Bayesian inference of multiscale diffusion processes

Assyr Abdulle* Giacomo Garegnani* Grigoris Pavlitois †

Abstract

Add abstract

AMS subject classifications.

Keywords.

1 Introduction

Add introduction.

2 Problem statement

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

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

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

$$\mathcal{G}^\varepsilon \colon \omega \times \vartheta^\varepsilon \mapsto \mathbf{x}^\varepsilon := \left((x_1^\varepsilon)^\top, (x_2^\varepsilon)^\top, \ldots, (x_N^\varepsilon)^\top\right)^\top,$$

where $x_k^\varepsilon = x^\varepsilon(t_k)$ and $t_1, t_2, \ldots, t_N$, $T = t_N$ 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

$$\mathcal{G}^0 \colon \omega \times \vartheta^\varepsilon \mapsto \mathbf{x}^0 := \left((x_1^0)^\top, (x_2^0)^\top, \ldots, (x_N^0)^\top\right)^\top,$$

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

†Department of Mathematics, Imperial College London

where $x^0_k = x^0(t_k)$. 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 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

$$\text{Find } \vartheta^\varepsilon \text{ given observations } \mathbf{y} = \mathcal{G}^\varepsilon(\vartheta^\varepsilon) + \eta, \tag{3}$$

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

$$p(y_k \mid \mathbf{x}^\varepsilon, \vartheta) = p(y_k \mid x_k). \tag{4}$$

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 (3). This problem has been analysed in the framework of elliptic partial differential equations in [1,2,11] (is there other literature?). The second inference problem can be summarized as

$$\text{Find } \vartheta^0 \text{ given observations } \mathbf{y} = \mathcal{G}^\varepsilon(\vartheta^\varepsilon) + \eta. \tag{5}$$

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 [12] (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 (3) and (5). 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

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

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

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

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

$$\begin{aligned} p^\varepsilon(\mathbf{y} \mid \vartheta^\varepsilon) &= \int_{\mathbb{R}^{Nd}} p^\varepsilon(\mathbf{y}, \mathbf{x} \mid \vartheta^\varepsilon) \, \mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^{Nd}} p^\varepsilon(\mathbf{x} \mid \vartheta^\varepsilon)\, p^\varepsilon(\mathbf{y} \mid \mathbf{x}, \vartheta^\varepsilon) \, \mathrm{d}\mathbf{x}. \end{aligned} \tag{6}$$

Let us now consider the two factors appearing in (6). The first can be factored due to the Markov property as

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

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 (4) can be factored as

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

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

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

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

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

In Section 3.1 we study the convergence of μ^ε to μ^0 in the limit for $\varepsilon \to 0$.

3 Convergence analysis

3.1 Inference of the multiscale equation

In this section, we consider the inverse problem (3).

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

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

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, 14]), 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 [13, Proposition 4.6], [10, Theorem 3.1] and [1, Theorem 5]

Theorem 1. *Under Assumption 1,*

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

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

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

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

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

and thus

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

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

$$\begin{aligned}
\mathcal{E}_1^\varepsilon &= \int_{\mathbb{R}^{Nd}} p(x_0)\,\Big(\prod_{k=0}^{N-1} p^0(x_{k+1} \mid x_k, \mathcal{H}(\vartheta^\varepsilon)) - \prod_{k=0}^{N-1} p^\varepsilon(x_{k+1} \mid x_k, \vartheta^\varepsilon)\Big)\prod_{k=1}^{N} p(y_k \mid x_k)\,\mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{R}^{Nd}} p(x_0)\,\mathcal{E}_2^\varepsilon \prod_{k=1}^{N} p(y_k \mid x_k)\,\mathrm{d}\mathbf{x},
\end{aligned} \tag{7}$$

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

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

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

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

where we adopted the convention

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

Therefore, we can write

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

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 (7) 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

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

which implies

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

Finally, since

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

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. □

3.2 Inference of the homogenized equation

In this section we consider problem (5).

4 Modelling error

Explain approach for both (3) and (5).

5 Numerical discretization

- Discretization of (1) is cheaper than (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.

6 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 (6), it is possible to employ the pseudo-marginal Metropolis–Hastings method (PMMH) [4]. Given an initial value $\vartheta^{(0)}$, the algorithm can be summarized as

a) compute $p(\vartheta^{(0)} \mid \mathbf{y})$

b) for $k = 1, \dots, L$

b).1) sample ϑ^* from $p(\cdot \mid \vartheta^{(k-1)})$,

b).2) compute the acceptance probability

$$\alpha\big(\vartheta^*, \vartheta^{(k-1)}\big) = \min\left\{1, \frac{\hat{p}(\vartheta^* \mid \mathbf{y})}{\hat{p}(\vartheta^{(k-1)} \mid \mathbf{y})} \frac{q(\vartheta^{(k-1)} \mid \vartheta^*)}{q(\vartheta^* \mid \vartheta^{(k-1)})}\right\},$$

b).3) set $\vartheta^{(k)} = \vartheta^*$ with probability $\alpha\big(\vartheta^*, \vartheta^{(k-1)}\big)$, $\vartheta^{(k)} = \vartheta^{(k-1)}$ otherwise.

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. Therefore, choosing L big enough it is possible to generate samples distributed as the targeted posterior. 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. [6]). A short overview of methods to produce unbiased estimators of the posterior ought to be found in the next section.

6.1 Unbiased estimators of the posterior density

An unbiased estimator of the posterior density $p(\vartheta \mid \mathbf{y})$ can be computed with a Monte Carlo approximation. We assume we can evaluate the prior density $p_{\mathrm{pr}}(\vartheta)$ up to a constant, and therefore focus only on the estimation of the likelihood term $p(\mathbf{y} \mid \vartheta)$. In particular, a Monte Carlo scheme to obtain such an estimator is summarized as

1) sample M particles $\{X_0^{(i)}\}_{i=1}^M$ from $p(x_0)$,
2) initialize the likelihood estimator $\hat{p}(\mathbf{y} \mid \vartheta) = 1$,
3) for $k = 0, \dots, N-1$,
    - 3).1) for $i = 1, \dots, M$, propagate the particles sampling $X_{k+1}^{(i)}$ from $p(\cdot \mid X_k^{(i)}, \vartheta)$,
    - 3).2) update the likelihood estimation

$$\hat{p}(\mathbf{y} \mid \vartheta) \leftarrow \hat{p}(\mathbf{y} \mid \vartheta)\, \frac{1}{M} \sum_{i=1}^{M} p(y_{k+1} \mid X_{k+1}^{(i)}).$$

The procedure above produces indeed an unbiased estimator of the likelihood. Nonetheless, if the noise driving the process $x(t)$ is not negligible, the variance of the estimator will render the estimation not employable in practice. In order to overcome this issue, particle filters ought to be employed for obtaining an unbiased estimation of the likelihood. The main idea of particle filters is introducing weights on the particles, updating the weights at each step according to the observations and replicate the particles with higher weights, thus discarding the particles with meaningless values.. Given an importance density function $p_{\mathrm{IS}}(\cdot \mid x, \vartheta)$, the particle filter proceeds as

i) sample M particles $\{X_0^{(i)}\}_{i=1}^M$ from $p(x_0)$, initialize weights $\{W^{(i)}\} = M^{-1}$,

ii) initialize the likelihood estimator $\hat{p}(\mathbf{y} \mid \vartheta) = 1$,

iii) for $k = 0, \dots, N-1$,

- *iii*).1) sample an index $i^* \in \{1, \dots, M\}$ from $p(i^*) = W^{(i^*)}$,
- *iii*).2) propagate the particles sampling $X_{k+1}^{(i)}$ from $p_{\mathrm{IS}}(\cdot \mid X_k^{(i^*)}, \vartheta)$,
- *iii*).3) compute the unnormalized weight

$$\widehat{W}^{(i)} = \frac{p(X_{k+1}^{(i)} \mid X_k^{(i^*)}, \vartheta)\, p(y_{k+1} \mid X_{k+1}^{(i)})}{p_{\mathrm{IS}}(X_{k+1}^{(i)} \mid X_k^{(i^*)}, \vartheta)}, \tag{8}$$

- *iii*).4) compute the normalized weights

$$W^{(i)} = \frac{\widehat{W}^{(i)}}{\sum_{i=1}^{M} \widehat{W}^{(i)}},$$

iii).5) update the likelihood estimation

$$\hat{p}(\mathbf{y} \mid \vartheta) \leftarrow \hat{p}(\mathbf{y} \mid \vartheta)\, \frac{1}{M} \sum_{i=1}^{M} \widehat{W}^{(i)}.$$

In the algorithm above, we omit for ease of notation at each line from *iii*).1) to *iii*).4) that the computations have to be carried for all $i = 1, \dots, M$. The estimator $\hat{p}(\mathbf{y} \mid \vartheta)$ is an unbiased estimator of $p(\mathbf{y} \mid \vartheta)$ independently of the choice of the importance density p_{IS}, provided that $p_{\mathrm{IS}}(x^* \mid x, \vartheta) = 0$ only if $p(x^* \mid x, \vartheta) p(y \mid x^*) = 0$, so that (8) is well-defined. Let us remark that the overall MCMC algorithm employing a particle filter for the likelihood estimation ought to be found in literature under the name of particle MCMC methods (PMCMC) [3]. The variance of the estimator $\hat{p}(\mathbf{y} \mid \vartheta)$ depends strongly on the choice of the importance density p_{IS} [3,8]. In fact, the choice which appears more natural is to propagate the particle following the transition probability $p(\cdot \mid x, \vartheta)$. This choice gives rise to the boostrap particle filter [9], which has the advantage that the transition density does not have to be evaluated. In fact, at the k-th step the unnormalized weights simplify to

$$\widehat{W}^{(i)} = p(y_{k+1} \mid X_{k+1}^{(i)}),$$

so that a particle is assigned a higher weight if it is closer to the observations. While the bootstrap particle filter provides with a direct and intuitive implementation, it has been verified in practice that a drastic variance reduction is obtained by choosing at the k-th step the importance density as

$$p_{\mathrm{IS}}(\cdot \mid x_k, \vartheta) \approx p(\cdot \mid y_{k+1}, x_k, \vartheta). \tag{9}$$

Heuristically, this conditioning allows the particles to be generated close to the next observation, so that in turn the phenomenon of particle degeneracy is avoided, the effective sample size is higher and the final likelihood estimation has a lower variance. Nevertheless, an expression for $p(\cdot \mid y_{k+1}, x_k, \vartheta)$ is in general unavailable. There exist though methods which provide approximations as in (9) [7,8]. Running numerical experiments, we noticed in practice a good improvement of the quality of the estimators using an approach based on Gaussian distributions, which goes under the name of diffusion bridges approximation and is described extensively in [8].

7 Numerical experiments

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] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

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

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