Electronic Journal of Statistics
Vol. 5 (2011) 1090–1122
ISSN: 1935-7524
DOI: 10.1214/11-EJS632

Particle-based likelihood inference in partially observed diffusion processes using generalised Poisson estimators

Jimmy Olsson and Jonas Ströjby

Center of Mathematical Sciences
Lund University
Lund, Sweden
e-mail: jimmy@maths.lth.se; strojby@maths.lth.se

Abstract: This paper concerns the use of the expectation-maximisation (EM) algorithm for inference in partially observed diffusion processes. In this context, a well known problem is that all except a few diffusion processes lack closed-form expressions of the transition densities. Thus, in order to estimate efficiently the EM intermediate quantity we construct, using novel techniques for *unbiased* estimation of diffusion transition densities, a random weight fixed-lag auxiliary particle smoother, which avoids the well known problem of particle trajectory degeneracy in the smoothing mode. The estimator is justified theoretically and demonstrated on a simulated example.

AMS 2000 subject classifications: Primary 62M09; secondary 65C05.
Keywords and phrases: Auxiliary particle filter, EM algorithm, exact algorithm, generalised Poisson estimator, partially observed diffusion process, sequential Monte Carlo.

Received August 2010.

Contents

1. Introduction

A *state space model* is a statistical model where a Markov process (the *state process*) is only partially observed through an *observation process*. The two processes are linked in that the values (the *states*) of the hidden Markov process govern the distribution of the corresponding observations, which are assumed to be conditionally independent given the states. In some cases the unobserved variables can simply not be measured, while in other cases measurement costs limit the information available (e.g. meteorological and environmental data). The use of hidden states makes this class of models to an outermost generic and powerful statistical modeling tool, and state space models are nowadays successfully applied within a variety of scientific disciplines such as genetics [6], neurophysiology [2], target tracking [27], and speech recognition [26]. In many examples, the state process can be considered as a continuous time Markov process with observations occurring at discrete time points only, and of special interest is the case where the state process is a diffusion process; we will refer to such models as *partially observed diffusion* (POD) *processes*.

In this paper we discuss the use of *sequential Monte Carlo* (SMC) methods (alternatively termed *particle methods*) for likelihood-based inference in PODs. Maximum likelihood inference in general state space models is a nontrivial task, and for PODs the problem is complicated further by the fact that most diffusion processes lack closed-form transition densities. In cases where the transitions of the latent process can be simulated it is possible to produce pointwise and consistent estimates of the likelihood function using the standard *bootstrap particle filter* [14], in which the particles are assigned importance weights determined completely by the known local likelihood function. In such a framework, the likelihood surface can be explored using grid-based methods [16, 22] or stochastic approximation [17]. However, simulating exactly the transitions a diffusion process is in general infeasible and we are most often referred to discretisation methods, such as the Euler scheme, imposing a bias of the particle estimates; see [7]. Moreover, proposing (or *mutating*), as in the bootstrap filter, the particles “blindly” without taking into account the information provided by the current observation will in general lead to serious degeneracy of the particle weights, especially for models where the observations are informative.

In this contribution we take an approach to likelihood-based inference in PODs that relies on a novel technique of estimating diffusion process transition densities via so-called *generalised Poisson estimators* (GPEs). More specifically, we show how the expectation-step (E-step) of the *expectation-maximisation* (EM) *algorithm* [8] can be approximated efficiently using a GPE-based random weight particle smoother. There are two main difficulties with applying the EM algorithm to PODs: firstly, as mentioned, the transition density of the diffusion process and, as a consequence, the complete data log-likelihood function lack analytic expressions in general; secondly, computing the *intermediate quantity* of the E-step involves the computation of expectations under the *smoothing distribution*, i.e. the conditional distribution of the hidden states at the time points of observation given the observed data, which is not—even in the case of a known transition density—available on closed-form. In this paper we address these problems by applying the GPE suggested (as a refinement of results obtained in [4]) in [12] in conjunction with SMC smoothing algorithms. Unfortunately, it has been observed by several authors (see e.g. [5, Chapter 8]) that applying standard SMC methods to smoothing may be unreliable for larger observation sample sizes n, since resampling systematically the particles leads to degeneracy of the particle paths. As a solution, we adapt the *fixed-lag smoother* proposed in [21] to the framework of PODs. This technique relies, in the spirit of [18], on *forgetting properties* of the conditional hidden chain; by this is meant that the hidden chain forgets its past when evolving, backwards as well as forwards, conditionally on the given observation sequence. The constructed algorithm avoids efficiently the problem of particle path degeneracy at the cost of a bias that may be controlled by a suitable choice of the introduced lag parameter.

In order to obtain a high performance of the particle smoother it is in general necessary to mutate the particles according to a kernel that incorporates the information provided by the current observation; however, such an improved mutation strategy is not straightforwardly adopted to PODs, since computing the resulting importance weights involves computing a ratio of the transition density of the hidden diffusion process (for which a closed-form expression is missing in general) to the transition density of the chosen proposal kernel. To cope with this, we follow [12] and replace each evaluation of the latent process transition density by a draw from the GPE. Thus, the GPE serves two purposes in our algorithm as it is used, firstly, for computing unbiased estimates of the importance weights of a particle filter based on a proposal kernel possibly different from the transition kernel of the hidden diffusion process and, secondly, for estimating the complete data log-likelihood function itself.

The proposed EM intermediate quantity estimator

- approximates efficiently the E-step in a *single sweep* of the data record, yielding an algorithm with a computational complexity of order $\mathcal{O}(nN)$;
- copes, as it is not based on any Euler discretisation or linearisation technique, efficiently with model nonlinearities;
- has only limited computer data storage requirements, which is essential in, e.g., high frequency applications where sometimes very long measurement sequences are considered;
- is provided with a rigorous convergence result describing its convergence to the true intermediate quantity. This result is derived via a convergence result, obtained under minimal assumptions, for the GPE-based particle smoother.

The paper is organised as follows: In Section 2 we recall the concept of PODs and discuss likelihood-based inference in such models via the EM-algorithm. GPEs are described in Section 2.1 (with additional details in Section B) and

Section 2.2 is devoted to SMC smoothing in general. In Section 2.3 we introduce the fixed-lag smoother and discuss how the fixed-lag approach can be used for state estimation in PODs via GPEs. A theoretical result describing the convergence of the fixed-lag-based estimator is found in Section 2.3.1 and in the implementation part, Section 3, we illustrate the method on two examples. In Section 4, the paper is concluded by some final conclusions and remarks. Proofs are found in Section A.

2. Preliminaries

In the following we assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathbb{E}$ denote expectations associated with $\mathbb{P}$. Denoting by $\mathbb{1}$ the indicator function and letting X be any random variable on $(\Omega, \mathcal{F})$, we will often make use of the short-hand notation $\mathbb{E}[X; A] = \mathbb{E}[X \mathbb{1}_A]$. Let $X \overset{\text{def}}{=} (X_t)_{t \geq 0}$ be continuous-time diffusion process taking values in some state space $(\mathsf{X}, \mathcal{X})$, with $\mathsf{X} \subseteq \mathbb{R}^{d_X}$. More specifically, the dynamics of the process is governed by the the stochastic differential equation

$$dX_t = \mu(X_t, \theta)\, dt + \sigma(X_t, \theta)\, dW_t \ , \tag{2.1}$$

where $W \overset{\text{def}}{=} (W_t)_{t \geq 0}$ is Brownian motion. We denote by $\mathbb{W}^{(x)}$ the law of W given that $W_0 = x$ and let $(\mathcal{F}_t)_{0 \leq t}$ be the filtration generated by W. The functions $\mu(\cdot, \theta)$ and $\sigma(\cdot, \theta)$ are assumed to satisfy regularity conditions (locally Lipschitz with a linear growth bound) that guarantee a weakly unique, global solution of (2.1). We will consider a framework where the process X is only partially observed at discrete time points $(t_k)_{k \geq 0}$ through the process $Y \overset{\text{def}}{=} (Y_k)_{k \geq 0}$ taking values in some measurable space $(\mathsf{Y}, \mathcal{Y})$. The observations of Y are assumed to be, conditionally on the latent process X, independent and such that the conditional distribution G_θ of Y_k given X depends on X_{t_k} only. In the following we write, in order to simplify the notation, X_k instead of X_{t_k}. The dynamics of the diffusion as well as the measurement process depend on some unknown model parameter θ which is assumed to belong to some compact parameter space $\Theta \subseteq \mathbb{R}^{d_\theta}$. Our main target is to estimate θ using the maximum likelihood method. For simplicity we assume that the observation time points are *equally spaced* and denote by Q_θ and χ the transition kernel and initial distribution, respectively, of the time homogeneous Markov chain $(X_k)_{k \geq 0}$. The family $(Q_\theta(x, \cdot); x \in \mathsf{X}, \theta \in \Theta)$ is assumed to be dominated by the Lebesque-measure λ with corresponding Radon-Nikodym derivatives $(q_\theta(x, \cdot); x \in \mathsf{X}, \theta \in \Theta)$. Moreover, suppose that G_θ has a density function g_θ with respect to some measure μ on $(\mathsf{Y}, \mathcal{Y})$ such that, for $k \geq 0$,

$$\mathbb{P}(Y_k \in A | X_k) = \int_A g_\theta(X_k, y)\, \mu(dy) \ , \quad A \in \mathcal{Y} \ .$$

Given a record $Y_{0:n} = (Y_0, Y_1, \ldots, Y_n)$ (this will be our generic notation for vectors) of observations, a consistent estimate of the parameter θ is ideally formed by maximising the *observed data likelihood function* $\ell_n(\theta; Y_{0:n}) \stackrel{\text{def}}{=} \log \mathrm{L}_n(\theta; Y_{0:n})$, where

$$\mathrm{L}_n(\theta; Y_{0:n}) \stackrel{\text{def}}{=} \int \cdots \int g_\theta(x_0, Y_0)\, \chi(dx_0) \prod_{k=1}^{n} g_\theta(x_k, Y_k)\, Q_\theta(x_{k-1}, dx_k) \; .$$

A problem with this approach is that we in general cannot compute L_n on closed-form, since this involves the evaluation of a high-dimensional integral over a complicated integrand. Since the partially observed diffusion model above is, like more general latent variable models, specified using conditional dependence relations, computation of parameter posterior distributions is facilitated significantly by maximising instead the *complete data* log-likelihood function by means of the EM algorithm. Thus, assume that we have at hand an initial estimate θ' of the parameter vector; in the EM algorithm an improved estimate is obtained by computing and maximising the intermediate quantity

$$\mathcal{Q}_n(\theta; \theta') \stackrel{\text{def}}{=} \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n-1} \log q_\theta(X_k, X_{k+1}) \middle| Y_{0:n} \right] + \mathbb{E}_{\theta'}\left[\sum_{k=0}^{n} \log g_\theta(X_k, Y_k) \middle| Y_{0:n} \right] , \tag{2.2}$$

with respect to θ'. Here we have written $\mathbb{E}_{\theta'}$ to stress that the expectations are taken under the dynamics determined by the initial parameter θ'. Under weak assumptions, repeating recursively this procedure yields a sequence of parameter estimates that converges to a stationary point $\hat{\theta}$ of the observed data log-likelihood [28]. As clear from (2.2), computing $\mathcal{Q}_n$ requires the computation of expectations under the smoothing distribution, i.e. the distribution of the state sequence $X_{0:n}$ conditionally on the observations $Y_{0:n}$, given by, for $A \in \mathcal{X}^{\otimes(n+1)}$,

$$\phi_n(A; \theta) \stackrel{\text{def}}{=} \frac{\int \cdots \int_A g_\theta(x_0, Y_0)\, \chi(dx_0) \prod_{k=1}^{n} g_\theta(x_k, Y_k)\, Q_\theta(x_{k-1}, dx_k)}{\mathrm{L}_n(\theta; Y_{0:n})} \; . \tag{2.3}$$

Of special interest is the *filter distribution*, i.e. the distribution of X_n conditionally on $Y_{0:n}$, given by the restriction $\phi_{n|n}(A) \stackrel{\text{def}}{=} \phi_n(\mathsf{X}^n \times A)$, $A \in \mathcal{X}$, of the smoothing distribution to the last component. It is easily shown that the flow $(\phi_k)_{k=0}^{\infty}$ satisfies the well-known *forward smoothing recursion*

$$\phi_{k+1}(A; \theta) = \frac{\mathrm{L}_k(\theta; Y_{0:k})}{\mathrm{L}_{k+1}(\theta; Y_{0:k+1})} \iint_A g_\theta(x_{k+1}, Y_{k+1})\, Q_\theta(x_k, dx_{k+1})\, \phi_k(dx_{0:k}; \theta) \; , \tag{2.4}$$

where $A \in \mathcal{X}^{\otimes(k+2)}$. By introducing the (non-Markovian) transition kernel

$$L_k(x_k, A; \theta) \stackrel{\text{def}}{=} \int_A g_\theta(x_{k+1}, Y_{k+1})\, Q_\theta(x_k, dx_{k+1}) \; ,$$

for $x_k \in \mathsf{X}$ and $A \in \mathcal{X}$, we may rewrite the recursion (2.4) as

$$\phi_{k+1}(A; \theta) = \frac{\iint_A L_k(x_k, dx_{k+1}; \theta)\, \phi_k(dx_{0:k}; \theta)}{\iint L_k(x_k, dx_{k+1}; \theta)\, \phi_k(dx_{0:k}; \theta)} \; . \tag{2.5}$$

Here the normalised (Markovian) kernel $L_k(x_k, A; \theta)/L_k(x, \mathsf{X}; \theta)$ is the so-called *optimal kernel* describing the distribution of X_{k+1} given $X_k = x_k$ *and* the new observation Y_{k+1}.

In general, a closed-form solution of the recursion (2.4) is not available. A standard approach is thus to apply some SMC smoothing algorithm (described in Section 2.2) to approximate the expectations in (2.2). Unfortunately, both the SMC smoother itself as well as the intermediate quantity (2.2) call for the transition density q_θ, which is usually unknown except in a few special cases. Nevertheless, results obtained by Beskos et al. [4] and Fearnhead et al. [12] offer a method for estimating this density *without bias*. A full treatment of this technique—which is a key ingredient of the estimation technique proposed here—is beyond the scope of this paper; nevertheless, the main framework and assumptions are described briefly in the next section. In addition, some more details can be found in Appendix B.

2.1. Generalised Poisson estimators

Define

$$\eta(u, \theta) \overset{\text{def}}{=} \int^u \frac{1}{\sigma(v, \theta)}\, dv$$

and set $\tilde{X}_t \overset{\text{def}}{=} \eta(X_t, \theta)$. Denote by f^{-1} the inverse of any invertable function f. By applying Itô's formula we obtain the stochastic differential equation

$$d\tilde{X}_t = \beta(\tilde{X}_t, \theta)\, dt + dW_t \;, \tag{2.6}$$

where

$$\beta(u, \theta) \overset{\text{def}}{=} \frac{\mu\{\eta^{-1}(u, \theta), \theta\}}{\sigma\{\eta^{-1}(u, \theta), \theta\}} + \frac{1}{2}\sigma'\{\eta^{-1}(u, \theta), \theta\} \;,$$

for the transformed process $\tilde{X} \overset{\text{def}}{=} (\tilde{X}_t)_{t \geq 0}$. Using again the notation $\tilde{X}_k = \tilde{X}_{t_k}$, let $\tilde{q}_\theta$ be the transition density (with respect to the Lebesgue measure λ) of $(\tilde{X}_k)_{k \geq 0}$. Then, straightforwardly,

$$q_\theta(x, x') = \tilde{q}_\theta(x, x') |\eta'(x', \theta)| \;. \tag{2.7}$$

Assume the following:

(A1) The process $(M_t)_{t \geq 0}$, with

$$M_t \overset{\text{def}}{=} \exp\left(\int_0^t \beta(\tilde{X}_s, \theta)\, d\tilde{X}_s + \int_0^t \beta^2(\tilde{X}_s, \theta)\, ds \right) \;,$$

is a martingale with respect to $\mathbb{W}^{(x)}$;

(A2) $\beta(\cdot, \theta)$ is continuously differentiable;

(A3) $\beta^2(\cdot, \theta) + \beta'(\cdot, \theta)$ is bounded from below by some function $l(\theta)$.

Under these conditions, which are satisfied for a relatively large class of diffusions, the GPE approach developed by [12] makes it possible to construct a kernel P_θ on $\mathsf{X}^2 \times \mathcal{B}(\mathbb{R}_+)$ such that $\int v \, P_\theta(x, x', dv) = \tilde{q}_\theta(x, x')$ for $(x, x') \in \mathsf{X}^2$. Consequently, and draw $\tilde{V}_\theta(x, x') \sim P_\theta(x, x', \cdot)$ is an unbiased estimate of $\tilde{q}_\theta(x, x')$. Then, letting $V_\theta(x, x') \overset{\text{def}}{=} \tilde{V}_\theta(x, x') |\eta'(x', \theta)|$ gives, by (2.7), an unbiased estimator also of $q_\theta(x, x')$. A full description of GPEs is beyond the scope of this paper; however, its main features are discussed in Appendix B. Similarly, using a related algorithm developed in [4], it is possible to construct a kernel $\bar{P}_\theta$ on $\mathsf{X}^2 \times \mathcal{B}(\mathbb{R})$ such that $\int v \bar{P}_\theta(x, x', dv) = \log q_\theta(x, x')$ for $(x, x') \in \mathsf{X}^2$, i.e. any draw $\bar{V}_\theta(x, x', \theta) \sim \bar{P}_\theta(x, x', \cdot)$ is an unbiased estimate of $\log q_\theta(x, x')$. Appealingly, it is in many cases (see Section 3 for examples) even possible to construct P_θ and $\bar{P}_\theta$ such that the functions $\theta \mapsto V_\theta(x, x')(\omega)$ and $\theta \mapsto \bar{V}_\theta(x, x')(\omega)$ are continuous for any fixed outcome $\omega \in \Omega$, yielding unbiased estimates of $q_\theta(x, x')$ and $\log q_\theta(x, x')$ *for all* $\theta \in \Theta$ *simultaneously*. This useful property makes, as we will see, the GPE approach well suited to numerical (log-)likelihood function optimisation.

2.2. *GPE-based particle smoothing*

Since we in this part deal with the problem of sampling $\phi_k(\cdot; \theta)$ for a given *fixed* parameter value, we will throughout this section expunge θ from the notation. To begin with, we assume that we know the transition kernel density q.

In order to describe precisely how SMC methods may be used for producing approximate solutions to the smoothing recursion (2.4), we suppose that we are given a weighted sample $(\xi^i_{0:k|k}, \omega^i_k)_{i=1}^N$ of particles and associated unnormalised weights, each particle $\xi^i_{0:k|k} = (\xi^i_{0|k}, \ldots, \xi^i_{k|k})$ being a random vector in X^{k+1}, approximating ϕ_k in the sense that

$$\phi^N_k(f) \overset{\text{def}}{=} \left(\Omega^N_k\right)^{-1} \sum_{i=1}^N \omega^i_k f(\xi^i_{0:k|k}) \approx \phi_k(f) \,, \tag{2.8}$$

where $\Omega^N_k \overset{\text{def}}{=} \sum_{\ell=1}^N \omega^\ell_k$ normalises the weights, for a large class of estimand functions f on X^{k+1}. Now, in order to form an updated particle sample approximating ϕ_{k+1} as a new observation Y_{k+1} becomes available, a natural approach is to replace ϕ_k in (2.5) by its particle approximation (2.8). This yields the mixture (recall the notation δ_a for a Dirac mass located at a)

$$\bar{\phi}^N_{k+1}(A) \overset{\text{def}}{=} \sum_{i=1}^N \frac{\omega^i_k L_k(\xi^i_{k|k}, \mathsf{X})}{\sum_{\ell=1}^N \omega^\ell_k L_k(\xi^\ell_{k|k}, \mathsf{X})} \iint_A \frac{L_k(\xi^i_{k|k}, dx_{k+1})}{L_k(\xi^i_{k|k}, \mathsf{X})} \delta_{\xi^i_{0:k|k}}(dx_{0:k}) \,,$$

for $A \in \mathcal{X}^{\otimes(k+2)}$. Now, the aim is to simulate a new set of particles from $\bar{\phi}^N_{k+1}$ and then repeat the whole procedure recursively to obtain particle samples approximating the smoothing distributions at all time steps. However, since we in general cannot neither simulate draws from the optimal kernel nor compute the mixture weights $L_k(\xi^i_{k|k}, \mathsf{X})$, we apply importance sampling and draw instead new particles from the instrumental mixture distribution

$$\pi^N_{k+1}(A) \overset{\text{def}}{=} \sum_{i=1}^{N} \frac{\omega^i_k \psi^i_k}{\sum_{\ell=1}^{N} \omega^\ell_k \psi^\ell_k} \iint_A \delta_{\xi^i_{0:k|k}}(dx_{0:k})\, R_k\left(\xi^i_{k|k}, dx_{k+1}\right) \ ,$$

for $A \in \mathcal{X}^{\otimes(k+2)}$, where R_k is a Markovian proposal kernel and $(\psi^i_k)_{i=1}^N$ are positive numbers referred to as *adjustment multiplier weights*. We will from now on assume that $\psi^i_k = \Psi_k(\xi^i_{0:k|k})$ for some nonnegative function $\Psi_k : \mathsf{X}^{k+1} \to \mathbb{R}^+$ and that each kernel R_k has a density r_k with respect to λ. Simulating a particle $\xi^i_{0:k+1|k+1}$ from π^N_{k+1} is easily done by, firstly, drawing, according to the probability distribution proportional to $(\omega^i_k \psi^i_k)_{i=1}^N$, a mixture component (or ancestor) index I^i_k among $\{1, \ldots, N\}$ and, secondly, extending the selected ancestor with a draw from the proposal kernel, i.e. letting $\xi^i_{0:k+1|k+1} \overset{\text{def}}{=} (\xi^{I^i_k}_{0:k|k}, \xi^i_{k+1|k+1})$ with $\xi^i_{k+1|k+1} \sim R_k(\xi^{I^i_k}_{k|k}, \cdot)$. After this, the drawn particle is assigned the importance weight

$$\omega^i_{k+1} \overset{\text{def}}{=} \Phi_{k+1}\left(\xi^i_{0:k+1|k+1}\right) \ , \tag{2.9}$$

where, for $x_{0:k+1} \in \mathsf{X}^{k+2}$,

$$\Phi_{k+1}(x_{0:k+1}) \overset{\text{def}}{=} \frac{g(x_{k+1}, Y_{k+1}) q(x_k, x_{k+1})}{\Psi_k(x_{0:k}) r_k(x_k, x_{k+1})} \ ,$$

implying $\omega^i_{k+1} \propto d\bar{\phi}^N_{k+1}/d\pi^N_{k+1}(\xi^i_{0:k+1|k+1})$. Finally, the weighted particle sample formed by the updated particles and weights is returned as an approximation of ϕ_{k+1}. Moreover, since the filter distribution is the marginal of the smoothing distribution with respect to the last component, an estimate of $\phi_{k+1|k+1}$ is formed by the marginal sample $(\xi^i_{k+1|k+1}, \omega^i_{k+1})_{i=1}^N$.

Proposing and selecting the particles according to the dynamics of the latent process, i.e. without making use of the information about the current state provided by the current observation, by letting $R_k \equiv Q$ and $\Psi_k \equiv \mathbf{1}$ for all k, corresponds to the bootstrap particle filter proposed by Gordon et al. [14].

The algorithm, which was developed gradually by, mainly, Handschin and Mayne [15], Gordon et al. [14], and Pitt and Shephard [25], will be referred to as the *auxiliary particle smoother* (APS). In the setting of a partially observed diffusion process we do not have access to a closed-form expression of the transition density q, which is needed when evaluating the importance weight function Φ_{k+1}. However, the GPE makes it possible to estimate this density without bias via the kernel P. This yields following algorithm, in following referred to as the *GPE-based particle smoother* (GPEPS), in which q in the weighting operation (2.9) is replaced by the Monte Carlo estimate

$$q^\alpha(x, x') \overset{\text{def}}{=} \frac{1}{\alpha} \sum_{\ell=1}^{\alpha} V^\ell(x, x') \ , \tag{2.10}$$

where the $V^\ell(x, x')$'s are drawn independently from $P(x, x', \cdot)$. Denote by

$$\Phi_{k+1}^\alpha(x_{0:k+1}) \overset{\text{def}}{=} \frac{g(x_{k+1}, Y_{k+1}) q^\alpha(x_k, x_{k+1})}{\Psi_k(x_{0:k}) r_k(x_k, x_{k+1})} , \tag{2.11}$$

the resulting estimated importance weight function. One iteration of the GPEPS is described in detail in the following scheme.

Algorithm 1
(∗ One iteration of GPEPS ∗)
Input: $(\xi^i_{0:k|k}, \omega^i_k)_{i=1}^N$

```
for $i \leftarrow 1$ to $N$
    simulate $I^i_k \sim (\omega^j_k \psi^j_k / \sum_{\ell=1}^N \omega^\ell_k \psi^\ell_k)_{j=1}^N$;
    simulate $\xi^i_{k+1|k+1} \sim R_k(\xi^{I^i_k}_{k|k}, \cdot)$;
    set $\xi^i_{0:k+1|k+1} \leftarrow (\xi^{I^i_k}_{0:k|k}, \xi^i_{k+1|k+1})$;
    simulate $V^{1:\alpha}(\xi^i_{k:k+1|k+1}) \sim P^{\otimes\alpha}(\xi^i_{k:k+1|k+1}, \cdot)$;
    compute $\Phi^\alpha_{k+1}$ via (2.11);
    set $\omega^i_{k+1} \leftarrow \Phi^\alpha_{k+1}(\xi^i_{k:k+1|k+1})$;
return $(\xi^i_{0:k+1|k+1}, \omega^i_{k+1})_{i=1}^N$.
```

Here we have used the notations $V^{1:\alpha}(x, x') \overset{\text{def}}{=} (V^1(x, x'), \dots, V^\alpha(x, x'))$ and $P^{\otimes\alpha}(x, x', \cdot) \overset{\text{def}}{=} P(x, x', \cdot) \otimes \cdots \otimes P(x, x', \cdot)$ (α times). Algorithm 1 is the smoothing mode formulation of the *random weight auxiliary particle filter* proposed by Fearnhead et al. [12]. Note that we have, in the scheme above, suppressed the dependence of the particles and the particle weights on α from the notation for clarity.

In Algorithm 1 selection is carried through by drawing indices $(I^i_k)_{i=1}^N$ multinomially with respect to weights $(\omega^j_k \psi^j_k / \sum_{\ell=1}^N \omega^\ell_k \psi^\ell_k)_{j=1}^N$. There are however alternative selection approaches, such as *deterministic plus residual multinomial resampling* proposed in [20] and described in detail in Section C. All theoretical results obtained in the following hold for multinomial resampling as well as residual plus multinomial resampling. In addition, our results can easily be extended to so-called *branching selection* (see Remark 2.1 below); however, since the number of drawn indices is random in this case, we omit these results for brevity.

2.2.1. Convergence of the GPEPS

We will describe the convergence, as N tends to infinity, of the self-normalised Monte Carlo approximations formed by weighted particle samples returned by Algorithm 1 using the concept of *consistency* adopted from [10] and defined in the following. Let $(\Xi, \mathcal{B}(\Xi))$ denote some given state space and $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$ a Ξ-valued particle sample.

Definition 2.1. A weighted sample $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$ is *consistent* for a probability measure μ and a set $\mathsf{C} \subseteq \mathsf{L}^1(\Xi, \mu)$ if, as $N \to \infty$,

$$\Omega_N^{-1} \sum_{i=1}^N \omega_{N,i} f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \mu(f) \,, \quad \text{for all } f \in \mathsf{C} \,, \tag{2.12}$$

and, additionally,

$$\Omega_N^{-1} \max_{1 \leq i \leq N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0 \,. \tag{2.13}$$

The following assumption is mild but essential when establishing consistency of the GPEPS scheme.

(A4) For all $0 \leq k \leq n$, $\Psi_k \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_k)$ *and* $L_k(\cdot, \mathsf{X}) \in \mathsf{L}^1(\mathsf{X}, \phi_k)$.

Proposition 2.1. *Assume (A1–3) (p. 5) and (A4). In addition, assume that the initial sample* $(\xi_0^i, \omega_0^i)_{i=1}^N$ *is consistent for* $(\phi_0, \mathsf{L}^1(\mathsf{X}, \phi_0))$. *Then, for all* $1 \leq k \leq n$, *each sample* $(\xi_{0:k|k}^i, \omega_k^i)_{i=1}^N$ *produced by Algorithm 1 is consistent for* $(\phi_k, \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_k))$. *The same holds when the multinomial selection schedule is replaced by deterministic plus residual multinomial selection.*

The proof of Proposition 2.1 is postponed to Appendix A.1. Proposition 2.1 provides a qualitative statement about the particle estimates produced by Algorithm 1 by establishing that, for all $k \geq 0$, $\phi_k^N(f)$ converges in probability to $\phi_k(f)$ for all ϕ_k-integrable target functions f. Remarkably, convergence is obtained for any *fixed* GPE sample size α.

Remark 2.1. Proposition 2.1 can, without any additional assumptions and with only very small changes of the proof, be extended to the cases where selection is based on *Poisson*, *binomial*, or *Bernoulli branching* (see [10] for a theoretical analysis of these branching algorithms). In these cases, the particle sample sizes are random at all time steps. On the other hand, if a constant number N of particles is targeted at each time step, it holds, for all k, that $N_k/N \xrightarrow{\mathbb{P}} 1$ as N tends to infinity, where N_k is the random particle sample size at time step k.

In a companion paper [23], Proposition 2.1 is complemented with a quantitative statement about the convergence of Algorithm 1 (when based on multinomial selection) in terms of a central limit theorem (CLT). More specifically, under the assumption that input particle sample $(\xi_{0:k|k}^i, \omega_k^i)_{i=1}^N$ of Algorithm 1 is asymptotically normal in the sense that

$$\sqrt{N}(\phi_k^N(f) - \phi_k(f)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_k^2(f)) \,, \quad \text{as } N \to \infty \,,$$

for some nonnegative functional σ_k^2 defined on some "large" class of target functions f, it can be shown inductively [the details are given in 23] that the updated particle sample obtained using Algorithm 1 satisfies the similar CLT

$$\sqrt{N}(\phi_{k+1}^N(f) - \phi_{k+1}(f)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_{k+1}^2(f)) \,, \quad \text{as } N \to \infty \,.$$

Here the asymptotic variance σ^2_{k+1} can be expressed as

$$\sigma^2_k(f) = \sigma^2_{\mathrm{APS},k+1}(f) + \varsigma^2_{k+1}(f)/\alpha \;, \tag{2.14}$$

where $\sigma^2_{\mathrm{APS},k+1}(f)$ is the asymptotic variance [obtained in 11, Theorem 3.2] of an ideal particle approximation obtained by updating the ancestor sample $(\xi^i_{0:k|k}, \omega^i_k)^N_{i=1}$ using a standard APS where the importance weight functions Φ_k are assumed to be known on closed-form; moreover,

$$\varsigma^2_{k+1}(f) \stackrel{\mathrm{def}}{=} \frac{\phi_k(\Psi_k) \iint \{ f\langle\phi_k\rangle^2(x')\sigma^2_{P_k}(x,x')/\Psi_k(x)\} \, R_k(x, dx') \, \phi_k(dx)}{[\phi_k L_k(\mathsf{X})]^2} \;,$$

with $f\langle\phi_k\rangle(x) \stackrel{\mathrm{def}}{=} f(x) - \phi_k(f)$ and

$$\sigma^2_{P_k}(x,x') \stackrel{\mathrm{def}}{=} \left(\frac{g(x', Y_{k+1})}{r_k(x,x')} \right)^2 \int \{v - q(x,x')\}^2 \, P_k(x,x',dv) \;, \quad (x,x') \in \mathsf{X}^2 \;,$$

being the conditional variance of the estimates of the Radon-Nikodym derivative dL_k/dR_k obtained with the GPE. Thus, ignoring its dependence on the multiplier weights and the target function, the quantity $\varsigma^2_{k+1}(f)$ can be viewed as the expected variance of the GPE under the asymptotic proposal distribution of the particle filter; consequently, the expression (2.14) can be interpreted as the standard decomposition of variance into expected conditional variance and variance of conditional expectation. The asymptotic variance (2.14) provides some guidance for how to select an optimal sample size α; indeed, for a sufficiently large number of particles,

$$\mathrm{Var}\left(\phi^N_k(f)\right) \approx \frac{1}{N}(\sigma^2_{\mathrm{APS},k}(f) + \varsigma^2_k(f)/\alpha) \;, \tag{2.15}$$

and assuming that it takes time τ_{GPE} to produce a draw from the GPE and time τ_{APS} to select and mutate a particle, the total cost for evolving the particle cloud one step using Algorithm 1 is $N(\tau_{\mathrm{APS}} + \alpha\tau_{\mathrm{GPE}})$. Assuming further that we have a fixed computational time τ available, we consider the constrained optimisation problem

$$\begin{cases} \min_{\alpha,N} \frac{1}{N}(\sigma^2_{\mathrm{APS},k}(f) + \varsigma^2_k(f)/\alpha) \;, \\ N(\tau_{\mathrm{APS}} + \alpha\tau_{\mathrm{GPE}}) = \tau \;, \\ \alpha > 0 \;, N > 0 \;, \end{cases}$$

where α and N are treated as continuous variables, having solution

$$\alpha_{\mathrm{opt}} = \sqrt{\frac{\varsigma^2_k(f)/\tau_{\mathrm{GPE}}}{\sigma^2_{\mathrm{APS},k}(f)/\tau_{\mathrm{APS}}}} \;.$$

The expression above is entirely in line with our expectations: the sample sizes α and N should be increased resp. decreased if the precision of the GPE is low relatively the precision of the (ideal) particle smoother (and vice versa) *or* if the computational cost of operating the GPE is low compared to the cost of mutating and selecting the particles.

2.3. Fixed-lag smoothing

Unfortunately, it has been observed by several authors that using standard SMC methods in the smoothing mode may be unreliable for larger observation sample sizes n, since resampling systematically the particles degenerates the particle paths [see e.g. 5, Section 8.3]. Indeed, when $k \ll n$, most (or possibly all) marginal particles $(\xi^i_{k|n})_{i=1}^N$ coincide with a large probability, inflicting large variance when estimating X_k conditionally to $Y_{0:n}$ using these particles. Especially, returning to the problem of estimating the intermediate quantity $\mathcal{Q}_n$ in (2.2), for any type of *additive functional* $t(x_{0:n}) \overset{\text{def}}{=} \sum_{k=0}^{n-1} s_k(x_{k:k+1})$, $(s_k)_{k=0}^{n-1}$ being a set of functions (cf. the two terms of (2.2)), we may expect that the estimator

$$(\Omega_n^N)^{-1} \sum_{k=0}^{n-1} \sum_{i=1}^{N} \omega_n^i s_k(\xi^i_{k:k+1|n}) \tag{2.16}$$

of $\mathbb{E}[t(X_{0:n})|Y_{0:n}]$ is poor when n is large. To compensate for this degeneracy the particle sample size N has to be increased drastically, yielding a computationally inefficient algorithm.

On the other hand, since we may expect that remote observations are only weakly dependent, it should hold that, for a large enough integer Δ_n,

$$\mathbb{E}\left[s_k(X_{k:k+1})|Y_{0:n}\right] \approx \mathbb{E}\left[s_k(X_{k:k+1})|Y_{0:k(\Delta_n)}\right] \ ,$$

where $k(\Delta_n) \overset{\text{def}}{=} \min\{k + \Delta_n, n\}$, yielding

$$\mathbb{E}[t(X_{0:n})|Y_{0:n}] = \sum_{k=0}^{n-1} \mathbb{E}\left[s_k(X_{k:k+1})|Y_{0:n}\right] \approx \sum_{k=0}^{n-1} \mathbb{E}\left[s_k(X_{k:k+1})|Y_{0:k(\Delta_n)}\right] \ . \tag{2.17}$$

Thus, as long as the approximation (2.17) is relatively precise for a Δ_n which is smaller than the average particle trajectory collapsing time, i.e. most marginal particles $(\xi^i_{k|k(\Delta_n)})_{i=1}^N$ are different for all k, we should replace (2.16) by the estimator

$$\sum_{k=0}^{n-1} \left(\Omega^N_{k(\Delta_n)}\right)^{-1} \sum_{i=1}^{N} \omega^i_{k(\Delta_n)} s_k\left(\xi^i_{k:k+1|k(\Delta_n)}\right) \ . \tag{2.18}$$

The lag-based approximation (2.18) may be computed recursively in a *single sweep* of the data with only limited computer data storage demands, and computing (2.18) is clearly not more computationally demanding than computing (2.16) (having $\mathcal{O}(nN)$ complexity); see Olsson et al. [21] for details. Finally, using (2.18) in conjunction with the kernel $\bar{P}_\theta$ for estimating $\log q_\theta$ gives us the following approximation of the intermediate quantity $\mathcal{Q}_n(\theta;\theta')$:

$$\mathcal{Q}_n^N(\theta;\theta') \overset{\text{def}}{=} \sum_{k=0}^{n-1} \left(\Omega^{N,\theta'}_{k(\Delta_n)}\right)^{-1} \sum_{i=1}^{N} \omega^{i,\theta'}_{k(\Delta_n)} s_k^{\bar{\alpha}}\left(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)};\theta\right) \ , \tag{2.19}$$

where, for $(x, x') \in \mathsf{X}^2$,

$$s_k^{\bar{\alpha}}(x, x'; \theta) \overset{\text{def}}{=} \bar{\alpha}^{-1} \sum_{\ell=1}^{\bar{\alpha}} \bar{V}_\theta^\ell(x, x') + \log g_\theta(x', Y_{k+1})$$

and

$$\bar{V}_\theta^{1:\bar{\alpha}}(x, x') \sim \bar{P}_\theta^{\otimes \bar{\alpha}}(x, x', \cdot) \; .$$

In (2.19) we have added θ' as an index to the particles as well as the associated weights to indicate that the particle system of the fixed-lag smoother is evolved under the dynamics determined by the initial parameter value.

2.3.1. Convergence of the intermediate quantity

Under weak assumptions on the functions Ψ_k, the kernels L_k and $\bar{P}$, and the local likelihoods functions $\log g_\theta(\cdot, Y_k)$ one may establish the convergence of the GPEPS-based fixed-lag approximation $\mathcal{Q}_n^N$ defined in (2.19). Define, for a given lag Δ_n and parameters (θ, θ'),

$$b_n(\Delta_n, \theta, \theta') \overset{\text{def}}{=} \sum_{k=0}^{n-1} \int s_k(x_{k:k+1}, \theta)\, \phi_{k(\Delta_n)}(dx_{k:k+1}, \theta') - \sum_{k=0}^{n-1} \int s_k(x_{k:k+1}, \theta)\, \phi_n(dx_{k:k+1}, \theta') \; ; \qquad (2.20)$$

then the following result, which is the main result of this section, establishes the pointwise convergence (in probability as N tends to infinity) of $\mathcal{Q}_n^N$ to a limit quantity that differs from the true intermediate quantity $\mathcal{Q}_n$ by b_n. Consequently, b_n is the bias imposed by the lag. Note that this convergence does not follow directly from Proposition 2.1, since the latter result states convergence of the GPEPS for given deterministic target functions only (whereas the terms of the complete data log-likelihood lack closed-form expressions and thus have to be replaced by random estimates obtained using the GPE).

Theorem 2.1. *Assume (A1–3). Let $n \geq 0$, $(\theta, \theta') \in \Theta^2$, and $(\Delta_n, \alpha, \bar{\alpha}) \in \mathbb{N}^3$. Suppose that (A4) holds for $\Psi_k(\cdot; \theta')$, $L_k(\cdot; \theta')$, and $\phi_k(\cdot; \theta')$ and that the initial sample $(\xi_0^{i,\theta'}, \omega_0^{i,\theta'})_{i=1}^N$ is consistent for $(\phi_0(\cdot; \theta'), \mathsf{L}^1(\phi_0(\cdot; \theta'), \mathsf{X}))$. Moreover, assume that the mappings $x_{0:k(\Delta_n)} \mapsto \log g_\theta(x_k, Y_k)$, $0 \leq k \leq n$, and $x_{0:k(\Delta_n)} \mapsto \int |v| \bar{P}_\theta(x_k, x_{k+1}, dv)$, $0 \leq k < n$, belong to $\mathsf{L}^1(\phi_{k(\Delta_n)}(\cdot; \theta'), \mathsf{X}^{k(\Delta_n)+1})$. Then, as $N \to \infty$,*

$$\mathcal{Q}_n^N(\theta, \theta') \xrightarrow{\mathbb{P}} \mathcal{Q}_n(\theta, \theta') + b_n(\Delta_n, \theta, \theta') \; ,$$

where the bias b_n is defined in (2.20).

The proof of Theorem 2.1 is found in Appendix A.2.

The bias term b_n, which was studied by [21], is controlled by the speed with which the hidden chain $(X_k)_{k\geq 0}$ forgets its initial distribution when evolving *conditionally* on the observations. Indeed, when the state space X is compact it can be shown [see 21, for details] that b_n is $\mathcal{O}(n\rho^{\Delta_n})$, where $0 < \rho < 1$ is the *uniform* (with respect to observation records $Y_{0:n}$ as well as initial distributions χ) mixing coefficient of the conditional chain. From this we deduce that the lag Δ_n should be increased with n at the minimum rate $c \log n$, $c > -1/\log \rho$, in order to keep the bias suppressed. Increasing Δ_n faster eliminates the bias and increases the variance of the approximation; see again Olsson et al. [21] for a detailed study of these issues. Since a similar forgetting property holds also in the case of a non-compact state space X [9], the same arguments can be applied for very general models; however, the analysis of the general case is significantly more involved since the mixing coefficient is neither uniform with respect to observation records nor initial distributions χ in this case.

Remarkably, the convergence stated in Theorem 2.1 holds for any *fixed* sample sizes $(\alpha, \bar{\alpha})$. In particular, nothing prevents us from setting $\alpha = \bar{\alpha} = 1$, yielding a fast algorithm; this is indeed the choice made in Section 3. However, more understanding of how to select optimally the sample size $\bar{\alpha}$ can be gained by proceeding as in the discussion following Proposition 2.1. Indeed, let $\mathcal{F}_n^N$ denote the σ-algebra generated by all random numbers of the GPEPS up to time n and write

$$\begin{aligned}
\operatorname{Var}\left(\mathcal{Q}_n^N(\theta;\theta')\right) &= \mathbb{E}\left[\operatorname{Var}\left(\mathcal{Q}_n^N(\theta;\theta') \middle| \mathcal{F}_n^N\right)\right] + \operatorname{Var}\left(\mathbb{E}\left[\mathcal{Q}_n^N(\theta;\theta') \middle| \mathcal{F}_n^N\right]\right) \\
&= \bar{\alpha}^{-1} \sum_{k=0}^{n-1} \mathbb{E}\left[\left(\Omega_{k(\Delta_n)}^{N,\theta'}\right)^{-2} \sum_{i=1}^{N} \left(\omega_{k(\Delta_n)}^{i,\theta'}\right)^2 \sigma_{\bar{P}_\theta}^2\left(\xi_{k:k+1|k(\Delta_n)}^{i,\theta'}\right)\right] \\
&\quad + \operatorname{Var}\left(\sum_{k=0}^{n-1} \int s_k\left(x_{k:k+1};\theta\right) \phi_{k(\Delta_n)}^N(dx_{k:k+1};\theta')\right) ,
\end{aligned} \tag{2.21}$$

where $\sigma_{\bar{P}_\theta}^2(x,x') \overset{\text{def}}{=} \int \{v - \log q_\theta(x,x')\}^2 \, \bar{P}_\theta(x,x',dv)$, $(x,x') \in \mathsf{X}^2$, is the conditional variance of the log-density estimator. Note that the second term in the expression above corresponds to the variance of an ideal fixed-lag approximation where the terms of the complete data log-likelihood is supposed to be known on closed-form (and thus involving $(s_k)_{k=0}^{n-1}$ instead of $(s_k^{\bar{\alpha}})_{k=0}^{n-1}$). To obtain the last equality in (2.21) we used the conditional unbiasedness and independence of the GPE draws estimating the latent chain log-density. By applying theory derived in the companion paper [23] it can be established (see Theorem 3.2 in the paper in question) that the first sum on the RHS of (2.21) behaves asymptotically like $\hat{\varsigma}^2/(\bar{\alpha}N)$ where $\hat{\varsigma}^2$ can be interpreted as expected GPE variance. Moreover, in a work in progress we establish a CLT for fixed-lag approximations of type (2.18), implying that the last term on the RHS of (2.21) behaves asymptotically like $\hat{\sigma}^2/N$, where $\hat{\sigma}^2$ is the asymptotic variance of a fixed-lag approximation of the intermediate quantity for a known complete data log-likelihood. Consequently, the arguments of the discussion following Proposition 2.1 apply immediately and we obtain, for a given available computational time τ, the optimal sample size

$$\bar{\alpha}_{\text{opt}} = \sqrt{\frac{\hat{\varsigma}^2/\tau_{\text{GPE}}}{\hat{\sigma}^2/\tau_{\text{GPEPS}}}} \, ,$$

where τ_{GPE} is the average computational time needed for producing a single unbiased estimate of the log-density and τ_{GPEPS} the average time needed for updating (in terms of selection, mutation, and random weight association) a single particle of the GPEPS.

3. Simulation study

3.1. Log-growth model

In the first example we estimate the parameters of the so-called *log-growth model*

$$dX_t = \kappa X_t(1 - X_t/\gamma)\, dt + \sigma X_t\, dW_t \tag{3.1}$$

(discussed in [4]) from simulated data. In our framework, we assume that we access only noisy observations $(Y_k)_{k \geq 0}$ of the process (3.1) according to

$$Y_k = X_{0.6k} + \epsilon_k\ , \tag{3.2}$$

where $(\epsilon_k)_{k \geq 0}$ are i.i.d. Gaussian random variables with zero mean and standard deviation 50. The noise sequence $(\epsilon_k)_{k \geq 0}$ is supposed to be independent of the Brownian motion $(W)_{t \geq 0}$ driving the latent process. Applying Itô's formula to the transformation $\tilde{X}_t = \eta(X_t, \sigma)$, with $\eta(x, \sigma) \overset{\text{def}}{=} -\log(x)/\sigma$, yields

$$d\tilde{X}_t = \beta(\tilde{X}_t) + dW_t\ , \tag{3.3}$$

where $\beta(x) \overset{\text{def}}{=} \sigma/2 - \kappa/\sigma + \kappa/(\sigma\gamma)\exp(-\sigma x)$. Since β is bounded from above, we are only required to simulate the minimum of the Brownian path and let $\tilde{W}_\beta^-$ be β evaluated at this minimum (see Section B for the meaning of $\tilde{W}_\beta^-$). The minimum of the Brownian bridge has a known law and conditionally on the minimum the bridge can be simulated retrospectively using Bessel bridges [see 4]. Our aim is to estimate all the unknown parameters $\theta \overset{\text{def}}{=} (\kappa, \gamma, \sigma)$ of (3.1) given a record $Y_{0:200}$ of observations obtained through simulation under the parameters $\theta^* = (0.1, 1000, 0.1)$.

To get an idea of how the quality of the proposed particle approximation of the EM intermediate quantity is influenced by the lag, we computed lag-based approximations of $\mathcal{Q}_{200}(\theta, \theta')$ evaluated on the diagonal $\theta = \theta' = (0.15, 1100, 0.15)$. Here the number of particles was fixed to $N = 400$, while the lag Δ_{200} was varied over the values $\{1, 2, 3, 4, 6, 20\}$. The outcome is displayed in the box-and-whisker diagram in Figure 1 together with a reference value (the dashed line) obtained as the arithmetic average of 10 values obtained by executing, equally many times, the smoother with the relatively large particle sample size $N = 10,000$ and lag $\Delta_{200} = 20$ (for which the bias is negligible). As seen in the figure, the lag-based approximation clearly suffers from bias for small lags; however, already at the lag $\Delta_{200} = 4$ the bias appears to be eliminated, and increasing the lag further only adds undesired variance to the estimates. This

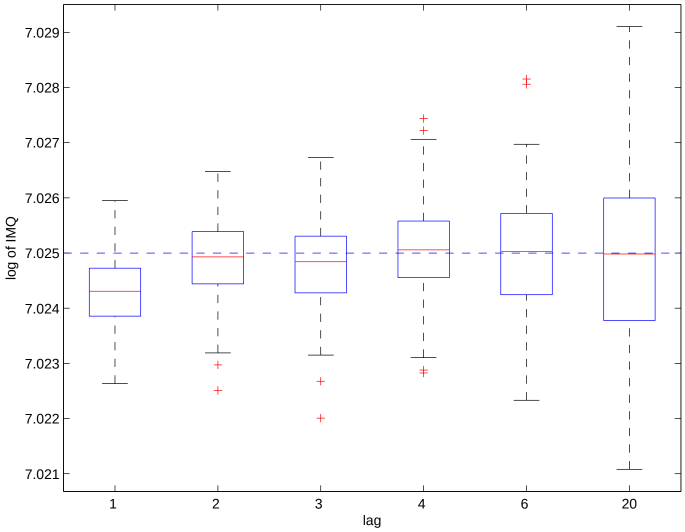

FIG 1. *Box-and-whisker diagram (on logarithmic scale) of estimates of $\mathcal{Q}_{200}(\theta, \theta')$ for the partially observed log-growth model* (3.1). *Here $\theta = \theta' = (0.15, 1100, 0.15)$ and the estimates are obtained using GPE-based fixed-lag smoothers with varying lags $\Delta_{200} = \{1, 2, 3, 4, 6, 20\}$. Each box (with adjuvant whisker) contains* 200 *estimates, where each estimate is based on $N = 400$ particles. The dotted line is a reference value obtained as the arithmetic average of the output of* 10 *independent smoothers using each $N = 10,000$ particles and lag $\Delta_n = 20$.*

indicates a quite strong forgetting in the model under consideration due to the relatively large distance between the observations.

Next, we implemented a full MCEM algorithm providing the likelihood estimate of θ for the same 200 observations. This algorithm combines the Monte Carlo-based E-step above with a subsequent maximization step and loops the two 50 times. At the E-step of each iteration, an approximation of the intermediate quantity was obtained using a fixed-lag smoother with lag $\Delta_{200} = 4$, i.e. the optimal lag obtained in the previous, and a particle sample size N_ℓ that was, in order to obtain convergence, increased with the iteration index ℓ according to $N_\ell = 100\sqrt{\ell + 1}$. To increase the Monte Carlo sample size at a polynomial rate is in line with the recommendation of [13] for the case of MCMC-based MCEM. The M-step was carried through using the *Nelder-Mead simplex algorithm* as implemented in MATLAB's `fminsearch`-command and in order to gain computational speed we used consequently $\alpha = \bar{\alpha} = 1$. At the mutation step the random weight fixed-lag smoother used the proposal

$$R_k(x, A) = \frac{1}{\sigma x} \int_A t_4 \left(\frac{x' - \kappa x(1 - x/\gamma)}{\sigma x} \right) dx' \ , \qquad (3.4)$$

obtained by discretising the hidden dynamics according to the Euler scheme. Here t_4 denotes the density of the student's t-distribution with 4 degrees of freedom. Further the adjustment multiplier weights are set to unity. This full MCEM algorithm, which was initialised with $\theta_0 = (0.7, 1500, 0.7)$, was executed

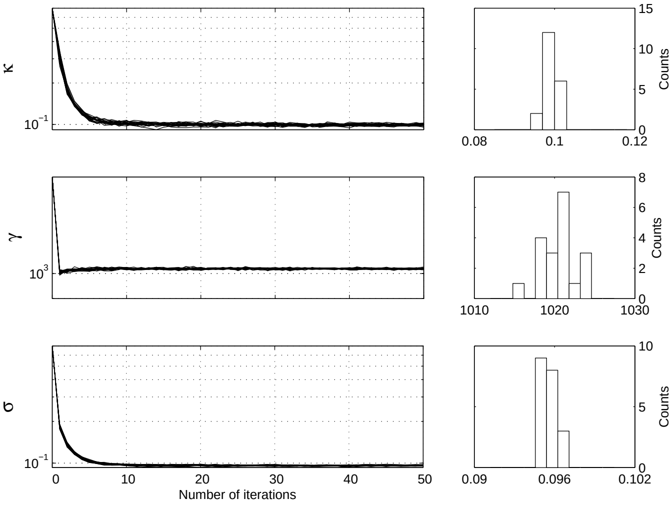

Fig 2. *EM learning curves (on logarithmic scale) for the parameters of the partially observed log-growth model* (3.1)*. The E-step of the MCEM algorithm was carried through by running the fixed-lag smoother with lag* $\Delta_{200} = 4$*. Each plot overlays* 20 *independent realizations and the histograms refer to the values of the last (i.e. the* 50*th) EM iteration.*

Table 1
Means and standard deviations of the MCEM parameter estimates plotted in the histograms of Figure 2 and Figure 3

Algorithm	κ	γ	σ
standard smoother	0.0990	1020	0.0959
	std. 0.0064	std. 3.5	std. 0.0031
fixed-lag smoother	0.0994	1020	0.0962
	std. 0.0018	std. 2.8	std. 0.0007

repeatedly 20 times, resulting in the bundle of EM learning trajectories plotted in Figure 2. We have chosen to plot the trajectories on logarithmic scale due to the large deviation of the initial values. The same figure also displays histograms of the parameter output pertaining to the last (i.e. the 50th) iteration. As seen in the figure, the learning trajectories converge smoothly around the parameter values $\hat{\theta} = (0.099, 1020, 0.096)$. For a comparison, the same experiment was repeated when the fixed-lag smoother was replaced by the standard random weight particle smoother using the complete genealogical history of the particles; see Figure 3. Table 1 displays means and standard deviations of the 20 parameter estimates obtained at the last iteration and evidently the fixed-lag method outperforms, in terms of standard deviation, the standard method by about a factor of 4 for the κ and σ parameters and a factor of 1.25 for the γ parameter. The arithmetic average of the final parameter estimates are more or less indistinguishable for the two approaches, re-confirming that the choice of the lag is indeed consistent with the speed of the forgetting of the model.

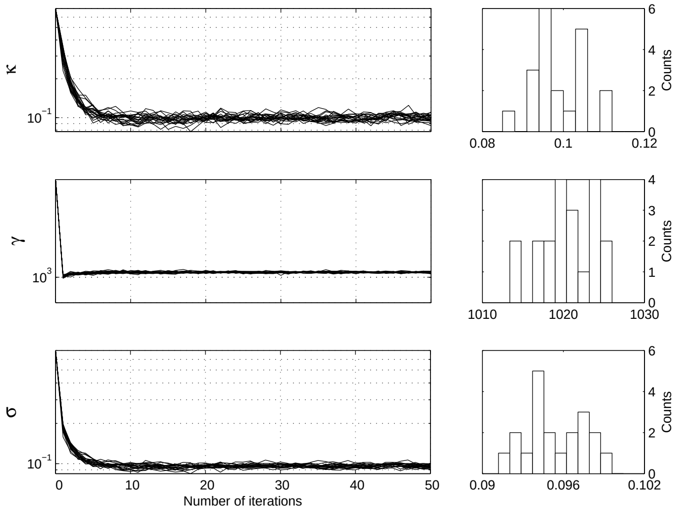

Fig 3. *The same experiment as in Figure 2 with the difference that the E-step of the MCEM algorithm was carried through using standard particle smoothing based on the full genealogical history of the particle trajectories.*

3.2. Genetics diffusion model

In a second example we consider noisy observations of the so-called *genetics diffusion model* presented in [19] and discussed in [3]. More specifically, we let the observations be generated according to

$$\begin{aligned} dV_t &= (\mu + \nu V_t)\, dt + \sigma V_t(1 - V_t)\, dW_t \ , \\ Y_k &= V_k + \epsilon_k \ , \end{aligned} \tag{3.5}$$

where the noise sequence $(\epsilon_k)_{k \geq 0}$ consists of i.i.d. Gaussian variables with zero mean and standard deviation 0.1. In this setting we used the proposed MCEM-algorithm to estimate the unknown parameters $\theta \overset{\text{def}}{=} (\mu, \nu, \sigma)$ given a record $Y_{0:1000}$ of observations obtained through simulation under the parameters $\theta^* = (0.05, 0.1, 1)$. Applying Itô's formula to the transformation $\tilde{X}_t = \eta(V_t, \sigma)$, where $\eta(v, \sigma) \overset{\text{def}}{=} (\log(v) - \log(1 - v))/\sigma$, allows for using the GPE for estimating the transition density of the latent process. In this case, the drift function β of the transformed process becomes more involved than in the previous example, and it is neither bounded from above nor below. Thus, we have to draw both $\tilde{W}_\beta^-$ and $\tilde{W}_\beta^+$ and a Brownian bridge $(\tilde{W}_s)_{s=0}^t$ such that $\tilde{W}_\beta^- \leq \beta(\tilde{W}_s) \leq \tilde{W}_\beta^+$ for all $0 \leq s \leq t$; see Section B for a justification of this. For this purpose we apply the method proposed in [3], which involves sampling first a maximum $\tilde{W}_{\text{id}}^+$ and a minimum $\tilde{W}_{\text{id}}^-$ and then a Brownian bridge such that $\tilde{W}_{\text{id}}^- \leq \tilde{W}_s \leq \tilde{W}_{\text{id}}^+$ for all $0 \leq s \leq t$. Since a linear transformation of a Brownian bridge is still a Brownian bridge, it suffices to consider the case when the path $(\tilde{W}_s)_{s=0}^t$ is conditioned to start and end in zero. Sampling a lower and upper bound can then be done by using rejection sampling in the following way: let $(a_i)_{i\geq 0}$ with $a_0 = 0$ be an increasing sequence and consider the intervals $(-a_i, a_i]$. Since the probability that a Brownian bridge stays in a specific interval $[-K, K]$ has a known expression (having the form of an infinite series), it is possible to calculate the probability that it is contained in $(-a_i, a_i]$ but not in $(-a_{i-1}, a_{i-1}]$; this means that either its maximum is contained in $(a_{i-1}, a_i]$ or its minimum is contained in $(-a_i, -a_{i-1}]$ or both. Thus, we first propose an interval $(a_{i-1}, a_i]$; given this interval, we then propose, with probability 1/2, a maximum conditioned to belong to $(a_{i-1}, a_i]$, otherwise a minimum in $(-a_i, -a_{i-1}]$. Since the distributions of the maximum and minimum are known on closed-form, this is easily done. Next, we propose a Brownian bridge by decomposing around the proposed maximum (minimum) as in the previous example. The resulting path $(\tilde{W}_s)_{s=0}^t$ is accepted, with a probability depending on the path in question, only if it remains in the interval; see [3] for details. Finally, we set $\tilde{W}_\beta^\pm \overset{\text{def}}{=} \beta(\tilde{W}_{\text{id}}^\pm)$.

For brevity, we do not repeat the extensive simulation study of the previous example in order to extract the optimal lag; instead we hedge—ad hoc—with the value $\Delta_n = 20$. Since the state space $\mathbb{R}(0, 1)$ is compact, we may propose the particles simply by using the uniform distribution over $(0, 1)$ as independent sampler. As in the previous example, we set $\alpha = \bar{\alpha} = 1$. The MCEM-algorithm, in which the M-step was again carried through using the Nelder-Mead simplex algorithm, was executed 25 iterations and as in the previous example the particle sample size was increased with the iteration index as $N_\ell = 100\sqrt{\ell + 1}$. Figure 4 displays the resulting EM learning curves.

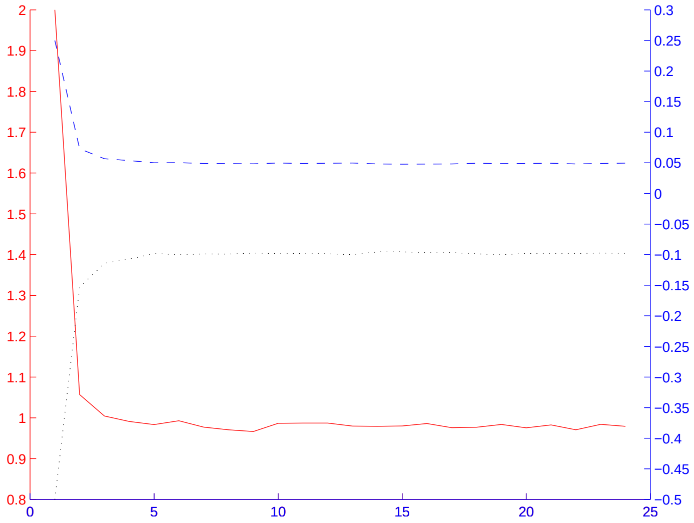

Fig 4. *Convergence of σ (continuous line, left y-axis), ν (dashed line, right y-axis) and μ (dotted line, right y-axis).*

4. Conclusion

We have proposed an EM-based method for estimating unknown parameters of PODs. The method combines recent approaches to efficient estimation of the joint smoothing distribution in hidden Markov models with recently proposed techniques of estimating, without bias, transition densities of a large class of diffusion processes via GPEs. Interestingly, the GPE provides a way of producing unbiased estimates of the transition densities *simultaneously* for all parameter values; this is critical when carrying through the maximisation-step of the resulting EM-algorithm. For models having forgetting properties, the degeneracy of the particle trajectories can be efficiently avoided by means of fixed-lag smoothing [18, 21]. The decrease of variance gained by the fixed-lag approximation is obtained at the cost of a bias; the bias is however easily controlled by increasing logarithmically the size of the lag with the size of the observation record, yielding an algorithm of $\mathcal{O}(N)$ computational complexity. We have provided a detailed study of the convergence of the GPE-based particle smoother as well as the full intermediate quantity of EM. The results were obtained under, what we believe, minimal assumptions and may, since we analyse separately the GPE-based mutation step (Lemma A.1), be extended to any selection schedule for which consistency has been established in the literature. In this way, our GPEPS convergence results differ significantly from that presented in [12]. The method was successfully demonstrated on two examples.

Finally we should mention that there exist alternative techniques, either Monte Carlo-based [see e.g. 24] or based on basis expansions [1], for approximating the transition density. Nevertheless, none of these approaches produce unbiased estimates. The former is, while quite general, computationally very demanding and the latter is only valid for very short time intervals (recall that the performance of the GPE is independent of the size of the time grid). Sometimes more direct numerical approaches, such as solving the Fokker-Plank equations or taking the Fourier inverse of the characteristic function of the SDE, are possible; however, these methods often tend to be computationally expensive. Anyway, the theoretical results obtained by us presume only unbiasedness of the transition density estimator, and thus other approximation schemes may be applicable within our framework.

Acknowledgements

The authors thank the anonymous referees for insightful comments that improved the presentation of the paper.

Appendix A: Proofs

The proofs of Proposition 2.1 and Theorem 2.1 rely on recent results on limit theorems for weighted samples obtained by [10]. Since we in this section deal exclusively with asymptotic properties of the sample as the sample size tends to infinity, we let, when not specified differently, the limit notation → refer to an *increasing number N of particles* only. In addition, we let also the particles and the associated weights be indexed by N for clearness. The following kernel notation will be useful in the following: Let μ be a measure on $(\Xi, \mathcal{B}(\Xi))$, f a measurable function on $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$, and K a kernel from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$; then we set

$$\mu K(A) \stackrel{\text{def}}{=} \int \mu(d\xi)\, K(\xi, A)$$

and

$$K(\xi, f) \stackrel{\text{def}}{=} \int f(\tilde{\xi})\, K(\xi, d\tilde{\xi}) \ .$$

The following definition specifies the structure that we want any class of estimand functions to have.

Definition A.1. A set C of measurable functions on Ξ is *proper* if the following holds.

(i) C is a linear space; that is, if f and g belong to C and $(\alpha, \beta) \in \mathbb{R}^2$, then $\alpha f + \beta g \in \mathsf{C}$;
(ii) if $g \in \mathsf{C}$ and f is measurable with $|f| \leq |g|$, then $f \in \mathsf{C}$;
(iii) for all $c \in \mathbb{R}$, the constant function $\xi \mapsto c$ belongs to C.

We will frequently make use of the following lemma obtained by Douc and Moulines [10]. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_{N,i})_{i=0}^N$, $N \geq 1$, a triangular array of sub-σ-fields of $\mathcal{F}$ such that $\mathcal{F}_{N,i-1} \subseteq \mathcal{F}_{N,i}$ for all $1 \leq i \leq N$ and $N \geq 1$. In addition, let $(U_{N,i})_{i=1}^N$, $N \geq 1$, be a triangular array of random variables such that each $U_{N,i}$ is $\mathcal{F}_{N,i}$-measurable.

Theorem A.1 ([10])**.** *Assume that* $\mathbb{E}\left[|U_{N,j}||\mathcal{F}_{N,j-1}\right] < \infty$, $\mathbb{P}$*-a.s., for all* $N \geq 1$ *and* $1 \leq j \leq N$*. Suppose that*

(i) as $\lambda \to \infty$,

$$\sup_{N \geq 1} \mathbb{P}\left(\sum_{j=1}^N \mathbb{E}\left[|U_{N,j}|\,|\, \mathcal{F}_{N,j-1}\right] \geq \lambda\right) \longrightarrow 0 \ ; \tag{A.1}$$

(ii) in addition, for all $\epsilon > 0$,

$$\sum_{j=1}^N \mathbb{E}\left[|U_{N,j}|; |U_{N,j}| \geq \epsilon \,|\, \mathcal{F}_{N,j-1}\right] \xrightarrow{\mathbb{P}} 0 \tag{A.2}$$

as $N \to \infty$*. Then*

$$\max_{1 \leq i \leq N} \left| \sum_{j=1}^i U_{N,j} - \sum_{j=1}^i \mathbb{E}\left[U_{N,j} \,|\, \mathcal{F}_{N,j-1}\right] \right| \xrightarrow{\mathbb{P}} 0 \ .$$

A.1. Proof of Proposition 2.1

Algorithm 1 is conveniently analysed within a more general framework of *random weight mutation* (RWM). Assume that we are given a $\boldsymbol{\Xi}$-valued, weighted particle sample $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$ which is consistent for some measure ν on $\mathcal{B}(\boldsymbol{\Xi})$ and let L be a finite transition kernel from $(\boldsymbol{\Xi}, \mathcal{B}(\boldsymbol{\Xi}))$ to $(\tilde{\boldsymbol{\Xi}}, \mathcal{B}(\tilde{\boldsymbol{\Xi}}))$. We wish to transform $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$ into another sample $(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})_{i=1}^N$ targeting the measure

$$\mu(A) = \frac{\nu L(A)}{\nu L(\tilde{\boldsymbol{\Xi}})} \;, \quad A \in \mathcal{B}(\tilde{\boldsymbol{\Xi}}) \;,$$

by means of the RWM operation described below. The input parameters are: a proposal kernel R such that $R(\xi, \cdot)$ dominates $L(\xi, \cdot)$ for all $\xi \in \boldsymbol{\Xi}$, a random weight kernel S from $(\boldsymbol{\Xi} \times \tilde{\boldsymbol{\Xi}}, \mathcal{B}(\boldsymbol{\Xi} \times \tilde{\boldsymbol{\Xi}}))$ to $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ targeting dL/dR in the sense that, for all $(\xi, \tilde{\xi}) \in \boldsymbol{\Xi} \times \tilde{\boldsymbol{\Xi}}$,

$$\int v \, S(\xi, \tilde{\xi}, dv) = \frac{dL(\xi, \cdot)}{dR(\xi, \cdot)}(\tilde{\xi}) \;,$$

and, finally, a Monte Carlo sample size $\alpha \in \mathbb{N}$.

```
Algorithm 2
(∗ random weight mutation ∗)
Input: $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$, $R$, $S$, $\alpha$
for $i \leftarrow 1$ to $N$
    do simulate $\tilde{\xi}_{N,i} \sim R(\xi_{N,i}, \cdot)$;
       simulate $V^{1:\alpha}(\xi_{N,i}, \tilde{\xi}_{N,i}) \sim S^{\otimes\alpha}(\xi_{N,i}, \tilde{\xi}_{N,i}, \cdot)$;
       $\tilde{\omega}_{N,i} \leftarrow \omega_{N,i} \alpha^{-1} \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i})$;
return $(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})_{i=1}^N$.
```

The sample $(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})_{i=1}^N$ returned by the algorithm is taken as an approximation of μ. In order to evaluate the quality of this sample, define the set

$$\tilde{\mathsf{C}} \overset{\text{def}}{=} \left\{ f \in \mathsf{L}^1(\mu, \tilde{\boldsymbol{\Xi}}) : L(\cdot, |f|) \in \mathsf{C} \right\} \;; \tag{A.3}$$

then the following result stating consistency for weighted samples produced by Algorithm 2 is instrumental when establishing Proposition 2.1.

Lemma A.1. *Assume the weighted sample $(\xi_{N,i}, \omega_{N,i})_{i=1}^N$ is consistent for (ν, C) and that the function $L(\cdot, \tilde{\boldsymbol{\Xi}})$ belongs to C. Then the set $\tilde{\mathsf{C}}$ defined in (A.3) and the weighted particle sample $(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})_{i=1}^N$ produced by Algorithm 2 are proper resp. $(\mu, \tilde{\mathsf{C}})$-consistent for any fixed $\alpha \in \mathbb{N}$.*

Proof. Properness of the set $\tilde{\mathsf{C}}$ is straightforwardly established: To check Property (i) in Definition A.1, suppose that f and g belong to $\tilde{\mathsf{C}}$ and let $(\alpha, \beta) \in \mathbb{R}^2$;

then

$$\begin{aligned}\iint |\alpha f(\tilde{\xi}) + \beta g(\tilde{\xi})| v \, S(\cdot, \tilde{\xi}, dv) \, R(\cdot, d\tilde{\xi}) \\ \leq |\alpha| \iint |f(\tilde{\xi})| v \, S(\cdot, \tilde{\xi}, dv) \, R(\cdot, d\tilde{\xi}) \\ + |\beta| \iint |g(\tilde{\xi})| v \, S(\cdot, \tilde{\xi}, dv) \, R(\cdot, d\tilde{\xi}) \\ = |\alpha| L(\cdot, |f|) + |\beta| L(\cdot, |g|) \, ,\end{aligned}$$

where the function on the right hand side belongs to C by construction of $\tilde{\mathsf{C}}$ and the fact that C is a linear space. That the integral on the left hand side belongs to C is now a consequence of Property (ii) in Definition A.1. Properties (ii) and (iii) are checked in a similar manner.

To establish Condition (2.12) in Definition 2.1 it is enough to show that, for all $f \in \tilde{\mathsf{C}}$,

$$\Omega_N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \nu L(f) \; ; \tag{A.4}$$

indeed, since $\tilde{\mathsf{C}}$ contains the unity mapping $\tilde{\xi} \mapsto 1$ (as $\tilde{\mathsf{C}}$ is proper), (A.4) implies that

$$\Omega_N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{N,i} \xrightarrow{\mathbb{P}} \nu L(\tilde{\boldsymbol{\Xi}}) \, , \tag{A.5}$$

from which Condition (2.12) in Definition 2.1 follows by Slutsky's lemma. Thus, we define the triangular array $U_{N,i} \overset{\text{def}}{=} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i}) / \Omega_N$, $N \geq 1$, $1 \leq i \leq N$, and sub-σ-fields $\mathcal{F}_N \overset{\text{def}}{=} \sigma\{(\xi_{N,i}, \omega_{N,i})_{i=1}^N\}$, $N \geq 1$. We then get, by applying the tower property of conditional expectations and the consistency of the ancestor sample,

$$\begin{aligned}\sum_{i=1}^{N} \mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_N \right] \\ = \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} \mathbb{E}\left[\mathbb{E}\left[\alpha^{-1} \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \middle| \tilde{\xi}_{N,i}, \mathcal{F}_N \right] f(\tilde{\xi}_{N,i}) \middle| \mathcal{F}_N \right] \\ = \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} \int f(\tilde{\xi}) \int v \, S(\xi_{N,i}, \tilde{\xi}, dv) \, R(\xi_{N,i}, d\tilde{\xi}) \\ = \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} L(\xi_{N,i}, f) \xrightarrow{\mathbb{P}} \nu L(f) \, ,\end{aligned}$$

since $L(\cdot, f) \leq L(\cdot, |f|) \in \mathsf{C}$. To show that $\sum_{i=1}^N U_{N,i}$ tends to $\sum_{i=1}^N \mathbb{E}[U_{N,i} | \mathcal{F}_N]$ in probability, implying (A.4), we apply Theorem A.1. In order to establish the first condition of that theorem we reuse the arguments above and use that

$L(\cdot, |f|) \in \mathsf{C}$, yielding the limit

$$\sum_{i=1}^{N} \mathbb{E}\left[|U_{N,i}| \,|\, \mathcal{F}_N \right] \xrightarrow{\mathbb{P}} \nu L(|f|) \; .$$

Now, since convergence in probability implies tightness, we conclude that Condition (i) in Theorem A.1 is fulfilled.

To verify (ii), define, for some $\epsilon > 0$, $A_N \overset{\text{def}}{=} \sum_{i=1}^{N} \mathbb{E}[|U_{N,i}|; |U_{N,i}| \geq \epsilon | \mathcal{F}_N]$. Since, as the ancestor sample is assumed to be consistent, $\max_{1 \leq i \leq N} \omega_{N,i}/\Omega_N$ vanishes in probability as N tends to infinity, the same holds for the product $A_N \mathbb{1}\{C \max_{1 \leq i \leq N} \omega_{N,i} > \epsilon \Omega_N\}$, where $C > 0$ is an arbitrary constant. On the other hand,

$$\begin{aligned} A_N \mathbb{1}\left\{ C \max_{1 \leq i \leq N} \omega_{N,i} \leq \epsilon \Omega_N \right\} & \\ \leq \sum_{i=1}^{N} \mathbb{E}\left[|U_{N,i}|; |f(\tilde{\xi}_{N,i})| \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C \right| \left. \mathcal{F}_N \right] & \\ = \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} \int |f(\tilde{\xi})| \int_{|f(\tilde{\xi})| \sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes \alpha}(\xi_{N,i}, \tilde{\xi}, dv_{1:\alpha})\, R(\xi_{N,i}, d\tilde{\xi}) \; . \end{aligned}$$

Now, since, for all $\xi \in \mathbf{\Xi}$,

$$\int |f(\tilde{\xi})| \int_{|f(\tilde{\xi})| \sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes \alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi}) \leq L(\xi, |f|) \; ,$$

where $L(\cdot, |f|) \in \mathsf{C}$, we conclude, using Property (ii) of Definition A.1, that the mapping

$$\xi \mapsto \int |f(\tilde{\xi})| \int_{|f(\tilde{\xi})| \sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes \alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi})$$

on $\mathbf{\Xi}$ belongs to C as well. Thus, consistency of the ancestor sample implies that

$$\begin{aligned} \sum_{i=1}^{N} \mathbb{E}\left[|U_{N,i}|; |f(\tilde{\xi}_{N,i})| \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C \right| \left. \mathcal{F}_N \right] & \\ \xrightarrow{\mathbb{P}} \iint |f(\tilde{\xi})| \int_{|f(\tilde{\xi})| \sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes \alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi})\, \nu(\xi) \; . & \qquad \text{(A.6)} \end{aligned}$$

In addition, since the constant C may be chosen arbitrarily large, the limit (A.6) can be made arbitrarily small by the dominated convergence theorem. We hence conclude that A_N tends to zero in probability as N tends to infinity. This establishes (A.4).

In order to establish (2.13) it is, by Slutsky's theorem and (A.5), enough to prove that

$$\Omega_N^{-1} \max_{1 \leq i \leq N} \tilde{\omega}_{N,i} \xrightarrow{\mathbb{P}} 0 \; . \qquad \text{(A.7)}$$

Thus, take again a constant $C > 0$ and write

$$\Omega_N^{-1} \max_{1\leq i\leq N} \tilde{\omega}_{N,i} \mathbb{1}\left\{\sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C\right\} \leq \Omega_N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{N,i} \mathbb{1}\left\{\sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C\right\} . \quad \text{(A.8)}$$

To prove that the right hand side of (A.8) converges, we introduce the triangular array $U_{N,i} \overset{\text{def}}{=} \tilde{\omega}_{N,i} \mathbb{1}\{\sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C\}/\Omega_N$, $N \geq 1$, $1 \leq i \leq N$, and let the sub-σ-fields $\mathcal{F}_N$, $N \geq 1$, be defined as above. Next, we use again Theorem A.1. To verify the first condition, take conditional expectation with respect to $\mathcal{F}_N$ and reuse (A.6) with f being the unity function; this yields

$$\sum_{i=1}^{N} \mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_N\right] \xrightarrow{\mathbb{P}} \iiint_{\sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes\alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi})\, \nu(d\xi) ,$$

implying (i). To verify (ii), take an $\epsilon > 0$ and define $A_N \overset{\text{def}}{=} \sum_{i=1}^{N} \mathbb{E}[|U_{N,i}|; |U_{N,i}| \geq \epsilon | \mathcal{F}_N]$. Then

$$A_N = \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} \mathbb{E}\left[V^1(\xi_{N,i}, \tilde{\xi}_{N,i}); \tilde{\omega}_{N,i} \geq \epsilon\Omega_N, \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha C \middle| \mathcal{F}_N\right],$$

implying that, for an arbitrary constant $C' > 0$, following the lines of (A.6),

$$\begin{aligned} &A_N \mathbb{1}\left\{C' \max_{1\leq i\leq N} \omega_{N,i} \leq \epsilon\Omega_N\right\} \\ &\quad \leq \Omega_N^{-1} \sum_{i=1}^{N} \omega_{N,i} \mathbb{E}\left[V^1(\xi_{N,i}, \tilde{\xi}_{N,i}); \sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) \geq \alpha(C \vee C') \middle| \mathcal{F}_N\right] \\ &\quad\quad \xrightarrow{\mathbb{P}} \iiint_{\sum_{\ell=1}^{\alpha} v_\ell \geq \alpha(C\vee C')} v_1 S^{\otimes\alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi})\, \nu(d\xi) . \end{aligned} \quad \text{(A.9)}$$

On the other hand,

$$\Omega_N^{-1} \max_{1\leq i\leq N} \tilde{\omega}_{N,i} \mathbb{1}\left\{\sum_{\ell=1}^{\alpha} V^{\ell}(\xi_{N,i}, \tilde{\xi}_{N,i}) < \alpha C\right\} \leq C\Omega_N^{-1} \max_{1\leq i\leq N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0 .$$

Thus, since the limit (A.9) can be made arbitrarily small by increasing C', we conclude that A_N tends to zero as N tends to infinity. This in turn implies that the upper bound in (A.8) tends to

$$\iiint_{\sum_{\ell=1}^{\alpha} v_\ell \geq \alpha C} v_1 S^{\otimes\alpha}(\xi, \tilde{\xi}, dv_{1:\alpha})\, R(\xi, d\tilde{\xi})\, \nu(d\xi) . \quad \text{(A.10)}$$

Finally, we complete the proof by noting that (A.10) can be made arbitrarily small by increasing C. □

We now use Lemma A.1 to prove consistency of Monte Carlo estimates produced by the GPEPS. For this purpose, let $\bar{\xi}^i_{0:k|k} \overset{\text{def}}{=} \xi^{I^i_k}_{0:k|k}$, $1 \leq i \leq N$, denote the selected particles obtained in Step (2) of Algorithm 1. Consequently, the sample $(\bar{\xi}^i_{0:k|k})_{i=1}^N$ is obtained by resampling the ancestor particles $(\xi^i_{0:k|k})_{i=1}^N$ multinomially with respect to the normalised adjusted weights $(\omega^j_k \psi^j_k / \sum_{\ell=1}^N \omega^\ell_k \psi^\ell_k)_{j=1}^N$. This operation will in the following be referred to as *selection*. Using this notation and terminology it is now possible to describe one iteration of the GPEPS by the following three transformations:

$$(\xi^i_{0:k|k}, \omega^i_k)_{i=1}^N \xrightarrow{\textbf{I}\text{: Weighting}} (\xi^i_{0:k|k}, \psi^i_k \omega^i_k)_{i=1}^N \to \\ \xrightarrow{\textbf{II}\text{: Selection}} (\bar{\xi}^i_{0:k|k}, 1)_{i=1}^N \xrightarrow{\textbf{III}\text{: Mutation}} (\xi^i_{0:k+1|k+1}, \omega^i_{k+1})_{i=1}^N \ .$$

Here the third operation refers to the random weight mutation procedure described in Algorithm 2.

To prove Proposition 2.1 we proceed by induction and assume that $(\xi^i_{0:k|k}, \omega^i_k)_{i=1}^N$ is consistent for $(\phi_k, \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_k))$. Next, we show how consistency is preserved through one iteration of the algorithm by analysing separately Steps (**I**–**III**).

Step **I**. Define the modulated smoothing measure

$$\phi_k \langle \Psi_k \rangle (A) \overset{\text{def}}{=} \frac{\phi_k(\Psi_k \mathbb{1}_A)}{\phi_k(\Psi_k)} \ , \quad A \in \mathcal{X}^{\otimes(n+1)} \ ;$$

then the weighting operation in Step **I** can be viewed as a transformation according Algorithm 2 with $\boldsymbol{\Xi} = \mathsf{X}^{n+1}$, $\tilde{\boldsymbol{\Xi}} = \mathsf{X}^{n+1}$, and

$$\begin{cases} \nu = \phi_k \ , \\ \mu = \phi_k \langle \Psi_k \rangle \ , \\ R(x_{0:k}, A) = \delta_{x_{0:k}}(A) \ , \\ L(x_{0:k}, A) = \Psi_k(x_{0:k}) \, \delta_{x_{0:k}}(A) \ , \\ S(x_{0:k}, x'_{0:k}, A) = \delta_{\Psi_k(x'_{0:k})}(A) \ . \end{cases}$$

Thus, by applying Lemma A.1 we conclude that $(\xi^i_{0:k|k}, \psi^i_k \omega^i_k)_{i=1}^N$ is consistent for $\phi_k \langle \Psi_k \rangle$ and the (proper) set

$$\left\{ f \in \mathsf{L}^1(\phi_k \langle \Psi_k \rangle, \mathsf{X}^{n+1}) : \Psi_k |f| \in \mathsf{L}^1(\phi_k, \mathsf{X}^{n+1}) \right\} = \mathsf{L}^1(\phi_k \langle \Psi_k \rangle, \mathsf{X}^{n+1}) \ .$$

Step **II**. Applying Theorem 3 in [10] gives immediately that $(\bar{\xi}^i_{0:k|k}, 1)_{i=1}^N$ is consistent for $[\phi_k \langle \Psi_k \rangle, \mathsf{L}^1(\phi_k \langle \Psi_k \rangle, \mathsf{X}^{n+1})]$ for both the selection schedules (C.1) and (C.2).

Step **III**. Also the third step is handled using Lemma A.1. In this case, we set $\boldsymbol{\Xi} = \mathsf{X}^{n+1}$, $\tilde{\boldsymbol{\Xi}} = \mathsf{X}^{n+2}$, and

$$\begin{cases} \nu = \phi_k \langle \Psi_k \rangle \ , \\ \mu = \phi_{k+1} \ , \\ R(x_{0:k}, A) = \int_A \delta_{x_{0:k}}(dx'_{0:k}) \, R_k(x'_k, dx'_{k+1}) \ , \\ L(x_{0:k}, A) = \int_A \Phi_k(x'_{0:k+1}) \, \delta_{x_{0:k}}(dx'_{0:k}) \, R_k(x'_k, dx'_{k+1}) \ , \\ S(x_{0:k}, x'_{0:k+1}, A) \\ \qquad = \int \mathbb{1}_A\{vg(x'_{k+1}, Y_{k+1})/[\Psi_k(x'_{0:k}) r_k(x'_k, x'_{k+1})]\} P(x'_k, x'_{k+1}, dv) \ , \end{cases}$$

where P is the GPE described in Section 2.1 (and in more detail in Appendix B). Thus, using Lemma A.1 yields that $(\xi^i_{0:k+1|k+1}, \omega^i_{k+1})_{i=1}^N$ is consistent for ϕ_{k+1} and the set

$$\left\{ f \in \mathsf{L}^1(\phi_{k+1}, \mathsf{X}^{k+2}) : L(\cdot, |f|) \in \mathsf{L}^1(\phi_k \langle \Psi_k \rangle, \mathsf{X}^{n+1}) \right\} = \mathsf{L}^1(\phi_{k+1}, \mathsf{X}^{k+2}) \ .$$

Finally, we complete the proof by noting that the induction hypothesis is fulfilled for $k = 0$ by assumption.

A.2. Proof of Theorem 2.1

Decompose the error according to

$$\begin{aligned} \mathcal{Q}_n^N(\theta, \theta') - \mathcal{Q}_n(\theta, \theta') \\ = \sum_{k=0}^{n-1} \Bigg[\left(\Omega^{N,\theta'}_{k(\Delta_n)} \right)^{-1} \sum_{i=1}^N \omega^{i,\theta'}_{k(\Delta_n)} s_k^{\bar{\alpha}} \left(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)}, \theta \right) \\ - \int s_k(x_{k:k+1}; \theta) \, \phi_{k(\Delta_n)} \, (dx_{k:k+1}; \theta') \Bigg] \\ + b_n(\Delta_n, \theta, \theta') \ , \end{aligned} \tag{A.11}$$

where the bracket terms are errors originating from the GPEPS and the second term b_n, defined in (2.20), is the cost of introducing the fixed lag. By combining Proposition 2.1 with Slutsky's theorem we conclude that

$$\begin{aligned} \sum_{k=0}^n \left(\Omega^{N,\theta'}_{k(\Delta_n)} \right)^{-1} \sum_{i=1}^N \omega^{i,\theta'}_{k(\Delta_n)} \log g_\theta \left(\xi^{i,\theta'}_{k|k(\Delta_n)}, Y_k \right) \\ \xrightarrow{\mathbb{P}} \sum_{k=0}^n \int \log g_\theta \left(x_k, Y_k \right) \, \phi_{k(\Delta_n)}(dx_k; \theta') \ , \end{aligned} \tag{A.12}$$

as $x_{0:k(\Delta_n)} \mapsto \log g_\theta(x_k, Y_k)$ belongs to $\mathsf{L}^1(\phi_{k(\Delta_n)}(\cdot; \theta'), \mathsf{X}^{k(\Delta_n)+1})$ by assumption. Thus, the second term of the intermediate quantity estimator (2.19) is consistent. In order to establish consistency of the complete estimator it remains to prove that

$$\sum_{k=0}^{n-1} \left(\bar{\alpha}\Omega^{N,\theta'}_{k(\Delta_n)}\right)^{-1} \sum_{i=1}^{N} \omega^{i,\theta'}_{k(\Delta_n)} \sum_{\ell=1}^{\bar{\alpha}} \bar{V}^{\ell}_{\theta}\left(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)}\right) \\ \xrightarrow{\mathbb{P}} \sum_{k=0}^{n-1} \int \log q_\theta\left(x_k, x_{k+1}\right) \phi_{k(\Delta_n)}(dx_{k:k+1};\theta') \;. \quad \text{(A.13)}$$

To do this, we define $\bar{U}_{N,i} \overset{\text{def}}{=} \omega^{i,\theta'}_{k(\Delta_n)} \sum_{\ell=1}^{\bar{\alpha}} \bar{V}^{\ell}_{\theta}(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)})/\bar{\alpha}\Omega^{N,\theta'}_{k(\Delta_n)}$ and $\bar{\mathcal{F}}_N \overset{\text{def}}{=} \sigma\{(\xi^{i,\theta'}_{0:k(\Delta_n)|k(\Delta_n)}, \omega^{i,\theta'}_{k(\Delta_n)})_{i=1}^{N}\}$ and appeal to Theorem A.1 and Proposition 2.1. Since $\log q_\theta(x_k, x_{k+1}) \le \int |v|\, \bar{P}_\theta(x_k, x_{k+1}, dv)$ for all $x_{k:k+1} \in \mathsf{X}^2$, the mapping $x_{0:k(\Delta_n)} \mapsto \log q_\theta(x_k, x_{k+1})$ belongs to $\mathsf{L}^1(\phi_{k(\Delta_n)}(\cdot;\theta'), \mathsf{X}^{k(\Delta_n)+1})$. Hence,

$$\sum_{i=1}^{N} \mathbb{E}\left[\bar{U}_{N,i}\,\middle|\bar{\mathcal{F}}_N\right] = \left(\Omega^{N,\theta'}_{k(\Delta_n)}\right)^{-1} \sum_{i=1}^{N} \omega^{i,\theta'}_{k(\Delta_n)} \log q_\theta\left(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)}\right) \\ \xrightarrow{\mathbb{P}} \int \log q_\theta(x_k, x_{k+1})\, \phi_{k(\Delta_n)}(dx_{k:k+1};\theta') \;, \quad \text{(A.14)}$$

from which we conclude that (A.13) may be established by verifying the two assumptions of Theorem A.1. Following (A.14) and using again that $x_{0:k(\Delta_n)} \mapsto \int |v| \bar{P}_\theta(x_k, x_{k+1}, dv)$ belongs to $\mathsf{L}^1(\phi_{k(\Delta_n)}(\cdot;\theta'), \mathsf{X}^{k(\Delta_n)+1})$ by assumption, we conclude that

$$\sum_{i=1}^{N} \mathbb{E}\left[\left|\bar{U}_{N,i}\right|\,\middle|\bar{\mathcal{F}}_N\right] \xrightarrow{\mathbb{P}} \iint |v|\, \bar{P}_\theta(x_k, x_{k+1}, dv)\, \phi_{k(\Delta_n)}(dx_{k:k+1};\theta') \;,$$

which verifies Assumption (i) (by tightness of sequences converging in probability). To verify (ii), let $\epsilon > 0$ and set $\bar{A}_N \overset{\text{def}}{=} \sum_{i=1}^{N} \mathbb{E}[|\bar{U}_{N,i}|; |\bar{U}_{N,i}| \ge \epsilon | \bar{\mathcal{F}}_N]$. Then, for any constant $C > 0$, by consistency of the particle sample,

$$\bar{A}_N \mathbb{1}\left\{C \max_{1\le i\le N} \omega^{i,\theta'}_{k(\Delta_n)} > \epsilon \Omega^{N,\theta'}_{k(\Delta_n)}\right\} \xrightarrow{\mathbb{P}} 0 \;. \quad \text{(A.15)}$$

On the other hand,

$$\bar{A}_N \mathbb{1}\left\{C \max_{1\le i\le N} \omega^{i,\theta'}_{k(\Delta_n)} \le \epsilon \Omega^{N,\theta'}_{k(\Delta_n)}\right\} \\ \le \sum_{i=1}^{N} \mathbb{E}\left[|\bar{U}_{N,i}|; \left|\sum_{\ell=1}^{\bar{\alpha}} \bar{V}^{\ell}_{\theta}\left(\xi^{i,\theta'}_{k:k+1|k(\Delta_n)}\right)\right| \ge C\bar{\alpha} \middle| \bar{\mathcal{F}}_N\right] \\ \le \left(\Omega^{N,\theta'}_{k(\Delta_n)}\right)^{-1} \sum_{i=1}^{N} \omega^{i,\theta'}_{k(\Delta_n)} \int_{|\sum_{\ell=1}^{\bar{\alpha}} v_\ell| \ge C\bar{\alpha}} |v_1|\, \bar{P}^{\otimes\bar{\alpha}}_{\theta}(x_k, x_{k+1}, dv_{1:\bar{\alpha}}) \;.$$

Now, since, for all $x_{k:k+1} \in \mathsf{X}^2$,

$$\int_{|\sum_{\ell=1}^{\bar{\alpha}} v_\ell| \geq C\bar{\alpha}} |v_1| \, \bar{P}_\theta^{\otimes \bar{\alpha}}(x_k, x_{k+1}, dv_{1:\bar{\alpha}}) \leq \int |v| \, \bar{P}_\theta(x_k, x_{k+1}, dv) \ ,$$

we get, using Proposition 2.1,

$$\begin{aligned} &\left(\Omega_{k(\Delta_n)}^{N,\theta'}\right)^{-1} \sum_{i=1}^{N} \omega_{k(\Delta_n)}^{i,\theta'} \int_{|\sum_{\ell=1}^{\bar{\alpha}} v_\ell| \geq C\bar{\alpha}} |v_1| \, \bar{P}_\theta^{\otimes \bar{\alpha}}(x_k, x_{k+1}, dv_{1:\bar{\alpha}}) \\ &\quad \xrightarrow{\mathbb{P}} \iint_{|\sum_{\ell=1}^{\bar{\alpha}} v_\ell| \geq C\bar{\alpha}} |v_1| \, \bar{P}_\theta^{\otimes \bar{\alpha}}(x_k, x_{k+1}, dv_{1:\bar{\alpha}}) \, \phi_{k(\Delta_n)}(dx_{k:k+1}; \theta') \ . \end{aligned} \tag{A.16}$$

We now note that the limit in (A.16) can be made arbitrarily small by increasing C. This verifies condition (ii) in Theorem A.1, which completes the proof of (A.13). Finally, combining (A.13) with (A.12) completes the proof of Theorem 2.1.

Appendix B: More on the GPE

The outline of this section follows Beskos et al. [4] and Fearnhead et al. [12], and we limit our scope to the one-dimensional case; multivariate extensions are treated by Beskos et al. [3]. Let $(C[0,t], \mathcal{C}[0,t])$ be the measurable space of continuous functions on $[0,t]$ and denote by $\mathbb{S}_\theta^{(x)}$ the law of $\tilde{X}$ on $(C[0,t], \mathcal{C}[0,t])$ for the initial condition $\tilde{X}_0 = W_0 = x$. Also, let $\mathbb{W}^{(t,x,x')}$ be the law, on the same space, of the Brownian bridge process $\tilde{W} = (\tilde{W}_s)_{0 \leq s \leq t}$ starting in x at time zero and ending in x' at time t. Similarly, denote by $\mathbb{S}_\theta^{(t,x,x')}$ the law of the *diffusion bridge* obtained when $\tilde{X}$ is conditioned to start at $\tilde{X}_0 = W_0 = x$ and to finish at $\tilde{X}_t = x'$. Recall the definition (2.1) of $\beta(\cdot, \theta)$ and let

$$A(u, \theta) \overset{\text{def}}{=} \int^u \beta(v, \theta) \, dv$$

be any antiderivative of $\beta(\cdot, \theta)$. The role of Assumptions *(A1–A3)* is to guarantee that $\mathbb{S}_\theta^{(t,x,x')}$ is absolutely continuous with respect to $\mathbb{W}^{(t,x,x')}$ with Radon-Nikodym derivative

$$\begin{aligned} &\frac{d\mathbb{S}_\theta^{(x,x',t)}}{d\mathbb{W}^{(x,x',t)}}(w) \\ &\quad = \frac{\mathcal{N}_t(x' - x)}{\tilde{q}_\theta(x, x', t)} \exp\left(A(x', \theta) - A(x, \theta) - \frac{1}{2} \int_0^t (\beta^2 + \beta')(w_s, \theta) \, ds \right) \ , \end{aligned} \tag{B.1}$$

where $w \in C[0,t]$ and $\mathcal{N}_t$ denotes the density function of the zero mean normal distribution with variance t. Now, define, for $u \in \mathbb{R}$, the *drift functional*

$$\phi(u, \theta) \overset{\text{def}}{=} \frac{\beta^2(u, \theta) + \beta'(u, \theta)}{2} - l(\theta) \ ,$$

where $l(\theta)$ is the lower bound given in Assumption *(A3)*. The transition density $\tilde{q}_\theta$ can, using (B.1), be expressed as

$$\tilde{q}_\theta(x, x', t) = \mathcal{N}_t(x' - x) \exp\left(A(x', \theta) - A(x, \theta) - l(\theta)t\right) \\ \times \int \exp\left(-\int_0^t \phi(w_s, \theta)\, ds\right) \mathbb{W}^{(t,x,x')}(dw) \;,$$

Accordingly, we wish to calculate expectations of the form

$$\int \exp\left(-\int_0^t f(w_s)\, ds\right) \mathbb{W}^{(t,x,x')}(dw) \;. \tag{B.2}$$

Now assume that it is possible to simulate simultaneously a pair $(\tilde{W}_f^-, \tilde{W}_f^+)$ of random variables and a trajectory $(\tilde{W}_s)_{s=0}^t$ such that

$$\tilde{W}_f^- \leq f(\tilde{W}_s) \leq \tilde{W}_f^+ \;, \quad \text{for all } s \in [0, t] \;;$$

in practice this will most often be carried through by first simulating a maximum and a minimum of the Brownian bridge process $\tilde{W}$ and hereafter interpolating, using Bessel bridges, the rest of the bridge conditionally on these. Let κ be a discrete random variable having, conditionally on $\tilde{W}_f^\pm$, probability distribution $p_t(\cdot|\tilde{W}_f^\pm)$. Then it is easily established that the GPE

$$\exp(-\tilde{W}_f^+ t) \frac{t^\kappa}{\kappa! p_t(\kappa|\tilde{W}_f^\pm)} \prod_{\ell=1}^{\kappa} [\tilde{W}_f^+ - f(\tilde{W}_{\psi_\ell})]$$

(associated with p_t) is an unbiased estimator of (B.2). Here $(\psi_\ell)_{\ell \geq 1}$ are mutually independent variables that are uniformly distributed over $[0, t]$ and independent of $\mathcal{F}_t$. Note that the distribution p_t can be chosen freely, yielding a *whole class* of GPEs, and an optimal choice is discussed by Fearnhead et al. [12]. In all applications considered in this paper we will use let κ be Poisson-distributed.

Using the Girsanov theorem, it can be shown that

$$\log \tilde{q}_t(x, x') = -\frac{1}{2}\log(2\pi t) - \frac{(x' - x)^2}{2t} \\ + A(x', \theta) - A(x, \theta) - l(\theta)t - \int \left(\int_0^t \phi(w_s, \theta)\, ds\right) \mathbb{S}^{(x,x',t)}(dw) \;. \tag{B.3}$$

Since the right hand side of (B.1) can be bounded from above and below, a rejection sampler producing samples from the diffusion bridge can be constructed. This is possible as the right hand side of (B.1) is proportional to the probability that a marked Poisson process on $[0, t] \times [0, 1]$ with intensity $r \overset{\text{def}}{=} \sup_x\{\phi(x); \tilde{W}_\phi^- < x < \tilde{W}_\phi^+\}$ is below the graph $s \mapsto \phi(\tilde{W}_s; \theta)/r$. However, while observing the path for all s is impossible, a finite construction can be devised by sampling the Brownian bridge at points specified by the marked Poisson process; we refer to Beskos et al. [4] for details. The algorithm is described by the following.

Algorithm 3

```
(∗ Sampling a skeleton of a diffusion bridge ∗)
1.  simulate an outcome $(\chi_\ell, \psi_\ell)_{\ell=1}^{\kappa}$ of the marked Poisson process with intensity $r$ and $\kappa \sim \mathrm{Po}(r)$;
2.  conditional on $\tilde{W}_{\phi}^{\pm}$, simulate $(\tilde{W}_{\chi_\ell})_{\ell=1}^{\kappa}$;
3.  if $\phi(\tilde{W}_{\chi_\ell})/r < \psi_\ell$
4.     then return $(\tilde{W}_{\chi_\ell})_{\ell=1}^{\kappa}$
5.     else  go to (1)
```

By interpolating the returned skeleton $(\tilde{W}_{\chi_\ell})_{\ell=1}^{\kappa}$, samples $\tilde{W}_u$, with $(\tilde{W}_s)_{s=0}^{t} \sim \mathbb{S}^{(x,x',t)}$, can be obtained for any $0 \leq u \leq t$. Given samples from the diffusion bridge, an unbiased estimator of (B.3) can be straightforwardly constructed in the following way. Let $\psi \sim \mathrm{Unif}(0,t)$ be independent of $\mathcal{F}_t$. Then $-t\phi(\tilde{W}_\psi, \theta)$ is an unbiased estimator of $\int(\int_0^t \phi(w_s,\theta)\,ds)\,\mathbb{S}^{(x,x',t)}(dw)$ since

$$\mathbb{E}\left[t\phi(\tilde{W}_\psi,\theta)\right] = \mathbb{E}\left[\mathbb{E}\left[t\phi(\tilde{W}_\psi,\theta)\middle|\,\mathcal{F}_t\right]\right] \\ = \mathbb{E}\int_0^t \phi(\tilde{W}_s,\theta)\,ds = \int\left(\int_0^t \phi(w_s,\theta)\,ds\right)\mathbb{S}^{(x,x',t)}(dw)\;.$$

Finally, plugging this estimator into (B.3) yields an unbiased estimator of $\log \tilde{q}_t$.

Appendix C: Residual resampling

In the selection operation in Step 2 of Algorithm 1, each particle index is drawn from the probability distribution formed by the weights $(\omega_k^j \psi_k^j / \sum_{\ell=1}^N \omega_k^\ell \psi_k^\ell)_{j=1}^N$. Consequently, letting M_k^i denote the number of times that index i was drawn, this selection operation may be alternatively expressed as

$$(M_k^1, \ldots, M_k^N) \sim \mathrm{Mult}\left(N, \left(\frac{\omega_k^j \psi_k^j}{\sum_{\ell=1}^N \omega_k^\ell \psi_k^\ell}\right)_{j=1}^N\right)\;. \tag{C.1}$$

In the deterministic plus residual multinomial resampling approach one sets instead $M_k^i \overset{\mathrm{def}}{=} \lfloor N\omega_k^i\psi_k^i / \sum_{\ell=1}^N \omega_k^\ell \psi_k^\ell \rfloor + H_k^i$ with

$$(H_k^1, \ldots, H_k^N) \\ \sim \mathrm{Mult}\left(\sum_{i=1}^N \left\langle \frac{N\omega_k^i\psi_k^i}{\sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell}\right\rangle, \left(\frac{\langle N\omega_k^i\psi_k^i/\sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell\rangle}{\sum_{j=1}^N \langle N\omega_k^j\psi_k^j/\sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell\rangle}\right)_{i=1}^N\right)\;, \tag{C.2}$$

where $\lfloor x \rfloor$ denotes the integer part of a real number x and $\langle x \rangle \overset{\mathrm{def}}{=} x - \lfloor x \rfloor$. In this selection schedule, index i is first copied $\lfloor N\omega_k^i\psi_k^i / \sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell \rfloor$ times; the remaining $\sum_{i=1}^N \langle N\omega_k^i\psi_k^i / \sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell \rangle$ indices are hereafter drawn multinomially with respect to weights proportional to the residuals $(\langle N\omega_k^i\psi_k^i/\sum_{\ell=1}^N \omega_k^\ell\psi_k^\ell\rangle)_{i=1}^N$. In [10] it is proved that deterministic plus residual resampling preserves consistency as well as asymptotic normality of the particle sample.

References

[1] Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Ann. Statist.*, 36(2):906–937. MR2396819

[2] Ball, F. G. and Rice, J. H. (1992). Stochastic models for ion channels: Introduction and bibliography. *Math. Biosci.*, 112:189–206.

[3] Beskos, A., Papaspiliopoulos, O., and Roberts, G. (2008). A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10(1):85–104. MR2394037

[4] Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. Roy. Statist. Soc. Ser. B*, 68(3):333–382. With discussion. MR2278331

[5] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer. MR2159833

[6] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers & Chemistry*, 16(2):107–115.

[7] Del Moral, P., Jacod, J., and Protter, P. (2001). The Monte-Carlo method for filtering with discrete-time observations. *Probability Theory and Related Fields*, 120:346–368. MR1843179

[8] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38. MR0501537

[9] Douc, R., Fort, G., Moulines, E., and Priouret, P. (2009). Forgetting the initial distribution for hidden markov models. *Stoch. Process. Appl.*, 119(4):1235–1256. MR2508572

[10] Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, 36(5):2344–2376. MR2458190

[11] Douc, R., Moulines, É., and Olsson, J. (2008). Optimality of the auxiliary particle filter. *Probab. Math. Statist.*, 29(1):1–28. MR2552996

[12] Fearnhead, P., Papaspiliopoulos, O., and Roberts, G. (2008). Particle filters for partially observed diffusions. *J. Roy. Statist. Soc. Ser. B*, 70(4):755–777. MR2523903

[13] Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259. MR2001649

[14] Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc. F, Radar signal Process.*, 140:107–113.

[15] Handschin, J. and Mayne, D. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *Int. J. Control*, 9:547–559. MR0246490

[16] Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximising the likelihood for a general state-space model. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, pages 159–175. Springer. MR1847791

[17] Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. A. (2011). Iterated filtering. *Ann. Statist.*, 39(3):1776–1802.

[18] Kitigawa, G. (1998). A self-organizing state-space-model. *J. Am. Statist. Assoc.*, 93(443):1203–1215.

[19] Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer.

[20] Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *J. Am. Statist. Assoc.*, 90(420):567–576.

[21] Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179. MR2401658

[22] Olsson, J. and Rydén, T. (2008). Asymptotic properties of the bootstrap particle filter maximum likelihood estimator for state space models. *Stoch. Process. Appl.*, 118:649–680. MR2394847

[23] Olsson, J. and Ströjby, J. (2010). Convergence of random weight particle filters. Appears in J. Ströjby's PhD thesis *On Inference in Partially Observed Markov Models using Sequential Monte Carlo Methods*, Centre for Mathematical Sciences, Lund University, 2010.

[24] Pedersen, A. R. (1995). Consistency and Asymptotic Normality of an Approximative Maximum Likelihood Estimator for Discretely Observed Diffusion Processes. *Bernoulli*, 1(3):257–279. MR1363541

[25] Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Am. Statist. Assoc.*, 87:493–499. MR1702328

[26] Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall.

[27] Ristic, B., Arulampalam, M., and Gordon, A. (2004). *Beyond Kalman Filters: Particle Filters for Target Tracking.* Artech House.

[28] Wu, C. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103. MR0684867