{z}_h = \widetilde{w}_h - \widetilde{u}_h/2 \in \widetilde{V}_h$. Then

$$\begin{aligned} \|u_h^+ - u_h\|^2 + \|u_h^+ - \widetilde{u}_h\|^2 &= \frac{1}{2} a(u_h - \widetilde{u}_h, u_h - \widetilde{u}_h) + a(u_h - \widetilde{u}_h, w_h - \frac{u_h}{2} - (\widetilde{w}_h - \frac{\widetilde{u}_h}{2})) \\ &= \frac{1}{2}\|u_h - \widetilde{u}_h\|^2 + a(u_h - \widetilde{u}_h, z_h - \widetilde{z}_h). \end{aligned} \tag{3.81}$$

Consider now the second term in the sum. Adding and subtracting $a(u_h^+, z_h - \widetilde{z}_h)$ and considering Galerkin orthogonality we obtain

$$a(u_h - \widetilde{u}_h, z_h - \widetilde{z}_h) = a(u_h - u_h^+, v_h - \widetilde{z}_h) + a(u_h^+ - \widetilde{u}_h, z_h - \widetilde{v}_h), \tag{3.82}$$

for all $v_h \in V_h$ and $\widetilde{v}_h \in \widetilde{V}_h$. Hence, applying Cauchy–Schwarz and Young's inequalities we obtain

$$\|u_h^+ - u_h\|^2 + \|u_h^+ - \widetilde{u}_h\|^2 \leq \|u_h - \widetilde{u}_h\|^2 + \inf_{v_h \in V_h} \|\widetilde{z}_h - v_h\|^2 + \inf_{\widetilde{v}_h \in \widetilde{V}_h} \|z_h - \widetilde{v}_h\|^2. \tag{3.83}$$

□

Moreover, since the perturbed mesh and the original mesh could switch their roles by changing the sign to the random perturbations, the same assumption as (3.73) should be imposed for the probabilistic solution, i.e.

$$\|\widetilde{u}_h - u_h^+\| \leq \widetilde{\gamma}\|u_h - \widetilde{u}_h\|. \tag{3.84}$$

Applying the triangular inequality, we get

$$\begin{aligned} (\gamma + \widetilde{\gamma})\|u_h - \widetilde{u}_h\| &\geq \|\widetilde{u}_h - u_h^+\| + \|u_h - u_h^+\| \\ &\geq \|u_h - \widetilde{u}_h\|, \end{aligned} \tag{3.85}$$

which implies that $(\gamma + \widetilde{\gamma}) \geq 1$. The duality in the roles of deterministic and probabilistic meshes implies that γ and $\widetilde{\gamma}$ should be in general approximately equal, at least asymptotically. Hence, the lower bound above guarantees that neither γ nor $\widetilde{\gamma}$ should tend to zero with $h \to 0$.

It is known [10] that under assumption 3.19 the estimate

$$\|u_h - u_h^+\|_a \leq \|u - u_h\|_a \leq \frac{1}{1-\beta}\|u_h - u_h^+\|_a, \tag{3.86}$$

holds almost surely. The quantity $\|u_h - u_h^+\|_a$ thus serves as an a posteriori error estimator for the error. However, computations involving the sum space V_h^+ are often intractable if the dimension $d > 1$. Hence, we further expand the upper bound thanks to (3.73) as

$$\|u - u_h\| \leq \frac{\gamma}{1-\beta}\|u_h - \widetilde{u}_h\|, \tag{3.87}$$

which means that the difference between the deterministic and the probabilistic solutions can be employed as an a posteriori upper bound for the error.

Remark 3.22. Let us remark that the value of β is influenced by the choice of p in assumption 3.3. Let us consider the limit case of $p \to \infty$. In this case, the spaces V_h and $\widetilde{V}_h$ coincide, and in turn coincide both with V_h^+. Hence, the space V_h^+ is in the limit not wider than V_h and one expects $\beta \to 1$. We hence postulate that $\beta = \beta(h,p)$ takes the form

$$\beta(h,p) = 1 - \beta_1 h^{\beta_2(p-1)}, \tag{3.88}$$

for some $0 < \beta_1 \leq 1$ and $\beta_2 > 0$. This is motivated by the fact that the two terms in (3.87) converge with the same rate $\mathcal{O}(h)$ in case $p = 1$ due to a priori error results. Hence, in this case, $\beta(h,1)$ is independent of h and equals a constant value β. Conversely, if $p > 1$, one gets on the right hand side a term of order $\mathcal{O}(h^{\beta_2(1-p)} h^{(p+1)/2})$, bounding a term of order $\mathcal{O}(h)$ on the left hand side. Hence, we impose

$$\beta_2(1-p) + \frac{p+1}{2} \leq 1, \tag{3.89}$$

which, since $p > 1$, gives $\beta_2 \geq 1/2$. Numerical experiments confirm the qualitative behaviour of the function $\beta(h,p)$ explained above, and lead to the good working practice of fixing $p = 1$.

A more robust estimator could be obtained by averaging a family of M probabilistic solutions $\widetilde{u}_h^{(i)}$, $i = 1,\dots,M$, obtained by M i.i.d. random perturbations of the original mesh. In particular, we have

$$\|u - u_h\|^2 \leq C\,\mathbb{E}\|u_h - \widetilde{u}_h\|^2 =: C\eta_h^2, \tag{3.90}$$

where we approximate the estimator η_h via Monte Carlo sampling as

$$\eta_h \approx \sqrt{\frac{1}{M}\sum_{i=1}^{M} \|u_h - \widetilde{u}_h^{(i)}\|^2}. \tag{3.91}$$

Taking the expectation over several realisations should in practice provide a sharper error estimator, as in case $p = 1$ a good portion of the domain D is explored by the vertices of several realisations of the random mesh. Let us consider for simplicity the case $\kappa \equiv 1$, so that $\|u\| = \|\nabla u\|_{L^2(D)}$ for all $u \in H_0^1(D)$. In this case, we have

$$\begin{aligned} \eta_h &= \int_K \mathbb{E}|\nabla(u_h - \widetilde{u}_h)|^2 \,\mathrm{d}x \\ &\approx \int_K \mathbb{E}|\mathbb{E}(\nabla u_h) - \nabla \widetilde{u}_h|^2 \,\mathrm{d}x \\ &= \int_K \mathrm{tr}(\mathrm{Var}\,\nabla u_h)\,\mathrm{d}x. \end{aligned} \tag{3.92}$$

Hence, following the probabilistic numerics canon, it is possible to interpret the error estimator as an integral measure of the statistical dispersion of numerical solutions over the domain.

We now consider the task of adapting the mesh. Given the error estimator derived above and a prescribed tolerance, we apply a standard technique for generating a sequence of meshes, which we briefly summarise in the following. Let us first split the estimator over the elements

Algorithm 1: Probabilistic mesh adaptivity.

Data: $\mathcal{T}_h^{(0)}$, tolerance ϵ, safety factors fac_1, fac_2, $M \in \mathbb{N}$.

Set $i = 0$;
while $\eta_h > \epsilon \|u_h\|$ **do**
  Compute u_h and $\|u_h\|$;
  Draw M random meshes and compute $\widetilde{u}_h^{(j)}$ for $j = 1, \dots, M$;
  for $K \in \mathcal{T}_h^{(i)}$ **do**
    Compute η_K ;
    if $\eta_K > \text{fac}_1 \epsilon \|u_h\| / \sqrt{N}$ **then**
      Mark element K for refinement ;
    else if $\eta_K < \text{fac}_2 \epsilon \|u_h\| / \sqrt{N}$ **then**
      Mark element K for coarsening ;
  Build $\mathcal{T}^{(i+1)}$;
  Set $i \leftarrow i + 1$;

of the original mesh as

$$\begin{aligned} \eta_h^2 &= \sum_{K \in \mathcal{T}_h} \mathbb{E} \int_K \kappa \nabla (u_h - \widetilde{u}_h) \cdot \nabla (u_h - \widetilde{u}_h) \, \mathrm{d}x \\ &= \sum_{K \in \mathcal{T}_h} \eta_K^2, \end{aligned} \tag{3.93}$$

where we consider η_K to be an indicator of the error at a local level. If we impose a tolerance level ϵ for the error, i.e.,

$$\|u - u_h\| \leq \epsilon \|u_h\|, \tag{3.94}$$

we obtain that a sufficient condition is given by

$$\eta_K \leq \frac{\epsilon \|u_h\|}{\sqrt{N}}, \tag{3.95}$$

where N is the number of elements in $\mathcal{T}_h$. Hence, we proceed iteratively by refining the mesh around elements which do not fulfil the error requirement until the required tolerance is attained. Coarsening of elements where the error indicator is small could be as well employed for saving computational power. The algorithm for mesh adaptation is given in **??** 1, where safety factors fac_1 and fac_2 are introduced.

3.5 Inverse problems

Probabilistic numerical methods are particularly helpful when inserted in the framework of Bayesian inverse problems (BIPs) involving differential equations, as studied in [4, 19] for ODEs, and in [17, 15] for PDEs. Furthermore, in [41] a theoretical basis is laid for ensuring the

well-posedness of probabilistic solutions to BIPs.

We consider the framework introduced in [61] and expanded in [23]. With the notation of (3.1), we consider the PDE

$$\begin{aligned} -\nabla\cdot(e^{\vartheta}\nabla u) &= f, \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{3.96}$$

where the conductivity field κ is transformed through an exponential function $\kappa = \exp(\vartheta)$ in order to ensure positivity and hence well-posedness of the solution. Moreover, we suppose that $u \in W^{2,\infty}(D)$ and we let $\mathscr{U}$ = addspace be the space of admissible log-conductivity fields ϑ. The BIP consists in retrieving the true value $\vartheta^\dagger$ of the field ϑ given prior information and corrupted observations $z \in \mathbb{R}^m$ given by

$$z = \mathscr{G}(\vartheta^\dagger) + \varepsilon, \tag{3.97}$$

where we assume that $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ is a Gaussian source of additive noise and $\mathscr{G}\colon \mathscr{U} \to \mathbb{R}^m$ is the forward operator. In particular, we can write $\mathscr{G} = \mathscr{O} \circ \mathscr{S}$, where $\mathscr{S}\colon \mathscr{U} \to W^{2,\infty}(D)$ is the solution operator, mapping any value of the field ϑ to the solution u of (3.96), and $\mathscr{O}\colon W^{2,\infty}(D) \to \mathbb{R}^m$ is the observation operator. In this work, we simply consider $\mathscr{O}$ to be defined by point-wise evaluations of the solution, i.e.,

$$\mathscr{O}\colon \vartheta \mapsto \begin{pmatrix} u(x_1) & u(x_2) & \dots & u(x_m) \end{pmatrix}^\top. \tag{3.98}$$

If the prior information is encoded by a prior measure μ_0 over the space $\mathscr{U}$, then the solution of the BIP is given by the posterior distribution μ such that its Radon–Nikodym derivative satisfies

$$\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\vartheta; z) = \frac{1}{Z}\exp(-\Phi(\vartheta; z)), \tag{3.99}$$

where $\Phi\colon (L^\infty)^d \times \mathbb{R}^m \to \mathbb{R}$ is referred to as the potential function and Z is a normalisation constant. Under the Gaussian assumption for the noise, we have

$$\Phi(\vartheta; z) = \frac{1}{2}\|z - \mathscr{G}(\vartheta)\|^2_{\Sigma_\varepsilon}, \tag{3.100}$$

where the norm $\|\cdot\|_{\Sigma_\varepsilon}$ is defined as

$$\|y\|_{\Sigma_\varepsilon} = \|\Sigma_\varepsilon^{-1/2} y\|_{\mathbb{R}^m}. \tag{3.101}$$

In the following, we will consider the observations z to be fixed and hence denote $\mu(\mathrm{d}\vartheta) = \mu(\mathrm{d}\vartheta; z)$ as well as $\Phi(\vartheta) = \Phi(\vartheta; z)$. Let us denote by $\mathscr{G}_h\colon \mathscr{U} \to \mathbb{R}^m$ the forward model obtained as $\mathscr{G}_h = \mathscr{O} \circ \mathscr{S}_h$, where $S_h\colon \mathscr{U} \to V_h$ is the solution operator given by the linear FEM and we still denote by $\mathscr{O}$ the restriction of $\mathscr{O}$ to V_h. Denoting by Φ_h the approximate potential, given by

$$\Phi_h(\vartheta) = \frac{1}{2}\|z - \mathscr{G}_h(\vartheta)\|^2_{\Sigma_\varepsilon}, \tag{3.102}$$

we obtain the approximate posterior measure μ_h as

$$\frac{\mathrm{d}\mu_h}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{Z_h}\exp(-\Phi_h(\vartheta)), \tag{3.103}$$

where Z_h is the normalisation constant. Stuart proved [61, Theorem 4.6] that under suitable assumptions $d_{\mathrm{Hell}}(\mu_h, \mu) \to 0$ for $h \to 0$, where $d_{\mathrm{Hell}}(\cdot,\cdot)$ is the Hellinger distance for probability measures. Hence, assuming an infinite computational budget is available it is possible to compute the posterior measure via approximate computations. This result has then been extended to more general priors than Gaussian [24, 62].

It has been shown empirically that under a fixed computational budget, employing a standard numerical method for the approximation of the solution operator $\mathscr{S}$ can lead to inaccurate results [4, 19, 17]. In particular, in case the variance Σ_ε of the observational noise is small with respect to the discretisation error, the posterior measure μ_h will be overconfident and peaked away from the true value of the unknown field. Probabilistic numerical methods can efficiently tackle this overconfidence issue thanks to the uncertainty quantification of numerical errors they naturally introduce. Given the probability space Ω on which the random variables defining the probabilistic scheme $\alpha_i \colon \Omega \to \mathbb{R}^d$ introduced in assumption 3.3 are defined, let us denote by $\widetilde{\mathscr{G}}_h \colon \Omega \times \mathscr{U} \to \mathbb{R}^m$ the random forward model obtained as $\widetilde{\mathscr{G}}_h = \mathscr{O} \circ \widetilde{\mathscr{S}}_h$, where $\widetilde{\mathscr{S}}_h \colon \Omega \times \mathscr{U} \to \widetilde{V}_h$ is the solution operator corresponding to the random FEM introduced in this work. Replacing $\mathscr{G}_h$ with $\widetilde{\mathscr{G}}_h$ in (3.102) we get a random potential $\widetilde{\Phi}_h$ and eventually a random posterior measure $\widetilde{\mu}_h$ defined by

$$\frac{\mathrm{d}\widetilde{\mu}_h}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{\widetilde{Z}_h}\exp(-\widetilde{\Phi}_h(\vartheta)), \tag{3.104}$$

where $\widetilde{Z}_h$ is the normalisation constant. In order to obtain an approximation of μ through $\widetilde{\mu}_h$, we need to take the expectation of the random probabilistic solution, which is viable in two different manners as explained in [41]. The first approach is to define the measure $\widetilde{\mu}_h^{\mathrm{fix}} = \mathbb{E}\,\widetilde{\mu}_h$. Otherwise, one could define a measure $\widetilde{\mu}_h^{\mathrm{var}}$ through

$$\frac{\mathrm{d}\widetilde{\mu}_h^{\mathrm{var}}}{\mathrm{d}\mu_0}(\vartheta) = \frac{1}{\mathbb{E}\,\widetilde{Z}_h}\mathbb{E}\exp(-\widetilde{\Phi}_h(\vartheta)), \tag{3.105}$$

which is already a deterministic measure. The choice of the names of these two approximation comes from their computation, which is in spirit slightly different. In the case of $\widetilde{\mu}_h^{\mathrm{fix}}$, for each event ω one evaluates the forward model and computes the value of the posterior. The expectation is then taken with the respect to the posterior itself, in practice via averaging techniques. Hence, for each ω we fix a perturbed mesh $\widetilde{\mathscr{T}}_h(\omega)$ and compute the posterior for several values of ϑ. Conversely, in the case of the measure $\widetilde{\mu}_h^{\mathrm{var}}$ the field ϑ is first fixed, and then the posterior is in practice obtained evaluating the forward model on several (variable) realisations of the random probabilistic solution.

We now need to prove the convergence of the posterior distributions $\widetilde{\mu}_h$ and $\widetilde{\mu}_h^{\mathrm{var}}$ towards the

true posterior μ with respect to the mesh size, which is granted by the following result under three regularity assumptions.

Theorem 3.23 (Theorem 3.9 of [41])**.** *With the notation above, if*

1. *there exists $q > 0$ such that $\exp(\Phi) \in L^q_{\mu_0}(\mathscr{U})$,*
2. *there exists a constant $C > 0$ such that*

$$\mathbb{E}_{\mu_0}[\widetilde{\Phi}_N] \leq C, \quad \text{almost surely in } \Omega, \tag{3.106}$$

3. *it holds*

$$\lim_{h\to 0} \| \left(\mathbb{E} \| \widetilde{\mathscr{G}}_h - \mathscr{G} \|^2 \right)^{1/2} \|_{L^s_{\mu_0}(\mathscr{U})} = 0, \tag{3.107}$$

where $s = 2q/(q-1)$ and q is given in 1,

then

$$\begin{aligned} \mathbb{E}\left[d_{\mathrm{Hell}}(\mu, \widetilde{\mu}_h)^2 \right]^{1/2} &\leq C \| \left(\mathbb{E} \| \widetilde{\mathscr{G}}_h - \mathscr{G} \|^4_{\mathbb{R}^m} \right)^{1/2} \|^{1/2}_{L^2_{\mu_0}(\mathscr{U})}, \\ d_{\mathrm{Hell}}(\mu, \widetilde{\mu}^{\mathrm{var}}_h) &\leq C \| \mathbb{E} \| \widetilde{\mathscr{G}}_h - \mathscr{G} \|^2_{\mathbb{R}^m} \|^{1/2}_{L^s_{\mu_0}(\mathscr{U})}. \end{aligned} \tag{3.108}$$

Let us remark that for a measure μ the spaces $L^q_\mu(\mathscr{U})$ are defined as

$$L^q_\mu(\mathscr{U}) = \left\{ f : \mathscr{U} \to \mathbb{R} : \int_{\mathscr{U}} f(\vartheta)^q \, \mu(\mathrm{d}\vartheta) < \infty \right\}, \tag{3.109}$$

with norm

$$\| f \|_{L^q_\mu(\mathscr{U})} = \left(\int_{\mathscr{U}} f(\vartheta)^q \, \mu(\mathrm{d}\vartheta) \right)^{1/q}. \tag{3.110}$$

theorem 3.23 gives in a general framework the convergence of posterior measures defined through approximate random forward models The following result now guarantees the convergence of the posterior distributions

3.6 Numerical experiments

3.6.1 Convergence

One-dimensional case

We consider (3.1) with $\kappa \equiv 1$ on $D = (0,1)$ and $f(x) = (x - 1/2)\chi_{(1/2,1)}(x)$, so that the solution u satisfies assumption 4.1. We verify the result of **??** by choosing $p \in \{1,2,3\}$ and by varying the mesh size h in the range $[9 \cdot 10^{-3}, 0.25]$. Moreover, we compute only one realisation of the random mesh for each couple $\{p,h\}$ as our bound holds almost surely. Results, shown in fig. 3.3, confirm the validity of the convergence estimates.

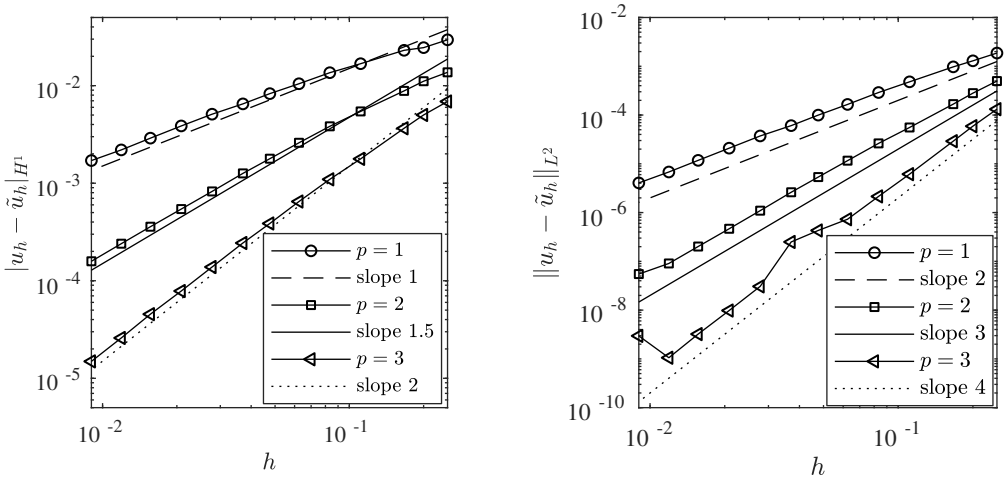

Figure 3.3 – Convergence rates in the H^1 semi-norm and the L^2 norm for the one-dimensional Poisson equation.

3.6.2 Error estimators

One-dimensional example

Consider

$$\begin{aligned} -u'' &= f, \quad \text{in } (0,1), \\ u(0) &= u(1) = 0, \end{aligned} \tag{3.111}$$

with f chosen such that $u(x) = -\sin(12\pi x)\exp(-100(x-1/2)^2)$ is the true solution. We consider the error estimations of presented in section 3.4, both in a local and global manner. Results, displayed in figs. 3.4 and 3.5 show that the estimates hold in practice for this case. In particular, in fig. 3.5 we can remark that the overall effectivity index $\eta_{\mathscr{X}}$, defined as

$$\eta_{\mathscr{X}} = \frac{\mathbb{E}\|u_h - \tilde{u}_h\|_{\mathscr{X}}}{\|u_h - u\|_{\mathscr{X}}}, \tag{3.112}$$

with $\mathscr{X} = H_0^1, L^2$, is in this case close to one for both norms. Errors are estimated employing $M = 10$ realisations of the probabilistic solution and with a Monte Carlo simulation.

Two-dimensional case

TO DO

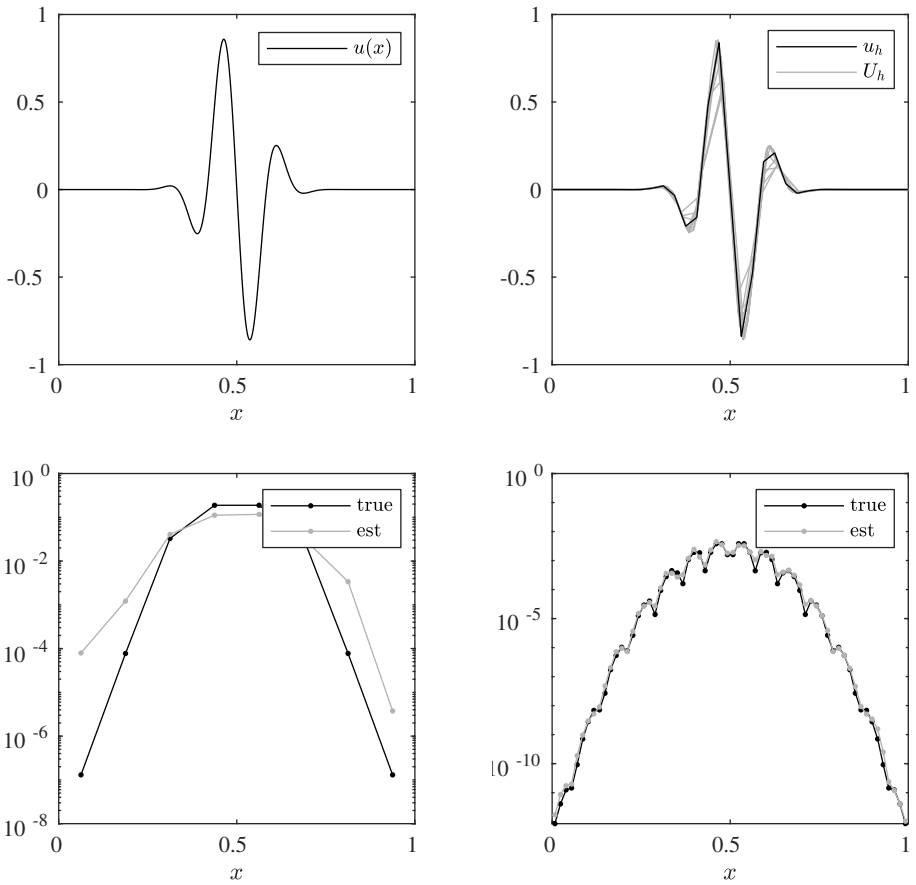

Figure 3.4 – Error estimation for the 1D problem with two different values of h – error in each element.

3.6.3 Mesh adaptivity

Two-dimensional case

See results fig. 3.6.

3.6.4 Bayesian inverse problems

Let us consider the following one-dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\kappa}\frac{\mathrm{d}u}{\mathrm{d}x}\right) &= f, \quad \text{in } (0,1), \\ u &= 0, \quad \text{on } \{0,1\}, \end{aligned} \tag{3.113}$$

and the inverse problem of retrieving the field $\kappa \in L^2(0,1)$ given synthetic noisy observations of the solution u corresponding to a true field κ^*. First, we consider a case where information

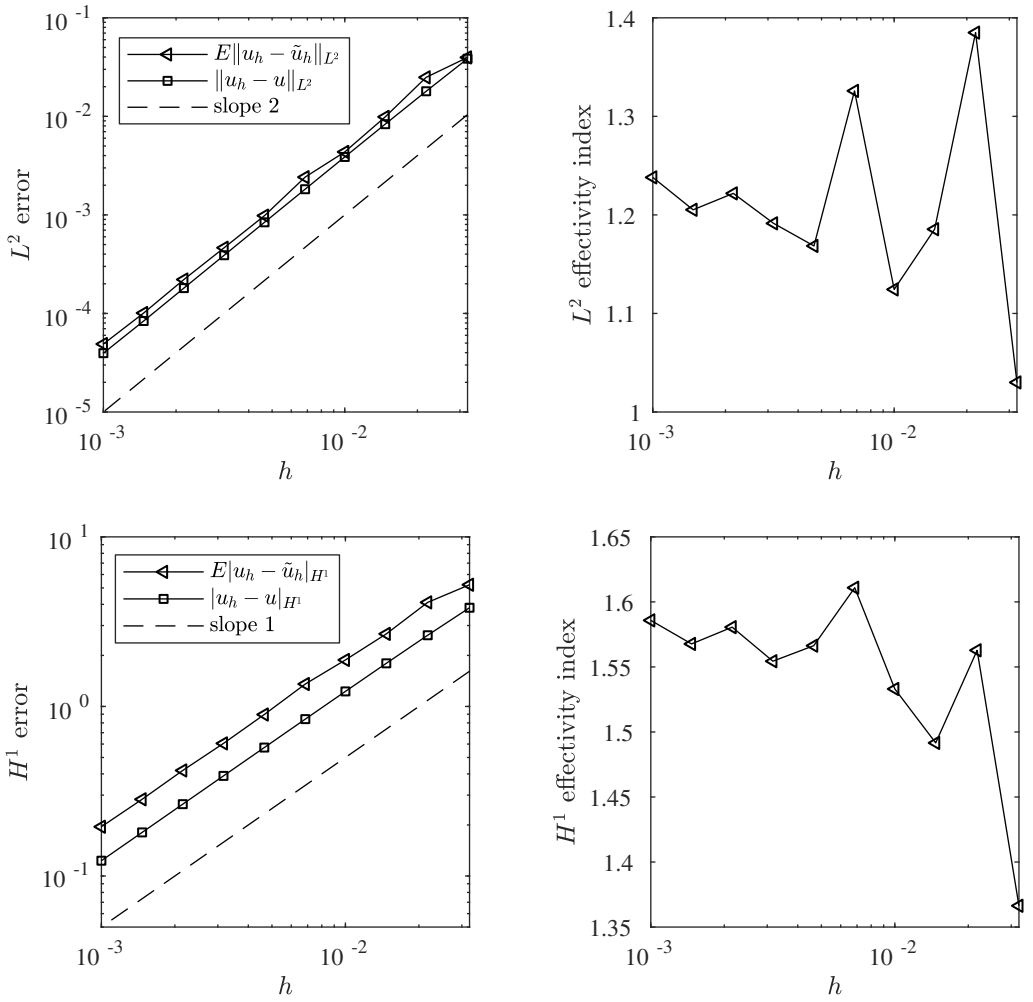

Figure 3.5 – Error estimation for the 1D problem with two different values of h – convergence of error estimators and effectivity indices

on κ is available beforehand. In particular, we assume that κ has the form

$$\kappa(x) = \begin{cases} \log(1+\kappa_1), & \text{if } x \in I_1, \\ \log(1+\kappa_2), & \text{if } x \in I_2, \\ 0 & \text{otherwise,} \end{cases} \tag{3.114}$$

where κ_1, κ_2 are real scalars and I_1, I_2 are the intervals $(0.2, 0.4)$ and $(0.6, 0.8)$ respectively. Fixing a standard Gaussian prior on both parameters κ_1 and κ_2 we are able to compute the posterior distribution corresponding to both the deterministic and probabilistic forward models. In particular, we vary the number of elements N in the set $\{20, 40, 80, 160\}$, thus studying the effects of numerical errors on the numerical posterior distribution. Observations are obtained from a reference solution evaluated at four equispaced points in the interior of

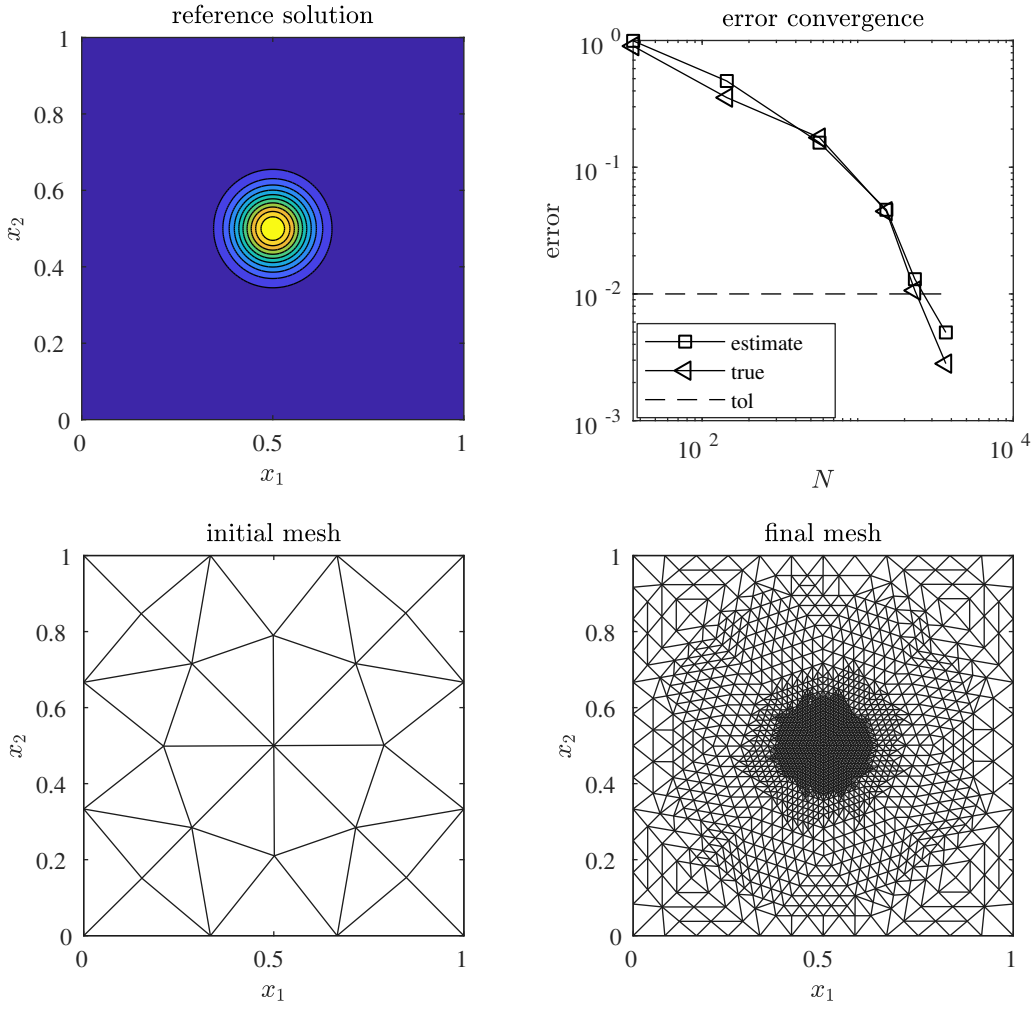

Figure 3.6 – Mesh adaptivity – two-dimensional case

$(0,1)$ each corrupted by an additive source of noise $\varepsilon \sim \mathcal{N}(0, 10^{-4})$. The posterior distributions are obtained with Metropolis–Hastings initialised near the true value of (κ_1, κ_2) and ran as explained in section 3.5, with 240 parallel chains employed for the probabilistic forward model. Results are shown in fig. 3.8, where TO DO

In a second experiment, we consider the same exact field κ^* and observation model, but without the additional information encoded in (3.114).

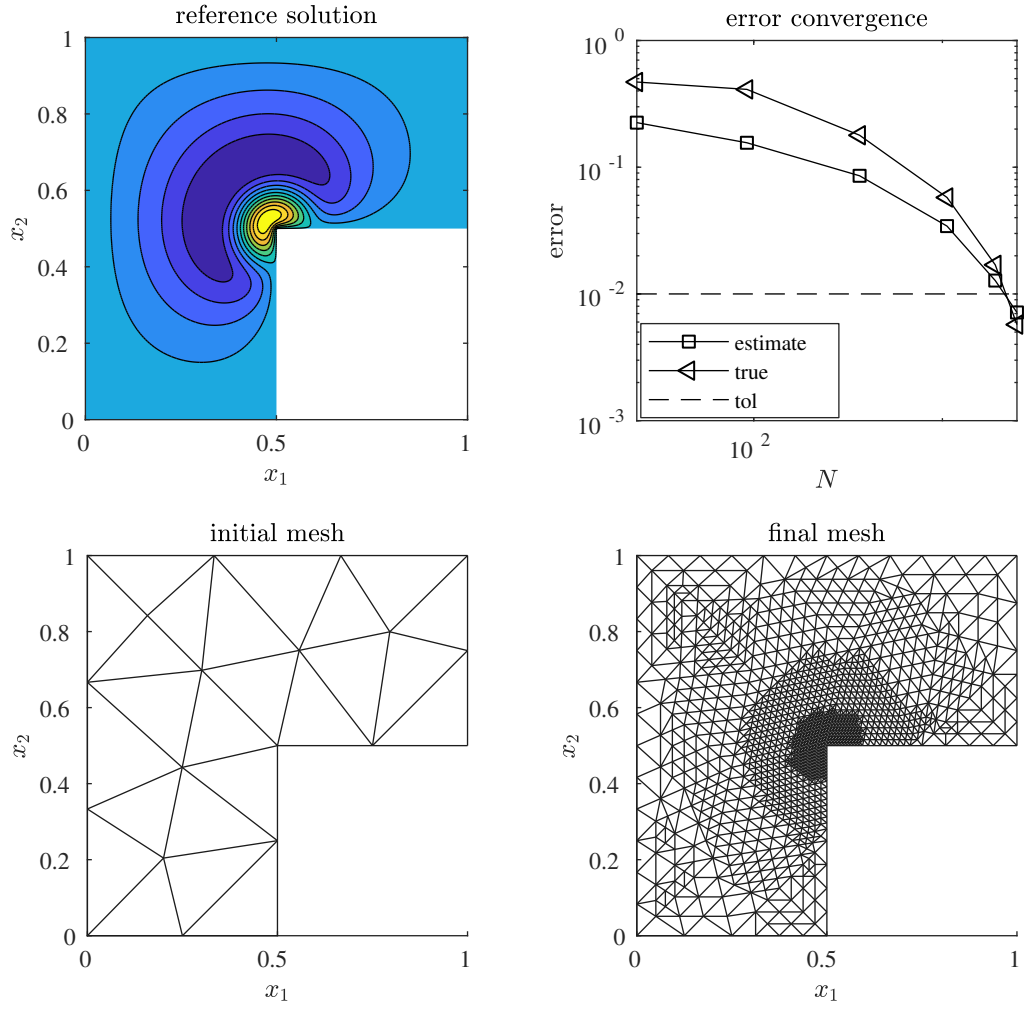

Figure 3.7 – Mesh adaptivity – two-dimensional case

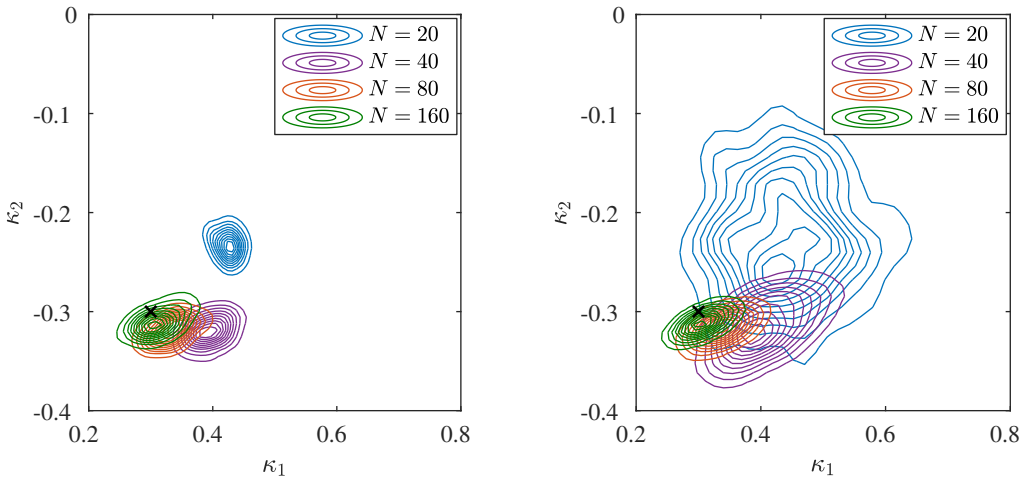

Figure 3.8 – Bayesian inverse problem – finite dimensional case.

4 Parameter inference of Multiscale Diffusions

4.1 Introduction

Data-driven approaches to the derivation of effective stochastic coarse-grained dynamics have become a standard tool in different areas such as molecular dynamics or atmosphere and ocean science [35, 20, 66]. When fitting data to a low dimensional coarse-grained model, one is faced with the problem of model misspecification; indeed, the data, coming from the full dynamics, are compatible with the coarse-grained model only at the time scales at which the effective dynamics is valid. This is related to the issue of high-frequency data that arises in econometrics [7, 8, 67, 48].

The issue of model misspecification in inverse problems with a multiscale structure has been treated in the context of partial differential equations (PDE). 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 asymptotically the correct result [46]. A series of papers [2, 3, 5] focuses on retrieving the full model when the multiscale coefficient is endowed with a specific parametrized structure. Being these problems ill-posed, the latter is achieved via Tikhonov regularization [3, 46], adopting a Bayesian approach [2, 46] or exploiting techniques of Kalman filtering [5]. In [2, 5], the authors highlight the need of accounting explicitly for the modelling error due to homogenization and apply statistical techniques taken from [13, 14].

For simple models in molecular dynamics, the effect of model misspecification was studied in a series of papers [53, 50, 52, 27, 28] under the assumption of scale separation. 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 [53, Theorem 3.4]. To be more precise, in the large sample size limit the MLE converges to the coefficients of the unhomogenized equation, rather than to those of the homogenized one. 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 [53, Theorems 3.5 and 3.6]. This is similar to the estimation of the integrated stochastic volatility in the presence of market micro-structure noise, where the data have to be subsampled at an

appropriate rate [8, 67]. The correct subsampling rate can be in some instances rather extreme with respect to the frequency of the data and lead to get rid of more than 99% of the data. As the intuition suggests, this increases significantly the bias of the estimator, which is usually taken care of with additional bias corrections and variance reduction procedures.

The necessity to subsample the data can be alleviated by using appropriate martingale estimators, as was done in [38, 35]. 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 [39].

In this paper, we bypass the issue of subsampling by implementing an appropriate filtering methodology. In particular, we show that smoothing the data coming from the multiscale model with an appropriate linear time-invariant filter of the exponential family allows to retrieve the drift coefficient of the homogenized equation. The methodology we present is not involved computationally, easy to implement in practice and presents two main advantages:

1. the filter's kernel depends on two parameters which can be tuned to obtain more robust results and which can be interpreted as analogous to the subsampling rate. Nevertheless, while the MLE is extremely sensitive with respect to the latter, changing the filter parameters does not have a strong influence on the result, therefore allowing our technique to be applied as a black-box tool for parameter estimation,
2. the entire stream of data is employed, which, in practice, enhances the quality of the filter-based MLE in terms of bias. Moreover, avoiding subsampling and thus discretising the data allows us to employ continuous-time theoretical tools.

It is natural to reinterpret the problem of inferring the drift parameter of multiscale diffusion processes under a Bayesian perspective [56, 55]. In particular, the form of the likelihood function guarantees, under a Gaussian prior hypothesis, that the posterior distribution is itself a Gaussian. Therefore, one can simply compute the mean and covariance of the posterior distribution to obtain a full uncertainty quantification of the inference process. The Bayesian approach has been adopted in several works concerning single and multiscale PDEs, too, and has proved itself to be versatile and efficient in this framework (see, e.g., [2, 5]). Moreover, infinite-dimensional inverse problems are intrinsically ill-posed, and interpreting them in a Bayesian manner is an efficient tool for regularizing them and guaranteeing their well-posedness [61, 24]. The regularization property of a Bayesian approach has been recently demonstrated for inverse problems involving multiscale PDEs [2]. In this work, we investigate the effects of inserting our filtering methodology in a Bayesian framework. In particular, we analyse the effects of modifying the likelihood function inserting the filtered trajectory, showing how this corrects the faulty asymptotic behaviour which would be present in case no pre-processing would be applied.

The rest of the paper is organised as follows. In Section 4.2 we introduce the problem and

lay the basis of our analysis setting the main assumptions and notation. In Section 4.3 we present our filtering methodology, with a particular focus on ergodic properties, on multiscale convergence and, naturally, on the properties of our estimators. In Section 4.4 we introduce the Bayesian framework and show how it can be enhanced with filtering techniques. Finally, in Section 4.5 we demonstrate the effectiveness of our methodology via a series of numerical experiments.

4.2 Problem setting

In this section, we introduce the class of diffusion processes which we treat in this paper and the classical methodology employed for the estimation of the drift. Let $\varepsilon > 0$ and let us consider the one-dimensional multiscale stochastic differential equation (SDE)

$$\mathrm{d}X_t^\varepsilon = -\alpha \cdot V'(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right)\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \tag{4.1}$$

where, given a positive integer N, we have that $\alpha \in \mathbb{R}^N$ and $\sigma > 0$ are the drift and diffusion coefficients respectively and W_t is a standard one-dimensional Brownian motion. The functions $V\colon \mathbb{R} \to \mathbb{R}^N$ and $p\colon \mathbb{R} \to \mathbb{R}$ correspond to the slow-scale and the fast-scale confining potentials. In particular, we assume

$$V(x) = \begin{pmatrix} V_1(x) & V_2(x) & \cdots & V_N(x) \end{pmatrix}^\top, \tag{4.2}$$

for smooth functions $V_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, \dots, N$. Moreover, we assume p to be smooth and periodic of period L. Theory of homogenization [12, Chapter 3] guarantees the existence of an SDE of the form

$$\mathrm{d}X_t = -A \cdot V'(X_t)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t, \tag{4.3}$$

where W_t is the same Brownian motion as in (4.1), such that $X_t^\varepsilon \to X_t$ for $\varepsilon \to 0$ in law as random variables in $\mathscr{C}^0([0,T];\mathbb{R})$. In particular, we have $A = K\alpha$ and $\Sigma = K\sigma$, where the coefficient K is given by the formula

$$K = \int_0^L (1 + \Phi'(y))^2\,\mu(\mathrm{d}y), \tag{4.4}$$

with

$$\mu(\mathrm{d}y) = \frac{1}{Z}e^{-p(y)/\sigma}\,\mathrm{d}y, \quad \text{where} \quad Z = \int_0^L e^{-p(y)/\sigma}\,\mathrm{d}y, \tag{4.5}$$

and where the function Φ is the unique solution with zero-mean with respect to the measure μ of the elliptic partial differential equation

$$-p'(y)\Phi'(y) + \sigma\Phi''(y) = p'(y), \quad 0 \le y \le L, \tag{4.6}$$

endowed with periodic boundary conditions.

We now briefly present the classical methodology for estimating the drift coefficient. Let $T > 0$ and let $X^\varepsilon := (X_t^\varepsilon, 0 \le t \le T)$ be a realization of the solution of (4.1) up to final time. Girsanov's change of measure formula applied to (4.3) allows to write the likelihood of X^ε given a drift coefficient A as

$$p(X^\varepsilon \mid A) = \exp\left(-\frac{I(X^\varepsilon \mid A)}{2\Sigma}\right), \tag{4.7}$$

where

$$I(X^\varepsilon \mid A) = \int_0^T A \cdot V'(X_t^\varepsilon)\,\mathrm{d}X_t^\varepsilon + \frac{1}{2}\int_0^T \left(A \cdot V'(X_t^\varepsilon)\right)^2 \mathrm{d}t. \tag{4.8}$$

Minimizing the functional $I(X^\varepsilon \mid A)$ with respect to A therefore gives the maximum likelihood estimator (MLE) of A, which can be formally computed in closed form as

$$\widehat{A}^\varepsilon(T) := \arg\min_{A \in \mathbb{R}^N} I(X^\varepsilon \mid A) = M^{-1} h, \tag{4.9}$$

where $M \in \mathbb{R}^{N\times N}$ and $h \in \mathbb{R}^N$ are defined as

$$M = \frac{1}{T}\int_0^T V'(X_t^\varepsilon) \otimes V'(X_t^\varepsilon)\,\mathrm{d}t, \quad h = \frac{1}{T}\int_0^T V'(X_t^\varepsilon)\,\mathrm{d}X_t^\varepsilon, \tag{4.10}$$

and where $\otimes$ denotes the outer product in $\mathbb{R}^N$. Let us now state the assumptions which will be employed throughout the rest of our work. In particular, we consider the same dissipative setting as [53, Assumption 3.1].

Assumption 4.1. The potentials p and V satisfy

1. $p \in \mathscr{C}^\infty(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and is L-periodic for some $L > 0$,
2. $V_i \in \mathscr{C}^\infty(\mathbb{R})$ for all $i = 1, \dots, N$ is polynomially bounded from above and bounded from below, and there exist $a, b > 0$ such that
$$-\sum_{i=1}^N V_i'(x)x \le a - bx^2. \tag{4.11}$$
3. V' is Lipschitz continuous, i.e. there exists a constant $C > 0$ such that
$$\|V'(x) - V'(y)\|_2 \le C|x - y|, \tag{4.12}$$
4. for all $T > 0$, the symmetric matrix M is positive definite and there exists $\bar{\lambda} > 0$ such that $\lambda_{\min}(M) \ge \bar{\lambda}$.

Under these assumptions, the MLE given in (4.9) is indeed the unique minimizer of the likelihood function, as shown in [56, Theorem 2.4].

Given the convergence of $X_t^\varepsilon \to X_t$ in the space of continuous stochastic processes, one would expect that the MLE (4.9) would be asymptotically unbiased for the drift coefficient A of the

homogenized equation (4.3). Instead, it is possible to prove that in the asymptotic limit for $T \to \infty$ and $\varepsilon \to 0$, the MLE tends to the drift coefficient α of the unhomogenized equation (4.1). We report here this result, whose proof can be found for the case $N = 1$ in [53, Theorem 3.4]. Let us remark that the proof for $N > 1$ follows directly from the one-dimensional case.

Theorem 4.2. *Let Assumption 4.1 hold and let X_0^ε be distributed according to the invariant measure of the process X^ε solution of* (4.1)*. Then*

$$\lim_{\varepsilon \to 0} \lim_{T \to \infty} \widehat{A}^\varepsilon(T) = \alpha, \quad a.s., \tag{4.13}$$

where α is the drift coefficient of equation (4.1)*.*

As anticipated in the introduction, the main tool for obtaining unbiased estimators in the literature is subsampling the data. In particular, let the dimension of the parameter $N = 1$, let $\delta > 0$ and let $T = n\delta$ with n a positive integer. Then, a subsampled estimator for A is given by

$$\widehat{A}_\delta^\varepsilon(T) = -\frac{\sum_{j=0}^{n-1} V'(X_{j\delta}^\varepsilon)\left(X_{(j+1)\delta}^\varepsilon - X_{j\delta}^\varepsilon\right)}{\delta \sum_{j=0}^{n-1} V'(X_{j\delta}^\varepsilon)^2}, \tag{4.14}$$

which is a discretized version of $\widehat{A}^\varepsilon(T)$. It is possible to show [53, Theorem 3.5] that choosing $\delta = \varepsilon^\zeta$ with $\zeta \in (0,1)$ and if T is sufficiently big, then $\widehat{A}_\delta^\varepsilon(T)$ is an asymptotically unbiased estimator of A with respect to $T \to \infty$ and $\varepsilon \to 0$, in probability. Despite being widely employed in practice, estimators based on subsampling present some drawbacks, as discussed in the introduction. In the following, we will introduce and analyse a novel approach for the drift estimation.

Remark 4.3. Let us remark that for enhancing the clarity of the exposition, in this article we chose to focus on the case of a multi-dimensional parameter in the setting of one-dimensional diffusion processes. In fact, all the theory we present in the following could be generalized to the case of d-dimensional SDEs equivalent to (4.1), which can be written as

$$\mathrm{d}X_t^\varepsilon = -\sum_{i=1}^{N} \alpha_i \nabla V_i(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon}\nabla p\left(\frac{X_t^\varepsilon}{\varepsilon}\right)\mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W_t, \tag{4.15}$$

where W_t is a standard d-dimensional Brownian motion. The proof of all results below should be slightly modified, but we verified that all arguments still hold true in the d-dimensional case.

4.3 The filtering approach

In this section, we introduce and analyse a novel filtering approach to solve the biasedness issue highlighted by Theorem 4.2. Let $\beta, \delta > 0$ and let us consider a family of exponential

kernel functions $k\colon \mathbb{R}^+ \to \mathbb{R}$ defined as

$$k(r) = C_\beta \delta^{-1/\beta} e^{-r^\beta/\delta}, \tag{4.16}$$

where C_β is a normalizing constant given by

$$C_\beta = \beta \Gamma(1/\beta)^{-1}, \tag{4.17}$$

and where $\Gamma(\cdot)$ is the gamma function. We consider the process $Z^\varepsilon := (Z_t^\varepsilon, 0 \le t \le T)$ defined by the weighted average

$$Z_t^\varepsilon := \int_0^t k(t-s) X_s^\varepsilon \, \mathrm{d}s. \tag{4.18}$$

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 (4.16) 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, only in case $\beta = 1$ a rigorous analysis is carried on. 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 4.4. 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".

In the rest of this section we will focus on the properties of the process Z^ε when it is considered together with the original process X^ε. In particular, we focus on ergodic properties and multiscale convergence. Finally, we present and analyse unbiased estimators obtained with the filtering process.

4.3.1 Ergodic properties of the filter

Let us consider the filtering kernel (4.16) with $\beta = 1$, i.e.,

$$k(r) = \frac{1}{\delta} e^{-r/\delta}. \tag{4.19}$$

In this case, Leibniz integral rule yields the equality

$$\mathrm{d}Z_t^\varepsilon = k(0) X_t^\varepsilon \, \mathrm{d}t + \int_0^t k'(t-s) X_s^\varepsilon \, \mathrm{d}s \, \mathrm{d}t = \frac{1}{\delta}\left(X_t^\varepsilon - Z_t^\varepsilon\right) \mathrm{d}t, \tag{4.20}$$

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

$$\begin{aligned} \mathrm{d}X_t^\varepsilon &= -\alpha \cdot V'(X_t^\varepsilon)\,\mathrm{d}t - \frac{1}{\varepsilon} p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right) + \sqrt{2\sigma}\,\mathrm{d}W_t, \\ \mathrm{d}Z_t^\varepsilon &= \frac{1}{\delta}\left(X_t^\varepsilon - Z_t^\varepsilon\right)\mathrm{d}t. \end{aligned} \tag{4.21}$$

The first ingredient for verifying 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.

Lemma 4.5. *Let $(X^\varepsilon, Z^\varepsilon)^\top$ be the solution of* (4.21) *and let μ_t^ε be the measure induced by the couple at time t. Then, the measure μ_t^ε admits a smooth density ρ_t^ε with respect to the Lebesgue measure.*

Proof. We have to show that the joint process solution to (4.21) is hypo-elliptic. Denoting as $f\colon \mathbb{R} \to \mathbb{R}$ the function

$$f(x) = -\alpha \cdot V'(x) - \frac{1}{\varepsilon} p'\left(\frac{x}{\varepsilon}\right), \tag{4.22}$$

the generator of the process $(X^\varepsilon, Z^\varepsilon)^\top$ is given by

$$\mathscr{L} = f\partial_x + \sigma\partial_{xx}^2 + \frac{1}{\delta}(x-z)\partial_z =: \mathscr{X}_0 + \sigma\mathscr{X}_1^2, \tag{4.23}$$

where

$$\mathscr{X}_0 = f\partial_x + \frac{1}{\delta}(x-z)\partial_z, \quad \mathscr{X}_1 = \partial_x. \tag{4.24}$$

The commutator $[\mathscr{X}_0, \mathscr{X}_1]$ applied to a test function v then gives

$$\begin{aligned} [\mathscr{X}_0, \mathscr{X}_1]v &= f\partial_x^2 v + \frac{1}{\delta}(x-z)\partial_x\partial_z v - \partial_x\left(f\partial_x v + \frac{1}{\delta}(x-z)\partial_z v\right) \\ &= -\partial_x f\partial_x v - \frac{1}{\delta}\partial_z v. \end{aligned} \tag{4.25}$$

Consequently,

$$\mathrm{Lie}\,(\mathscr{X}_1, [\mathscr{X}_0, \mathscr{X}_1]) = \mathrm{Lie}\left(\partial_x, -\partial_x f\partial_x - \frac{1}{\delta}\partial_z\right), \tag{4.26}$$

which spans $T_x\mathbb{R}^2$. The desired result then follows from Hörmander's theorem (see e.g. [51, Chapter 6]). □

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, we are interested in the limiting properties of this law. In particular, we know that the process X^ε alone is geometrically ergodic [43, Theorem 4.4], and we wish the couple $(X^\varepsilon, Z^\varepsilon)^\top$ to be endowed with the same property. The following Lemma guarantees that the couple is indeed geometrically ergodic exploiting an argument of dissipativity.

Lemma 4.6. *Let Assumption 4.1 hold and let $b > 0$ be given in Assumption 4.12. Then, if $\delta > 1/(4b)$, the process $(X^\varepsilon, Z^\varepsilon)^\top$ solution of (4.21) is geometrically ergodic, i.e., there exists $C, \lambda > 0$ such that for all measurable $f : \mathbb{R}^2 \to \mathbb{R}$ such that for some integer $q > 0$*

$$f(x,z) \le 1 + \| \begin{pmatrix} x & z \end{pmatrix}^\top \|_2^q, \tag{4.27}$$

it holds

$$|\mathbb{E} f(X_t^\varepsilon, Z_t^\varepsilon) - \int_{\mathbb{R}} \int_{\mathbb{R}} f(x,z) \rho^\varepsilon(x,z) \, \mathrm{d}x \, \mathrm{d}z| \le C \left(1 + \| \begin{pmatrix} X_0^\varepsilon & Z_0^\varepsilon \end{pmatrix}^\top \|_2^q \right) e^{-\lambda t}, \tag{4.28}$$

where $\mathbb{E}$ denotes expectation with respect to the Wiener measure and ρ^ε is the solution to the stationary Fokker–Planck equation

$$\sigma \partial_{xx}^2 \rho^\varepsilon(x,z) + \partial_x \left(\left(\alpha \cdot V'(x) + \frac{1}{\varepsilon} p'\left(\frac{x}{\varepsilon}\right) \right) \rho^\varepsilon(x,z) \right) + \frac{1}{\delta} \partial_z \left((z-x) \rho^\varepsilon(x,z) \right) = 0. \tag{4.29}$$

Proof. Lemma 4.5 guarantees that the Fokker–Planck equation can be written directly from the system (4.21). For geometric ergodicity, let

$$\mathscr{S}(x,z) := \begin{pmatrix} -\sum_{i=1}^N V_i'(x) \\ \frac{1}{\delta}(x-z) \end{pmatrix} \cdot \begin{pmatrix} x \\ z \end{pmatrix} = -\sum_{i=1}^N V_i'(x) x + \frac{1}{\delta}(xz - z^2). \tag{4.30}$$

Due to Assumption 4.12 and Young's inequality, we then have for all $\gamma > 0$

$$\mathscr{S}(x,z) \le a + \left(\frac{1}{2\gamma\delta} - b \right) x^2 + \frac{1}{\delta} \left(\frac{\gamma}{2} - 1 \right) z^2. \tag{4.31}$$

We choose $\gamma = \gamma^* := 1 - b\delta + \sqrt{1 + (1-b\delta)^2} > 0$ so that

$$C(\gamma^*) := -\frac{1}{2\gamma^*\delta} + b = -\frac{1}{\delta} \left(\frac{\gamma^*}{2} - 1 \right), \tag{4.32}$$

and we notice that $C(\gamma^*) > 0$ if $\delta > 1/(4b)$. In this case, we have

$$\mathscr{S}(x,z) \le a - C(\gamma^*) \| \begin{pmatrix} x & z \end{pmatrix}^\top \|^2, \tag{4.33}$$

and problem (4.21) is dissipative. The result then follows from [43, Theorem 4.4]. □

Remark 4.7. Let us remark that for the multiscale equation (4.1) one should check the condition above for the potential $V(x) + p(x/\varepsilon)$. Nevertheless, for functions p satisfying Assumption 4.1, the result holds trivially.

Remark 4.8. The condition $\delta > 1/(4b)$ is not very restrictive. Let the parameter dimension $N = 1$ and let $V(x) \propto x^{2p}$ for an integer $p > 1$. Then, Assumption 4.12 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 $N > 1$.

Example 4.9. A closed form solution of (4.29) can be obtained in a simple homogenized case with the dimension of the parameter $N = 1$. Let $p(y) = 0$ and the parameters α, σ be replaced

respectively by A and Σ. Then, if $V(x) = x^2/2$, equation (4.29) has the analytical solution

$$\rho^0(x,z) = \frac{1}{C_{\rho^0}} \exp\left(-\frac{A}{\Sigma}\frac{x^2}{2} - \frac{1}{\delta\Sigma}\frac{(x-(1+A\delta)z)^2}{2}\right), \tag{4.34}$$

where

$$C_{\rho^0} = \int_{\mathbb{R}}\int_{\mathbb{R}} \exp\left(-\frac{A}{\Sigma}\frac{x^2}{2} - \frac{1}{\delta\Sigma}\frac{(x-(1+A\delta)z)^2}{2}\right) \mathrm{d}x\,\mathrm{d}z. \tag{4.35}$$

This is the density of a multivariate normal distribution $\mathcal{N}(0,\Gamma)$, where the covariance matrix is given by

$$\Gamma = \frac{\Sigma}{A(1+A\delta)}\begin{pmatrix} 1+A\delta & 1 \\ 1 & 1 \end{pmatrix}. \tag{4.36}$$

Let us remark that this distribution can be obtained from direct computations involving Gaussian processes. In particular, it is known that $X \sim \mathcal{GP}(\mu_t, \mathscr{C}(t,s))$, where at stationarity $\mu_t = 0$ and

$$\mathscr{C}(t,s) = \frac{\Sigma}{A} e^{-A|t-s|}. \tag{4.37}$$

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.

In a general case, it is not possible to find an explicit solution to (4.29). Nevertheless, it is possible to show some relevant properties of the solution itself, which are summarized in the following Lemma.

Lemma 4.10. *Under the assumptions of Lemma 4.6, let ρ^ε be the solution of* (4.29) *and let us write*

$$\rho^\varepsilon(x,z) = \varphi^\varepsilon(x)\psi^\varepsilon(z)R^\varepsilon(x,z), \tag{4.38}$$

where φ^ε and ψ^ε are the marginal densities of X^ε and Z^ε respectively, i.e.,

$$\varphi^\varepsilon(x) = \int_{\mathbb{R}} \rho^\varepsilon(x,z)\,\mathrm{d}z, \quad \psi^\varepsilon(z) = \int_{\mathbb{R}} \rho^\varepsilon(x,z)\,\mathrm{d}x. \tag{4.39}$$

Then, it holds

$$\varphi^\varepsilon(x) = \frac{1}{C_{\varphi^\varepsilon}} \exp\left(-\frac{1}{\sigma}\alpha\cdot V(x) - \frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)\right), \tag{4.40}$$

where

$$C_{\varphi^\varepsilon} = \int_{\mathbb{R}} \exp\left(-\frac{1}{\sigma}\alpha\cdot V(x) - \frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)\right) \mathrm{d}x. \tag{4.41}$$

Moreover, it holds

$$\sigma\delta\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z = \mathbb{E}^{\rho^\varepsilon}[((X^\varepsilon)^2 - (Z^\varepsilon)^2)V''(Z^\varepsilon)]. \tag{4.42}$$

Proof. Integrating equation (4.29) with respect to z we obtain the stationary Fokker–Planck

equation for the process X^ε, i.e.

$$\sigma(\varphi^\varepsilon)''(x) + \frac{\mathrm{d}}{\mathrm{d}x}\left(\left(\alpha\cdot V'(x) + \frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)\right)\varphi^\varepsilon(x)\right) = 0, \tag{4.43}$$

whose solution is given by

$$\varphi^\varepsilon(x) = \frac{1}{C_{\varphi^\varepsilon}}\exp\left(-\frac{1}{\sigma}\alpha\cdot V(x) - \frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)\right), \tag{4.44}$$

and which proves (4.40). By integrating equation (4.29) with respect to x and integrating by parts we obtain

$$\psi^\varepsilon(z) + z(\psi^\varepsilon)'(z) = \frac{\mathrm{d}}{\mathrm{d}z}\int_{\mathbb{R}} x\rho^\varepsilon(x,z)\,\mathrm{d}x, \tag{4.45}$$

which can be written as

$$\frac{\mathrm{d}}{\mathrm{d}z}(z\psi^\varepsilon(z)) = \frac{\mathrm{d}}{\mathrm{d}z}\int_{\mathbb{R}} x\rho^\varepsilon(x,z)\,\mathrm{d}x. \tag{4.46}$$

Now, since ψ^ε is the density of a probability distribution, this implies that

$$z\psi^\varepsilon(z) = \int_{\mathbb{R}} x\rho^\varepsilon(x,z)\,\mathrm{d}x, \tag{4.47}$$

which, replacing the decomposition (4.38), yields

$$z = \int_{\mathbb{R}} x\varphi^\varepsilon(x)R^\varepsilon(x,z)\,\mathrm{d}x. \tag{4.48}$$

In view of (4.38) and (4.43), equation (4.29) can be rewritten as

$$\sigma\varphi^\varepsilon\psi^\varepsilon\partial^2_{xx}R^\varepsilon + \frac{1}{\delta}\varphi^\varepsilon\psi^\varepsilon R^\varepsilon + \sigma(\varphi^\varepsilon)'\psi^\varepsilon\partial_x R^\varepsilon + \frac{1}{\delta}(z-x)\varphi^\varepsilon((\psi^\varepsilon)'R + \psi^\varepsilon\partial_z R^\varepsilon) = 0. \tag{4.49}$$

We now multiply the equation above by x and integrate with respect to x. Let us consider some simplifications explicitly. First, an integration by parts yields

$$\sigma\psi^\varepsilon\int_{\mathbb{R}} x\varphi^\varepsilon\partial^2_{xx}R^\varepsilon\,\mathrm{d}x + \sigma\psi^\varepsilon\int_{\mathbb{R}} x(\varphi^\varepsilon)'\partial_x R^\varepsilon\,\mathrm{d}x = -\sigma\psi^\varepsilon\int_{\mathbb{R}}\varphi^\varepsilon\partial_x R^\varepsilon\,\mathrm{d}x. \tag{4.50}$$

Then, applying (4.48), we have

$$\frac{1}{\delta}\psi^\varepsilon\int_{\mathbb{R}} x\varphi^\varepsilon R^\varepsilon\,\mathrm{d}x = \frac{1}{\delta}z\psi^\varepsilon. \tag{4.51}$$

Moreover, again applying (4.48), we can compute

$$\frac{1}{\delta}(\psi^\varepsilon)'\int_{\mathbb{R}}(z-x)x\varphi^\varepsilon R\,\mathrm{d}x = \frac{1}{\delta}(\psi^\varepsilon)'\left(z^2 - \int_{\mathbb{R}} x^2\varphi^\varepsilon R^\varepsilon\,\mathrm{d}x\right). \tag{4.52}$$

Finally, we compute the last term always applying (4.48) obtaining

$$\frac{1}{\delta}\psi^\varepsilon \int_{\mathbb{R}} (z-x)x\varphi^\varepsilon \partial_z R \,\mathrm{d}x = \frac{1}{\delta}\psi^\varepsilon \left(z - \int_{\mathbb{R}} x^2 \varphi^\varepsilon \partial_z R^\varepsilon \,\mathrm{d}x \right). \tag{4.53}$$

Replacing the equalities (4.50), (4.51), (4.52) and (4.53) into (4.49) we obtain

$$(\psi^\varepsilon)' \left(z^2 - \int_{\mathbb{R}} x^2 \varphi^\varepsilon R^\varepsilon \,\mathrm{d}x \right) + \psi^\varepsilon \left(2z - \int_{\mathbb{R}} x^2 \varphi^\varepsilon \partial_z R^\varepsilon \,\mathrm{d}x - \delta\sigma \int_{\mathbb{R}} \varphi^\varepsilon \partial_x R^\varepsilon \,\mathrm{d}x \right) = 0. \tag{4.54}$$

We rewrite the equality above as

$$\begin{aligned} \delta\sigma\psi^\varepsilon(z) \int_{\mathbb{R}} \varphi^\varepsilon(x)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x = (\psi^\varepsilon)'(z)z^2 - (\psi^\varepsilon)'(z) \int_{\mathbb{R}} x^2 \varphi^\varepsilon(x) R^\varepsilon(x,z)\,\mathrm{d}x \\ + 2\psi^\varepsilon(x)z - \psi^\varepsilon \int_{\mathbb{R}} x^2 \varphi^\varepsilon(x)\partial_z R^\varepsilon(x,z)\,\mathrm{d}x. \end{aligned} \tag{4.55}$$

Multiplying by $V'(z)$, integrating with respect to z and integrating by parts, we obtain the following identity in $\mathbb{R}^N$

$$\begin{aligned} \sigma\delta \int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z &= \int_{\mathbb{R}}\int_{\mathbb{R}} (x^2 - z^2)V''(z)\rho_\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \mathbb{E}^{\rho^\varepsilon}[((X^\varepsilon)^2 - (Z^\varepsilon)^2)V''(Z^\varepsilon)], \end{aligned} \tag{4.56}$$

which is the desired result. □

4.3.2 Multiscale convergence

We now investigate the convergence of the couple $(X^\varepsilon, Z^\varepsilon)^\top$ with respect to the multiscale parameter $\varepsilon \to 0$. In particular, it is known that the invariant measure of X^ε converges weakly to the invariant measure of X solution of the homogenized equation (4.3). The following result guarantees the same kind of convergence for the couple $(X^\varepsilon, Z^\varepsilon)^\top$.

Lemma 4.11. *Under Assumption 4.1, let μ^ε be the invariant measure of the couple $(X^\varepsilon, Z^\varepsilon)^\top$ and let $(X_0^\varepsilon, Z_0^\varepsilon)^\top \sim \mu^\varepsilon$. Then, the measure μ^ε converges weakly to the measure $\mu^0(\mathrm{d}x,\mathrm{d}z) = \rho^0(x,z)\,\mathrm{d}x\,\mathrm{d}z$, whose density ρ^0 is the unique solution of the Fokker–Planck equation*

$$\Sigma\partial^2_{xx}\rho^0(x,z) + \partial_x\left(A\cdot V'(x)\rho^0(x,z)\right) + \frac{1}{\delta}\partial_z\left((z-x)\rho^0(x,z)\right) = 0, \tag{4.57}$$

where A and Σ are the coefficients of the homogenized equation (4.3).

Proof. Let $(X,Z)^\top := \left((X_t,Z_t)^\top, 0 \le t \le T\right)$ be the solution of

$$\begin{aligned} \mathrm{d}X_t &= -A\cdot V'(X_t)\,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W_t, \\ \mathrm{d}Z_t &= \frac{1}{\delta}(X_t - Z_t)\,\mathrm{d}t, \end{aligned} \tag{4.58}$$

with $(X_0, Z_0)^\top \sim \mu^0$. The arguments of Section 4.3.1 can be repeated to conclude that the invariant measure of $(X, Z)^\top$ admits a smooth density ρ^0 which satisfies (4.57). Moreover, standard homogenization theory (see e.g. [12, Chapter 3]) guarantees that $(X^\varepsilon, Z^\varepsilon)^\top \to (X, Z)^\top$ for $\varepsilon \to 0$ in law as random variables with values in $\mathscr{C}^0([0,T];\mathbb{R}^2)$. The Portmanteau theorem can be employed to conclude that the measure μ^ε converges weakly to μ^0 for $\varepsilon \to 0$. □

We conclude this section presenting an analogous result to Lemma 4.10 for the limit distribution.

Corollary 4.12. *Let ρ^0 be the solution of* (4.57) *and let us write*

$$\rho^0(x,z) = \varphi^0(x)\psi^0(z)R^0(x,z), \tag{4.59}$$

where φ^0 and ψ^0 are the marginal densities, i.e.,

$$\varphi^0(x) = \int_{\mathbb{R}} \rho^0(x,z)\,\mathrm{d}z, \quad \psi^0(z) = \int_{\mathbb{R}} \rho^0(x,z)\,\mathrm{d}x. \tag{4.60}$$

Then, if A and Σ are the coefficients of the homogenized equation (4.3)*, it holds*

$$\varphi^0(x) = \frac{1}{C_{\varphi^0}} \exp\left(-\frac{1}{\Sigma} A \cdot V(x)\right), \quad \textit{where} \quad C_{\varphi^0} = \int_{\mathbb{R}} \exp\left(-\frac{1}{\Sigma} A \cdot V(x)\right) \mathrm{d}x. \tag{4.61}$$

Moreover, it holds

$$\Sigma\delta \int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^0(x)\psi^0(z)\partial_x R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z = \mathbb{E}^{\rho^0}[(X^2 - Z^2)V''(Z)]. \tag{4.62}$$

Proof. The proof is directly obtained from Lemma 4.10 replacing $p(y) = 0$ and α, σ by A, Σ respectively. □

4.3.3 The filtering-based estimator

We now consider the inference problem and propose our filtering-based estimator of the drift, which is formally given by the formula

$$\widehat{A}^\varepsilon_k(T) = -\widetilde{M}^{-1}\widetilde{h}, \tag{4.63}$$

where we employ the subscript k for reference to the filter's kernel in (4.16), and where

$$\widetilde{M} = \frac{1}{T}\int_0^T V'(Z^\varepsilon_t) \otimes V'(X^\varepsilon_t)\,\mathrm{d}t, \qquad \text{and} \qquad \widetilde{h} = \frac{1}{T}\int_0^T V'(Z^\varepsilon_t)\,\mathrm{d}X^\varepsilon_t. \tag{4.64}$$

Let us remark that the formula above is obtained from (4.9) by replacing only one instance of X^ε_t with Z^ε_t in both M and h. In particular, it is fundamental for proving unbiasedness to keep in the definition of h the differential of the original process $\mathrm{d}X^\varepsilon_t$. Let us furthermore remark

that $\widehat{A}_k^\varepsilon(T)$ need not be the minimizer of some filtering-based likelihood function. In fact, if one were to replace Z_t^ε directly in (4.7), the symmetric part of the matrix $\widetilde{M}$ would appear and $\widehat{A}_k^\varepsilon(T)$ would not be the minimizer. Therefore, the estimator $\widehat{A}_k^\varepsilon(T)$ has to be thought of as a perturbation of $\widehat{A}^\varepsilon(T)$, directly at the level of estimators and after the maximization procedure. The only theoretical guarantee which is still needed for the well-posedness of $\widehat{A}_k^\varepsilon(T)$ is for $\widetilde{M}$ to be invertible, which we assume to be true and which we observed to hold in practice.

We are now able to present the main result of unbiasedness of the filtering-based estimator.

Theorem 4.13. *Let the assumptions of Lemma 4.6 and Lemma 4.11 hold, and let $\widehat{A}_k^\varepsilon(T)$ be defined in* (4.63)*. If $\widetilde{M}$ is invertible, then*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty}\widehat{A}_k^\varepsilon(T) = A, \quad a.s., \tag{4.65}$$

where A is the drift coefficient of the homogenized equation (4.1)*.*

Proof. Replacing the expression of $\mathrm{d}X_t^\varepsilon$ into (4.64), we get for $\widetilde{h}$

$$\widetilde{h} = -\widetilde{M}\alpha - \frac{1}{T}\int_0^T \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right)V'(Z_t^\varepsilon)\,\mathrm{d}t + \frac{\sqrt{2\sigma}}{T}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}W_t. \tag{4.66}$$

Therefore, we have

$$\begin{aligned}\widehat{A}_k^\varepsilon(T) &= \alpha + \frac{1}{T}\widetilde{M}^{-1}\int_0^T \frac{1}{\varepsilon}p'\left(\frac{X_t^\varepsilon}{\varepsilon}\right)V'(Z_t^\varepsilon)\,\mathrm{d}t - \frac{\sqrt{2\sigma}}{T}\widetilde{M}^{-1}\int_0^T V'(Z_t^\varepsilon)\,\mathrm{d}W_t \\ &=: \alpha + I_1^\varepsilon(T) - I_2^\varepsilon(T).\end{aligned} \tag{4.67}$$

We study the terms $I_1^\varepsilon(T)$ and $I_2^\varepsilon(T)$ separately. The ergodic theorem applied to $I_1^\varepsilon(T)$ yields

$$\lim_{T\to\infty} I_1^\varepsilon(T) = \mathbb{E}^{\rho^\varepsilon}[V'(Z^\varepsilon)\otimes V'(X^\varepsilon)]^{-1}\,\mathbb{E}^{\rho^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(Z^\varepsilon)\right], \quad \text{a.s.} \tag{4.68}$$

Due to Lemma 4.10 and integrating by parts, we have

$$\begin{aligned}\mathbb{E}^{\rho^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(Z^\varepsilon)\right] &= \int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\frac{1}{\varepsilon}p'\left(\frac{x}{\varepsilon}\right)\frac{1}{C_{\varphi^\varepsilon}}e^{-\frac{1}{\sigma}\alpha\cdot V(x)}e^{-\frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)}\psi^\varepsilon(z)R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= -\sigma\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-\frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)}\right)\frac{1}{C_{\varphi^\varepsilon}}e^{-\frac{1}{\sigma}\alpha\cdot V(x)}V'(z)\psi^\varepsilon(z)R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \sigma\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{1}{C_{\varphi^\varepsilon}}e^{-\frac{1}{\sigma}p\left(\frac{x}{\varepsilon}\right)}\partial_x\left(e^{-\frac{1}{\sigma}\alpha\cdot V(x)}R^\varepsilon(x,z)\right)V'(z)\psi^\varepsilon(z)\,\mathrm{d}x\,\mathrm{d}z,\end{aligned} \tag{4.69}$$

which implies

$$\mathbb{E}^{\rho^\varepsilon}\left[\frac{1}{\varepsilon}p'\left(\frac{X^\varepsilon}{\varepsilon}\right)V'(Z^\varepsilon)\right] = -\left(\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\otimes V'(x)\rho^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z\right)\alpha \tag{4.70}$$

$$+\sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z$$

$$= -\mathbb{E}^{\rho^\varepsilon}[V'(Z^\varepsilon)\otimes V'(X^\varepsilon)]\alpha + \sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z. \tag{4.71}$$

Replacing the equality above into (4.68), we obtain

$$\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \mathbb{E}^{\rho^\varepsilon}[V'(Z^\varepsilon)\otimes V'(X^\varepsilon)]^{-1}\sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^\varepsilon(x)\psi^\varepsilon(z)\partial_x R^\varepsilon(x,z)\,\mathrm{d}x\,\mathrm{d}z, \quad \text{a.s.} \tag{4.72}$$

Due to Lemma 4.10, we therefore have

$$\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \frac{1}{\delta}\mathbb{E}^{\rho^\varepsilon}[V'(Z^\varepsilon)\otimes V'(X^\varepsilon)]^{-1}\,\mathbb{E}^{\rho^\varepsilon}[((X^\varepsilon)^2-(Z^\varepsilon)^2)V''(Z^\varepsilon)], \quad \text{a.s.} \tag{4.73}$$

We now pass to the limit as ε goes to zero and

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + \frac{1}{\delta}\mathbb{E}^{\rho^0}[V'(Z)\otimes V'(X)]^{-1}\,\mathbb{E}^{\rho^0}[(X^2-Z^2)V''(Z)], \quad \text{a.s.,} \tag{4.74}$$

where the function ρ^0 is the density of the limit invariant distribution for $\varepsilon\to 0$, as in Lemma 4.11. Due to Corollary 4.12, we have

$$\frac{1}{\delta}\mathbb{E}^{\rho^0}[(X^2-Z^2)V''(Z)] = \Sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\varphi^0(x)\psi^0(z)\partial_x R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z, \tag{4.75}$$

and moreover, an integration by parts yields

$$\begin{aligned}\frac{1}{\delta}\mathbb{E}^{\rho^0}[(X^2-Z^2)V''(Z)] &= -\Sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)(\varphi^0)'(x)\psi^0(z)R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z\\ &= -\Sigma\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{C_{\varphi^0}}e^{-\frac{1}{\Sigma}A\cdot V(x)}\right)\psi^0(z)R^0(x,z)\,\mathrm{d}x\,\mathrm{d}z\\ &= \left(\int_{\mathbb{R}}\int_{\mathbb{R}} V'(z)\otimes V'(x)\rho^0(x,z)\,\mathrm{d}x\,\mathrm{d}z\right)A\\ &= \mathbb{E}^{\rho^0}[V'(Z)\otimes V'(X)]\,A.\end{aligned} \tag{4.76}$$

We can therefore conclude that

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I_1^\varepsilon(T) = -\alpha + A, \quad \text{a.s.} \tag{4.77}$$

We now consider the second term $I_2^\varepsilon(T)$, and rewrite it as

$$I_2^\varepsilon(T) = \sqrt{2\sigma}\,I_{2,1}^\varepsilon(T)\,I_{2,2}^\varepsilon(T), \tag{4.78}$$

where

$$\begin{aligned} I^{\varepsilon}_{2,1}(T) &:= \left(\frac{1}{T}\int_0^T V'(Z^{\varepsilon}_t)\otimes V'(X^{\varepsilon}_t)\,\mathrm{d}t\right)^{-1}\left(\frac{1}{T}\int_0^T V'(Z^{\varepsilon}_t)\otimes V'(Z^{\varepsilon}_t)\,\mathrm{d}t\right), \\ I^{\varepsilon}_{2,2}(T) &:= \left(\frac{1}{T}\int_0^T V'(Z^{\varepsilon}_t)\otimes V'(Z^{\varepsilon}_t)\,\mathrm{d}t\right)^{-1}\left(\frac{1}{T}\int_0^T V'(Z^{\varepsilon}_t)\,\mathrm{d}W_t\right). \end{aligned} \tag{4.79}$$

The ergodic theorem yields

$$\lim_{T\to\infty} I^{\varepsilon}_{2,1}(T) = \mathbb{E}^{\rho^{\varepsilon}}\left[V'(Z^{\varepsilon})\otimes V'(X^{\varepsilon})\right]^{-1}\mathbb{E}^{\rho^{\varepsilon}}\left[V'(Z^{\varepsilon})\otimes V'(Z^{\varepsilon})\right] =: R^{\varepsilon}, \tag{4.80}$$

where R^{ε} is bounded uniformly in ε. Moreover, the strong law of large numbers for martingales implies

$$\lim_{T\to\infty} I^{\varepsilon}_{2,2}(T) = 0, \quad \text{a.s.}, \tag{4.81}$$

independently of ε. Therefore

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} I^{\varepsilon}_{2}(T) = 0, \quad \text{a.s.}, \tag{4.82}$$

which, together with (4.77) and (4.67), proves the desired result. □

4.3.4 A second filtering-based estimator

In this section we consider a modification to the estimator (4.63), which is defined as

$$\widetilde{A}^{\varepsilon}_k(T) = -M^{-1}\widetilde{h}, \tag{4.83}$$

where the matrix M is taken from (4.10) and the vector $\widetilde{h}$ is taken from (4.64). For this estimator we will be able to prove a result of unbiasedness under a condition on the filtering parameter δ. Let us remark that for $\widehat{A}^{\varepsilon}_k(T)$ we do not need any condition on this parameter. Nevertheless, this second estimator is the minimizer of a likelihood function unlike the first, and will therefore be used in the Bayesian framework. Let us start with an estimate concerning the distance between X^{ε} and Z^{ε}.

Lemma 4.14. *Under Assumption 4.1, let the couple $(X^{\varepsilon}, Z^{\varepsilon})^{\top}$ be distributed as its invariant measure μ^{ε}. Then, if $\delta \le 1$, it holds for any integer $p \ge 1$*

$$\mathbb{E}^{\rho^{\varepsilon}}|X^{\varepsilon} - Z^{\varepsilon}|^{2p} \le C\left(\varepsilon^{2p} + \delta^{p}\right), \tag{4.84}$$

for a constant $C > 0$ independent of ε and δ.

Proof. Let us first remark that the filtering kernel satisfies

$$\int_0^t k(t-s)\,\mathrm{d}s = 1 - e^{-t/\delta}, \tag{4.85}$$

which implies that the measure $\kappa_t(\mathrm{d}s)$ on $(0,t)$ defined as

$$\kappa_t(\mathrm{d}s) := \frac{k(t-s)}{1-e^{-t/\delta}}\,\mathrm{d}s, \tag{4.86}$$

is a probability measure. Let X_t^ε be at stationarity with respect to its invariant measure, which we recall having density denoted as φ^ε. Let Z_t^ε be the corresponding filtered process. By definition of Z_t^ε and due (4.85) we now have

$$\begin{aligned} X_t^\varepsilon - Z_t^\varepsilon &= X_t^\varepsilon - \int_0^t k(t-s)X_s^\varepsilon\,\mathrm{d}s \\ &= \int_0^t k(t-s)\left(X_t^\varepsilon - X_s^\varepsilon\right)\mathrm{d}s + e^{-t/\delta}X_t^\varepsilon \\ &= (1-e^{-t/\delta})\int_0^t \left(X_t^\varepsilon - X_s^\varepsilon\right)\kappa_t(\mathrm{d}s) + e^{-t/\delta}X_t^\varepsilon. \end{aligned} \tag{4.87}$$

Therefore, Jensen's inequalities yields for a constant $C>0$ depending only on p

$$\begin{aligned} \mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - Z_t^\varepsilon|^{2p} &\le C(1-e^{-t/\delta})^{2p}\int_0^t \mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - X_s^\varepsilon|^{2p}\kappa_t(\mathrm{d}s) + Ce^{-2pt/\delta}\,\mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon|^{2p} \\ &\le C\int_0^t \mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - X_s^\varepsilon|^{2p}k(t-s)\,\mathrm{d}s + Ce^{-2pt/\delta}\,\mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon|^{2p}, \end{aligned} \tag{4.88}$$

where we replaced the definition of $\kappa(\mathrm{d}s)$ and exploited the fact that $1-e^{-t/\delta}\le 1$ for all $t\ge 0$. Due to [53, Corollary 5.4], we know that X_t^ε has bounded moments of all orders with respect to its invariant measure. Moreover, the estimate

$$\mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - X_s^\varepsilon|^{2p} \le C\left(\varepsilon^{2p} + (t-s)^{2p} + (t-s)^p\right), \tag{4.89}$$

holds for all $t>s$ due to [53, Lemma 6.1]. Finally, it is possible to verify by induction on r that for all integers r it holds

$$\int_0^t k(t-s)(t-s)^r\,\mathrm{d}s \le r!\delta^r, \tag{4.90}$$

which implies that there exists a constant $C>0$ depending only on p such that

$$\mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - Z_t^\varepsilon|^{2p} \le C\left(\varepsilon^{2p} + \delta^p + e^{-2pt/\delta}\right). \tag{4.91}$$

Let us now remark that this result holds for X_t^ε being at stationarity and for Z_t^ε be its filtered process, and not for a couple $(X^\varepsilon, Z^\varepsilon)^\top \sim \mu^\varepsilon$. In order to conclude, we remark that due to Lemma 4.6 we have for all $t\ge 0$ and for a constant C depending on the initial condition and p

$$\mathbb{E}^{\rho^\varepsilon}|X^\varepsilon - Z^\varepsilon|^{2p} \le \mathbb{E}^{\varphi^\varepsilon}|X_t^\varepsilon - Z_t^\varepsilon|^{2p} + Ce^{-\lambda t}, \tag{4.92}$$

which, for t sufficiently big, yields

$$\mathbb{E}^{\rho^\varepsilon}|X^\varepsilon - Z^\varepsilon|^{2p} \le C\left(\varepsilon^{2p} + \delta^p\right), \tag{4.93}$$

which is the desired result. □

Let us consider a couple $(X^\varepsilon, Z^\varepsilon)^\top$. In the following, we will adopt the notation

$$\mathscr{M} = \mathbb{E}^{\rho^\varepsilon}[V'(X^\varepsilon) \otimes V'(X^\varepsilon)] \quad \text{and} \quad \widetilde{\mathscr{M}} = \mathbb{E}^{\rho^\varepsilon}[V'(Z^\varepsilon) \otimes V'(X^\varepsilon)], \tag{4.94}$$

which due to the ergodic theorem satisfy

$$\lim_{T\to\infty} M = \mathscr{M}, \quad \text{and} \quad \lim_{T\to\infty} \widetilde{M} = \widetilde{\mathscr{M}}, \tag{4.95}$$

almost surely. In the following Lemma, we consider the difference between the two matrices $\mathscr{M}$ and $\widetilde{\mathscr{M}}$.

Lemma 4.15. *Let the assumptions of Lemma 4.14 hold. Then the matrices $\mathscr{M}$ and $\widetilde{\mathscr{M}}$ defined in* (4.94) *satisfy*

$$\|\mathscr{M} - \widetilde{\mathscr{M}}\|_2 \le C\left(\varepsilon + \delta^{1/2}\right), \tag{4.96}$$

for a constant $C > 0$ independent of ε and δ.

Proof. Applying Jensen's and Cauchy–Schwarz inequalities we have

$$\begin{aligned} \|\mathscr{M} - \widetilde{\mathscr{M}}\|_2 &\le \mathbb{E}^{\rho^\varepsilon} \|\left(V'(Z^\varepsilon) - V'(X^\varepsilon)\right) \otimes V'(X^\varepsilon)\|_2 \\ &\le \left(\mathbb{E}^{\rho^\varepsilon} \|V'(Z^\varepsilon) - V'(X^\varepsilon)\|_2^2\right)^{1/2} \left(\mathbb{E}^{\rho^\varepsilon} \|V'(X^\varepsilon)\|_2^2\right)^{1/2}. \end{aligned} \tag{4.97}$$

The Lipschitz condition on V together with the boundedness of the moments of X^ε and Lemma 4.14 yield for a constant $C > 0$

$$\|\mathscr{M} - \widetilde{\mathscr{M}}\|_2 \le C\left(\mathbb{E}^{\rho^\varepsilon} |Z^\varepsilon - X^\varepsilon|^2\right)^{1/2} \le C\left(\varepsilon + \delta^{1/2}\right), \tag{4.98}$$

which is the desired result. □

We can finally prove the unbiasedness of the estimator $\widetilde{A}_k^\varepsilon(T)$.

Theorem 4.16. *Let the assumptions of Theorem 4.13 hold. Then, if δ in* (4.16) *satisfies $\delta = \delta(\varepsilon)$ with $\delta(\varepsilon) \to 0$ for $\varepsilon \to 0$, it holds*

$$\lim_{\varepsilon\to 0} \lim_{T\to\infty} \widetilde{A}_k^\varepsilon(T) = A, \quad a.s., \tag{4.99}$$

for $\widetilde{A}_k^\varepsilon(T)$ defined in (4.83).

Proof. We first consider the difference between the two estimators $\widetilde{A}_k^\varepsilon(T)$ and $\widehat{A}_k^\varepsilon(T)$. In

particular, the ergodic theorem and an algebraic equality imply

$$\begin{aligned}\lim_{T\to\infty}\left(\widetilde{A}_k^\varepsilon(T)-\widehat{A}_k^\varepsilon(T)\right) &= \left(\mathscr{M}^{-1}-\widetilde{\mathscr{M}}^{-1}\right)\lim_{T\to\infty}\widetilde{h}\\ &= \mathscr{M}^{-1}\left(\mathscr{M}-\widetilde{\mathscr{M}}\right)\widetilde{\mathscr{M}}^{-1}\lim_{T\to\infty}\widetilde{h}\\ &= \mathscr{M}^{-1}\left(\mathscr{M}-\widetilde{\mathscr{M}}\right)\lim_{T\to\infty}\widehat{A}_k^\varepsilon(T),\end{aligned} \tag{4.100}$$

almost surely. Therefore, due to Assumption 4.1 which allows controlling the norm of $\mathscr{M}^{-1}$ and due to Lemma 4.15 we have for a constant $C>0$

$$\lim_{T\to\infty}\|\widetilde{A}_k^\varepsilon(T)-\widehat{A}_k^\varepsilon(T)\|_2 \le C\left(\varepsilon+\delta^{1/2}\right). \tag{4.101}$$

Let us remark that $\widehat{A}_k^\varepsilon(T)$ has a bounded norm for ε sufficiently small due to Theorem 4.13. Now, the triangle inequality yields

$$\|\widetilde{A}_k^\varepsilon(T)-A\|_2 \le \|\widetilde{A}_k^\varepsilon(T)-\widehat{A}_k^\varepsilon(T)\|_2+\|\widehat{A}_k^\varepsilon(T)-A\|_2. \tag{4.102}$$

Therefore, due to Theorem 4.13, the inequality (4.101) and since $\delta\to 0$ for $\varepsilon\to 0$, the desired result holds. □

4.4 The Bayesian setting

In this section we present a Bayesian reinterpretation of the inference procedure, which, given the structure of the problem, allows to get at a full uncertainty quantification with a low computational effort.

Let us fix a Gaussian prior $\mu_0=\mathscr{N}(A_0,C_0)$ on A, where $A_0\in\mathbb{R}^N$ and $C_0\in\mathbb{R}^{N\times N}$ is symmetric positive definite. Then, given a final time $T>0$, the posterior distribution $\mu_{T,\varepsilon}$ admits a density $p(A\mid X^\varepsilon)$ with respect to the Lebesgue measure which satisfies

$$p(A\mid X^\varepsilon)=\frac{1}{Z^\varepsilon}\,p(X^\varepsilon\mid A)\,p_0(A), \tag{4.103}$$

where Z^ε is the normalization constant, p_0 is the density of μ_0, and where the likelihood $p(X^\varepsilon\mid A)$ is given in (4.7). The log-posterior density is therefore given by

$$\log p(A\mid X^\varepsilon)=-\log Z^\varepsilon-\frac{T}{2\Sigma}A\cdot h-\frac{T}{4\Sigma}A\cdot MA-\frac{1}{2}(A-A_0)\cdot C_0^{-1}(A-A_0), \tag{4.104}$$

where M and h are defined in (4.10). 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. We denote by $m_{T,\varepsilon}$ and $C_{T,\varepsilon}$ the mean and covariance matrix, respectively.

Completing the squares in the log-posterior density, we formally obtain

$$\begin{aligned} C_{T,\varepsilon}^{-1} &= C_0^{-1} + \frac{T}{2\Sigma} M, \\ C_{T,\varepsilon}^{-1} m_{T,\varepsilon} &= C_0^{-1} A_0 - \frac{T}{2\Sigma} h. \end{aligned} \tag{4.105}$$

Under Assumption 4.1, one can show that the posterior at time $T > 0$ is indeed given by $\mu_{T,\varepsilon} = \mathcal{N}(m_{T,\varepsilon}, C_{T,\varepsilon})$. We are now interested in the asymptotic limit of the posterior distribution for $T \to \infty$ and for $\varepsilon \to 0$.

We can now state the main result for asymptotic convergence of the posterior distribution.

Theorem 4.17. *Under Assumption 4.1, the posterior measure $\mu_{T,\varepsilon} = \mathcal{N}(m_{T,\varepsilon}, C_{T,\varepsilon})$ satisfies*

$$\lim_{\varepsilon \to 0} \lim_{T \to \infty} \mu_{T,\varepsilon}(\cdot) = \delta(\cdot - \alpha), \quad a.s., \tag{4.106}$$

weakly in the space of probability measures on $\mathbb{R}^N$, where δ is the Dirac measure on $\mathbb{R}^N$.

Proof. Let us first consider the covariance matrix. Hua's identity yields

$$C_{T,\varepsilon} = \frac{2\Sigma}{T} \left(M^{-1} - Q^{-1} \right), \tag{4.107}$$

where

$$Q = M + \frac{T}{2\Sigma} M C_0 M. \tag{4.108}$$

Let us first remark that due to the hypothesis on M and the ergodic theorem it holds for all $T > 0$

$$\| M^{-1} \|_2 \leq \frac{1}{\bar{\lambda}}. \tag{4.109}$$

We now have that for generic symmetric positive definite matrices R and S it holds

$$\| (R+S)^{-1} \|_2 \leq \| S^{-1} \|_2. \tag{4.110}$$

Applying this inequality to Q^{-1}, we obtain

$$\| Q^{-1} \|_2 \leq \frac{2\Sigma}{T} \| (M C_0 M)^{-1} \|_2 \leq \frac{2\Sigma}{T} \| M^{-1} \|_2^2 \| C_0^{-1} \|_2 = \frac{2\Sigma}{T\bar{\lambda}^2} \| C_0^{-1} \|_2, \tag{4.111}$$

which implies

$$\lim_{T \to \infty} \| Q^{-1} \|_2 = 0, \tag{4.112}$$

and due to the triangle inequality

$$\lim_{T \to \infty} \| C_{T,\varepsilon} \|_2 = 0. \tag{4.113}$$

We proved that in the limit for $T \to \infty$ the covariance shrinks to zero independently of ε. We

now consider the mean. First, we remark that the triangle inequality yields

$$\|m_{T,\varepsilon} - \alpha\|_2 \leq \|m_{T,\varepsilon} - \widehat{A}^\varepsilon(T)\|_2 + \|\widehat{A}^\varepsilon(T) - \alpha\|_2. \tag{4.114}$$

For the second term, Theorem 4.2 implies

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \|\widehat{A}^\varepsilon(T) - \alpha\|_2 = 0, \quad \text{a.s.} \tag{4.115}$$

Let us now consider the first term. Replacing the expression of the maximum likelihood estimator (4.9) and due to the Cauchy–Schwarz and triangle inequalities, we obtain

$$\begin{aligned}\|m_{T,\varepsilon} - \widehat{A}^\varepsilon(T)\|_2 &= \frac{2\Sigma}{T}\|M^{-1}C_0^{-1}A_0 - Q^{-1}\left(C_0^{-1}A_0 - \frac{T}{2\Sigma}h\right)\|_2 \\ &\leq \frac{2\Sigma}{T\bar{\lambda}}\|C_0^{-1}\|_2\left(\|A_0\|_2 + \frac{1}{\bar{\lambda}}\|h\|_2 + \frac{2\Sigma}{T\bar{\lambda}}\|C_0^{-1}\|_2\|A_0\|_2\right).\end{aligned} \tag{4.116}$$

Moreover, the ergodic theorem and the strong law of large numbers for martingales guarantee that $\|h\|_2$ is bounded a.s. for $T\to\infty$. Therefore

$$\lim_{T\to\infty} \|m_{T,\varepsilon} - \widehat{A}^\varepsilon(T)\|_2 = 0, \quad \text{a.s.}, \tag{4.117}$$

independently of ε. Finally,

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \|m_{T,\varepsilon} - \alpha\|_2 = 0, \quad \text{a.s.}, \tag{4.118}$$

which, together with (4.113), implies the desired result. □

Remark 4.18. The result above has the same consequences in the Bayesian setting as Theorem 4.2 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 unhomogenized equation (4.1). Moreover, a partial result which can be deduced from the proof is that in the limit for $T\to\infty$ and for a positive value $\varepsilon > 0$ the Bayesian and the MLE approaches are equivalent. In particular, we have for all $\varepsilon > 0$

$$\begin{aligned}&\lim_{T\to\infty} \|C_{T,\varepsilon}\|_2 = 0, \\ &\lim_{T\to\infty} \|m_{T,\varepsilon} - \widehat{A}^\varepsilon(T)\|_2 = 0,\end{aligned} \tag{4.119}$$

i.e., the weak limit of the posterior $\mu_{T,\varepsilon}$ for $T\to\infty$ is the Dirac delta concentrated on the limit of $\widehat{A}^\varepsilon(T)$ for $T\to\infty$.

4.4.1 The filtering approach

In this section, we present how to correct the faulty behaviour highlighted by Theorem 4.17 with filtering techniques. In particular, we employ the same strategy that led to the estimator

$\widetilde{A}_k^\varepsilon(T)$ presented in Section 4.3.4. Therefore, we modify the likelihood function as

$$\widetilde{p}(X^\varepsilon \mid A) = \exp\left(-\frac{\widetilde{I}(X^\varepsilon \mid A)}{2\Sigma}\right), \tag{4.120}$$

where

$$\widetilde{I}(X^\varepsilon \mid A) = \int_0^T A \cdot V'(Z_t^\varepsilon)\,\mathrm{d}X_t^\varepsilon + \int_0^T \left(A \cdot V'(X_t^\varepsilon)\right)^2 \mathrm{d}t, \tag{4.121}$$

so that the estimator $\widetilde{A}_k^\varepsilon(T)$ introduced in (4.83) is the MLE based on the new likelihood $\widetilde{p}(X^\varepsilon \mid A)$. With this likelihood function, we obtain the modified posterior $\widetilde{\mu}_{T,\varepsilon} = \mathcal{N}(\widetilde{m}_{T,\varepsilon}, C_{T,\varepsilon})$, whose parameters are given by

$$\begin{aligned} C_{T,\varepsilon}^{-1} &= C_0^{-1} + \frac{T}{2\Sigma} M, \\ C_{T,\varepsilon}^{-1}\widetilde{m}_{T,\varepsilon} &= C_0^{-1} A_0 - \frac{T}{2\Sigma}\widetilde{h}. \end{aligned} \tag{4.122}$$

Let us remark that the posterior $\widetilde{\mu}_{T,\varepsilon}$ has the same covariance as $\mu_{T,\varepsilon}$. The following result now holds.

Theorem 4.19. *Let the Assumptions of Theorem 4.16 hold. Then, the modified posterior measure $\widetilde{\mu}_{T,\varepsilon} = \mathcal{N}(\widetilde{m}_{T,\varepsilon}, C_{T,\varepsilon})$ satisfies*

$$\lim_{\varepsilon\to 0}\lim_{T\to\infty} \widetilde{\mu}_{T,\varepsilon}(\cdot) = \delta(\cdot - A), \quad a.s., \tag{4.123}$$

weakly in the space of probability measures on $\mathbb{R}^N$, where δ is the Dirac measure on $\mathbb{R}^N$.

Proof. The proof follows from the proof of Theorem 4.17 and from Theorem 4.16. □

4.5 Numerical experiments

In this section we show numerical experiments confirming our theoretical findings and showing the potentiality of the filtering approach.

4.5.1 Parameters of the filter

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 multiscale equation (4.1) with $\varepsilon = 0.1$ and the fast potential $p(y) = \cos(y)$. We generate data X^ε for $0 \le t \le T$ and $T = 10^3$ employing the Euler–Maruyama method with time step $\Delta_t = \varepsilon^3$.

As a first experiment, we consider the effect of setting different values of the parameter δ of the filter (4.16) on the estimator. As a comparison and for similarity with the subsampling approach, we consider $\delta = \varepsilon^\zeta$ and vary $\zeta \in [0,1]$. We employ the same values of δ for the subsampling, too, as the theoretical results guarantee convergence only for these values.

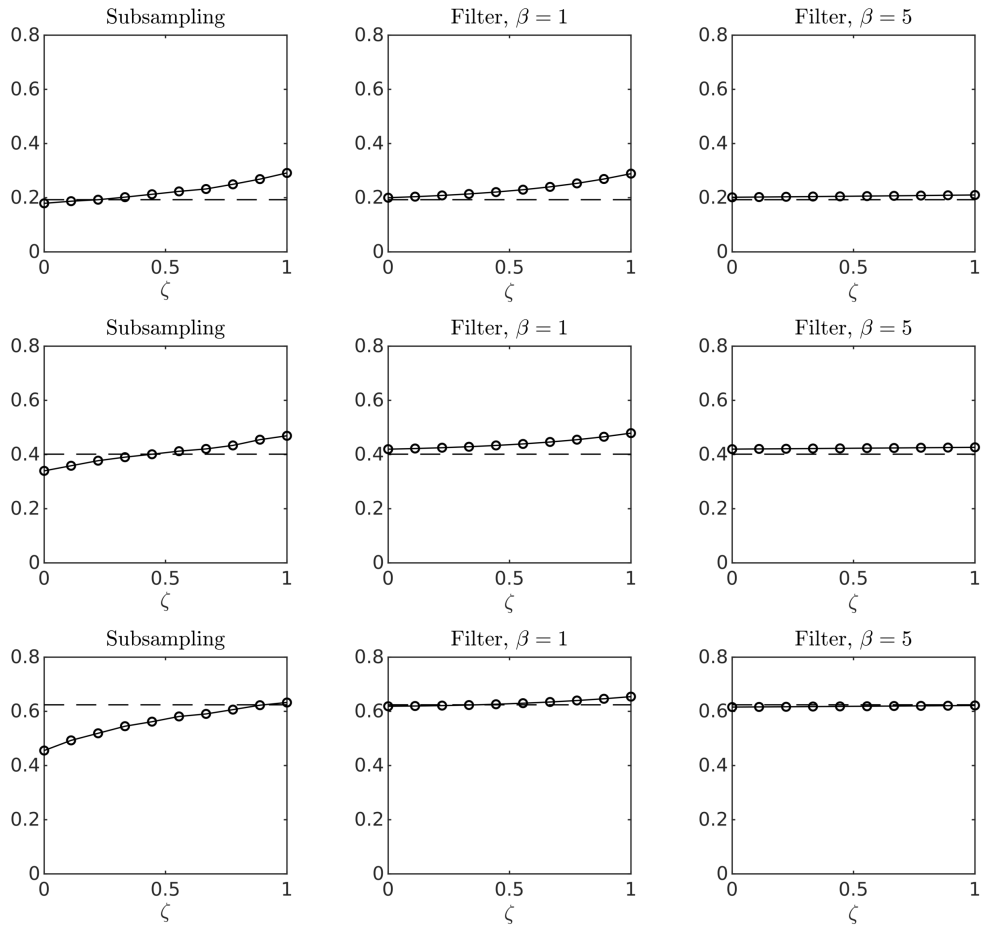

Figure 4.1 – Results for the Ornstein–Uhlenbeck process. The axis of abscissae ζ represents the value of $\delta = \varepsilon^{\zeta}$ for both subsampling and the filter. The three rows correspond to $\sigma = 0.5, 0.7, 1.0$ from top to bottom.

Moreover, we repeat the experiments fixing $\beta = 1,5$ and for three different values of the diffusion, i.e., we choose $\sigma = 0.5, 0.7, 1$. We report in Figure 4.1 the experimental results. Let us remark that

1. 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 filtering approach seems better than subsampling,
2. in general, choosing a higher value of β seems beneficial for the quality of the estimator,
3. 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, as predicted in Theorem 4.13. This can be observed for a higher value of β but we have no theoretical guarantee in this case.

As a second experiment, we therefore test the variability of the estimator with respect to β

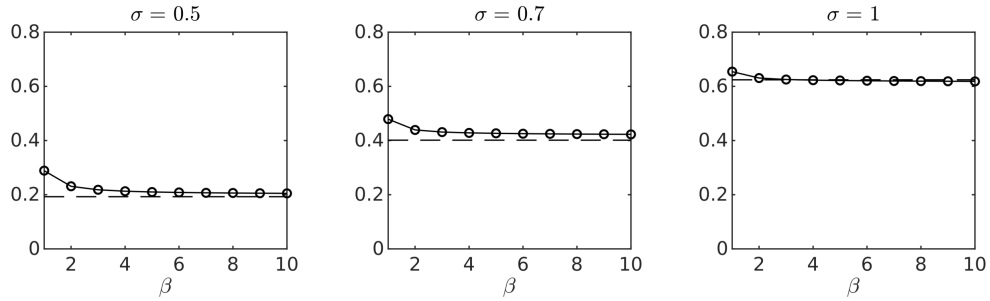

Figure 4.2 – Dependence of the estimator given by the filter for the Ornstein–Uhlenbeck process on the parameter β in (4.16). From left to right we consider different values of σ.

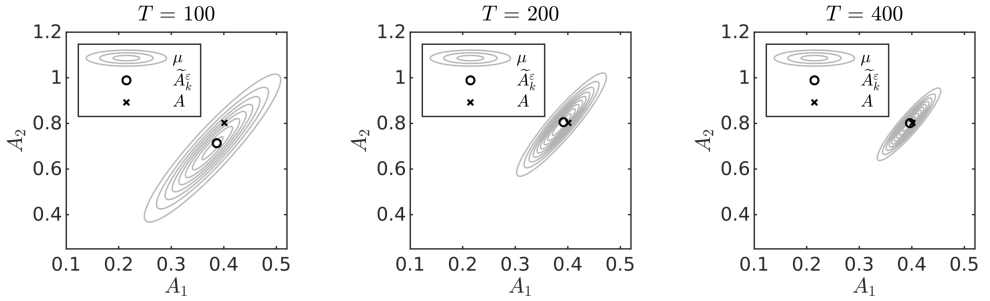

Figure 4.3 – Posterior distributions over the parameter $A = (A_1, A_2)^\top$ for the bistable potential obtained with the filtering approach. The figures refer to final time $T = 100, 200, 400$ from left to right, respectively. The MLE $\widetilde{A}_k^\varepsilon(t)$ is represented with a circle, while the true value A of the drift coefficient of the homogenized equation is represented with a cross.

in (4.16). We repeat the numerical experiment in Figure 4.1 and we choose to 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 4.2, show empirically that the estimator is fast stable with respect to β.

4.5.2 The Bayesian approach: bistable potential

In this numerical experiment we consider $N = 2$ and the bistable potential, i.e., the function V defined as

$$V(x) = \begin{pmatrix} \dfrac{x^4}{4} & -\dfrac{x^2}{2} \end{pmatrix}^\top, \tag{4.124}$$

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 $\widetilde{\mu}_{T,\varepsilon}$ obtained with the filtering approach introduced in Section 4.4.1. The parameters of the filter are set to $\beta = 1$ and $\delta = \varepsilon$ in (4.16). 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

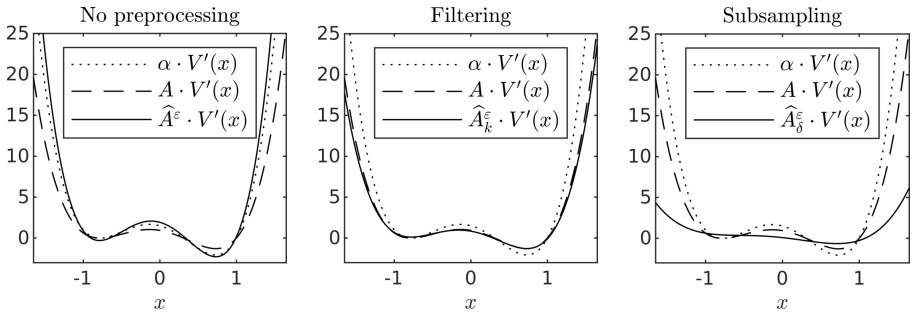

Figure 4.4 – Results for the four-dimensional parameter. From left to right the potential function estimated with the data itself, the filter, subsampled data.

Coefficient	Truth A	No preprocessing $\widehat{A}^{\varepsilon}$	Filtering $\widehat{A}^{\varepsilon}_{k}$	Subsampling $\widehat{A}^{\varepsilon}_{\delta}$
A_1	-0.62	-0.92	-0.70	-0.59
A_2	-0.31	-0.70	-0.27	0.05
A_3	0.31	0.55	0.31	0.14
A_4	0.62	1.22	0.57	0.13

Table 4.1 – Numerical results for the four-dimensional parameter rounded to the second significant figure.

of Σ via the coefficient K and the theory of homogenization, but let us remark that Σ could be estimated employing the subsampling technique of [53]. We stop computations at times $T = 100, 200, 400$ in order to observe the shrinkage of the Gaussian posterior towards the MLE $\widetilde{A}^{\varepsilon}_{k}(T)$ with respect to time. In Figure 4.3, 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.

4.5.3 Multi-dimensional parameter

Let us consider the Chebyshev polynomials of the first kind, i.e., the polynomials $T_i : \mathbb{R} \to \mathbb{R}$, $i = 0, 1, \ldots$, defined by the recurrence relation

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x). \tag{4.125}$$

We consider the potential function $V(x)$ as in (4.2) with

$$V_i(x) = T_i(x), \quad i = 1, \ldots, 4. \tag{4.126}$$

This potential function satisfies Assumption 4.1 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 $\Delta_t = \epsilon^3$. We estimate the drift coefficient $A \in \mathbb{R}^4$ with the estimators

1. $\widehat{A}^\epsilon(T)$ based on the data X^ϵ itself,
2. $\widehat{A}^\epsilon_\delta(T)$ based on subsampled data with subsampling parameter $\delta = \epsilon^{2/3}$,
3. $\widehat{A}^\epsilon_k(T)$ 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 [53]. Results, given in Figure 4.4 and Table 4.1, 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 4.2 holds in practice, i.e., with no pre-processing the estimator $\widehat{A}^\epsilon(T)$ tends to the drift coefficient α of the unhomogenized equation.

4.6 Conclusion

In this work we considered a novel methodology based on filtering for the estimation of the drift of multiscale diffusion processes. We proved results of ergodicity, convergence and, most importantly, unbiasedness of estimators drawn from our methodology. Moreover, we combined a Bayesian approach and our new technique to guarantee a robust uncertainty quantification of the inference procedure. Numerical experiments demonstrate how the filtering-based estimator requires less knowledge of the characteristic time-scales of the multiscale equation and how it can be employed as a black-box tool for parameter estimation on a range of academic examples. We believe this work gives way to several further developments. In particular, we believe it would be relevant to

1. analyse the filtering methodology for $\beta > 1$ in (4.16), which seems to give more robust results in practice,
2. extend the analysis to the non-parametric framework most likely by means of Bayesian regularization techniques,
3. consider multiscale models for which the homogenized equation presents multiplicative noise,
4. test the filtering methodology against real-world data,
5. apply similar filtering methodologies to correct faulty behaviour of non likelihood-based estimators.

5 Ensemble Kalman filter for multiscale inverse problems

5.1 Introduction

In this work we consider the application of techniques derived from the Kalman filter to inverse problems involving multiscale phenomena which can be modelled by means of partial differential equations (PDEs). Inverse problems arise in many fields, such as seismography, meteorology and tomography, all physical domains with a multiscale nature. Our reference mathematical model is given by multiscale elliptic PDEs of the form

$$\begin{cases} -\nabla \cdot (A_u^\varepsilon \nabla p^\varepsilon) = f, & \text{in } \Omega, \\ p^\varepsilon = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.1}$$

where $\Omega \subset \mathbb{R}^d$ is the physical domain, A_u^ε is a tensor oscillating with an amplitude described by the parameter ε and u is a possibly infinite-dimensional unknown which parametrizes the tensor A_u^ε. We are then interested in the solution of inverse problems involving the retrieval of the parameter u given noisy observations derived from the solution p^ε.

Multiscale inverse problems of this form have been recently introduced in [46] and analysed extensively in [2, 3]. In particular, in [3] Abdulle and Di Blasio build a coarse-graining approach to solve the inverse problem regularized with a Tikhonov technique. The main idea is replacing the computationally expensive solution of the highly-oscillating multiscale problem with an homogenized surrogate, which eliminates the fast variables and is therefore cheaper. In particular, the theory of homogenization guarantees under certain assumptions, which will be specified throughout this work, that there exists a PDE of the form

$$\begin{cases} -\nabla \cdot (A_u^0 \nabla p^0) = f, & \text{in } \Omega, \\ p^0 = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.2}$$

such that the solution p^0 is the weak limit of the functions p^ε in the vanishing limit for ε, and such that A_u^0 is independent of ε. In [**?**], the authors showed that employing this homogenized model to the multiscale inverse problem guarantees a good approximation to its solution if a

Tikhonov regularization is employed. This framework has been successively enlarged by the same authors to the Bayesian case in [**?**], where the analysis involves posterior distributions arising from both the multiscale and the homogenized model. In the same work, a technique for estimating the modelling error which was developed in [14, 13] is successfully applied to multiscale inverse problems to account for the homogenization and discretization errors.

The ensemble Kalman filter (EnKF), first introduced in [**?**], is an algorithm which is widely employed in the engineering community for the estimation of the state of partially-observed dynamical systems whose dynamics are governed by a nonlinear agent. In particular, Kalman filters have long been used successfully in meteorology, oceanography and automation applications. In [**?**], Iglesias et al. propose the application of the EnKF method to obtain a point-wise solution to inverse problems involving PDEs, and an extension of their analysis giving a Bayesian interpretation of the filtering solution is presented in [**?**].

In this work, we present a combination of the well-established techniques of homogenization and filtering to build a novel scheme for solving multiscale inverse problems in an efficient and reliable manner. In the same spirit of [**?**, 2], we prove that it is possible to eliminate the fast scales from the PDE appearing in the inverse problem relying on the theory of homogenization, thus obtaining a solution which is accurate in the vanishing limit for the multiscale parameter ε. In our analysis, we both consider point-wise estimations as in [**?**] and Bayesian solutions as in [**?**], thus showing convergence results which are endowed with decay rates under special assumptions on the problem. Inspired by [14, 13, 2], we then consider offline and online techniques for estimating the modelling error and prove a novel result indicating the computational cost which is required for such an estimation for any given multiscale problem.

We identified two main advantages of a filtering approach as the one provided by the EnKF method with respect to other approaches. First, a Bayesian interpretation of the solution to the inverse problem is obtained from the algorithm without any additional cost with respect to a point-wise estimation. A distribution on the unknown provides in fact with a deeper insight and a full uncertainty quantification on the solution, which is therefore in turn more interpretable and robust. Secondly, the EnKF can be simply divided in sequential parallel runs, which we verified in practice to allow a faster computation of the solution of rather complex inverse problems with respect to standard Monte Carlo approaches, such as the Metropolis–Hastings algorithm.

The outline of the work is the following. In Section 5.2 we introduce the setting of the problem in a rigorous manner, as well as the notation which will be employed throughout this work. Then, in Section 5.3, we briefly summarize the techniques introduced in [**?**, **?**] to solve inverse problems both in a point-wise and in a Bayesian spirit employing the EnKF method. In Section 5.4 we present the results of convergence of the EnKF scheme in the multiscale setting, which is the main contribution of this work. Then, Section 5.5 is dedicated to the estimation of the modelling error, and to a novel theoretical results which strengthens its value in practice.

Finally, Section 5.6 is devoted to numerical experiments which corroborate our analysis.

5.2 Problem setting

Given a positive parameter ε, let us consider the multiscale inverse problem

$$\text{find } u \in X \text{ given observations } y = \mathcal{G}^\varepsilon(u) + \eta \in Y, \tag{5.3}$$

where the parameter space X and the observation space Y are Hilbert spaces, the multiscale forward operator $\mathcal{G}^\varepsilon\colon X \to Y$ maps the unknown to the observation space, and $\eta \in Y$ is a source of additive noise, which we assume to be distributed as a Gaussian $\mathcal{N}(0,\Gamma)$, where Γ is a positive definite covariance operator. We assume that the forward operator $\mathcal{G}^\varepsilon$ can be written as $\mathcal{G}^\varepsilon = \mathcal{O}\circ\mathcal{S}^\varepsilon$, where $\mathcal{O}\colon H_0^1(\Omega)\to Y$ is the observation operator and $\mathcal{S}^\varepsilon\colon X \to H_0^1(\Omega)$ is the solution operator of a multiscale elliptic partial differential equation (PDE). Letting Ω be a bounded open domain in $\mathbb{R}^d$, the operator $\mathcal{S}^\varepsilon$ maps the unknown u to the solution p^ε of

$$\begin{cases} -\nabla\cdot(A_u^\varepsilon \nabla p^\varepsilon) = f, & \text{in } \Omega, \\ p^\varepsilon = 0, & \text{on } \partial\Omega. \end{cases} \tag{5.4}$$

Let us introduce a regularity assumption for the observation operator.

Assumption 5.1. The observation operator $\mathcal{O}\colon H_0^1(\Omega)\to Y$ satisfies for all $p_1, p_2 \in H_0^1(\Omega)$

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_Y \le m\|p_1 - p_2\|_{L^2(\Omega)},$$

where m is a positive constant.

Note that since $\mathcal{O}$ is defined on $H_0^1 \subset L^2$, Assumption 5.1 is stronger than Lipschitz continuity. The tensor A_u^ε belongs to the class of parametrized multiscale tensors which admit explicit scale separation between slow and fast spatial variables, i.e.,

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

where the map $(t,x)\to A(t,x/\varepsilon)$ is assumed to be known and A is periodic in its second argument. If ε is small, a fine discretization is needed to resolve the smallest scale and thus evaluate the forward operator $\mathcal{G}^\varepsilon$, which in turn leads to a high computational cost. Considering also that the PDE has to be solved several times in the framework of inverse problems, this procedure can be infeasible.

In order to approach the multiscale problem in a more efficient manner we therefore apply the theory of homogenization (see e.g. [**?**]), which ensures the existence of an homogenized tensor A^0, such that for $\varepsilon\to 0$ the solution p^ε of (5.4) tends weakly in $H^1(\Omega)$ to the solution p^0

of the problem

$$\begin{cases} -\nabla \cdot (A_u^0 \nabla p^0) = f, & \text{in } \Omega, \\ p^0 = 0, & \text{on } \partial\Omega. \end{cases} \tag{5.5}$$

Hence, the function p^0 can be assumed to be a good approximation of p^ε when the multiscale parameter ε is small and thus, in this case, the multiscale problem (5.4) can be replaced by its homogenized version (5.5). Let us denote by $\mathcal{G}_h^0 \colon \mathcal{O} \circ \mathcal{S}_h^0$ the forward operator which maps the unknown parameter into the solution of (5.5) computed with the finite element method (FEM) with discretization parameter $h > 0$. Inspired by [46, **?**, 2], we employ the homogenized operator $\mathcal{G}_h^0$, which is cheaper to evaluate numerically, to solve the inverse problem (**??**) Finally, let us introduce a regularity assumption on the multiscale and homogenzed tensors A^ε and A^0.

Assumption 5.2. The tensors A^ε and A^0 in (5.4) and (5.5) respectively satisfy for all $u, u_1, u_2 \in X$ and $\xi \in \mathbb{R}^d$

$$\begin{aligned} &\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \le M\|u_1 - u_2\|_X, \\ &\|A^0(u_1) - A^0(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \le M\|u_1 - u_2\|_X, \\ &A^\varepsilon(u)\xi\cdot\xi \ge \alpha_0\|\xi\|_2^2, \quad A^0(u)\xi\cdot\xi \ge \alpha_0\|\xi\|_2^2, \end{aligned}$$

where M and α_0 are positive constants.

5.3 A Kalman filter solution to inverse problems

In this section, we present a technique to solve an inverse problem of the form (5.3) based on the ensemble Kalman filter (EnKF). Further details about the EnKF and its applications to inverse problems ought to be found in [**?**, **?**], where the authors develop a framework for inverse problems involving single-scale PDE models. Inverse problems are often ill-posed and hence require regularization, which can be granted by, for example, variational and Bayesian techniques. The EnKF method achieves regularization by searching for the solution of the inverse problem in a finite dimensional and compact subset $\mathcal{A}$ of X, which incorporates prior knowledge of u. In this section, we consider a general forward map $\mathcal{G} \colon X \to Y$ and the observations to be given by $y = \mathcal{G}(u) + \eta$ for an unknown $u \in X$ and a Gaussian noise $\eta \sim \mathcal{N}(0, \Gamma)$.

The traditional theory of Kalman filters involves the estimation of the state of a dynamical system which is observed in a partial and noisy manner. Therefore, in order to approximate the unknown by means of the Kalman filter theory, we need to introduce artificial dynamics based on state augmentation. Given the space $Z = X \times Y$, let us define the map $\Xi \colon Z \to Z$ as

$$\Xi(z) = \begin{bmatrix} u \\ \mathcal{G}(u) \end{bmatrix}, \quad \text{for} \quad z = \begin{bmatrix} u \\ v \end{bmatrix} \in Z, \tag{5.6}$$

which is employed to construct artificial dynamics as

$$z_{n+1} = \Xi(z_n). \tag{5.7}$$

We assume that data related to the artificial dynamics has the form

$$y_{n+1} = Hz_{n+1} + \eta_{n+1},$$

where $H\colon Z \to Y$ is a projection operator defined by $H = \begin{bmatrix} 0 & I \end{bmatrix}$ and $\{\eta_n\}_{n\in\mathbb{N}}$ is an i.i.d. sequence of random variables distributed as $\eta_n \sim \mathcal{N}(0,\Gamma)$, i.e., with the same distribution as the noise in (5.3). Consequently, we get

$$y_{n+1} = H\Xi(z_n) + \eta_{n+1} = \mathcal{G}(u_n) + \eta_{n+1}.$$

The EnKF method, which we briefly describe here, proceeds by propagating an ensemble $\{z_n^{(j)}\}_{j=1}^J \subset Z$ of J particles for $n = 0,\dots,N$, following the classical Kalman update formulae. Let $\mathcal{A} \subset X$ be such that $\dim(\mathcal{A}) \leq J$, and let $\{\psi^{(j)}\}_{j=1}^J \subset \mathcal{A}$. The initial ensemble $\{z_0^{(j)}\}_{j=1}^J$ is then built as

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

Each iteration of the EnKF method can be split in two parts, the prediction and analysis steps. At the n-th step of the algorithm, the current ensemble of particles $\{z_n^{(j)}\}_{j=1}^J$ is first mapped forward through the augmented dynamics (5.7), i.e., for all $j = 1,\dots,J$

$$\widehat{z}_{n+1}^{(j)} = \Xi(z_n^{(j)}). \tag{5.8}$$

Let us remark that this step introduces information on the forward model due to how the second component of the map Ξ is defined, which implies that the partially-updated ensemble $\{\widehat{z}_{n+1}^{(j)}\}_{j=1}^J$ can be interpreted as a prior estimate. In the analysis step, the ensemble is updated comparing the prior estimate (5.8) with versions of the data perturbed with noise $\{y_{n+1}^{(j)}\}_{j=1}^J$, where $y_{n+1}^{(j)} = y + \eta_{n+1}^{(j)}$, via the standard Kalman update formula

$$z_{n+1}^{(j)} = \widehat{z}_{n+1}^{(j)} + K_{n+1}(y_{n+1}^{(j)} - H\widehat{z}_{n+1}^{(j)}), \tag{5.9}$$

where K_{n+1}, the Kalman gain, is defined as

$$K_{n+1} = C_{n+1}H^*(HC_{n+1}H^* + \Gamma)^{-1}, \tag{5.10}$$

and is employed to weigh the prior guess provided by Ξ and the information carried by the observations. In the definition (5.10), the operator H^* is the adjoint of H and the matrix C_{n+1} is the empirical covariance matrix of the partially-updated ensemble $\{\widehat{z}_{n+1}^{(j)}\}_{j=1}^J$. At the N-th and final step, the EnKF estimator is obtained by averaging over the ensemble of particles

projected on the space X, i.e.,

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

where $H^{\perp}: Z \to X$ is defined by $H^{\perp} = \begin{bmatrix} I & 0 \end{bmatrix}$.

The last detail needed to run the EnKF algorithm is the definition of the initial ensemble, which is closely related to the choice of the space $\mathscr{A}$. In particular, the space $\mathscr{A}$ should incorporate all the prior knowledge on the solution of the inverse problem. Let us assume that the prior knowledge can be summarized as a probability measure on X denoted by μ_0. A possible choice for initialization is then to generate the set $\{\psi^{(j)}\}_{j=1}^{J}$ as J draws from μ_0, and to fix the set $\mathscr{A}$ as

$$\mathscr{A} = \text{span}\,\{\psi^{(j)}\}_{j=1}^{J}.$$

Remark 5.3. The cost of the EnKF method can be measured in terms of the number of evaluations of the forward operator, which is indeed dominating the other operations in terms of computational time. Therefore, the complexity of the algorithm is $\mathscr{O}(JN)$, where J is the dimension of the ensemble and N is the number of iterations. Nonetheless, let us remark that the ensemble Kalman method can be easily parallelized, since at each iteration the forward operator is applied independently to each particle. Hence, if the number of computing threads N_{comp} available is such that $N_{\text{comp}} = \mathscr{O}(J)$, we have that the overall cost is of order $\mathscr{O}(N)$.

In [**?**], the authors show that the EnKF admits a Bayesian interpretation, which we briefly summarise here. Given a prior distribution μ_0 on X, the posterior distribution μ of the unknown given the data is defined as (see e.g. [61])

$$\mu(du) = \frac{1}{Z} e^{-\Phi(u;y)} \mu_0(du),$$

where Z is the normalization constant and $\Phi(u;y)$ is the least squares functional

$$\Phi(u;y) = \frac{1}{2}\|\Gamma^{-1/2}(y - \mathscr{G}(u))\|_2^2.$$

The map from the prior distribution μ_0 to the posterior μ can be divided in N sub-steps through the intermediate measures

$$\mu_n(du) = \frac{1}{Z_n} e^{-n\Delta\Phi(u;y)} \mu_0(du),$$

where $\Delta = 1/N$. Note that $\mu_N = \mu$ is the desired final measure and that

$$\mu_{n+1}(du) = \frac{Z_n}{Z_{n+1}} e^{-\Delta\Phi(u;y)} \mu_n(du).$$

The distribution μ_n can then be approximated by the discrete probability measure induced by

the particles of the EnKF method at the n-th step, i.e.,

$$\mu_n \simeq \frac{1}{J}\sum_{j=1}^{J} \delta_{u_n^{(j)}}. \tag{5.11}$$

The particles at time n are still mapped into the the particles at time $n+1$ with the ensemble Kalman filter update formula described above, with the slight modification that Γ has to be replaced by $\Delta^{-1}\Gamma$ (see [**?**] for the details).

To conclude this section, we introduce an assumption on the algorithm which will be employed in the analysis.

Assumption 5.4. The algorithm is stable, in the sense that all the particles in the ensemble at each iteration lie in the ball $B_R(u^*)$ for some $R > 0$ sufficiently big, where u^* is the true value of the unknown.

5.4 Convergence analysis

In this section we show the convergence of the ensemble of particles generated by the EnKF algorithm which employs the multiscale forward operator $\mathcal{G}^\varepsilon$ in its update formulae to the ensemble which is given by the application of the same algorithm with the FEM solution to the homogenized problem, i.e., by the forward operator $\mathcal{G}_h^0$, in the limit $\varepsilon, h \to 0$. Moreover, from the Bayesian perspective, we show the convergence of the associated posterior distributions, thus providing under further assumptions a rate of convergence. 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. Summarizing, the forward operators involved are $\mathcal{G}^\varepsilon \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}^\varepsilon = \mathcal{O} \circ \mathcal{S}^\varepsilon$ and $\mathcal{G}_h^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}_h^0 = \mathcal{O} \circ \mathcal{S}_h^0$. We also introduce the forward operator $\mathcal{G}^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G}^0 = \mathcal{O} \circ \mathcal{S}^0$, where $\mathcal{S}^0$ is the solution operator associated to (5.5), which maps the unknown u to the exact solution p^0. Convergence is shown in the ensemble norm, which we define for an ensemble of particles $u = \{u^{(j)}\}_{j=1}^J$ as

$$\|u\| := \frac{1}{J}\sum_{j=1}^{J} \|u^{(j)}\|_2, \tag{5.12}$$

and which is indeed a norm. Given a scalar α, let us finally define the linear combination $w = u + \alpha v$ between two ensembles u and v with the same number of particles as $\{w^{(j)} = u^{(j)} + \alpha v^{(j)}\}_{j=1}^J$.

5.4.1 Convergence of the point estimate

We first consider the convergence of the particle ensembles, which can be summarized by the following theorem.

Theorem 5.5. *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 for the forward operators $\mathcal{G}_h^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, under Assumption 5.1, Assumption 5.2 and Assumption 5.4, we have

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

In particular, if the exact solution p^0 of the homogenized problem (5.5) *is in $H^{q+1}(\Omega)$ with $q \geq 1$, $A^0 \in C^q(\Omega; \mathbb{R}^{N\times N})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$, and we employ polynomials of degree r for the finite element basis, then*

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

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

It is clear from the statement of Theorem 5.5 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

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^{J} \|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\|_2, \tag{5.13}$$

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

$$\widetilde{e}(h, u) = \frac{1}{J}\sum_{j=1}^{J} \|\mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)})\|_2. \tag{5.14}$$

Before proving the main theorem, we introduce some preliminary results. In particular, Lemma 5.6 and Lemma 5.7 are linear algebra results. In Lemma 5.8 we prove Lipschitz continuity of forward operators which involve a Lipschitz observation operator and the solution of an elliptic PDE, and in Lemma 5.9 we show that the homogenization error defined in (5.13) vanishes in the limit $\varepsilon \to 0$. Finally, in Lemma 5.10 we consider the particle empirical covariances of ensembles given by the EnKF algorithm, thus proving their boundedness and Lipschitz continuity. The proof of all the results above can be found in the Appendix.

Lemma 5.6. *Let A and B be square invertible matrices, then*

$$\|A^{-1} - B^{-1}\|_2 \leq \|A^{-1}\|_2 \|B^{-1}\|_2 \|A - B\|_2.$$

Lemma 5.7. *Let A and B be square, symmetric matrices, such that A is positive semidefinite and B is positive definite, then*

$$\|(A + B)^{-1}\|_2 \leq \|B^{-1}\|_2.$$

Lemma 5.8. *Let $\mathcal{G}\colon \mathbb{R}^M \to \mathbb{R}^L$, $\mathcal{G} = \mathcal{O} \circ \mathcal{S}$ be a forward operator such that $\mathcal{O}\colon H_0^1(\Omega) \to \mathbb{R}^L$ is Lipschitz and $\mathcal{S}\colon \mathbb{R}^M \to H_0^1(\Omega)$, $\mathcal{S}\colon u \mapsto p$ is defined by the solution of*

$$\begin{cases} -\nabla \cdot (A(u)\nabla p) = f, & \text{in } \Omega, \\ p = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.15}$$

where $\Omega \subset \mathbb{R}^d$ is an open bounded set, the right hand side $f \in L^2(\Omega)$ and the tensor $A(u) \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ satisfies

$$\|A(u_1) - A(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \le M\|u_1 - u_2\|_2, \qquad \text{for all } u_1, u_2 \in \mathbb{R}^M, \tag{5.16}$$

and

$$A(u)\xi \cdot \xi \ge \alpha\|\xi\|_2^2 \qquad \text{for all } \xi \in \mathbb{R}^d, \tag{5.17}$$

where M and α are positive constants. Then $\mathcal{G}$ is Lipschitz.

Lemma 5.9. *Let e be defined as in* (5.13)*. Under Assumption 5.1, 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 (5.5) *is sufficiently smooth independently of u, namely $p^0 \in H^2(\Omega)$, then there exists $K > 0$ independent of ε and u such that*

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

Lemma 5.10. *Let $C^{up}(u) \in \mathbb{R}^{M\times L}$ and $C^{pp}(u) \in \mathbb{R}^{L\times L}$ be defined as*

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

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 L. Then, there exist four constants $C_i > 0$, $i = 1, \dots, 4$, such that

$$\begin{aligned} &\|C^{up}(u)\|_2 \le C_1, \qquad &&\|C^{up}(u_1) - C^{up}(u_2)\|_2 \le C_3\|u_1 - u_2\|, \\ &\|C^{pp}(u)\|_2 \le C_2, \qquad &&\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \le C_4\|u_1 - u_2\|, \end{aligned}$$

for all ensembles $u, u_1, u_2 \in \mathcal{U}_{J,M}$ which are stable in the sense of Assumption 5.4.

In order to clarify the exposition, we first consider the amplification 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 5.11. *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 last iteration of the iterative ensemble Kalman filter for the forward operators $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ respectively. Then, under Assumption 5.1, Assumption 5.2 and Assumption 5.4, there exist positive constants α and γ such that*

$$\mathbb{E}\left[\|u_{n+1}^\varepsilon - u_{n+1}^0\|\right] \leq \alpha \mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right] + \gamma \mathbb{E}\left[e(\varepsilon, u_n^0)\right], \tag{5.18}$$

where $e(\varepsilon, u)$ is given in (5.13).

Proof. First, due to Assumption 5.1 and the Poincaré inequality with constant C_p we have

$$\|\mathcal{O}(p_1) - \mathcal{O}(p_2)\|_2 \leq m\|p_1 - p_2\|_{L^2(\Omega)} \leq mC_p\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)},$$

which shows that $\mathcal{O}$ is Lipschitz with constant mC_p. Therefore, applying Lemma 5.8, we deduce that both $\mathcal{G}^0$ and $\mathcal{G}^\varepsilon$ are Lipschitz with constant $L_\mathcal{G}$ independent of ε. The Kalman update formulae (5.9) restricted to the u variable read (see [**?**])

$$u_{n+1}^{\varepsilon^{(j)}} = u_n^{\varepsilon^{(j)}} + C^{up}(u_n^\varepsilon)(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})), \tag{5.19}$$

$$u_{n+1}^{0^{(j)}} = u_n^{0^{(j)}} + C^{up}(u_n^0)(C^{pp}(u_n^0) + \Gamma)^{-1}(y_{n+1} - \mathcal{G}^0(u_n^{0^{(j)}})). \tag{5.20}$$

Combining (5.19) and (5.20), we have

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

and using the triangle inequality we obtain

$$\mathbb{E}\left[\|u_{n+1}^\varepsilon - u_{n+1}^0\|\right] \leq \mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right] + S_1 + S_2 + S_3, \tag{5.21}$$

where

$$S_1 = \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\|C^{up}(u_n^\varepsilon) - C^{up}(u_n^0)\|_2\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1}\|_2\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2\right], \tag{5.22}$$

$$S_2 = \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\|C^{up}(u_n^0)\|_2\|(C^{pp}(u_n^\varepsilon) + \Gamma)^{-1} - (C^{pp}(u_n^0) + \Gamma)^{-1}\|_2\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2\right], \tag{5.23}$$

$$S_3 = \frac{1}{J}\sum_{j=1}^J \mathbb{E}\left[\|C^{up}(u_n^0)\|_2\|(C^{pp}(u_n^0) + \Gamma)^{-1}\|_2\|\mathcal{G}^0(u_n^{0^{(j)}}) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2\right]. \tag{5.24}$$

Let us first consider S_1. Due to Lemma 5.10, we have

$$\|C^{up}(u_n^\varepsilon) - C^{up}(u_n^0)\|_2 \leq C_3\|u_n^\varepsilon - u_n^0\|,$$

and due to Lemma 5.7

$$\|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1}\|_2 \le \|\Gamma^{-1}\|_2.$$

Moreover, due to the definition of $y_{n+1}^{(j)}$ 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, we obtain the bound

$$\begin{aligned}\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 &= \|y+\xi_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 \\ &\le \|\mathcal{G}^\varepsilon(u^*) - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 + \|\xi_{n+1}^{(j)}+\eta\|_2,\end{aligned}$$

which, since $\mathcal{G}^\varepsilon$ is Lipschitz, yields

$$\begin{aligned}\|y_{n+1}^{(j)} - \mathcal{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 &\le L_{\mathcal{G}}\|u^* - u_n^{\varepsilon^{(j)}}\|_2 + \|\xi_{n+1}^{(j)}+\eta\|_2 \\ &\le L_{\mathcal{G}}R + \|\xi_{n+1}^{(j)}+\eta\|_2.\end{aligned}$$

Thus (5.22) can be bounded by

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

and, since the noise is i.i.d. and independent of the ensembles, we obtain

$$C_3\|\Gamma^{-1}\|_2(L_{\mathcal{G}}R + \mathbb{E}[\|\xi+\eta\|_2])\mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right].$$

Moreover, the random variable $\zeta := \xi+\eta$ is distributed by independence as $\zeta \sim \mathcal{N}(0, 2\Gamma)$, and therefore we get

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

and defining $\alpha_1 := C_3\|\Gamma^{-1}\|_2(L_{\mathcal{G}}R + \sqrt{2\mathrm{tr}(\Gamma)})$, the final bound for S_1 reads

$$S_1 \le \alpha_1\mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right]. \tag{5.25}$$

Let us now consider the second term S_2. Due to Lemma 5.10, we have

$$\|C^{up}(u_n^0)\|_2 \le C_1, \tag{5.26}$$

and an application of Lemma 5.6, Lemma 5.7 and Lemma 5.10 gives

$$\begin{aligned}&\|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1} - (C^{pp}(u_n^0)+\Gamma)^{-1}\|_2 \\ &\le \|(C^{pp}(u_n^\varepsilon)+\Gamma)^{-1}\|_2\|(C^{pp}(u_n^0)+\Gamma)^{-1}\|_2\|C^{pp}(u_n^\varepsilon) - C^{pp}(u_n^0)\|_2 \\ &\le C_4\|\Gamma^{-1}\|_2^2\|u_n^\varepsilon - u_n^0\|.\end{aligned}$$

The third factor appearing in (5.23) is equal to the third factor of (5.22), thus finally S_2 satisfies

$$S_2 \leq \frac{1}{J} C_1 C_4 \|\Gamma^{-1}\|_2^2 \sum_{j=1}^{J} \mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|(L_{\mathscr{G}} R + \|\xi_{n+1}^{(j)} + \eta\|_2)\right],$$

and, as above, defining $\alpha_2 \coloneqq C_1 C_4 \|\Gamma^{-1}\|_2^2 (L_{\mathscr{G}} R + \sqrt{2\mathrm{tr}(\Gamma)})$, we obtain

$$S_2 \leq \alpha_2 \mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right]. \tag{5.27}$$

We now consider the last term S_3. The first factor appearing in (5.24) can be bounded as in (5.26) and for the second part we use Lemma 5.7, thus obtaining

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

Regarding the third factor of (5.24), we apply the triangle inequality and the Lipschitz continuity of the forward operator $\mathscr{G}^\varepsilon$, which yield

$$\begin{aligned}\|\mathscr{G}^0(u_n^{0^{(j)}}) - \mathscr{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 &\leq \|\mathscr{G}^0(u_n^{0^{(j)}}) - \mathscr{G}^\varepsilon(u_n^{0^{(j)}})\|_2 + \|\mathscr{G}^\varepsilon(u_n^{0^{(j)}}) - \mathscr{G}^\varepsilon(u_n^{\varepsilon^{(j)}})\|_2 \\ &\leq \|\mathscr{G}^0(u_n^{0^{(j)}}) - \mathscr{G}^\varepsilon(u_n^{0^{(j)}})\|_2 + L_{\mathscr{G}} \|u_n^{0^{(j)}} - u_n^{\varepsilon^{(j)}}\|_2.\end{aligned}$$

Hence, a bound for S_3 is given by

$$S_3 \leq C_1 \|\Gamma^{-1}\|_2 \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J} \|\mathscr{G}^0(u_n^{0^{(j)}}) - \mathscr{G}^\varepsilon(u_n^{0^{(j)}})\|_2 + L_{\mathscr{G}} \frac{1}{J}\sum_{j=1}^{J} \|u_n^{0^{(j)}} - u_n^{\varepsilon^{(j)}}\|_2\right],$$

which is equivalent to

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

and defining $\alpha_3 = C_1 \|\Gamma^{-1}\|_2 L_{\mathscr{G}}$ and $\gamma = C_1 \|\Gamma^{-1}\|_2$ we have the final bound for S_3, i.e.,

$$S_3 \leq \alpha_3 \mathbb{E}\left[\|u_n^0 - u_n^\varepsilon\|\right] + \gamma \mathbb{E}\left[e(\varepsilon, u_n^0)\right]. \tag{5.28}$$

Therefore, using the results (5.21), (5.25), (5.27) and (5.28), we obtain

$$\mathbb{E}\left[\|u_{n+1}^\varepsilon - u_{n+1}^0\|\right] \leq (1 + \alpha_1 + \alpha_2 + \alpha_3)\mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right] + \gamma \mathbb{E}\left[e(\varepsilon, u_n^0)\right],$$

and defining $\alpha = 1 + \alpha_1 + \alpha_2 + \alpha_3$ we have

$$\mathbb{E}\left[\|u_{n+1}^\varepsilon - u_{n+1}^0\|\right] \leq \alpha \mathbb{E}\left[\|u_n^\varepsilon - u_n^0\|\right] + \gamma \mathbb{E}\left[e(\varepsilon, u_n^0)\right], \tag{5.29}$$

which is the desired result. □

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

Proposition 5.12. *Under the notation and assumptions of Lemma 5.11, letting $u_0^\varepsilon = u_0^0$ be the*

same initial ensemble, we have

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

Moreover, if the solution of the homogenized problem (5.5) *is sufficiently regular, namely* $p^0 \in H^2(\Omega)$*, then there exists* $K_1 > 0$ *independent of* ε *such that*

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

Proof. Iterating Lemma 5.11 and taking into account that $u_0^\varepsilon = u_0^0$, after N iterations, at the end of the EnKF algorithm, we get

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

Applying Lemma 5.9, we have $e(\varepsilon, u_i^0) \to 0$ for all $i = 0, \dots, N-1$, hence as $\varepsilon \to 0$

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

Moreover, if the solution of the homogenized problem p^0 belongs to $H^2(\Omega)$, then, by Lemma 5.9, we have the estimate

$$e(\varepsilon, u_i^0) \le K\varepsilon.$$

Therefore we obtain

$$\mathbb{E}\left[\|u_N^\varepsilon - u_N^0\|\right] \le \gamma \left(\sum_{i=0}^{N-1} \alpha^i \right) K\varepsilon = \gamma \frac{\alpha^N - 1}{\alpha - 1} K\varepsilon,$$

and defining $K_1 = \gamma(\alpha^N - 1)K/(\alpha - 1)$ we have

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

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 5.9 in the context of numerical convergence and whose proof is given in the Appendix.

Lemma 5.13. *Let* $\widetilde{e}$ *be defined in* (5.14)*. Under Assumption 5.1 and if the exact solution* p^0 *of the homogenized problem* (5.15) *is in* $H^{q+1}(\Omega)$*,* $A^0 \in C^q(\Omega; \mathbb{R}^{d\times d})$ *independently of* u*,* $f \in H^{q-1}(\Omega)$*,* $\partial\Omega \in C^{q+1}$*, and we employ polynomials of degree* r *for the finite element basis, then*

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

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 5.14. *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 5.1, Assumption 5.2, Assumption 5.4 and if the exact solution p^0 of the homogenized problem* (5.15) *is in $H^{q+1}(\Omega)$, $A^0 \in C^q(\Omega;\mathbb{R}^{d\times d})$, $f \in H^{q-1}(\Omega)$, $\partial\Omega \in C^{q+1}$ and we use polynomials of degree r for the finite element basis, we have*

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

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

Proof. The proof of Proposition 5.14 is identical to the proof of Proposition 5.12, 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 5.9 for the error e has to be replaced by Lemma 5.13 for the error $\widetilde{e}$. □

Applying Proposition 5.12 and Proposition 5.14, we finally prove Theorem 5.5.

Proof of Theorem 5.5. An application of the triangle inequality yields

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

The two addends can be bounded applying Proposition 5.12 and Proposition 5.14 respectively as

$$\begin{aligned}\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] &\leq K_1\varepsilon + K_2 h^{s+1} \\ &\leq \max\{K_1, K_2\}(\varepsilon + h^{s+1}).\end{aligned}$$

Finally, we define $C = \max\{K_1, K_2\}$ and obtain

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

which is the desired result. □

Remark 5.15. Note that if the exact solution of the homogenized problem in (5.15) $p^0 \in H^2(\Omega)$ and we use polynomials of degree $r = 1$ for the finite element basis, then we have

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \leq C(\varepsilon + h^2).$$

Therefore, in order to balance the two sources of error, the discretization parameter h for the

FEM approximation of the homogenized problem should be chosen as

$$h = O(\varepsilon^{1/2}), \tag{5.30}$$

which guarantees linear convergence with respect to ε. With this choice for h, the computational cost is drastically reduced with respect to the solution of the multiscale problem, for which h^ε needs to be chosen as

$$h^\varepsilon \ll \varepsilon,$$

in order to be able to resolve the oscillations of the multiscale solution.

5.4.2 Convergence of the posterior distributions

We now consider the Bayesian interpretation of the EnKF method. In particular, we consider the multiscale posterior distribution obtained after N steps of the EnKF algorithm, i.e.,

$$\mu^\varepsilon = \frac{1}{J}\sum_{j=1}^{J} \delta_{u_N^{\varepsilon(j)}},$$

and the posterior corresponding to the FEM solution of the homogenized problem, which reads

$$\mu_h^0 = \frac{1}{J}\sum_{j=1}^{J} \delta_{u_{N,h}^{0(j)}},$$

and we study the convergence of μ^ε to μ_h^0 as $\varepsilon, h \to 0$. Due to the discrete nature of the distributions above, we study convergence with respect to the Wasserstein metrics. 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 $s \in [1,\infty]$. We now report the definition of the Wasserstein distances in the metric space $(B_R(u^*), \|\cdot\|_s)$, which can be found, e.g., in [**?**].

Definition 5.16. 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

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{B_R(u^*)\times B_R(u^*)} \|u-v\|_s^p \, d\gamma(u,v)\right)^{1/p}, \tag{5.31}$$

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 5.17. 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 (5.31) can be written as

$$W_{p,s}(\mu,\nu) = \left(\inf_{\gamma\in\mathbb{R}^{K_1\times K_2}} \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \|u_i - v_j\|_s^p \gamma_{ij}\right)^{1/p}, \tag{5.32}$$

where the matrix γ has to satisfy the following constrains

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

Remark 5.18. The Wasserstein distance with $p = 1$ can be written in an equivalent formulation using its duality representation

$$W_{1,s}(\mu,\nu) = \sup_{\varphi\in\Phi} \left\{ \int_{B_R(u^*)} \varphi d(\mu-\nu) \right\}, \tag{5.33}$$

where Φ is the set of all continuous functions $\varphi\colon B_R(u^*) \to \mathbb{R}$ with minimal Lipschitz constant $L \le 1$ with respect to the norm $\|\cdot\|_s$.

In Lemma 5.19 we show that $W_{1,2}$ is bounded by the distance induced by the ensemble norm defined in 5.12. This result will be crucial later to deduce the convergence of the posterior distribution μ_h^0 to μ^ε from Theorem 5.5.

Lemma 5.19. *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 distributions defined as sum of Dirac masses*

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

Then

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

and, in particular,

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

Proof. Take γ^* defined as

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

which satisfies the constraints

$$\sum_{i=1}^J \gamma^*(u_1^{(j)}, u_2^{(i)}) = \mu_1(u_1^{(j)}) = \frac{1}{J} \qquad \text{for all } j = 1,\dots J,$$
$$\sum_{j=1}^J \gamma^*(u_1^{(j)}, u_2^{(i)}) = \mu_2(u_2^{(i)}) = \frac{1}{J} \qquad \text{for all } i = 1,\dots J,$$

and note that

$$\sum_{j=1}^{J}\sum_{i=1}^{J}\|u_1^{(j)}-u_2^{(i)}\|_s^p\gamma^*(u_1^{(j)},u_2^{(i)})=\frac{1}{J}\sum_{j=1}^{J}\|u_1^{(j)}-u_2^{(j)}\|_s^p.$$

Therefore, by definition of Wasserstein distance for discrete distributions on finite spaces (5.32), we deduce that

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

which is the desired result. Finally, taking $p=1$ and $s=2$ and recalling the ensemble norm defined in (5.12), we obtain

$$W_{1,2}(\mu_1,\mu_2)\leq\frac{1}{J}\sum_{j=1}^{J}\|u_1^{(j)}-u_2^{(j)}\|_2=\|u_1-u_2\|,$$

which concludes the proof. □

Let us remark that posterior distributions of the form (5.11) are random probability measures. Indeed, the EnKF algorithm randomizes data at each step, and therefore the resulting ensemble is not deterministic. Therefore, we need to introduce a notion of convergence for random probability measures, and analyse its connection with the Wasserstein distances.

Definition 5.20. Let $(\Omega,\mathscr{A},P)$ be a probability space. A sequence of random probability measures $\{\mu_n\}_{n\in\mathbb{N}}$ dependent on a random variable ξ on $(\Omega,\mathscr{A},P)$ is said to weakly converge in $L^1(\Omega)$ to a random probability measure μ if for all bounded continuous functions $f\in C_B^0$ we have

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

In this case we write

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

In Lemma 5.21, whose proof is given in the Appendix, we show that convergence with respect to the expectation of the Wasserstein distances implies weak L^1 convergence of random probability measures. The fact that convergence with respect to the Wasserstein distances implies weak convergence of distribution was proved in [**?**] for non-random measures, but here we extend the result to random probability measures.

Lemma 5.21. *Let $(\Omega,\mathscr{A},P)$ 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 $(\Omega,\mathscr{A},P)$. If*

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

then

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

Finally, applying Theorem 5.5, we show the convergence of the posterior distribution μ^ε to μ_h^0 as the multiscale and discretization parameters ε, h vanish.

Theorem 5.22. *Let the hypotheses of Theorem 5.5 be satisfied. Define the posterior random probability measures*

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

then as $\varepsilon, h \to 0$

$$\mu^\varepsilon - \mu_h^0 \xrightarrow{L^1} 0.$$

Proof. By Theorem 5.5 we know that the average of the ensemble norm of the difference of u_N^ε and $u_{N,h}^0$ vanishes as ε and h go to zero

$$\mathbb{E}[\|u_N^\varepsilon - u_{N,h}^0\|] \to 0,$$

and applying Lemma 5.19 we deduce that

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

Note that the only difference in the update step is that Γ is replaced by $\Delta^{-1}\Gamma$ where $\Delta = 1/N$. The constants of the proof of Theorem 5.5 depend on $\|\Gamma^{-1}\|_2$, which is now replaced by $\|(\Delta^{-1}\Gamma)^{-1}\|_2$, which can be bounded by $\|\Gamma^{-1}\|_2$ as

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

Finally, by Lemma 5.21, we obtain

$$\mu^\varepsilon - \mu_h^0 \xrightarrow{L^1} 0,$$

which is the desired result. □

5.5 Modelling error

In this section, we consider the effects of model misspecification due to the homogenization and discretization error. All the results presented above deal with the asymptotic case $h, \varepsilon \to 0$, which is unrealistic in applications. Let us recall that the original inverse problem involves predicting the exact unknown u^* from observations originated by the model

$$y = \mathcal{G}^\varepsilon(u^*) + \eta, \tag{5.34}$$

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 (5.34) as

$$y = \mathcal{G}_h^0(u^*) + \mathcal{E}(u^*) + \eta, \tag{5.35}$$

where

$$\mathscr{E}(u^*) := \mathscr{G}^{\varepsilon}(u^*) - \mathscr{G}_h^0(u^*).$$

The quantity $\mathscr{E}(u^*)$ represents the error introduced by misspecification of the forward model. Equation (5.35) 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 modelling error. This formulation of modelling error was originally presented in [14], and then applied to multiscale inverse problems in [2]. Following [14, 2], we assume that the modelling error is a Gaussian random variable independent of the noise η, so that $\mathscr{E} \sim \mathcal{N}(m, \Sigma)$ for all u, and write

$$y = \mathscr{G}_h^0(u^*) + m + \zeta + \eta, \tag{5.36}$$

where $\zeta \sim \mathcal{N}(0, \Sigma)$. Then we define

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

and, from (5.36), we obtain

$$\widetilde{y} = \mathscr{G}_h^0(u^*) + \widetilde{\eta}. \tag{5.37}$$

Therefore, if the mean m and covariance Σ of the modelling error are known, a more reliable approximation of the unknown u^* can be obtained applying the EnKF to (5.37). The modelling error distribution, by assumption fully determined by its mean and covariance, is approximated offline. We sample $N_{\mathscr{E}}$ unknowns $\{u_i\}_{i=1}^{N_{\mathscr{E}}}$ from μ_0 and, for all $i = 1, \dots, N_{\mathscr{E}}$, we apply both the forward operators $\mathscr{G}^{\varepsilon}(u_i)$ and $\mathscr{G}_h^0(u_i)$. Then we compute

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

and the mean m and the covariance Σ are then obtained as the empirical mean and covariance of the sample $\{\mathscr{E}_i\}_{i=1}^{N_{\mathscr{E}}}$. This procedure is computationally involved due to the multiple evaluations of $\mathscr{G}^{\varepsilon}$, but it has to be performed only once and can then be applied to different sets of observations and true values u^*. Moreover, we remarked in practice via numerical experiments that a small number $N_{\mathscr{E}}$ can be employed to obtain a satisfactory approximation of the modelling error. A theoretical insight of this property is provided by Proposition 5.24.

In order to obtain a more reliable approximation of the distribution of the modelling 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 [13]. In particular, we sequentially apply the ensemble Kalman method for $\mathscr{L}$ levels and, at each level, we update the distribution of the modelling error. We denote by $\nu^{\ell} = \mathcal{N}(m^{\ell}, \Sigma^{\ell})$ for any $\ell = 1, \dots, \mathscr{L}$ the approximated distribution at level ℓ. In particular, let

$$\mu_n^{\ell} = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n^{\ell(j)}}$$

be the approximation of the distribution of the particles at iteration n at level ℓ, $\mu_0^{\ell+1} = \mu_{N^{\ell}}^{\ell}$

and $\mu_0^1 = \mu_0$, where N^ℓ is the number of iterations at level ℓ. At the beginning of each level ℓ, we approximate the distribution ν^ℓ by sampling $N_{\mathscr{E}}^\ell$ particles $\{u_i^\ell\}_{i=1}^{N_{\mathscr{E}}^\ell}$ from the distribution μ_0^ℓ, thus computing the mean m^ℓ and the covariance Σ^ℓ as the empirical mean and covariance of the sample. This approach provides indeed a better approximation of the modelling error as instead of taking the samples from the prior distribution, they are drawn from distributions which are progressively closer to the true posterior. On the other hand, this procedure has to be done online and it is computationally expensive because it requires the resolution of $N_{\mathscr{E}} = \sum_{\ell=1}^{\mathscr{L}} N_{\mathscr{E}}^\ell$ full multiscale problems.

Remark 5.23. Let us remark that on the one hand, due to the theory of homogenization, the modelling error can be considered negligible when ε is very small, and the expensive estimation of $\mathscr{E}$ 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 $\mathscr{E}$ is required. One may rightfully argue that in case $\varepsilon = \mathscr{O}(1)$, it is possible to evaluate the forward operator $\mathscr{G}^\varepsilon$ without a large computational effort. Hence, the techniques presented in this section are relevant for mid-range values of ε, for which $\mathscr{E}$ is significant with respect to η.

Finally, we are interested in studying whether the simple offline method for estimating the modelling error provides indeed a good approximation, at least in the mean sense. In this direction, we give in Proposition 5.24 a criterion on how to choose the number $N_{\mathscr{E}}$ of full multiscale problems which has to be solved in order to have a reliable approximation of the true mean m^* of the modelling error with respect to ε and h. Before stating Proposition 5.24, let us recall the Hoeffding's inequality, which will be used in the proof. Let $\{X_i\}_{i=1}^N$ be independent random variables with values in $[a,b]$, and let $\bar{X}$ be the sample average of $\{X_i\}_{i=1}^N$. Then, for all $t \in \mathbb{R}$ it holds

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

Proposition 5.24. *Let $\alpha \in (0,1)$, $t > 0$ and $C_{\mathscr{E}} = \max\{K, \widetilde{K}\}$, where K and $\widetilde{K}$ are the constants of Lemma 5.9 and Lemma 5.13. Let $\{\mathscr{E}_i\}_{i=1}^{N_{\mathscr{E}}} \subset \mathbb{R}^L$ be given by*

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

for a sample of realizations $\{u_i\}_{i=1}^{N_{\mathscr{E}}}$ from the standard normal distribution $\mathscr{N}(0, I)$, let m be the sample mean of $\{\mathscr{E}_i\}_{i=1}^{N_{\mathscr{E}}}$ and $m^* = \mathbb{E}[\mathscr{E}_i]$. If

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

where s is given by Lemma 5.13, then

$$\mathbb{P}\left(\|m - m^*\|_2 \leq t\right) \geq 1 - \alpha.$$

Proof. First, note that the modelling error is bounded, indeed by Lemma 5.9 and Lemma 5.13, we have for each $i = 1, \dots, N_{\mathscr{E}}$

$$\|\mathscr{E}_i\|_2 = \|\mathscr{G}^{\varepsilon}(u_i) - \mathscr{G}_h^0(u_i)\|_2 \leq \|\mathscr{G}^{\varepsilon}(u_i) - \mathscr{G}^0(u_i)\|_2 + \|\mathscr{G}^0(u_i) - \mathscr{G}_h^0(u_i)\|_2 \leq K\varepsilon + \widetilde{K}h^{s+1},$$

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

$$|(\mathscr{E}_i)_l| \leq \|\mathscr{E}_i\|_2 \leq K\varepsilon + \widetilde{K}h^{s+1} \leq C_{\mathscr{E}}(\varepsilon + h^{s+1}). \tag{5.38}$$

Observe that if

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

then

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

which implies that

$$\mathbb{P}(\|m - m^*\|_2 \leq t) \geq \mathbb{P}\left(|m_l - m_l^*| \leq \frac{t}{\sqrt{L}} \quad \forall\, l = 1, \dots, L\right). \tag{5.39}$$

Using (5.38), applying Hoeffding's inequality and Young's inequality we have

$$\begin{aligned} \mathbb{P}\left(|m_l - m_l^*| \geq \frac{t}{\sqrt{L}}\right) &\leq 2\exp\left\{-\frac{2t^2 N_{\mathscr{E}}^2}{4LN_{\mathscr{E}}C_{\mathscr{E}}^2(\varepsilon + h^{s+1})^2}\right\} \\ &\leq 2\exp\left\{-\frac{t^2 N_{\mathscr{E}}}{4LC_{\mathscr{E}}^2(\varepsilon^2 + h^{2(s+1)})}\right\}. \end{aligned} \tag{5.40}$$

Define the events $A_l = \left\{|m_l - m_l^*| \leq \frac{t}{\sqrt{L}}\right\}$ for each $l = 1, \dots, L$, then we have

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

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

$$\mathbb{P}\left(\bigcap_{l=1}^{L} A_l\right) = 1 - \mathbb{P}\left(\left(\bigcap_{l=1}^{L} A_l\right)^C\right) = 1 - \mathbb{P}\left(\bigcup_{l=1}^{L} A_l^C\right) \geq 1 - \sum_{l=1}^{L} \mathbb{P}(A_l^C). \tag{5.41}$$

Therefore, thanks to (5.39), (5.40) and (5.41), we have

$$\begin{aligned} \mathbb{P}(\|m - m^*\|_2 \leq t) &\geq 1 - L\mathbb{P}\left(|m_l - m_l^*| \geq \frac{t}{\sqrt{L}}\right) \\ &\geq 1 - 2L\exp\left\{-\frac{t^2 N_{\mathscr{E}}}{4LC_{\mathscr{E}}^2(\varepsilon^2 + h^{2(s+1)})}\right\}, \end{aligned}$$

and, if $N_{\mathscr{E}}$ satisfies the hypothesis, we obtain

$$\mathbb{P}(\|m - m^*\|_2 \leq t) \geq 1 - 2L\exp\left\{-\log\left(\frac{2L}{\alpha}\right)\right\} = 1 - \alpha,$$

which is the desired result. □

Remark 5.25. Note that, in Proposition 5.24, as expected, the number $N_{\mathscr{E}}$ of full multiscale problems tends to infinity if we require no error between the sample and the true mean ($t \to 0$) or certainty that the error is below a certain value ($\alpha \to 0$). Moreover, note that for any given accuracy the number of samples required $N_{\mathscr{E}}$ is a increasing function of ε and h, so that if the model $\mathscr{G}_h^0$ is a good approximation of $\mathscr{G}$, only a few samples are needed.

5.6 Numerical experiments

In this section, using the setting of [2], we present some numerical experiments to illustrate the iterative ensemble Kalman method to solve multiscale inverse problems.
Let Ω be a bounded open domain. We consider a class of parametrized multiscale locally periodic tensors of the type $A^{\varepsilon}_{\sigma^*}(x) = A(\sigma^*(x), x/\varepsilon)$, where $\sigma^* : \Omega \to \mathbb{R}$. We assume to know the map $(t,x) \to A(t,x/\varepsilon)$ for all $x \in \Omega$ and $t \in \mathbb{R}$ and we want to estimate the function σ^* given measurements computed from the model

$$\begin{cases} -\nabla \cdot (A^{\varepsilon}_{\sigma^*} \nabla p^{\varepsilon}) = 0 & \text{in } \Omega, \\ p^{\varepsilon} = g & \text{on } \partial\Omega. \end{cases} \tag{5.42}$$

Remark 5.26. Note that the theory has been developed for Dirichlet homogeneous boundary conditions, but it can be applied to the non-homogeneous case by considering an extension of the function at the boundary and slightly modifying the PDE. For more details we refer to [**?**, Remark 8.10].

For the unknown σ^* we consider the following admissible set

$$\Sigma = \{\sigma \in L^{\infty}(\Omega) : \sigma^- \leq \sigma(x) \leq \sigma^+\},$$

where σ^- and σ^+ are two given values.
The measurements, which we take into account, are the integrals of the normal flux multiplied by some functions with compact support in a portion of the boundary of the domain. More precisely, we consider $I \in \mathbb{N}$ disjoint portions of Ω, which we denote by $\Gamma_i \in \partial\Omega$, $i = 1, \dots, I$, $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$, and I functions $\varphi_i \in H^{1/2}(\partial\Omega)$ with compact support supp $(\varphi_i) \subset \Gamma_i$ for all $i = 1, \dots, I$. Moreover, we solve (5.42) for $K \in \mathbb{N}$ Dirichlet data, denoted by g_k with $k = 1, \dots, K$. Then we define the multiscale operator $\mathscr{F}^{\varepsilon} : \Sigma \to \mathbb{R}^L$ where $L = IK$ by components

$$\mathscr{F}^{\varepsilon}(\sigma)_{ik} = \mathscr{F}^{\varepsilon}(\sigma)_l = \int_{\Gamma_i} A^{\varepsilon} \nabla p^{\varepsilon}_k \cdot \nu \varphi_i \, ds, \qquad i = 1, \dots, I, k = 1, \dots, K. \tag{5.43}$$

where p_k^ε is the solution of problem (5.42) with Dirichlet boundary condition g_k and ν is the exterior unit normal vector to $\partial\Omega$. The final vector of observations y is given by the sum of the operator $\mathscr{F}^\varepsilon$ and a noise

$$y = \mathscr{F}^\varepsilon(\sigma^*) + \eta,$$

where $\eta \sim \mathcal{N}(0,\Gamma)$ and Γ is a given covariance matrix, which, in our experiments, is a multiple of the identity $\Gamma = \gamma^2 I$ and γ is a given value. Observations are computed with a refined Finite Element Method (FEM) with mesh size $h_{\text{obs}} \ll \varepsilon$, while the homogenized version of problem (5.42) is solved using a macro mesh size $h \gg h_{\text{obs}}$. We call $\mathscr{T}_h$ the macro triangulation and N_h the total number of nodes defining $\mathscr{T}_h$. We assume that the prior distribution for the discretization of the unknown σ^* on the macro triangulation $\mathscr{T}_h$ is given by $\mathcal{N}(\sigma_0, C)$, where σ_0 is a given discretization of a function in Σ and $C \in \mathbb{R}^{N_h \times N_h}$ is defined by

$$C_{ij} = \delta \exp\left(-\frac{\|x_i - x_j\|_2}{\lambda}\right)$$

where $\delta, \lambda \in \mathbb{R}^+$ and $\{x_i\}_{i=1}^{N_h}$ are the nodes of the macro triangulation $\mathscr{T}_h$. 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.

In order to reduce the dimensionality of the unknown we use a truncated Karhunen-Loève expansion. Any sample from the prior distribution $\mathcal{N}(\sigma_0, C)$ can be represented as

$$\sigma = \sigma_0 + \sum_{m=1}^{N_h} \sqrt{\lambda_m}\, u_m \varphi_m, \tag{5.44}$$

where $\{\varphi_m\}_{m=1}^{N_h}$ is an orthonormal set of eigenvectors of C with corresponding eigenvalues $\{\lambda_m\}_{m=1}^{N_h}$ in decreasing order, and $\{u_m\}_{m=1}^{N_h}$ is an i.i.d sequence with $u_m \sim \mathcal{N}(0,1)$. Note that the Karhunen-Loève expansion works also in the infinite dimensional setting, where $\sigma_0 \in \Sigma$, C is a covariance operator and $\{\lambda_m, \varphi_m\}_{m=1}^{\infty}$ is an orthonormal set of eigenvalues-eigenfunctions with respect to the scalar product in $L^2(\Omega)$. Then the truncated Karhunen-Loève expansion of the discretization of σ consists of taking the first M components of the series in (5.44)

$$\sigma \simeq \sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m}\, u_m \varphi_m, \tag{5.45}$$

and the actual unknown becomes the vector $u \in \mathbb{R}^M$, whose components are the coefficients u_m in (5.45). Then we define the multiscale forward operator $\mathscr{G}^\varepsilon : \mathbb{R}^M \to \mathbb{R}^L$ as the composition of $\mathscr{F}^\varepsilon$ with the truncated Karhunen-Loève expansion

$$\mathscr{G}^\varepsilon(u) = \mathscr{F}^\varepsilon\left(\sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m}\, u_m \varphi_m\right).$$

On the other hand, in the iterative ensemble Kalman method we do not compute the exact solution of problem (5.42), but we solve its homogenized version numerically using the

macro triangulation $\mathcal{T}_h$, therefore we obtain the homogenized discrete solution p_h^0. The problem is solved applying the Finite Element Heterogeneous Multiscale Method (FE-HMM), which is described in [**?**]. Hence, analogously to the multiscale case, we define the discrete homogenized operator $\mathcal{F}_h^0 \colon \Sigma \to \mathbb{R}^L$ as

$$\mathcal{F}_h^0(\sigma)_l = \mathcal{F}_h^0(\sigma)_{ik} = \int_{\Gamma_i} A^0 \nabla p_{h_k}^0 \cdot \nu \varphi_i \, ds, \qquad i = 1, \dots, I, k = 1, \dots, K, \tag{5.46}$$

and the discrete homogenized forward operator $\mathcal{G}_h^0 \colon \mathbb{R}^M \to \mathbb{R}^L$, which is actually used in the algorithm, as

$$\mathcal{G}_h^0(u) = \mathcal{F}_h^0\left(\sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_m \varphi_m\right).$$

Finally, we call u_{EnKF} the solution of the iterative ensemble Kalman algorithm and the estimated σ_{EnKF} is obtained from the truncated Karhunen-Loève expansion

$$\sigma_{\text{EnKF}} = \sigma_0 + \sum_{m=1}^{M} \sqrt{\lambda_m} u_{\text{EnKF}_m} \varphi_m.$$

5.6.1 Data

In the numerical results presented in the following section the computational domain is the unit square

$$\Omega = (0,1)^2 \subset \mathbb{R}^2.$$

For the discretization parameters we set $\varepsilon = 1/64$ and $h_{\text{obs}} = 1/4096$ and for the forward homogenized problem we use a macro mesh size $h = 1/32$, which is much larger than h_{obs} and reduces the computational cost significantly.
We solve the problem for $K = 3$ Dirichlet conditions $\{g_k\}_{k=1}^3$ and $g_k = \sqrt{\mu_k}\psi_k$ where $\{(\mu_k, \psi_k)\}_{k=1}^3$ are couples of eigenvalues and eigenfunctions of the one dimensional discrete Laplacian operator corresponding to the first $K = 3$ smallest eigenvalues. For each g_k we consider its restriction to the boundary $\partial\Omega$ in order to obtain a Dirichlet condition. These functions are orthonormal with respect to the scalar product in $L^2(\Omega)$ and this ensures that each function gives independent information.
To compute the boundary integrals in (5.43) and (5.46), we consider $I = 12$ boundary portions, three for each side of the square Ω. In particular, for each side, all Γ_i have length equal to 0.2 and they consist of the intervals $(0.1, 0.3)$, $(0.4, 0.6)$ and $(0.7, 0.9)$. The functions $\{\varphi_i\}_{i=1}^{12}$ are hat functions with $\text{supp}(\varphi_i) = \Gamma_i$, which take value one at the midpoint and value 0 at the extremes of Γ_i. Then the parameter of the noise, which perturbs the observations, is $\gamma = 0.01$. Moreover, regarding the prior distribution for the unknown, we consider $\sigma_0 = 0$ and the parameters of the covariance matrices are $\delta = 0.05$ and $\lambda = 0.5$. In the truncated Karhunen-Loève expansion we take $M = 100$.
Finally, about the ensemble Kalman method, we consider $J = 1000$ particles for each ensemble and 500 iterations.

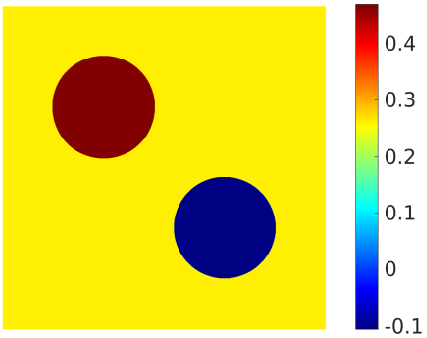

Figure 5.1 – Plot of the exact unknown σ^*

.

The exact tensor $A^{\varepsilon}_{\sigma^*}$ is given by

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

where

$$\sigma^*(x) = \log(1.3 + 0.3\mathbb{1}_{D_1} - 0.4\mathbb{1}_{D_2}),$$

and

$$
\begin{aligned}
D_1 &= \left\{x=(x_1,x_2):\ \left(x_1-\frac{5}{16}\right)^2+\left(x_2-\frac{11}{16}\right)^2 \le 0.025\right\},\\
D_2 &= \left\{x=(x_1,x_2):\ \left(x_1-\frac{11}{16}\right)^2+\left(x_2-\frac{5}{16}\right)^2 \le 0.025\right\}.
\end{aligned}
$$

Figure 5.1 shows the exact unknown σ^*. Note that σ^* is a non-continuous function, but, in order to approximate it, we are using a truncated Karhunen-Loève expansion, where the eigenfunctions are smooth.

One can verify that the tensor A^{ε}_{σ} satisfies Assumption 5.2. In particular, for $\xi \in \mathbb{R}^2$ we have

$$A^{\varepsilon}_{\sigma}\xi\cdot\xi = a_{1,1}\left(\sigma(x), \frac{x}{\varepsilon}\right)\xi_1^2 + a_{2,2}\left(\sigma(x), \frac{x}{\varepsilon}\right)\xi_2^2 \ge e^{\sigma(x)}(\xi_1^2+\xi_2^2) = e^{\sigma^-}\|\xi\|_2^2.$$

Moreover, since the EnKF algorithm estimates the coefficients u of the truncated Karhunen-Loève expansion, we can show that for all $u_1, u_2 \in \mathbb{R}^M$

$$\|A^{\varepsilon}(u_1) - A^{\varepsilon}(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} \le M\|u_1-u_2\|_2,$$

where $A^\varepsilon(u) = A^\varepsilon_{\sigma_u}$ and $\|\cdot\|_2$ stands for the 2-norm of a vector in $\mathbb{R}^M$. We have

$$\begin{aligned}\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &= \sup_{x\in\Omega}\sup_{\xi\in\mathbb{R}^2,\|\xi\|_2=1} \|(A^\varepsilon(u_1) - A^\varepsilon(u_2))\xi\|_2 \\ &\leq \sup_{x\in\Omega}\sqrt{\left(a_{1,1}\left(\sigma_{u_1}(x),\frac{x}{\varepsilon}\right) - a_{1,1}\left(\sigma_{u_2}(x),\frac{x}{\varepsilon}\right)\right)^2 + \left(a_{2,2}\left(\sigma_{u_1}(x),\frac{x}{\varepsilon}\right) - a_{2,2}\left(\sigma_{u_2}(x),\frac{x}{\varepsilon}\right)\right)^2},\end{aligned}$$

which implies

$$\begin{aligned}\|A^\varepsilon(u_1) - A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &\leq \sup_{x\in\Omega}\sqrt{13\left(e^{\sigma_{u_1}(x)} - e^{\sigma_{u_2}(x)}\right)^2} \\ &\leq \sup_{x\in\Omega}\sqrt{13}e^{\sigma^+}|\sigma_{u_1}(x) - \sigma_{u_2}(x)|.\end{aligned}$$

Using the truncated Karhunen-Loève expansion we obtain

$$\begin{aligned}\|A^\varepsilon(u_2)\|_{L^\infty(\Omega,\mathbb{R}^{d\times d})} &\leq \sup_{x\in\Omega}\sqrt{13}e^{\sigma^+}\left|\sum_{m=1}^M \sqrt{\lambda_m}\varphi_m(x)(u_{1,m} - u_{2,m})\right| \\ &\leq \sup_{x\in\Omega}\sqrt{13}e^{\sigma^+}\left(\sum_{m=1}^M \lambda_m\varphi_m(x)^2\right)^{1/2}\left(\sum_{m=1}^M (u_{1,m} - u_{2,m})^2\right)^{1/2} \\ &= \sup_{x\in\Omega}\sqrt{13}e^{\sigma^+}\left(\sum_{m=1}^M \lambda_m\varphi_m(x)^2\right)^{1/2}\|u_1 - u_2\|_2,\end{aligned}$$

and defining $M = \sup_{x\in\Omega}\sqrt{13}e^{\sigma^+}\left(\sum_{m=1}^M \lambda_m\varphi_m(x)^2\right)^{1/2}$ we get the result.

5.6.2 Results

In Figure 5.2 we plot the estimation σ_{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, indeed we do not note a significant difference between the last two plots. We point out that we obtain a quite good approximation of the real unknown σ^* , indeed we are trying to recover a non-continuous function in the whole domain given only some observations at the boundary.

We perform a sensitivity analysis with respect to the dimension of the ensemble and the multiscale parameter ε. In Figure 5.3 we vary the number of particles J and we compare the results obtained at the end of the algorithm after 500 iterations. As expected, the approximation becomes better when the ensemble contains more particles. In particular, note that if the number of particles is too small, e.g. $J = 10$, then the approximation is completely different from the true unknown.

In Figure 5.4 we compare the results obtained for different values of the multiscale parameter ε, in particular we take $\varepsilon = 1/4, 1/8, 1/32, 1/64$. We notice that the approximation becomes worse when ε is bigger, indeed the homogenized problem becomes too different with respect

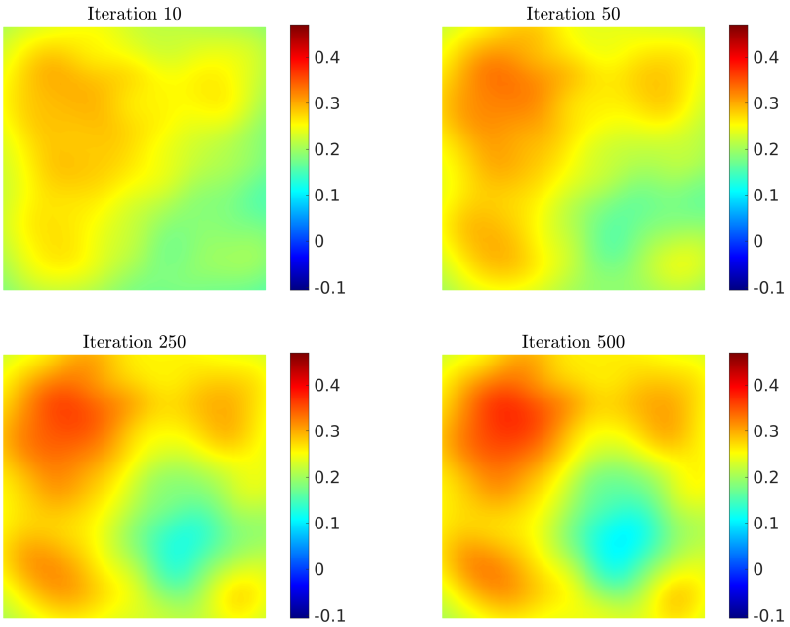

Figure 5.2 – Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 10, 50, 250 and 500 iterations.

to the multiscale one and, if ε is too big, the solution does not approximate the true unknown.

Moreover, in order to obtain good results even in case ε is not close to the asymptotic limit $\varepsilon \to 0$, in Figure 5.5 we apply offline modelling error estimation with $N_{\mathscr{E}} = 20$ and we plot the solution of the inverse problem (5.37) for different values of the multiscale parameter ε. Comparing these plots with the ones in Figure 5.4, in particular for $\varepsilon = 1/4$, we observe that the modelling error estimation significantly improves the results.

Finally, in Figure 5.6 we show the results obtained by applying the ensemble Kalman method with dynamic updating of the modelling error distribution with $\mathscr{L} = 5$ levels, $N_{\mathscr{E}}^{\ell} = 4$ samples and $N^{\ell} = 100$ iterations at each level $\ell = 1, \dots, \mathscr{L}$. The number of resolutions of the full multiscale problem is 20 and the total number of iterations is 500, which are equal to the previous approach, where the distribution of the modelling error was approximated offline. Comparing these plots with the ones in Figure 5.5, we note that updating the distribution of the modelling error dynamically still improves the results.

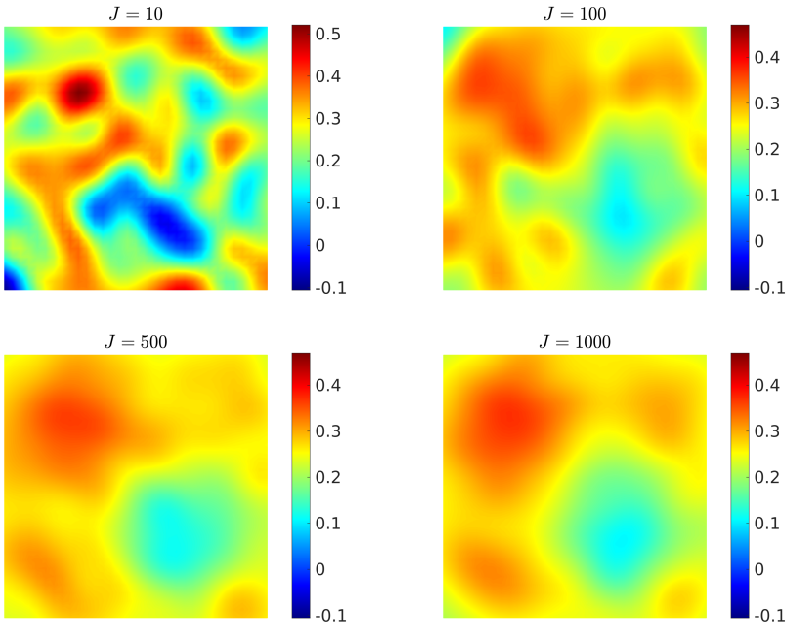

Figure 5.3 – Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different numbers of particles per ensemble $J = 10, 100, 500, 1000$.

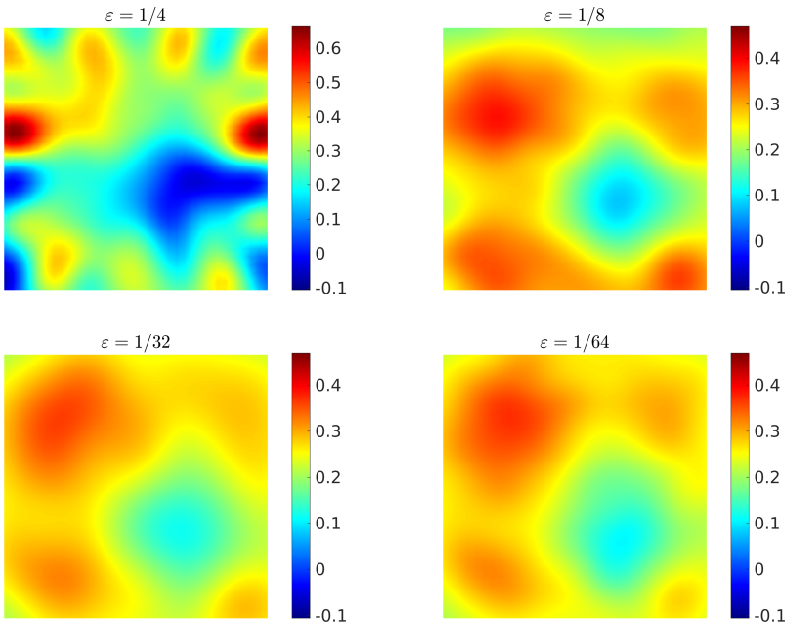

Figure 5.4 – Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

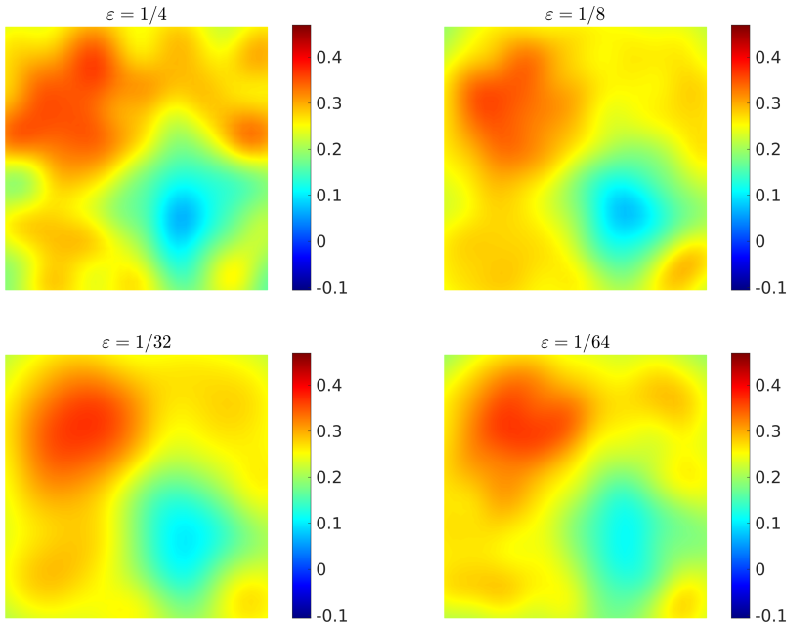

Figure 5.5 – Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

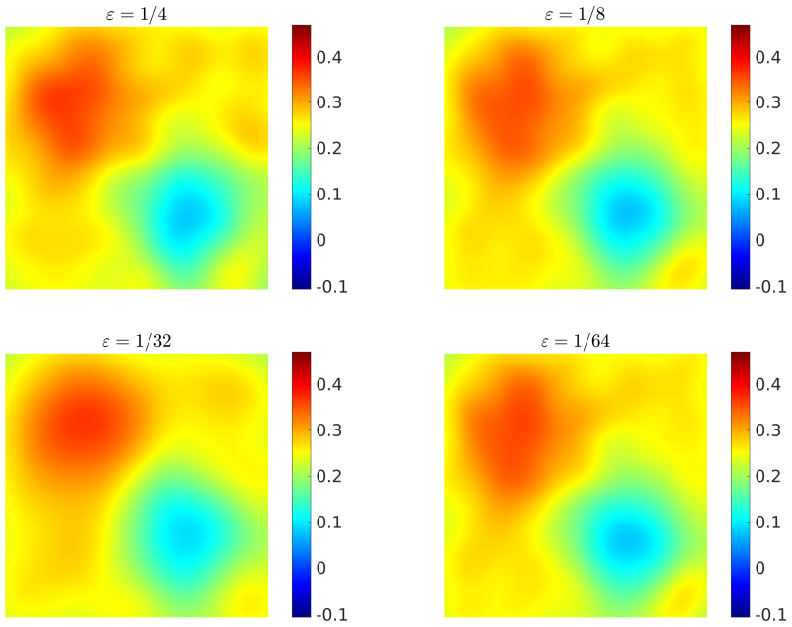

Figure 5.6 – Plot of the unknown σ_{EnKF} estimated by the iterative ensemble Kalman method with dynamic updating of the model error estimation after 500 iterations for different values of the multiscale parameter $\varepsilon = 1/4, 1/8, 1/32, 1/64$.

Appendix

Proof of Lemma 5.6

Note that

$$A^{-1} - B^{-1} = A^{-1}(I - AB^{-1}) = A^{-1}(B - A)B^{-1},$$

therefore we have

$$\|A^{-1} - B^{-1}\|_2 \leq \|A^{-1}\|_2 \|B - A\|_2 \|B^{-1}\|_2,$$

which is the desired result. □

Proof of Lemma 5.7

Let n be the dimension of the matrices, since A is symmetric positive semidefinite and B is symmetric positive definite, then $A + B$ is symmetric positive definite, and the eigenvalues of $A + B$ and B are real and positive, thus they can be written

$$0 < \lambda_1(\cdot) \leq \lambda_2(\cdot) \leq \cdots \leq \lambda_n(\cdot),$$

counted with their multiplicity. First, notice that, using the Rayleigh quotient and the fact that $x^T Ax \geq 0$ for all x, we have

$$\lambda_1(A + B) = \min_{x \neq 0} \frac{x^T(A + B)x}{x^T x} = \min_{x \neq 0} \frac{x^T Ax + x^T Bx}{x^T x} \geq \min_{x \neq 0} \frac{x^T Bx}{x^T x} = \lambda_1(B),$$

which implies

$$\|(A + B)^{-1}\|_2 = \frac{1}{\lambda_1(A + B)} \leq \frac{1}{\lambda_1(B)} = \|B^{-1}\|_2,$$

which is the desired result. □

Proof of Lemma 5.8

Let $u_1, u_2 \in \mathbb{R}^M$, and $p_1 = \mathscr{S}(u_1)$, $p_2 = \mathscr{S}(u_2)$ Then, writing the weak formulations of (5.15) we get

$$\int_\Omega \big(A(u_1)\nabla p_1 - A(u_2)\nabla p_2\big) \cdot \nabla v = 0,$$

for all $v \in H_0^1(\Omega)$. Adding and subtracting $A(u_1)\nabla p_2$ to the first factor inside the integral and rearranging terms yields

$$\int_\Omega A(u_1)(\nabla p_1 - \nabla p_2) \cdot \nabla v = -\int_\Omega (A(u_1) - A(u_2))\nabla p_2 \cdot \nabla v.$$

Choosing $v = p_1 - p_2$, using the hypotheses on A and the Hölder inequality, we obtain

$$\begin{aligned}\alpha\|\nabla p_1 - \nabla p_2\|^2_{L^2(\Omega;\mathbb{R}^d)} &\leq \int_\Omega \left|(A(u_1) - A(u_2))\nabla p_2 \cdot (\nabla p_1 - \nabla p_2)\right| \\ &\leq \|A(u_1) - A(u_2)\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})}\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)}\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)} \\ &\leq M\|u_1 - u_2\|_2\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)}\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)},\end{aligned}$$

which implies

$$\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)} \leq \frac{M}{\alpha}\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^d)}\|u_1 - u_2\|_2. \tag{5.47}$$

It remains to bound $\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)}$, which can be achieved with a standard coercivity argument. In particular, we have

$$\|\nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \leq \frac{C_p}{\alpha}\|f\|_{L^2(\Omega)}, \tag{5.48}$$

where C_p is the Poincaré constant associated to the domain Ω. Replacing in (5.47), we obtain

$$\|\nabla p_1 - \nabla p_2\|_{L^2(\Omega;\mathbb{R}^N)} \leq \frac{MC_p}{\alpha^2}\|f\|_{L^2(\Omega)}\|u_1 - u_2\|_2 = L_{\mathscr{S}}\|u_1 - u_2\|_2,$$

which shows that $\mathscr{S}$ is Lipschitz with constant

$$L_{\mathscr{S}} = \frac{MC_p}{\alpha^2}\|f\|_{L^2(\Omega)}.$$

Finally, $\mathscr{G}$ is the composition of two Lipschitz operators, so it is Lipschitz of constant $L_{\mathscr{G}} = L_{\mathscr{O}}L_{\mathscr{S}}$, which concludes the proof. □

Proof of Lemma 5.9

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

$$\begin{aligned}\|\mathscr{G}^\varepsilon(u^{(j)}) - \mathscr{G}^0(u^{(j)})\|_2 &= \|\mathscr{O}(\mathscr{S}^\varepsilon(u^{(j)})) - \mathscr{O}(\mathscr{S}^0(u^{(j)}))\|_2 &(5.49)\\ &= \|\mathscr{O}(p^\varepsilon(u^{(j)})) - \mathscr{O}(p^0(u^{(j)}))\|_2 &(5.50)\\ &\leq m\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\|_{L^2(\Omega)}, &(5.51)\end{aligned}$$

where we write explicitly the dependence of p^ε and p^0 on the particle it is generated by. By homogenization theory, we know that $p^\varepsilon(u^{(j)}) \rightharpoonup p^0(u^{(j)})$ in $H^1_0(\Omega)$, and therefore we have $p^\varepsilon(u^{(j)}) \to p^0(u^{(j)})$ in $L^2(\Omega)$, which implies

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^J \|\mathscr{G}^\varepsilon(u^{(j)}) - \mathscr{G}^0(u^{(j)})\|_2 \to 0.$$

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

estimate, which can be found in [**?**]

$$\|p^\varepsilon(u^{(j)}) - p^0(u^{(j)})\|_{L^2(\Omega)} \leq C\varepsilon,$$

hence we finally obtain

$$e(\varepsilon, u) = \frac{1}{J}\sum_{j=1}^{J} \|\mathcal{G}^\varepsilon(u^{(j)}) - \mathcal{G}^0(u^{(j)})\|_2 \leq mC\varepsilon,$$

which proves the result for $K = mC$. □

Proof of Lemma 5.10

Let L be the Lipschitz constant of $\mathcal{G}$. For all $x \in B_R(u^*)$ we have

$$\begin{aligned}
&\|\mathcal{G}(x)\|_2 \leq \|\mathcal{G}(x) - \mathcal{G}(u^*)\|_2 + \|\mathcal{G}(u^*)\|_2 \leq L\|x - u^*\|_2 + \|\mathcal{G}(u^*)\|_2 \leq LR + G,\\
&\|x\|_2 \leq \|x - u^*\|_2 + \|u^*\|_2 \leq R + g,
\end{aligned}$$

and we define the bounds $M = LR + G$ and $m = R + g$. The same bounds can be deduced for the mean values

$$\|\bar{u}\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \|u^{(j)}\|_2 \leq \frac{1}{J} Jm = m, \tag{5.52}$$

$$\|\bar{\mathcal{G}}\|_2 \leq \frac{1}{J}\sum_{j=1}^{J} \|\mathcal{G}(u^{(j)})\|_2 \leq \frac{1}{J} JM = M. \tag{5.53}$$

By definition of 2-norm of a matrix we have

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

and using (5.52) and (5.53) and the fact that $\|x\|_2 = 1$ we obtain

$$\begin{aligned}
\|C^{up}(u)\|_2 &\leq \frac{1}{J}\sum_{j=1}^{J} \left(\|\mathcal{G}(u^{(j)})\|_2 + \|\bar{\mathcal{G}}\|_2 \right) \left(\|u^{(j)}\|_2 + \|\bar{u}\|_2 \right)\\
&\leq \frac{1}{J} J(M + M)(m + m) = 4Mm,
\end{aligned}$$

and we define $C_1 = 4Mm$. The procedure is similar for the matrix $C^{pp}(u)$, where we have

$$\begin{aligned}\|C^{pp}(u)\|_2 &= \sup_{x\in\mathbb{R}^L:\,\|x\|_2=1}\left\|\frac{1}{J}\sum_{j=1}^{J}(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}})(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}})^T x\right\|_2\\ &\le \sup_{x\in\mathbb{R}^L:\,\|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left|(\mathcal{G}(u^{(j)})-\bar{\mathcal{G}})^T x\right|\left\|\mathcal{G}(u^{(j)})-\bar{\mathcal{G}}\right\|_2\\ &\le \sup_{x\in\mathbb{R}^L:\,\|x\|_2=1}\frac{1}{J}\sum_{j=1}^{J}\left\|\mathcal{G}(u^{(j)})-\bar{\mathcal{G}}\right\|_2^2\|x\|_2,\end{aligned}$$

and using bound (5.53) and the fact that $\|x\|_2 = 1$ we obtain

$$\begin{aligned}\|C^{up}(u)\|_2 &\le \frac{1}{J}\sum_{j=1}^{J}\left(\|\mathcal{G}(u^{(j)})\|_2+\|\bar{\mathcal{G}}\|_2\right)^2\\ &\le \frac{1}{J}J(M+M)^2 = 4M^2,\end{aligned}$$

and we define $C_2 = 4M^2$.
Before proving the last two results of the lemma we need the following estimates for the ensemble of particles u_1 and u_2

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

By definition of 2 norm of a matrix and using the triangle inequality we have

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

which implies

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

Using bounds (5.52) and (5.53) and the fact that $\mathcal{G}$ is Lipschitz with constant L, we obtain

$$\begin{aligned}
&\|C^{up}(u_1) - C^{up}(u_2)\|_2 \\
&\quad \leq \frac{1}{J}\sum_{j=1}^{J} \left\{ (\|u_1^{(j)}\|_2 + \|\bar{u}_1\|_2)(L\|u_1^{(j)} - u_2^{(j)}\|_2 + L\|u_1 - u_2\|) \right\} \\
&\qquad + \frac{1}{J}\sum_{j=1}^{J} \left\{ (\|u_1^{(j)} - u_2^{(j)}\|_2 + \|\bar{u}_2 - \bar{u}_1\|_2)(\|\mathcal{G}(u_2^{(j)})\|_2 + \|\bar{\mathcal{G}}_2\|_2) \right\} \\
&\quad \leq \frac{1}{J}\sum_{j=1}^{J} \{2m(LJ\|u_1 - u_2\| + L\|u_1 - u_2\|) + (J\|u_1 - u_2\| + \|u_1 - u_2\|)2M\} \\
&\quad \leq 2(J+1)\max\{mL, M\}\|u_1 - u_2\|,
\end{aligned}$$

and we define $C_3 = 2(J+1)\max\{mL, M\}$. The computation is similar for the last point of the statement, for which we have

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

which implies

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

Using bounds (5.52) and (5.53) and the fact that $\mathcal{G}$ is Lipschitz with constant L, we obtain

$$\begin{aligned}
&\|C^{pp}(u_1) - C^{pp}(u_2)\|_2 \\
&\quad\leq \frac{1}{J}\sum_{j=1}^{J} \{2M(LJ\|u_1 - u_2\| + L\|u_1 - u_2\|) + (LJ\|u_1 - u_2\| + L\|u_1 - u_2\|)2M\} \\
&\quad= 4ML(J+1)\|u_1 - u_2\|,
\end{aligned}$$

and we define $C_4 = 4ML(J+1)$. □

Proof of Lemma 5.13

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

$$\begin{aligned}
\|\mathcal{G}_h^0(u^{(j)}) - \mathcal{G}^0(u^{(j)})\|_2 &= \|\mathcal{O}(\mathcal{S}_h^0(u^{(j)})) - \mathcal{O}(\mathcal{S}^0(u^{(j)}))\|_2 &(5.54)\\
&= \|\mathcal{O}(p_h^0(u^{(j)})) - \mathcal{O}(p^0(u^{(j)}))\|_2 &(5.55)\\
&\leq m\|p_h^0(u^{(j)}) - p^0(u^{(j)})\|_{L^2(\Omega)}, &(5.56)
\end{aligned}$$

where we write explicitly the dependence of p^0 and p_h^0 on the particle it is generated by. We now consider the standard a priori error estimates of FEM, (see e.g. [16, Theorem 3.2.5]), which reads

$$\|p_h^0(u^{(j)}) - p^0(u^{(j)})\|_{L^2(\Omega)} \leq C|p(u^{(j)})|_{H^{s+1}(\Omega)} h^{s+1}.$$

Moreover, higher order boundary regularity results for elliptic partial differential equations (see e.g. [**?**, Theorem 6.3.5]), imply for a constant $\widetilde{C} > 0$ and for all $j = 1,\dots,J$

$$\|p_h^0(u^{(j)}) - p^0(u^{(j)})\|_{L^2(\Omega)} \leq \|f\|_{H^{q-1}(\Omega)} h^{s+1}.$$

Finally, we obtain

$$\widetilde{e}(h,u) = \frac{1}{J}\sum_{j=1}^{J} \|\mathscr{G}_h^0(u^{(j)}) - \mathscr{G}^0(u^{(j)})\|_2 \tag{5.57}$$

$$\leq mC\widetilde{C}\|f\|_{H^{q-1}(\Omega)} h^{s+1}, \tag{5.58}$$

which proves the result for $\widetilde{K} = mC\widetilde{C}\|f\|_{H^{q-1}(\Omega)}$. □

Proof of Lemma 5.21

We recall the duality formula (5.33) for the Wasserstein distance $W_{1,s}$

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

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

$$W_{1,s}(\mu_n,\mu) = \sup_{\varphi\in\Phi}\left\{\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right\} = \sup_{\varphi\in\Phi}\left\{\left|\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right|\right\}. \tag{5.59}$$

Indeed we have

$$\int_{B_R(u^*)} \varphi d(\mu_n-\mu) \leq \left|\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right|,$$

which implies the first inequality

$$\sup_{\varphi\in\Phi}\left\{\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right\} \leq \sup_{\varphi\in\Phi}\left\{\left|\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right|\right\}.$$

On the other hand, we also have

$$A = \left\{\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\colon \varphi\in\Phi\right\} \supseteq \left\{\left|\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right| : \varphi\in\Phi\right\} = A', \tag{5.60}$$

because if $c\in A'$, which means that there exists $\varphi\in\Phi$ such that

$$c = \left|\int_{B_R(u^*)} \varphi d(\mu_n-\mu)\right|,$$

then we can take $\widetilde{\varphi}\in\Phi$ defined as

$$\widetilde{\varphi} = \begin{cases} \varphi & \text{if } \int_{B_R(u^*)} \varphi d(\mu_n-\mu) > 0 \\ -\varphi & \text{if } \int_{B_R(u^*)} \varphi d(\mu_n-\mu) < 0, \end{cases}$$

and note that that

$$c = \int_{B_R(u^*)} \widetilde{\varphi} d(\mu_n-\mu),$$

which implies that $c \in A$. Therefore, by (5.60), we deduce the opposite inequality

$$\sup_{\varphi\in\Phi}\left\{\int_{B_R(u^*)}\varphi d(\mu_n-\mu)\right\} \geq \sup_{\varphi\in\Phi}\left\{\left|\int_{B_R(u^*)}\varphi d(\mu_n-\mu)\right|\right\}.$$

Then, thanks to (5.59), we have

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

and the right hand side vanishes by hypothesis, so we obtain

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

Hence

$$\mathbb{E}_\xi\left[\left|\int_{B_R(u^*)}\varphi d\mu_n-\int_{B_R(u^*)}\varphi d\mu\right|\right]\to 0 \qquad \text{for all } \varphi\in\Phi. \tag{5.61}$$

It remains to show that (5.61) holds true for all functions $f \in C^0(B_R(u^*))$. First, we consider any Lipschitz function ψ with Lipschitz constant L. We define $\varphi = \psi/L$, then $\varphi \in \Phi$, indeed

$$|\varphi(x)-\varphi(y)| = \left|\frac{1}{L}\psi(x)-\frac{1}{L}\psi(y)\right| = \frac{1}{L}|\psi(x)-\psi(y)| \leq \frac{1}{L}L\|x-y\|_s = \|x-y\|_s.$$

Therefore we have

$$\mathbb{E}_\xi\left[\left|\int_{B_R(u^*)}\psi d\mu_n-\int_{B_R(u^*)}\psi d\mu\right|\right] = L\mathbb{E}_\xi\left[\left|\int_{B_R(u^*)}\varphi d\mu_n-\int_{B_R(u^*)}\varphi d\mu\right|\right]\to 0. \tag{5.62}$$

By density, any continuous bounded function $f \in C^0(B_R(u^*))$ can be approximated by a sequence of Lipschitz functions $\{\psi_k\}_{k\in\mathbb{N}}$ such that $\|\psi_k\|_{L^\infty(B_R(u^*))} \leq C$ for all $k \in \mathbb{N}$ where C is a constant dependent on f and $\|\psi_k - f\|_{L^\infty(B_R(u^*))} \to 0$ as $k \to \infty$. Thanks to (5.62) we have

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

and, applying Lebesgue dominated convergence theorem, we can pass to the limit as $k \to \infty$. We can exchange the limit with the expectation and the integral because the integrand functions are bounded by C and the measures μ_n and μ are finite, since they are probability measures. Thus we obtain

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

for all bounded continuous functions $f \in C^0(B_R(u^*))$ which means

$$\mu_n \xrightarrow{L^1} \mu,$$

which is the desired result. □

Bibliography

[1] A. ABDULLE, *Fourth order Chebyshev methods with recurrence relation,* SIAM J. Sci. Comput., 23 (2002), pp. 2041–2054.

[2] A. ABDULLE AND A. DI BLASIO, *A Bayesian numerical homogenization method for elliptic multiscale inverse problems.* Accepted for publication in SIAM-ASA J Uncertain, 2018.

[3] ——, *Numerical homogenization and model order reduction for multiscale inverse problems,* Multiscale Model. Simul., 17 (2019), pp. 399–433.

[4] A. ABDULLE AND G. GAREGNANI, *Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration,* Stat. Comput., (2020).

[5] A. ABDULLE, G. GAREGNANI, AND A. ZANONI, *Ensemble Kalman filter for multiscale inverse problems.* arXiv preprint arXiv:1908.05495, 2019.

[6] A. ABDULLE AND A. A. MEDOVIKOV, *Second order Chebyshev methods based on orthogonal polynomials,* Numer. Math., 90 (2001), pp. 1–18.

[7] Y. AÏT-SAHALIA AND J. JACOD, *High-frequency financial econometrics,* Princeton University Press, 2014.

[8] 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.

[9] C. ANDRIEU AND G. O. ROBERTS, *The pseudo-marginal approach for efficient Monte Carlo computations,* Ann. Statist., 37 (2009), pp. 697–725.

[10] R. E. BANK AND R. K. SMITH, *A posteriori error estimates based on hierarchical bases,* SIAM J. Numer. Anal., 30 (1993), pp. 921–935.

[11] 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.

[12] A. BENSOUSSAN, J.-L. LIONS, AND G. PAPANICOLAOU, *Asymptotic analysis for periodic structures,* North-Holland Publishing Co., Amsterdam, 1978.

[13] D. CALVETTI, M. DUNLOP, E. SOMERSALO, AND A. STUART, *Iterative updating of model error for Bayesian inversion*, Inverse Problems, 34 (2018), pp. 025008, 38.

[14] D. CALVETTI, O. ERNST, AND E. SOMERSALO, *Dynamic updating of numerical model discrepancy using sequential sampling*, Inverse Problems, 30 (2014), pp. 114019, 19.

[15] 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.

[16] P. G. CIARLET, *The finite element method for elliptic problems.*, vol. 40 of Classics Appl. Math., SIAM, Philadelphia, 2002.

[17] 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.

[18] ——, *Bayesian probabilistic numerical methods*, SIAM Rev., 61 (2019), pp. 756–789.

[19] P. R. CONRAD, M. GIROLAMI, S. SÄRKKÄ, A. STUART, AND K. ZYGALAKIS, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput., 27 (2017), pp. 1065–1082.

[20] C. J. COTTER AND G. A. PAVLIOTIS, *Estimating eddy diffusivities from noisy Lagrangian observations*, Commun. Math. Sci., 7 (2009), pp. 805–838.

[21] 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.

[22] T. R. CURRY AND L. BANKOFF, *Problems and Solutions: Solutions of Elementary Problems: E1861*, Amer. Math. Monthly, 74 (1967), pp. 724–725.

[23] 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.

[24] ——, *The Bayesian Approach to Inverse Problems*, in Handbook of Uncertainty Quantification, Springer, 2016, pp. 1–118.

[25] P. E. FARRELL AND J. R. MADDISON, *Conservative interpolation between volume meshes by local Galerkin projection*, Comput. Methods Appl. Mech. Engrg., 200 (2011), pp. 89–100.

[26] P. E. FARRELL, M. D. PIGGOTT, C. C. PAIN, G. J. GORMAN, AND C. R. WILSON, *Conservative interpolation between unstructured meshes via supermesh construction*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 2632–2642.

[27] S. GAILUS AND K. SPILIOPOULOS, *Statistical inference for perturbed multiscale dynamical systems*, Stochastic Process. Appl., 127 (2017), pp. 419–448.

[28] ——, *Discrete-time statistical inference for multiscale diffusions*, Multiscale Model. Simul., 16 (2018), pp. 1824–1858.

[29] E. HAIRER, *Variable time step integration with symplectic methods*, Appl. Numer. Math., 25 (1997), pp. 219–227.

[30] 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.

[31] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations I. Nonstiff Problems*, vol. 8, Springer Verlag Series in Comput. Math., Berlin, 1993.

[32] E. HAIRER AND G. WANNER, *Solving ordinary differential equations II. Stiff and differential-algebraic problems*, Springer-Verlag, Berlin and Heidelberg, 1996.

[33] P. HENNIG, M. A. OSBORNE, AND M. GIROLAMI, *Probabilistic numerics and uncertainty in computations*, Proc. A., 471 (2015), pp. 20150142, 17.

[34] M. HÉNON AND C. HEILES, *The applicability of the third integral of motion: Some numerical experiments*, Astronom. J., 69 (1964), pp. 73–79.

[35] 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.

[36] 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.

[37] H. KERSTING, T. J. SULLIVAN, AND P. HENNIG, *Convergence rates of Gaussian ODE filters*. arXiv preprint arXiv:1807.09737, 2018.

[38] S. KRUMSCHEID, G. A. PAVLIOTIS, AND S. KALLIADASIS, *Semiparametric drift and diffusion estimation for multiscale diffusions*, Multiscale Model. Simul., 11 (2013), pp. 442–473.

[39] 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.

[40] 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.

[41] 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.

[42] E. N. LORENZ, *Deterministic nonperiodic flow*, J. Atmos. Sci., 20 (1963), pp. 130–141.

[43] J. C. MATTINGLY, A. M. STUART, AND D. J. HIGHAM, *Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise*, Stochastic processes and their applications, 101 (2002), pp. 185–232.

[44] G. N. MILSTEIN AND M. V. TRETYAKOV, *Stochastic numerics for mathematical physics*, Scientific Computing, Springer-Verlag, Berlin and New York, 2004.

[45] 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.

[46] 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.

[47] C. J. OATES AND T. J. SULLIVAN, *A modern retrospective on probabilistic numerics*, Stat. Comput., 29 (2019), pp. 1335–1351.

[48] 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.

[49] L. F. OLSEN, *An enzyme reaction with a strange attractor*, Phys. Lett. A, 94 (1983), pp. 454 – 457.

[50] A. PAPAVASILIOU, G. A. PAVLIOTIS, AND A. M. STUART, *Maximum likelihood drift estimation for multiscale diffusions*, Stochastic Process. Appl., 119 (2009), pp. 3173–3210.

[51] 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.

[52] 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.

[53] G. A. PAVLIOTIS AND A. M. STUART, *Parameter estimation for multiscale diffusions*, J. Stat. Phys., 127 (2007), pp. 741–781.

[54] G. A. PAVLIOTIS AND A. M. STUART, *Multiscale methods: averaging and homogenization*, vol. 53 of Texts in Applied Mathematics, Springer, New York, 2008.

[55] 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.

[56] Y. POKERN, A. M. STUART, AND E. VANDEN-EIJNDEN, *Remarks on drift estimation for diffusion processes*, Multiscale Model. Simul., 8 (2009), pp. 69–95.

[57] A. QUARTERONI, *Numerical Models for Differential Problems*, vol. 2 of Modeling, Simulation & Applications, Springer, 2009.

[58] 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.

[59] R. D. SKEEL AND C. W. GEAR, *Does variable step size ruin a symplectic integrator?*, Physica, 60 (1992), pp. 311–313.

[60] C. STÖRMER, *Sur les trajectoires des corpuscules électrisés*, Arch. sci. phys. nat. Genève, 24 (1907), pp. 5–18, 113–158, 221–247.

[61] A. M. STUART, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[62] T. J. SULLIVAN, *Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors*, Inverse Probl. Imaging, 11 (2017), pp. 857–874.

[63] 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.

[64] R. VERFÜRTH, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[65] L. VERLET, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Physical Review, 159 (1967), pp. 98–103.

[66] Y. YING, J. MADDISON, AND J. VANNESTE, *Bayesian inference of ocean diffusivity from Lagrangian trajectory data*, Ocean Model., 140 (2019).

[67] 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.

6 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: Prof. A. Abdulle.

2014 - 2017 **MSc in Computational Science and Engineering**
École Polytechnique Fédérale de Lausanne, Switzerland.
Thesis advisor: Prof. A. Abdulle.

2014 - 2017 **BSc in Mathematical Engineering**
Politecnico di Milano, Italy.
Thesis advisor: Prof. F. Tomarelli.

PhD Publications

[1] A. ABDULLE, G. GAREGNANI, *Probabilistic methods for elliptic partial differential equations: a posteriori error estimators with randomized meshes*, in preparation (2020)

[2] ———, *Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration*, Stat. Comput. (2020)

[3] A. ABDULLE, G. GAREGNANI, G.A. PAVLIOTIS, A.M. STUART AND A. ZANONI, *Drift Estimation of Multiscale Diffusion via Filtering*, in preparation (2020)

[4] A. ABDULLE, G. GAREGNANI AND A. ZANONI, *Ensemble Kalman filter for multiscale inverse problems*, preprint (2019)

Other Publications

[1] 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)

[2] 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*, Microelectronics Reliability, 63 (2016), pp. 90 – 96

Presentations

- SIAM CONFERENCE ON UNCERTAINTY QUANTIFICATION (Garching, Germany, March 2020);
  Talk: *Model Misspecification And Uncertainty Quantification For Drift Estimation In Multiscale Diffusion Processes.*

- IMPERIAL COLLEGE (London, UK, February 2020);
  Seminar: *A pre-processing technique for asymptotically correct drift estimation in multiscale diffusion processes.*

- CALIFORNIA INSTITUTE OF TECHNOLOGY (Pasadena, USA, August 2019);
  Seminar: *Bayesian inference of multiscale differential equations.*

- MATHICSE RETREAT (Champéry, Switzerland, June 2019);
  Short 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, Tawian, 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);
  Short talk: *Probabilistic Runge–Kutta methods for ODEs*.