Bayesian parameter inference of multiscale diffusion processes

Assyr Abdulle[*] Giacomo Garegnani[*] Grigoris Pavlitois [†]

Abstract

AMS subject classifications.

Keywords.

1 Introduction

2 Problem statement

Let $(\Omega, \mathcal{A}, P)$ be a probability space, ε, α and σ be positive real numbers and $V_0 \colon \mathbb{R}^d \to \mathbb{R}^d$, $V_1 \colon \mathbb{R}^d \to \mathbb{R}^d$. Let us consider the autonomous SDE on $(\Omega, \mathcal{A}, P)$

$$\begin{aligned} \mathrm{d}x^\varepsilon(t) &= -\alpha \nabla V_0(x^\varepsilon(t)) \,\mathrm{d}t - \frac{1}{\varepsilon} \nabla V_1\left(\frac{x^\varepsilon(t)}{\varepsilon}\right) \mathrm{d}t + \sqrt{2\sigma}\,\mathrm{d}W(t), \quad 0 < t \leq T, \\ x^\varepsilon(0) &= x_0, \end{aligned} \tag{1}$$

where $W(t)$ is a standard Brownian motion and x_0 is a random variable with bounded moments of all orders. Theory of homogenization guarantees that there exists a SDE of the form

$$\begin{aligned} \mathrm{d}x^0(t) &= -A \nabla V_0(x^0(t)) \,\mathrm{d}t + \sqrt{2\Sigma}\,\mathrm{d}W(t), \quad 0 < t \leq T, \\ x^0(0) &= x_0, \end{aligned} \tag{2}$$

where $W(t)$ is the same Brownian motion, such that $x^\varepsilon(t)$ converges to $x(t)$ in law. In particular, we have $A = K\alpha$ and $\Sigma = K\sigma$, where the value of K is given by (introduce theory of homogenization).

Let us denote by $\vartheta^\varepsilon = (\alpha, \sigma)$ the parameters appearing in (1) and by $\vartheta^0 = (A, \Sigma)$ the parameters of (2). We denote by Θ the domain of definition of both ϑ^ε and ϑ^0. Moreover, let us introduce the multiscale forward operator $\mathcal{G}^\varepsilon \colon \Omega \times \Theta \to \mathbb{R}^{Nd}$, which is defined by

$$\mathcal{G}^\varepsilon \colon \omega \times \vartheta^\varepsilon \mapsto \mathbf{x}^\varepsilon \coloneqq \left((x_1^\varepsilon)^\top, (x_2^\varepsilon)^\top, \dots, (x_N^\varepsilon)^\top\right)^\top, \tag{3}$$

where $x_k^\varepsilon = x^\varepsilon(t_k)$ and $t_1, t_2, \dots, t_N$, $T = t_N$ is an increasing sequence of time instants. We can write $\mathcal{G}^\varepsilon = \mathcal{O} \circ \mathcal{S}^\varepsilon$, where $\mathcal{O} \colon \mathcal{C}((0,T), \mathbb{R}^d) \to \mathbb{R}^{Nd}$ is the observation operator, mapping a continuous function with values in $\mathbb{R}^d$ into pointwise evaluations, and where $\mathcal{S}^\varepsilon \colon \Omega \times \Theta \to C((0,T), \mathbb{R}^d)$ is the random multiscale solution operator, mapping a pair $(\omega, \vartheta^\varepsilon)$ into the solution of (1). Analogously, we denote by $\mathcal{G}^0 \colon \Omega \times \Theta \to \mathbb{R}^{Nd}$ the homogenized forward operator, which is defined by

$$\mathcal{G}^0 \colon \omega \times \vartheta^\varepsilon \mapsto \mathbf{x}^0 \coloneqq \left((x_1^0)^\top, (x_2^0)^\top, \dots, (x_N^0)^\top\right)^\top, \tag{4}$$

where $x_k^0 = x^0(t_k)$. Evaluating $\mathcal{G}^0$ involves the computation of the homogenized coefficient, as well as solving of (2). Therefore, we can write $\mathcal{G}^0 = \mathcal{O} \circ \mathcal{S}^0 \circ \mathcal{H}$, where $\mathcal{H} \colon \Omega \times \Theta \to \Omega \times \Theta$ is the

[*] Institute of Mathematics, École Polytechnique Fédérale de Lausanne ({assyr.abdulle, giacomo.garegnani}@epfl.ch)

[†] Department of Mathematics, Imperial College London

homogenization operator and summarizes the operations necessary for computing the homogenized SDE (2) from the multiscale SDE (1) and $\mathcal{S}^0\colon \Omega \times \Theta \to \mathcal{C}((0,T),\mathbb{R}^d)$ is the solution operator associated to (2). Let us remark that since the same Brownian motion is employed in (1) and (2), the map $(\omega,\cdot) \mapsto \mathcal{H}(\omega,\cdot)$ is the identity. In the following, for ease of notation and clarity, we will omit the dependence on $\omega \in \Omega$ of the operators introduced above.

We are interested in two distinct inference problems. The first can be summarized as

$$\text{Find } \vartheta^\varepsilon \text{ given observations } \mathbf{y} = \mathcal{G}^\varepsilon(\vartheta^\varepsilon) + \eta, \tag{5}$$

where η is a random variable with density $p_\eta(\cdot)$ representing a source of additive noise. Here, both the parameter we wish to retrieve and the observations belong to the multiscale model. We assume noise at one time instant to be independent of all the other time instants and in general η to be independent of ϑ, i.e., we have that

$$p(y_k \mid \mathbf{x}^\varepsilon, \vartheta) = p(y_k \mid x_k). \tag{6}$$

In the Gaussian case $\eta \sim \mathcal{N}(0,\Gamma)$, this is equivalent to assuming a block-diagonal structure on the covariance matrix Γ. It is interesting to study the effect of employing $\mathcal{G}^0$ instead of $\mathcal{G}^\varepsilon$ on the solution of (5). This problem has been analysed in the framework of elliptic partial differential equations in [1,2,12] (is there other literature?). The second inference problem can be summarized as

$$\text{Find } \vartheta^0 \text{ given observations } \mathbf{y} = \mathcal{G}^\varepsilon(\vartheta^\varepsilon) + \eta. \tag{7}$$

In this case, we want to fit a homogenized model to observations coming from a multiscale equation. Let us remark that solving this inverse problem does not require, a priori, the knowledge of the functional form of the multiscale equation (1). This problem has been considered in [13] (introduce ideas – limitations of the analysis: asymptotic results, need of subsampling, non-Bayesian – other references/ideas in the literature?).

In this work, we consider the Bayesian interpretation of problems (5) and (7). In the Bayesian framework, the goal is computing a probability distribution over the parameter, the posterior, given observations of the state and a prior distribution μ_{pr} with density $p_{\mathrm{pr}}(\cdot)$. Since the parameter we consider is finite-dimensional, in the following we assume that all the distributions admit a probability density with respect to the Lebesgue measure. Employing the multiscale forward map $\mathcal{G}^\varepsilon$ or the homogenized map $\mathcal{G}^0$ gives rise to two different posterior distributions. In particular, we denote as $\mu^\varepsilon(\vartheta \mid \mathbf{y})$ the posterior whose probability density function $p^\varepsilon(\vartheta^\varepsilon \mid \mathbf{y})$ satisfies due to Bayes' rule

$$p^\varepsilon(\vartheta^\varepsilon \mid \mathbf{y}) = \frac{1}{Z^\varepsilon} p_{\mathrm{pr}}(\vartheta^\varepsilon)\, p^\varepsilon(\mathbf{y} \mid \vartheta), \tag{8}$$

where $p^\varepsilon(y \mid \vartheta)$ is the likelihood associated to the data and Z^ε is the normalization constant given by

$$Z^\varepsilon = \int_\Theta p_{\mathrm{pr}}(\vartheta^\varepsilon)\, p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)\, \mathrm{d}\vartheta^\varepsilon. \tag{9}$$

The likelihood function can be expressed as the marginal distribution of the random vector $(\mathbf{y}, \mathbf{x}^\varepsilon \mid \vartheta)$, which is given by

$$\begin{aligned} p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon) &= \int_{\mathbb{R}^{Nd}} p^\varepsilon(\mathbf{y}, \mathbf{x} \mid \vartheta^\varepsilon)\, \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^{Nd}} p^\varepsilon(\mathbf{x} \mid \vartheta^\varepsilon)\, p^\varepsilon(\mathbf{y} \mid \mathbf{x}, \vartheta^\varepsilon)\, \mathrm{d}\mathbf{x}. \end{aligned} \tag{10}$$

Let us now consider the two factors appearing in (10). The first can be factored due to the Markov property as

$$p^\varepsilon(\mathbf{x} \mid \vartheta^\varepsilon) = p(x_0) \prod_{k=0}^{N-1} p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon), \tag{11}$$

where $p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon)$ is the density function of the transition probability of the solution of (1) and $p(x_0)$ is the density of the distribution of the initial condition. The second, due to the independence assumption (6) can be factored as

$$p^\varepsilon(\mathbf{y} \mid \mathbf{x}, \vartheta^\varepsilon) = \prod_{k=1}^{N} p(y_k \mid x_k). \tag{12}$$

In particular, we remark that $p^\varepsilon(\mathbf{y} \mid \mathbf{x}, \vartheta^\varepsilon)$ is independent of ε and therefore we will write it as $p(\mathbf{y} \mid \mathbf{x}, \vartheta^\varepsilon)$ in the following. Summarizing, the posterior distribution is given by

$$p^\varepsilon(\vartheta^\varepsilon \mid \mathbf{y}) = \frac{1}{Z^\varepsilon} p_{\mathrm{pr}}(\vartheta^\varepsilon) \int_{\mathbb{R}^{Nd}} p(x_0) \prod_{k=0}^{N-1} p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon) \prod_{k=1}^{N} p(y_k \mid x_k) \,\mathrm{d}\mathbf{x}. \tag{13}$$

Replacing $\mathcal{G}^\varepsilon$ by $\mathcal{G}^0$ does not modify the structure of the posterior. The two modifications that occur are given by the different transition probabilities of the solution of (2) with respect to the solution of (1), and by the normalization constant. Therefore, we denote by $\mu^0(\vartheta^\varepsilon \mid \mathbf{y})$ the posterior distribution whose density $p^0(\vartheta^\varepsilon \mid \mathbf{y})$ satisfies

$$p^0(\vartheta^\varepsilon \mid \mathbf{y}) = \frac{1}{Z^0} p_{\mathrm{pr}}(\vartheta^\varepsilon) \int_{\mathbb{R}^{Nd}} p(x_0) \prod_{k=0}^{N-1} p^0(x_{k+1} \mid x_k, \mathcal{H}(\vartheta^\varepsilon)) \prod_{k=1}^{N} p(y_k \mid x_k) \,\mathrm{d}\mathbf{x}. \tag{14}$$

In Section 3 we study the convergence of μ^ε to μ^0 in the limit for $\varepsilon \to 0$.

3 Convergence analysis

In this section, we consider the inverse problem (5).

Definition 1. Let μ and ν be probability measures which admit densities f and g with respect to Lebesgue measure respectively. The Hellinger distance $d_{\mathrm{Hell}}(\mu, \nu)$ between μ and ν is defined as

$$2 d_{\mathrm{Hell}}(\mu, \nu)^2 := \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}x \tag{15}$$

Assumption 1. The density $p^\varepsilon(\cdot \mid x, \vartheta^\varepsilon)$ of the transition probability of the solution of (1) as well as the density $p^0(\cdot \mid x, \vartheta^0)$ of the transition probability of the solution of (2) are asymptotically equicontinuous (specify what this means – prove that it holds? Is it possible?).

Lemma 1. *Under Assumption 1, $p^\varepsilon(\cdot \mid x, \vartheta) \to p^0(\cdot \mid x, \mathcal{H}(\vartheta))$ pointwise for $\varepsilon \to 0$.*

Proof. Theory of homogenization guarantees that $x^\varepsilon \to x^0$ in law in $\mathcal{C}((0,T), \mathbb{R}^d)$. A converse of Scheffé's theorem holds under Assumption 1 (see e.g. [5, 16]), so that the desired result holds. □

The following theorem proves the convergence of the multiscale posterior towards the homogenized posterior in in the limit $\varepsilon \to 0$. The proof is inspired by [15, Proposition 4.6], [11, Theorem 3.1] and [1, Theorem 5]

Theorem 1. *Under Assumption 1,*

$$d_{\mathrm{Hell}}(\mu^\varepsilon(\cdot \mid \mathbf{y}), \mu^0(\cdot \mid \mathbf{y})) \to 0 \tag{16}$$

for $\varepsilon \to 0$ independently of $\mathbf{y}$.

Proof. In the following, we denote by C a positive constant which can change value from line to line. By definition of $d_{\mathrm{Hell}}(\cdot, \cdot)$, replacing and since for real numbers a, b it holds $(a+b)^2 \leq 2a^2 + 2b^2$ we have

$$
\begin{aligned}
2d_{\mathrm{Hell}}(\mu^\varepsilon(\cdot \mid \mathbf{y}), \mu^0(\cdot \mid \mathbf{y})) &= \int_\Theta p(\vartheta^\varepsilon)\Big(\sqrt{\frac{p^0(\mathbf{y} \mid \vartheta^\varepsilon)}{Z^0}} - \sqrt{\frac{p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)}{Z^\varepsilon}}\Big)^2 \,\mathrm{d}\vartheta^\varepsilon \\
&\leq 2\int_\Theta p(\vartheta^\varepsilon)\Big(\sqrt{\frac{1}{Z^0}} - \sqrt{\frac{1}{Z^\varepsilon}}\Big)^2 p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)\,\mathrm{d}\vartheta^\varepsilon \\
&\quad + \frac{2}{Z_0}\int_\Theta p(\vartheta^\varepsilon)\Big(\sqrt{p^0(\mathbf{y} \mid \vartheta^\varepsilon)} - \sqrt{p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)}\Big)^2 \,\mathrm{d}\vartheta^\varepsilon \\
&=: I_1^\varepsilon + I_2^\varepsilon.
\end{aligned} \tag{17}
$$

Let us first consider I_2^ε. For positive real numbers a and b it holds

$$
(a-b)^2 \leq \frac{(a^2-b^2)^2}{a^2+b^2}, \tag{18}
$$

and thus

$$
\Big(\sqrt{p^0(\mathbf{y} \mid \vartheta^\varepsilon)} - \sqrt{p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)}\Big)^2 \leq \frac{\big(p^0(\mathbf{y} \mid \vartheta^\varepsilon) - p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)\big)^2}{p^0(\mathbf{y} \mid \vartheta^\varepsilon) + p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)}. \tag{19}
$$

Let us consider the difference $\mathcal{E}_1^\varepsilon \coloneqq p^0(\mathbf{y} \mid \vartheta^\varepsilon) - p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon)$. We have

$$
\begin{aligned}
\mathcal{E}_1^\varepsilon &= \int_{\mathbb{R}^{Nd}} p(x_0)\,\Big(\prod_{k=0}^{N-1} p^0(x_{k+1} \mid x_k, \mathcal{H}(\vartheta^\varepsilon)) - \prod_{k=0}^{N-1} p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon)\Big)\prod_{k=1}^{N} p(y_k \mid x_k)\,\mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^{Nd}} p(x_0)\,\mathcal{E}_2^\varepsilon \prod_{k=1}^{N} p(y_k \mid x_k)\,\mathrm{d}\mathbf{x},
\end{aligned} \tag{20}
$$

where $\mathcal{E}_2^\varepsilon$ is defined as

$$
\mathcal{E}_2^\varepsilon \coloneqq \prod_{k=0}^{N-1} p^0(x_{k+1} \mid x_k, \mathcal{H}(\vartheta^\varepsilon)) - \prod_{k=0}^{N-1} p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon). \tag{21}
$$

For sequences of real numbers a_k, b_k, $k = 0, \dots, N-1$, a telescopic sum argument yields

$$
\prod_{k=0}^{N-1} a_k - \prod_{k=0}^{N-1} b_k = \sum_{l=0}^{N-1}\Big(\prod_{j=0}^{l-1} a_j\Big)(a_l - b_l)\Big(\prod_{j=l+1}^{N-1} b_j\Big), \tag{22}
$$

where we adopted the convention

$$
j < i \implies \prod_{l=i}^{j} a_l = 1. \tag{23}
$$

Therefore, we can write

$$
\begin{aligned}
\mathcal{E}_2^\varepsilon = \sum_{l=0}^{N-1}&\Big(\prod_{j=0}^{l-1} p^0(x_{j+1} \mid x_j, \mathcal{H}(\vartheta^\varepsilon))\Big)\big(p^0(x_{l+1} \mid x_l, \mathcal{H}(\vartheta^\varepsilon)) - p^\varepsilon(x_{l+1} \mid x_l, \vartheta^\varepsilon)\big) \\
&\times\Big(\prod_{j=l+1}^{N-1} p^\varepsilon(x_{j+1} \mid x_j, \vartheta^\varepsilon)\Big).
\end{aligned} \tag{24}
$$

Due to Lemma 1, we have $p^0(x_{l+1} \mid x_l, \vartheta^\varepsilon) - p^\varepsilon(x_{l+1} \mid x_l, \vartheta^\varepsilon) \to 0$ for $\varepsilon \to 0$, which implies $\mathcal{E}_2^\varepsilon \to 0$ for $\varepsilon \to 0$. Replacing $\mathcal{E}_2^\varepsilon$ in (20) and applying Lebesgue dominated convergence theorem (find bound), we have $\mathcal{E}_1^\varepsilon \to 0$ for $\varepsilon \to 0$. Therefore, another application of Lebesgue dominated convergence theorem (find bound) gives $I_2^\varepsilon \to 0$ for $\varepsilon \to 0$. Let us now consider I_1^ε. We can rewrite

$$
I_1^\varepsilon = \Big(\sqrt{\frac{1}{Z^0}} - \sqrt{\frac{1}{Z^\varepsilon}}\Big)^2 Z^\varepsilon, \tag{25}
$$

Algorithm 1: Pseudo-marginal Metropolis–Hastings

```
Data: initial guess $\vartheta^0 \in \Theta$, $M \in \mathbb{N}$, proposal distribution $q \colon \Theta \times \Theta \to \mathbb{R}$ ;
compute $\hat{p}^{(0)} = \hat{p}(\vartheta^{(0)} \mid \mathbf{y})$ ;
for $k = 0, \dots, M$ do
    sample $\vartheta^\star \sim q(\cdot \mid \vartheta^{(k)})$, compute $\hat{p}^\star = \hat{p}(\vartheta^\star \mid \mathbf{y})$ ;
    compute
        $\alpha\big(\vartheta^*, \vartheta^{(k)}\big) = \min\left\{1, \frac{\hat{p}^\star}{\hat{p}^{(k)}} \frac{q(\vartheta^{(k)} \mid \vartheta^\star)}{q(\vartheta^\star \mid \vartheta^{(k)})}\right\}$;
    with probability $\alpha\big(\vartheta^\star, \vartheta^{(k)}\big)$ set $\vartheta^{(k+1)} = \vartheta^\star$, $\hat{p}^{(k+1)} = \hat{p}^\star$ ;
    otherwise set $\vartheta^{(k+1)} = \vartheta^{(k)}$, $\hat{p}^{(k+1)} = \hat{p}^{(k)}$ ;
```

which implies

$$\begin{aligned} \frac{1}{Z^\varepsilon} I_1^\varepsilon &\leq C \max\{(Z^0)^{-3}, (Z^\varepsilon)^{-3}\}(Z^0 - Z^\varepsilon)^2 \\ &\leq C(Z^0 - Z^\varepsilon)^2. \end{aligned} \tag{26}$$

Finally, since

$$Z^0 - Z^\varepsilon = \int_\Theta p(\vartheta^\varepsilon) \left(\int_{\mathbb{R}^{Nd}} \prod_{k=1}^N p(y_k \mid x_k) \, \mathcal{E}_2^\varepsilon \, \mathrm{d}\mathbf{x} \right) \mathrm{d}\vartheta^\varepsilon, \tag{27}$$

by dominated convergence theorem (twice, find bound) and since $\mathcal{E}_2^\varepsilon \to 0$, we have $Z^0 - Z^\varepsilon \to 0$ for $\varepsilon \to 0$ and therefore $I_1^\varepsilon \to 0$, which concludes the proof. □

4 Sampling from the posterior

In order to obtain samples from the posterior distributions $p^\varepsilon(\vartheta \mid \mathbf{y})$ and $p^0(\vartheta \mid \mathbf{y})$, it is necessary to recur to Monte Carlo simulations. In particular, let us neglect in this section the difference between multiscale and homogenized posteriors and refer to a general posterior $p(\vartheta \mid \mathbf{y})$, where $\mathbf{y}$ is a set of observations coming from a generic Markov chain parametrized by ϑ and characterized by a transition probability with density $p(\cdot \mid x, \vartheta)$. In the context of Bayesian inference problems, it is frequent to employ algorithms of the family of the Markov chain Monte Carlo methods (MCMC). These algorithms proceed by generating a Markov chain over the space Θ from a proposal distribution, whose density we denote by $q(\cdot \mid \vartheta)$, and by tuning the probability of accepting a new sample so that samples are indeed generated from the posterior. These sampling schemes require the evaluation of the posterior distribution for each new sample. In our setting, in which it is unfeasible to evaluate the posterior due to the complex structure of the likelihood function (10), it is possible to employ the pseudo-marginal Metropolis–Hastings method (PMMH) [4], which is given in Algorithm 1 and which requires only an estimator of the posterior. In particular, if for each ϑ the estimator $\hat{p}(\vartheta \mid \mathbf{y})$ is unbiased, it is possible to prove [4] that the Markov chain generated by the PMMH algorithm has the posterior $p(\vartheta \mid \mathbf{y})$ as unique invariant distribution. The performances of the PMMH algorithm strongly depend on the quality of the unbiased estimator $\hat{p}(\vartheta \mid \mathbf{y})$. In particular, high values for the variance result in Markov chains with a degenerate behaviour, i.e., an extremely low acceptance ratio, regardless of the choice of the proposal distribution (see e.g. [7]). It is therefore fundamental to compute estimators of the posterior, i.e., of the likelihood function, which are unbiased and whose variance is relatively small. A popular choice in this framework is provided by particle filters, which we briefly describe below. Let us finally remark that the version of PMMH with a particle filter estimator for the posterior density is referred to in literature as Particle Markov chain Monte Carlo (PMCMC) [3].

```
Algorithm 2: Particle filter
Data: $M \in \mathbb{N}$, initial ensemble $\{x_0^{(j)}\}_{j=1}^M \sim p_x(\cdot \mid \vartheta)$ ;
For $j = 1, \dots, M$ initialize $w^{(j)} = 1/M$, set $\mathbf{w} = \{w^{(j)}\}_{j=1}^M$, set $\hat{p}(y_{1:K} \mid \vartheta) = 1$;
for $k = 1, \dots, K$ do
    For $j = 1, \dots, M$ sample $I_j \sim \mathcal{F}_M(\cdot \mid \mathbf{w})$ ;
    For $j = 1, \dots, M$ sample $x_k^{(j)} \sim q_x(\cdot \mid x_{k-1}^{(I_j)}, y_{k+1})$, set $x_{1:k}^{(j)} = \big(x_{1:k-1}^{(I_j)}, x_k^{(j)}\big)$ ;
    For $j = 1, \dots, M$ compute the weight
        $\widetilde{w}^{(j)} = \frac{p_x(x_k^{(j)} \mid x_{k-1}^{(I_j)}; \vartheta)\, p_y(y_k \mid x_k^{(j)})}{q_x(x_k^{(j)} \mid x_{k-1}^{(I_j)}, y_{k+1}; \vartheta)};$    (28)
    For $j = 1, \dots, M$ compute the normalized weight
        $w^{(j)} = \frac{\widetilde{w}^{(j)}}{\sum_{i=1}^M \widetilde{w}^{(i)}};$
    Update $\mathbf{w} = \{w^{(j)}\}_{j=1}^M$ and
        $\hat{p}(y_{1:K} \mid \vartheta) \leftarrow \hat{p}(y_{1:K} \mid \vartheta) \frac{1}{M} \sum_{i=1}^M \widetilde{w}^{(i)};$
Output: Estimators $\hat{p}(y_{1:K} \mid \vartheta)$ and $\hat{p}_M(x_{1:K} \mid y_{1:K}) = \sum_{j=1}^M w^{(j)} \delta(x_{1:K} - x_{1:K}^{(j)})$ ;
```

4.1 Particle filters

Particle filters are a popular method for Bayesian inference in the context of hidden Markov models. Let us consider the general setting of a homogeneous Markov chain $\{x_k\}_{k=0}^K$ over $\mathbb{R}^d$ whose transition probability has a known density, denoted by p_x, such that $x_{k+1} \sim p_x(\cdot \mid x_k; \vartheta)$, where $\vartheta \in \Theta$ is a given parameter. Moreover, let us consider an observed process $\{y_k\}_{k=1}^K$ given by the observation model $y_k \sim p_y(\cdot \mid x_k)$, and such that observations are conditionally independent. We adopt the notation $x_{0:j} = \{x_k\}_{k=0}^j$ and equivalently for the observations. A particle filter can in turn be employed for the estimation of the probability $p(x_{0:K} \mid y_{1:K}; \vartheta)$ and for the likelihood function $p(y_{1:K} \mid \vartheta)$. The basis of the algorithm lays on the filtering recursion

$$p(x_k \mid y_{1:k}; \vartheta) = \frac{p_y(y_k \mid x_k)\, p(x_k \mid y_{1:k-1}; \vartheta)}{p(y_k \mid y_{1:k-1}; \vartheta)}, \tag{29a}$$

$$p(y_k \mid y_{1:k-1}; \vartheta) = \int p_y(y_k \mid x_k)\, p(x_k \mid y_{1:k-1}; \vartheta)\, \mathrm{d}x_k, \tag{29b}$$

$$p(x_{k+1} \mid y_{1:k}; \vartheta) = \int p_x(x_{k+1} \mid x_k; \vartheta)\, p(x_k \mid y_{1:k})\, \mathrm{d}x_k. \tag{29c}$$

The densities above are approximated via an ensemble of trajectories, or particles, which is rejuvenated at each iteration via resampling procedures. Moreover, importance sampling techniques can be employed by noticing that (29c) can be rewritten as

$$p(x_{k+1} \mid y_{1:k}; \vartheta) = \int q_x(x_{k+1} \mid x_k; \vartheta, y_{k+1}) \frac{p_x(x_{k+1} \mid x_k; \vartheta)}{q_x(x_{k+1} \mid x_k, y_{k+1}; \vartheta)}\, p(x_k \mid y_{1:k})\, \mathrm{d}x_k, \tag{30}$$

where $q_x(x \mid x_k, y_{k+1})$ is an appropriately chosen valid importance density. In particular, let us remark that q_x depends on y_{k+1}, i.e., the following observation. A technique for forming a robust importance density is given by the so-called diffusion bridge approach, which is presented in [8, 9]. The final numerical procedure is summarized in Algorithm 2, where we introduce the notation $\mathcal{F}_M(\cdot \mid \mathbf{w})$ for the discrete distribution over the set $\{1, \dots, M\}$ with weights $\mathbf{w} = \{w^{(j)}\}_{j=1}^M$. Let us finally remark that the estimator $\hat{p}(y_{1:K} \mid \vartheta)$ is unbiased independently of the choice of the importance density [14]. The second output of the particle filter is an approximation $\hat{p}_M(x_{1:K} \mid y_{1:K})$ of the density $p(x_{1:K} \mid y_{1:K})$, which is consistent in the sense that it tends to the truth for $M \to \infty$ (add ref).

Let us consider the case $d = 1$ and $x_k = x(t_k)$ for an equispaced grid $0 = t_0 < t_1 < \ldots < t_K = T$, where x is the solution of the SDE

$$\mathrm{d}x(t) = f_\vartheta(t, x(t))\,\mathrm{d}t + g_\vartheta(t, x(t))\,\mathrm{d}W(t), \tag{31}$$

and $y_k = x_k + \eta_k$, where $\eta_k \sim p_\eta(\cdot)$ are i.i.d. random variables, so that $p_y(y_k \mid x_k) = p_\eta(y_k - x_k)$. For generic drift and diffusion functions f_ϑ and g_ϑ the transition density $p_x(x_{k+1} \mid x_k; \vartheta)$ does not admit a closed form, and cannot therefore be evaluated in Algorithm 2. Nonetheless, if the spacing $h = t_k - t_{k-1}$ between the time points where observations are obtained is small enough, a good approximation of the transition density is given by numerical integrators such as the Euler–Maruyama method, which reads

$$x_{k+1} = x_k + f_\vartheta(t_k, x_k)h + g_\vartheta(t_k, x_k)\sqrt{h}Z_k, \tag{32}$$

where $Z_k \sim \mathcal{N}(0, 1)$. The transition probability is therefore given by $\mathcal{N}(x_k + f_\vartheta(t_k, x_k)h, g_\vartheta(t_k, x_k)^2\sqrt{h})$, whose density can be evaluated.

5 Modelling error

In section 3, we considered the asymptotic limit $\varepsilon \to 0$, in which the forward model $\mathcal{G}^0$ is a good weak approximation to $\mathcal{G}^\varepsilon$. In case $\varepsilon > 0$ is a fixed value in the non-asymptotic regime, it is necessary to estimate the modelling error given by the replacement of $\mathcal{G}^\varepsilon$ by $\mathcal{G}^0$ as a forward model in order to solve the inverse problem correctly. The approach of [6] (add references) is well-suited for this purpose, as it has been demonstrated by means of experiments in [1].

Let us consider the multiscale model $x^\varepsilon_{k+1} \sim p^\varepsilon_x(\cdot \mid x^\varepsilon_k; \vartheta)$, where p^ε_x is the transition density of (1) and $x_k = x(t_k)$, and the homogenized model $x^0_k \sim p^0_x(\cdot \mid x^0_k; \vartheta)$, where p^0_x is the transition density of (2). Moreover, we consider the observation model to be given by $y_k = x^\varepsilon_k + \eta_k$, with $\eta_k \sim p_\eta$ are i.i.d. random variables, and we denote the modelling error $m_k \coloneqq x^\varepsilon_k - x^0_k$. We can therefore write the observation equation as

$$y_k = x^0_k + m_k + \eta_k. \tag{33}$$

In the following, we assume that m_k is independent of the hidden state x_k, of the parameter ϑ and of the noise η_k, so that $p_y(y_k \mid m_k, x_k) = p_\eta(y_k - m_k - x_k)$ and by marginalization

$$p_y(y_k \mid x_k) = \int p_\eta(y_k - m_k - x_k)\, p(m_k)\,\mathrm{d}m_k, \tag{34}$$

which allows to rewrite the filtering recursion (29), and in particular Bayes' rule (29a) as

$$p(x_k \mid y_{1:k}; \vartheta) = \frac{p(x_k \mid y_{1:k-1}; \vartheta)}{p(y_k \mid y_{1:k-1}; \vartheta)} \int p_\eta(y_k - m_k - x_k)\, p(m_k)\,\mathrm{d}m_k. \tag{35}$$

The evident task is now determining or approximating the distributions of the modelling error $\mu^m_k(\mathrm{d}m_k) = p(m_k)\,\mathrm{d}m_k$. Let $X^\varepsilon_{1:K}$ be the stochastic process defined by $X^\varepsilon_k = (x^\varepsilon_k, m_k)^\top$. It is indeed possible to sample from the numerical approximation of the transition density $p_{X^\varepsilon}(X^\varepsilon_{k+1} \mid X^\varepsilon_k; \vartheta)$, which involve the discretization of the SDEs (1) and (2). Moreover, we have the observation model $y_k = HX^\varepsilon_k + \eta$, where $H = (0, I)^\top$. Therefore, we can indeed run a particle filter with importance density $q_{X^\varepsilon} = p_{X^\varepsilon}$, i.e., a bootstrap particle filter [10], and obtain an approximation

$$p(m_{1:K} \mid y_{1:K}) \approx \hat{p}_M(m_{1:K} \mid y_{1:K}) = \sum_{i=1}^{M} w^{(i)}_m\, \delta(m_{1:K} - m^{(i)}_{1:K}) \tag{36}$$

Algorithm 3: Sampling from the posterior

```
Data: $L \in \mathbb{N}$, prior $\mu_{\mathrm{pr}}$ on $\vartheta$ ;
Set $\mu^0 = \mu_{\mathrm{pr}}$ ;
for $l = 1, \dots, L$ do
    Approximate $p^l(m_{1:K})$ running a bootstrap particle filter with parameter $\bar{\vartheta} = \mathbb{E}_{\mu^{l-1}}(\vartheta)$ ;
    Sample with PMCMC from posterior $\mu^l$, employing $p^l(m_{1:K})$ in (35) ;
```

where M indicates the number of particles and $\{w_m^{(i)}\}$, for $i = 1, \dots, M$ are the weights of each particle. The formula above can be replaced in (35) to obtain the approximation

$$p(x_k \mid y_{1:k}; \vartheta) \approx \frac{p(x_k \mid y_{1:k-1}; \vartheta)}{p(y_k \mid y_{1:k-1}; \vartheta)} \sum_{i=1}^{M} p_\eta(y_k - m_k^{(i)} - x_k) w_m^{(i)}. \tag{37}$$

This approximation can therefore be employed in a particle filter of the form of Algorithm 2, replacing the weight update (28) with

$$\widetilde{w}^{(j)} = \frac{p_x(x_k^{(j)} \mid x_{k-1}^{(I_j)}; \vartheta)}{q_x(x_k^{(j)} \mid x_{k-1}^{(I_j)}, y_{k+1}; \vartheta)} \sum_{i=1}^{M} p_\eta(y_k - m_k^{(i)} - x_k) w_m^{(i)}. \tag{38}$$

Conditioning the modelling to the observations partially solves the issue of the assumption of m_k being independent of x_k^ε. The second assumption can be solved partially with an approach similar to the one presented in [6]. In particular, a PMCMC algorithm with update formula for the inner particle filter given by (38) can be run, and the resulting point estimate of the parameter can be employed to compute a better approximation of the modelling error. Iterating this idea allows to progressively approximate the modelling error with values taken closer to the true posterior. The complete procedure is summarized in Algorithm 3.

6 Numerical discretization

- Discretization of (1): $h \propto \varepsilon^2 \implies \mathcal{G}^0$ cheap to evaluate.
- Sparse data – resulting from subsampling or access to a subset or observation period long, i.e., time between observations $\delta t > h$ where h integration time step.

References

[1] A. Abdulle and A. Di Blasio, *A Bayesian numerical homogenization method for elliptic multiscale inverse problems*. Submitted to SIAM UQ, 2018.

[2] ———, *Numerical homogenization and model order reduction for multiscale inverse problems*. Accepted in SIAM MMS, 2018.

[3] C. Andrieu, A. Doucet, and R. Holenstein, *Particle Markov chain Monte Carlo methods*, J. R. Stat. Soc. Ser. B. Stat. Methodol., (2010), pp. 269 – 342.

[4] C. Andrieu and G. O. Roberts, *The pseudo-marginal approach for efficient Monte Carlo computations*, Ann. Statist., 37 (2009), pp. 697–725.

[5] D. D. Boos, *A converse to Scheffé's theorem*, Ann. Statist., 13 (1985), pp. 423–427.

[6] D. Calvetti, M. Dunlop, E. Somersalo, and A. Stuart, *Iterative updating of model error for Bayesian inversion*, Inverse Problems, 34 (2018), pp. 025008, 38.

[7] A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn, *Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator*, Biometrika, (2015), pp. 1 – 19.

[8] A. Golightly and D. J. Wilkinson, *Markov chain Monte Carlo algorithms for SDE parameter estimation*, Learning and Inference for Computational Systems Biology, (2010), pp. 253–276.

[9] ——, *Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo*, Interface focus, 1 (2011), pp. 807–820.

[10] N. J. Gordon, D. J. Salmond, and A. F. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, in IEE proc F (radar and signal processing), vol. 140, IET, 1993, pp. 107–113.

[11] 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.

[12] 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.

[13] G. A. Pavliotis and A. M. Stuart, *Parameter estimation for multiscale diffusions*, J. Stat. Phys., 127 (2007), pp. 741–781.

[14] M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn, *On some properties of Markov chain Monte Carlo simulation methods based on the particle filter*, J. Econometrics, 171 (2012), pp. 134–151.

[15] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[16] T. J. Sweeting, *On a converse to Scheffé's theorem*, Ann. Statist., 14 (1986), pp. 1252–1256.