Journal of Machine Learning Research 25 (2024) 1-66 Submitted 3/23; Revised 11/23; Published 1/24

On Unbiased Estimation for Partially Observed Diffusions

Jeremy Heng HENG@ESSEC.EDU
ESSEC Business School

Jeremie Houssineau JEREMIE.HOUSSINEAU@NTU.EDU.SG
Division of Mathematical Sciences
Nanyang Technological University

Ajay Jasra AJAYJASRA@CUHK.EDU.CN
School of Data Science
Chinese University of Hong Kong, Shenzhen

Editor: Alexandre Bouchard

Abstract

We consider a class of diffusion processes with finite-dimensional parameters and partially observed at discrete time instances. We propose a methodology to unbiasedly estimate the expectation of a given functional of the diffusion process conditional on parameters and data. When these unbiased estimators with appropriately chosen functionals are employed within an expectation-maximization algorithm or a stochastic gradient method, this enables statistical inference using the maximum likelihood or Bayesian framework. Compared to existing approaches, the use of our unbiased estimators allows one to remove any time-discretization bias and Markov chain Monte Carlo burn-in bias. Central to our methodology is a novel and natural combination of the multilevel randomization schemes developed by Mcleish (2011); Rhee and Glynn (2015) and the unbiased Markov chain Monte Carlo methods of Jacob et al. (2020a,b), and the development of new couplings of multiple conditional particle filters of Andrieu et al. (2010). We establish under assumptions that our estimators are unbiased and have finite variance. We illustrate various aspects of our method on an Ornstein–Uhlenbeck model, a logistic diffusion model for population dynamics, and a neural network model for grid cells.

Keywords: diffusions, unbiased estimation, particle filters, coupling, stochastic gradient methods

1. Introduction

1.1 Model and Observations

We consider a diffusion process $(X_t)_{t\geq 0}$ in $\mathbb{R}^d$, defined as the solution of the stochastic differential equation (SDE)

$$dX_t = a_\theta(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x_\star \in \mathbb{R}^d, \tag{1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. We will assume that the drift function $a : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ depends on a vector of unknown parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, but the diffusion coefficient $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ does not. When the diffusion coefficient is parameter dependent, one can deal with this in some cases by finding a suitable transformation of the process. For example, if σ is diagonal matrix of coefficients to be inferred, then a simple rescaling would yield a transformed process with unit diffusion coefficient. Tackling the general case requires a more involved treatment; we discuss how to adapt our proposed methodology in Section A of the supplementary material. The drift and diffusion coefficients are assumed to be regular enough for (1) to admit a (weakly) unique solution for all $t > 0$; more precise regularity conditions needed in our analysis will be stated. The use of SDE models is ubiquitous in engineering, finance, machine learning, and many areas of science. We model observations $Y_{t_1}, \ldots, Y_{t_P} \in \mathbb{R}^{d_y}$ at a collection of time instances $0 \leq t_1 < \cdots < t_P = T$ as conditionally independent given the latent diffusion process $X = (X_t)_{0 \leq t \leq T}$, with conditional density $g : \Theta \times \mathbb{R}^d \times \mathbb{R}^{d_y} \to \mathbb{R}^+$, i.e. $Y_t \sim g_\theta(\cdot|X_t)$ for all $t \in \{t_1, \ldots, t_P\}$, that can be evaluated. For notational ease, we assume that observations are given at unit times, i.e. $t_p = p$ for $p \in \{1, \ldots, P\}$ and $P = T$, which covers the case of regularly observed data by a time rescaling. Irregularly observed data can also be accommodated with minor modifications to our presentation and considered for an application in Section 5.2. We note that this *partially observed* setting is related to but distinct from the *discretely observed* case where the diffusion process is observed without error (Sørensen, 2004). Extension of our methodology to continuous-time observation models is straightforward and illustrated on an application in Section 5.3.

Given a realization $y_{1:T} = (y_t)_{t=1}^T$ of the observation process, the marginal likelihood is

$$p_\theta(y_{1:T}) = \mathbb{E}_\theta \left[\prod_{t=1}^{T} g_\theta(y_t|X_t) \right], \tag{2}$$

where $\mathbb{E}_\theta$ denotes expectation with respect to (w.r.t.) the probability measure $\mathbb{P}_\theta$, induced by the solution of (1) on $\mathcal{C}_d([0, T])$, the space of continuous mappings from $[0, T]$ to $\mathbb{R}^d$. For most problems of practical interest, statistical inference for diffusion models is challenging for two main reasons. Firstly, as most diffusion processes do not have analytically tractable transition densities, one has to resort to time-discretization. Secondly, even if transition densities are available or a time-discretization scheme is employed, the expectation over the latent process (conditional on observations) is usually intractable, and one has to rely on Monte Carlo approximations. Existing approaches to these two issues (Beskos et al., 2006b, 2009; Fearnhead et al., 2008) are mostly based on the exact algorithm of Beskos and Roberts (2005); Beskos et al. (2006a) that allows exact simulation of diffusion sample paths without any time-discretization bias. Although these simulation techniques are very elegant, they require a user to know various properties of the diffusion process, which may not always be the case in practice.

1.2 Unbiased Estimation and Parameter Inference

In this paper, we propose to deal with the above-mentioned difficulties using a novel computational framework by bridging the randomized multilevel Monte Carlo (MLMC) schemes developed by Mcleish (2011); Rhee and Glynn (2015) to remove the time-discretization bias with the unbiased Markov chain Monte Carlo (MCMC) methods of Jacob et al. (2020a,b) to eliminate the MCMC burn-in bias. Our proposed methodology allows one to unbiasedly estimate conditional expectations of the form

$$S(\theta, \theta_\star) = \mathbb{E}_{\theta_\star}\left[G_\theta(X) \mid y_{1:T}\right], \tag{3}$$

for a given functional G_θ. We will write $S(\theta) = S(\theta, \theta)$ and choose G_θ according to the parameter inference algorithm of interest. We consider two types of algorithms in which (3) can be used: the expectation-maximization (EM) algorithm (Dempster et al., 1977) and gradient-based methods for maximum likelihood estimation and Bayesian inference.

EM algorithm. Within an EM algorithm to compute the maximum likelihood estimator (MLE) $\hat{\theta} \in \arg\max_{\theta\in\Theta} p_\theta(y_{1:T})$, the expectation step corresponds to $S(\theta, \theta_\star)$ with G_θ given by the complete-data log-likelihood, and the expectation is w.r.t. the law of X conditioned on current parameters $\theta_\star$ and observations $y_{1:T}$. Approximating the expectation step with unbiased estimators would yield a Monte Carlo EM algorithm (Wei and Tanner, 1990).

Gradient-based methods. To employ gradient-based algorithms, we shall consider approximations of the score function $\nabla_\theta \log p_\theta(y_{1:T})$, where ∇_θ denotes the gradient w.r.t. the parameter θ. We will show that the score can be represented as $S(\theta)$ with G_θ given by the gradient of the complete-data log-likelihood. If unbiased estimators $\widehat{S}(\theta)$ of $S(\theta)$ can be constructed, one can compute the MLE with stochastic gradient ascent (SGA)

$$\theta_m = \theta_{m-1} + \varepsilon_m \widehat{S}(\theta_{m-1}), \quad m = 1, 2, \ldots, \tag{4}$$

where $(\varepsilon_m)_{m=1}^{\infty}$ is a sequence of learning rates. Similarly, in the Bayesian framework, one can sample from the posterior distribution $p(d\theta|y_{1:T}) \propto p(\theta)p_\theta(y_{1:T})d\theta$ using the stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011)

$$\theta_m = \theta_{m-1} + \frac{1}{2}\varepsilon_m\left(\nabla_\theta \log p(\theta_{m-1}) + \widehat{S}(\theta_{m-1})\right) + \varepsilon_m^{1/2}\eta_m, \quad m = 1, 2, \ldots, \tag{5}$$

where $(\eta_m)_{m=1}^{\infty}$ is a sequence of independent standard Gaussian random vectors in $\mathbb{R}^{d_\theta}$. The use of unbiased estimators of the score function within stochastic gradient methods is particularly appealing. Under suitable assumptions and an appropriate choice of learning rates $(\varepsilon_m)_{m=1}^{\infty}$, unbiased scores ensures convergence of the SGA iterates (4) to a local maxima of the likelihood $p_\theta(y_{1:T})$ (Kushner and Yin, 2003), and convergence of the SGLD and its ergodic averages to the posterior distribution and posterior expectations, respectively (Teh et al., 2016). In contrast, the use of biased gradient estimates leads to asymptotic biases that have to be studied (Tadić and Doucet, 2017).

1.3 Proposed Methodology

Our approach first relies on the idea of debiasing by randomizing over the level of the time-discretization (Mcleish, 2011; Rhee and Glynn, 2015). Unbiased estimators based on randomized MLMC can also attain better convergence rates than Monte Carlo approaches based on the finest discretization level, and can be made the same as standard MLMC under simple modifications (Vihola, 2018). When employing randomized MLMC schemes, we shall assume access to unbiased estimates of differences between successive levels of discretization, i.e. $S_l - S_{l-1}$ where

$$S_l(\theta, \theta_\star) = \mathbb{E}_{\theta_\star}^{l}\left[G_\theta^{l}(X_{0:T}) \mid y_{1:T}\right] \tag{6}$$

denotes an approximation of (3) at discretization level $l \in \mathbb{N}$. In (6), G^l_θ is the finite-dimensional approximation of G_θ and $X_{0:T}$ is distributed according to $\pi^l_{\theta_\star}$, the law of the time-discretized process $X_{0:T}$ at level l conditioned on $\theta_\star$ and $y_{1:T}$. The specific definitions of G^l_θ and $\pi^l_{\theta_\star}$ will be given in Section 2.2. Unbiased estimators of each increment $I_l = S_l - S_{l-1}$ are not easily obtained as exact sampling from the posterior distributions $\pi^{l-1}_{\theta_\star}$ and $\pi^l_{\theta_\star}$ is not straightforward. Although MCMC methods targeting these posteriors can be considered, their routine application will lead to biased estimators of I_l due to MCMC burn-in biases. By employing recent advances in Jacob et al. (2020a,b) that builds on earlier work by Glynn and Rhee (2014), one can remove the MCMC burn-in bias and estimate S_l unbiasedly by simulating a pair of MCMC chains targeting $\pi^l_{\theta_\star}$ that are coupled in such a way that they meet exactly after some random number of iterations, and remain faithful after meeting. For the randomized MLMC schemes to return estimators of S with finite variance, the variance of estimated increments has to decrease sufficiently fast to zero with the discretization level l. Hence simply applying unbiased MCMC to estimate S_{l-1} and S_l independently is inadequate, as this would lead to estimators of S with infinite variance. This motivated us to propose an extension the unbiased MCMC framework to allow the terms S_{l-1} and S_l in I_l to be estimated in a dependent manner, and employ the estimated increments within randomized MLMC. This requires simulating a quadruple of MCMC chains, with a pair targeting $\pi^{l-1}_{\theta_\star}$ and another pair targeting $\pi^l_{\theta_\star}$, which are appropriately coupled so that the pair of chains for each discretization level can meet and remain faithful, and the pairs of chains between successive discretization levels are close for large levels l (see Figure 1 for a schematic illustration). After meeting has occurred for both pairs of chains on discretization levels $l-1$ and l, our framework reduces to having a single pair of coupled chains targeting $\pi^{l-1}_{\theta_\star}$ and $\pi^l_{\theta_\star}$, as one would have in a MLMC approach (Jasra et al., 2017). Therefore we believe our work provides a natural combination of the randomized MLMC and the unbiased MCMC frameworks. It extends the application of randomized MLMC (Mcleish, 2011; Rhee and Glynn, 2015) to compute expectations with respect to infinite-dimensional laws that are *conditioned on data*, and unbiased MCMC (Jacob et al., 2020a,b) to handle *infinite-dimensional models* that require discretization. Combining these two computational paradigms has also been recently explored by Wang and Wang (2023) for other statistical problems.

While the focus of this article is partially observed diffusions, our computational framework is more generally applicable to other settings where it is necessary to discretize a model during implementation; see Heng et al. (2023) for an application to Bayesian inverse problems. After selecting an appropriate MCMC method for the problem of interest, the main challenges are to then design a suitable coupling of a quadruple of MCMC chains and prove that it has the above-mentioned properties. We will focus on the conditional particle filter (CPF) of Andrieu et al. (2010), which is a MCMC algorithm targeting $\pi^l_{\theta_\star}$ with good ergodicity properties (Lindsten et al., 2015; Andrieu et al., 2018). Each iteration of CPF involves simulating a particle filter conditioned such that the current trajectory survives all resampling steps; we refer readers who are unfamiliar with particle filtering to the recent textbook by Chopin and Papaspiliopoulos (2020) on the subject. In addition to our proposed estimation framework, the construction of an appropriate coupling for CPF chains is another key methodological contribution. This involves couplings of time-discretizations of the diffusion process and couplings of resampling steps that are necessary to prevent weight

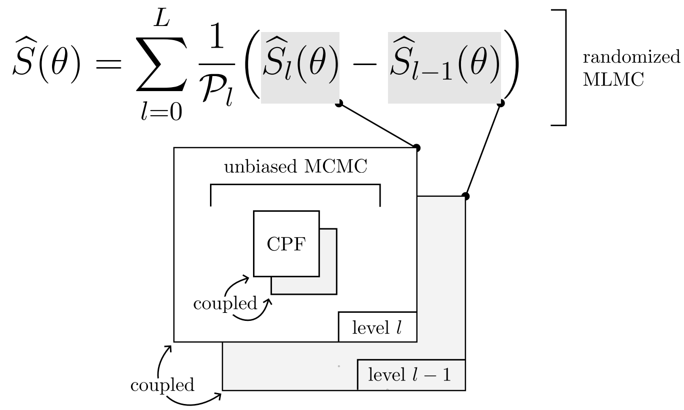


Figure 1: Schematic illustration of the proposed 4-CCPF algorithm. In the displayed equation, L is an independent random variable with probability mass function $(P_l)_{l=0}^{\infty}$ supported on $\mathbb{N}_0$, $\mathcal{P}_l = \sum_{k=l}^{\infty} P_k$ denotes the cumulative tail probability, and $\widehat{S}_l$ is an unbiased estimator of S_l.

degeneracy within the CPFs. While the former can be achieved using common Brownian increments which is standard in MLMC (Giles, 2008), the latter requires novel schemes to induce adequate dependencies between a quadruple of CPF chains as simple strategies based on common random variables will be inadequate. We provide a thorough analysis to establish, under assumptions, that our proposed methodology provides unbiased estimators of S with finite variance. This analysis involves a detailed study of the dependencies between the CPF chains induced by our proposed couplings and constitutes one of our main theoretical contributions.

This article is structured as follows. In Section 2.1, we begin by relating the expectation step of an EM algorithm and the score function needed in gradient-based algorithms to the problem of computing conditional expectations of the form (3). We then introduce discrete-time approximations (6) in Section 2.2. Our proposed methodology to unbiasedly estimate (3) is presented in Section 3; Section 3.1 outlines our overall framework, and the rest of Section 3 contains specific details of our coupling construction for CPF. We establish various properties of our unbiased estimators under assumptions in Section 4 and illustrate multiple aspects of our methodology on three applications in Section 5. The proof of our results are detailed in the supplementary material. An R package to reproduce all numerical results can be found at `https://github.com/jeremyhengjm/UnbiasedScore`.

2. Parameter Inference and Conditional Expectations

2.1 Continuous-Time Representation

We first relate the expectation step of an EM algorithm and the score function required in gradient-based algorithms with the task of computing conditional expectations (3) for an appropriate choice of functional G_θ. In the following, we write A^* to denote the transpose of A and $\|x\|_p$ to denote the $\mathbb{L}_p$-norm of a vector $x \in \mathbb{R}^d$. As the law $\mathbb{P}_\theta$ depends on the parameter θ, we consider a change of measure to the law $\mathbb{Q}$ induced by $dX_t = \sigma(X_t)dW_t$ with $X_0 = x_\star$ on the space $\mathcal{C}_d([0,T])$. Since $\mathbb{P}_\theta$ and $\mathbb{Q}$ are equivalent, by Girsanov theorem (Rogers and Williams, 2000, p. 79) the corresponding Radon–Nikodym derivative is

$$\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) = \exp\left\{-\frac{1}{2}\int_0^T \|b_\theta(X_t)\|_2^2 dt + \int_0^T b_\theta(X_t)^* \Sigma(X_t)^{-1}\sigma(X_t)^* dX_t\right\}, \tag{7}$$

where $\Sigma(x) = \sigma(x)\sigma(x)^*$ and $b_\theta(x) = \Sigma(x)^{-1}\sigma(x)^* a_\theta(x)$ for $x \in \mathbb{R}^d$. Therefore we can write the complete-data likelihood under $\mathbb{Q}$ as

$$\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X)\prod_{t=1}^T g_\theta(y_t|X_t). \tag{8}$$

The purpose of $\mathbb{Q}$ is to act as a reference measure to allow us to define densities; it will not play a role in the simulation algorithms that we will later introduce.

EM algorithm. At each iteration of an EM algorithm with current parameters $\theta_\star$, one determines the parameters in the next iteration by computing $\arg\max_{\theta\in\Theta} Q(\theta,\theta_\star)$, where $Q(\theta,\theta_\star)$ denotes the expectation of the complete-data log-likelihood w.r.t. the law of X conditioned on $\theta_\star$ and $y_{1:T}$. By defining $S(\theta,\theta_\star)$ in (3) with the functional

$$G_\theta(X) = \log\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) + \sum_{t=1}^T \log g_\theta(y_t|X_t), \tag{9}$$

the EM algorithm is equivalent to computing $\arg\max_{\theta\in\Theta} S(\theta,\theta_\star)$ as it follows from (8) that $S(\theta,\theta_\star)$ and $Q(\theta,\theta_\star)$ differ by a term that does not depend on θ.

Gradient-based methods. Next we consider the score function $\nabla_\theta \log p_\theta(y_{1:T})$. Under the above change of measure, the marginal likelihood in (2) can be written as

$$p_\theta(y_{1:T}) = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X)\prod_{t=1}^T g_\theta(y_t|X_t)\right]. \tag{10}$$

We will assume throughout the article that $\theta \mapsto a_\theta(x)$ and $\theta \mapsto g_\theta(y|x)$ are differentiable for each $(x,y) \in \mathbb{R}^d \times \mathbb{R}^{d_y}$. Under mild regularity conditions and thanks to the fact that $\mathbb{Q}$ does not depend on θ, we may differentiate (10) and, using Fisher's identity (Cappé et al., 2006, p. 353), we obtain

$$\begin{aligned}
\nabla_\theta \log p_\theta(y_{1:T}) &= \nabla_\theta \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X)\prod_{t=1}^T g_\theta(y_t|X_t)\right] \Big/ p_\theta(y_{1:T}) \\
&= \mathbb{E}_{\mathbb{Q}}\left[\left\{\nabla_\theta \log\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) + \sum_{t=1}^T \nabla_\theta \log g_\theta(y_t|X_t)\right\}\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X)\prod_{t=1}^T g_\theta(y_t|X_t)\right] \Big/ p_\theta(y_{1:T}) \\
&= \mathbb{E}_\theta\left[\left\{\nabla_\theta \log\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) + \sum_{t=1}^T \nabla_\theta \log g_\theta(y_t|X_t)\right\} \mid y_{1:T}\right].
\end{aligned} \tag{11}$$

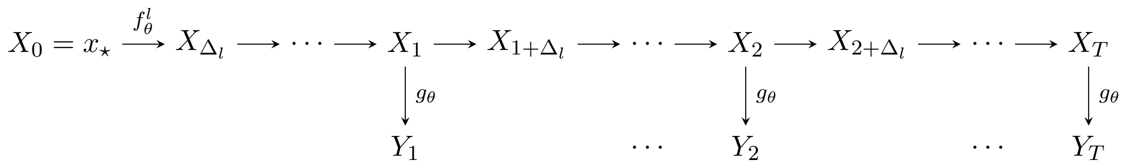


Figure 2: State space model with time-discretization step-size $\Delta_l = 2^{-1}$.

Hence we can represent the score function as $S(\theta) = S(\theta, \theta)$ in (3) with the functional

$$G_\theta(X) = \nabla_\theta \log \frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) + \sum_{t=1}^{T} \nabla_\theta \log g_\theta(y_t|X_t). \tag{12}$$

By differentiating (7), the first gradient term can be written as

$$\nabla_\theta \log \frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X) = -\int_0^T \{\nabla_\theta a_\theta(X_t)\}^* \Sigma(X_t)^{-1} a_\theta(X_t) dt + \int_0^T \{\nabla_\theta a_\theta(X_t)\}^* \Sigma(X_t)^{-1} dX_t,$$

where $\nabla_\theta a_\theta(x) \in \mathbb{R}^{d\times d_\theta}$ denotes the Jacobian matrix w.r.t. θ.

The above relations also hold in more general settings under small modifications to the functionals (9) and (12). Firstly, although we have assumed a deterministic initial condition of $X_0 = x_\star$ to simplify our presentation, random initializations can also be accommodated. Assuming that $X_0 \sim \mu_\theta$ is initialized from a distribution μ_θ on $\mathbb{R}^d$ that admits a differentiable density (w.r.t. the d-dimensional Lebesgue measure), the terms $\log \mu_\theta(X_0)$ and $\nabla_\theta \log \mu_\theta(X_0)$ should be added to the expressions in (9) and (12) respectively. This follows by conditioning on the value of X_0 and applying the above arguments. Secondly, our methodology can also handle continuous-time observation models. In this case, the sums $\sum_{t=1}^{T} \log g_\theta(y_t|X_t)$ in (9) and $\sum_{t=1}^{T} \nabla_\theta \log g_\theta(y_t|X_t)$ in (12) would be replaced by the conditional log-likelihood $\log p_\theta(y_{1:T}|X)$ and its gradient $\nabla_\theta \log p_\theta(y_{1:T}|X)$; see Section 5.3 for an application where the observational model is given by an inhomogenous Poisson point process.

When the diffusion coefficient also depends on unknown parameters, it may be possible to find an invertible transformation $\Psi : \mathbb{R}^d \to \mathbb{R}^d$, such that the transformed process $X_t = \Psi(Z_t)$ satisfies an SDE with a diffusion coefficient that is not parameter-dependent. We will illustrate two examples of this principle in Section 5 using the Lamperti transformation and a simple rescaling of each component of the diffusion process. In more general cases where such transformations are not available, one has to seek a different representation of the expectation step of EM and the score function as the above change of measure is no longer applicable; see Section A of the supplementary material.

2.2 Discrete-Time Approximation

As alluded to in the introduction, we will rely on time-discretizations of the diffusion process (1). We will employ a hierarchy of discretizations of the time interval $[0, T]$, indexed by a level parameter $l \in \mathbb{N}_0$ that determines the temporal resolution. Higher levels with finer time resolutions will require increased algorithmic cost. For each level l, let $0 = s_0 <$

$\cdots < s_{K_l} = T$ denote a dyadic uniform discretization of $[0, T]$, defined as $s_k = k\Delta_l$ for $k \in \{0, 1, \dots, K_l\}$, where $\Delta_l = 2^{-l}$ is the step-size and $K_l = 2^l T$ is the number of time steps. Note that, by construction, $(s_k)_{k=0}^{K_l}$ contains the unit observation times $(t_p)_{p=1}^{P}$. We consider the Euler–Maruyama scheme (Kloeden and Platen, 2013) which defines a time-discretized process $X_{0:T} = (X_{s_k})_{k=0}^{K_l}$ on the path space $\mathsf{X}^l = (\mathbb{R}^d)^{K_l+1}$ according to the following recursion for $k \in \{1, \dots, K_l\}$:

$$X_{s_k} = X_{s_{k-1}} + a_\theta(X_{s_{k-1}})\Delta_l + \sigma(X_{s_{k-1}})(W_{s_k} - W_{s_{k-1}}), \quad X_0 = x_\star. \tag{13}$$

In the following, we write $\mathcal{N}_d(\mu, \Sigma)$ to denote a d-dimensional Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, and its density as $x \mapsto \mathcal{N}_d(x; \mu, \Sigma)$. The notation $\mathcal{N}_d(0_d, I_d)$ refers to the standard Gaussian distribution, i.e. with zero mean vector $0_d \in \mathbb{R}^d$ and identity covariance matrix $I_d \in \mathbb{R}^{d\times d}$. Let $f_\theta^l(dx_{s_k}|x_{s_{k-1}}) = \mathcal{N}_d(x_{s_k}; x_{s_{k-1}} + a_\theta(x_{s_{k-1}})\Delta_l, \Delta_l\Sigma(x_{s_{k-1}}))dx_{s_k}$ denotes the Gaussian transition kernel corresponding to (13). To simplify our notation, we omit the dependence of the time grid and time-discretized process on the level parameter (until this distinction is necessary in Section 3.3), and write the transition in (13) as $X_{s_k} = F_\theta^l(X_{s_{k-1}}, V_{s_k})$, where $V_{s_k} = W_{s_k} - W_{s_{k-1}}$ denotes the Brownian increment. Higher-order schemes such as the Milstein method could be employed but would be difficult to implement for problems with dimension $d \geq 2$. Future work could consider the antithetic truncated Milstein method of Giles and Szpruch (2014) for such settings.

Under time-discretization, the joint distribution of the latent process and observations is given by a state space model (see Figure 2)

$$p_\theta^l(dx_{0:T}, y_{1:T}) = \prod_{t=1}^{T} g_\theta(y_t|x_t)\delta_{x_\star}(dx_0) \prod_{k=1}^{K_l} f_\theta^l(dx_{s_k}|x_{s_{k-1}}),$$

where $\delta_{x_\star}(dx_0)$ refers to the Dirac measure at the deterministic initial condition $x_\star$, and the marginal likelihood is $p_\theta^l(y_{1:T}) = \int_{\mathsf{X}^l} p_\theta^l(dx_{0:T}, y_{1:T})$. Hence the resulting finite-dimensional approximation of the law of X conditioned on θ and $y_{1:T}$ is

$$\pi_\theta^l(dx_{0:T}) = p_\theta^l(dx_{0:T}|y_{1:T}) = \frac{p_\theta^l(dx_{0:T}, y_{1:T})}{p_\theta^l(y_{1:T})}. \tag{14}$$

Following the literature on state space modelling, we will refer to (14) as a *smoothing distribution* (Chopin and Papaspiliopoulos, 2020, ch. 12). Particle smoothing algorithms (Briers et al., 2010; Fearnhead et al., 2010; Douc et al., 2011) can be used to approximate (14), but are not suited to our framework which is based on MCMC.

Next we define time-discretized approximations of the functional G_θ. For (9), we have

$$\begin{aligned} G_\theta^l(X_{0:T}) = &-\frac{1}{2}\sum_{k=1}^{K_l} \|b_\theta(X_{s_{k-1}})\|_2^2\Delta_l + \frac{1}{2}\sum_{k=1}^{K_l} b_\theta(X_{s_{k-1}})^*\Sigma(X_{s_{k-1}})^{-1}\sigma(X_{s_{k-1}})^*(X_{s_k} - X_{s_{k-1}}) \\ &+ \sum_{t=1}^{T} \log g_\theta(y_t|X_t), \end{aligned}$$

and for (12), we have

$$G_\theta^l(X_{0:T}) = -\sum_{k=1}^{K_l} \left\{\nabla_\theta a_\theta(X_{s_{k-1}})\right\}^* \Sigma(X_{s_{k-1}})^{-1} a_\theta(X_{s_{k-1}})\Delta_l$$
$$+ \sum_{k=1}^{K_l} \left\{\nabla_\theta a_\theta(X_{s_{k-1}})\right\}^* \Sigma(X_{s_{k-1}})^{-1} (X_{s_k} - X_{s_{k-1}}) + \sum_{t=1}^{T} \nabla_\theta \log g_\theta(y_t|X_t). \quad (15)$$

With G_θ^l and π_θ^l in place, we then define S_l using (6), which forms our approximation of S in (12) at discretization level l. In Section 4, under appropriate regularity conditions, we will study the rate at which the discrete-time approximation S_l converges to the desired S as the level l tends to infinity. The following section concerns numerical approximations of these conditional expectations.

3. Unbiased Estimation

3.1 Unbiased Estimation Framework

We begin this section by outlining our novel framework to construct unbiased estimators of the conditional expectation in (3) with a given functional $G_\theta : \mathcal{C}_d([0,T]) \to \mathbb{R}^{d_g}$. To simplify our exposition, we will henceforth consider $S(\theta) = S(\theta, \theta)$ as extension to the case $\theta \neq \theta_\star$ follows straightforwardly. For each $\theta \in \Theta$, convergence of time-discretized approximation $S_l(\theta)$ in (6) as $l \to \infty$ allows us to write

$$S(\theta) = \lim_{L\to\infty} \sum_{l=0}^{L} I_l(\theta), \quad (16)$$

where $I_l(\theta) = S_l(\theta) - S_{l-1}(\theta)$ denotes the increment at level $l \in \mathbb{N}_0$ (with $S_{-1} = 0$). We will design algorithms that allow us to construct unbiased estimators $\widehat{I}_l(\theta)$ of $I_l(\theta)$ independently for any $l \in \mathbb{N}_0$, i.e.

$$\mathbb{E}\left[\widehat{I}_l(\theta)\right] = I_l(\theta), \quad l \in \mathbb{N}_0, \quad (17)$$

where $\mathbb{E}$ denotes expectation w.r.t. all random variables generated in our algorithm. The key insight of the randomized MLMC schemes proposed by Mcleish (2011); Rhee and Glynn (2015) is to perform a random truncation of the sum in (16) at level L, so that unbiasedness is retained when unbiased estimators of the increments are employed. The relationships between the six different algorithms that will be used to compute $\widehat{I}_l(\theta)$ for $l \in \mathbb{N}_0$ is illustrated in Figure 3.

Let $(P_l)_{l=0}^\infty$ be a given probability mass function (PMF) with support on $\mathbb{N}_0$ and define its cumulative tail probabilities as $\mathcal{P}_l = \sum_{k=l}^\infty P_k$ for $l \in \mathbb{N}_0$. Suppose that

$$\sum_{l=0}^{\infty} \mathcal{P}_l^{-1} \left\{\mathrm{Var}\left[\widehat{I}_l(\theta)^j\right] + \left(S_l(\theta)^j - S(\theta)^j\right)^2\right\} < \infty \quad (18)$$

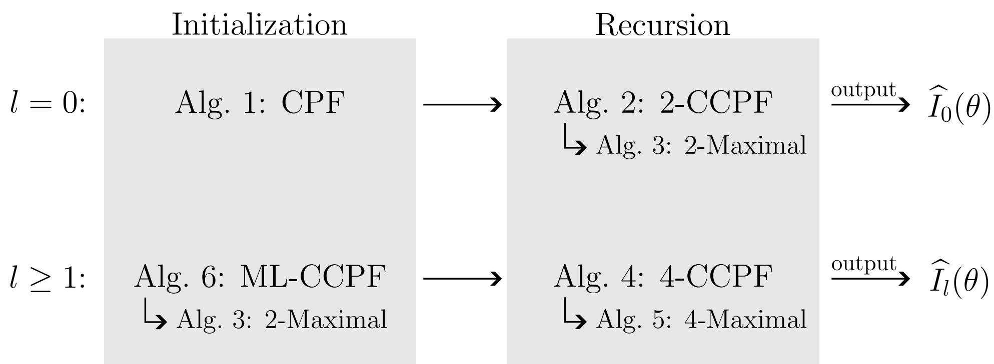


Figure 3: Relationships between Algorithms 1 to 6.

for all $j \in \{1, \dots, d_g\}$, where Var denotes variance under $\mathbb{E}$ and x^j refers to the j^{th}-component of the vector x. If we sample L from $(P_l)_{l=0}^{\infty}$ independently of $(\widehat{I}_l(\theta))_{l=0}^{\infty}$, it follows from Rhee and Glynn (2015) that the *independent sum* estimator

$$\widehat{S}(\theta) = \sum_{l=0}^{L} \frac{\widehat{I}_l(\theta)}{\mathcal{P}_l} \tag{19}$$

is an unbiased estimator of S with finite variance, i.e. $\mathbb{E}[\widehat{S}(\theta)] = S(\theta)$ and the entries of the covariance matrix $\text{Var}[\widehat{S}(\theta)]$ are finite. Alternatives to the estimator in (19) such as the *single term* estimator of Rhee and Glynn (2015) could also be considered here. Inspection of (18) reveals that we have to understand how fast S_l converges to S as $l \to \infty$, which will be established in Theorem 1. Moreover, for (17) and (18) to hold, we have to compute an unbiased estimator of $I_0(\theta) = S_0(\theta)$ at level $l = 0$ with finite variance, and construct unbiased estimators of the increments $I_l(\theta)$, whose variance vanishes sufficiently fast as $l \to \infty$ relative to the tails of $(P_l)_{l=0}^{\infty}$. Developing a methodology that meets these two requirements will be the focus of Sections 3.2 and 3.3 respectively. To establish that these requirements are satisfied, we provide a detailed analysis of our proposed methodology (in supplementary material) and summarize the key results in Theorem 2.

3.2 Unbiased Estimation under Time-Discretization

This section considers unbiased estimation of $S_l(\theta)$ at discretization level $l \in \mathbb{N}_0$; the case $l = 0$ will be employed to construct the first summand $\widehat{I}_0(\theta)$ in (19). Our basic algorithmic building block is the CPF of Andrieu et al. (2010), which defines a Markov kernel on the space of trajectories $X_{0:T} \in \mathsf{X}^l$ that admits the smoothing distribution π_θ^l as its invariant distribution. A detailed description for our application is given in Algorithm 1, which has a complexity of $\mathcal{O}(NK_l)$. The CPF involves simulating $N \geq 2$ trajectories under the time-discretized model dynamics (Steps 2a & 2b), weighting samples according to the conditional density g_θ (Step 2c), and resampling from the weighted particle approximation (Step 2d). We will consider multinomial resampling, in which case, $\mathcal{R}(w_t^{1:N})$ refers to the categorical distribution on $\{1, \dots, N\}$ with probabilities $w_t^{1:N} = (w_t^n)_{n=1}^N$. The main difference to a standard bootstrap particle filter (BPF) (Gordon et al., 1993) is that the input trajectory

---

Algorithm 1 Conditional particle filter (CPF) at parameter $\theta \in \Theta$ and discretization level $l \in \mathbb{N}_0$

Input: a trajectory $X^\star_{0:T} = (X^\star_{s_k})_{k=0}^{K_l} \in \mathsf{X}^l$. For time step $k = 0$

(1a) Set $X_0^n = x_\star$ for $n \in \{1, \ldots, N\}$.

(1b) Set $w_0^n = N^{-1}$ and $A_0^n = n$ for $n \in \{1, \ldots, N\}$.

For time step $k \in \{1, \ldots, K_l\}$

(2a) Sample Brownian increment $V^n_{s_k} \sim \mathcal{N}_d(0_d, \Delta_l I_d)$ independently for $n \in \{1, \ldots, N-1\}$.

(2b) Set $X^n_{s_k} = F^l_\theta(X^{A^n_{s_{k-1}}}_{s_{k-1}}, V^n_{s_k})$ for $n \in \{1, \ldots, N-1\}$, and $X^N_{s_k} = X^\star_{s_k}$.

If there is an observation at time $t = s_k \in \{1, \ldots, T\}$

(2c) Compute normalized weight $w_t^n \propto g_\theta(y_t|X_t^n)$ for $n \in \{1, \ldots, N\}$.

(2d) If $t < T$, sample ancestor $A_t^n \sim \mathcal{R}(w_t^{1:N})$ independently for $n \in \{1, \ldots, N-1\}$ and set $A_t^N = N$.

Else

(2e) Set $A^n_{s_k} = n$ for $n \in \{1, \ldots, N\}$.

After the terminal step

(3a) Sample particle index $B_T \sim \mathcal{R}(w_T^{1:N})$.

(3b) Set particle index $B_{s_k} = A^{B_{s_{k+1}}}_{s_k}$ for $k \in \{0, 1, \ldots, K_l - 1\}$.

Output: a trajectory $X^\circ_{0:T} = (X^{B_{s_k}}_{s_k})_{k=0}^{K_l} \in \mathsf{X}^l$.

$X^\star_{0:T}$ is conditioned to survive all resampling steps. The algorithm outputs a trajectory $X^\circ_{0:T}$ by sampling the particle indexes $(B_{s_k})_{k=0}^{K_l}$ after the terminal step (Steps 3a & 3b). We will write $X^\circ_{0:T} \sim M^l_\theta(\cdot|X^\star_{0:T})$ to denote a trajectory generated by the CPF kernel at parameter $\theta \in \Theta$ and discretization level $l \in \mathbb{N}_0$.

We will initialize the CPF Markov chain using the law of the time-discretized dynamics $\nu^l_\theta(dx_{0:T}) = \delta_{x_\star}(dx_0) \prod_{k=1}^{K_l} f^l_\theta(dx_{s_k}|x_{s_{k-1}})$. One could also consider the law of a trajectory sampled from a BPF, which provides a good approximation of π^l_θ for sufficiently large N (Andrieu et al., 2010, Theorem 1). Under mild assumptions, the Markov chain $(X_{0:T}(i))_{i=0}^\infty$ generated by

$$X_{0:T}(0) \sim \nu^l_\theta, \quad X_{0:T}(i) \sim M^l_\theta(\cdot|X_{0:T}(i-1)), \tag{20}$$

for $i \geq 1$, is uniformly ergodic (Chopin and Singh, 2015; Lindsten et al., 2015; Andrieu et al., 2018). Hence one can adopt a standard MCMC approach to approximate $S_l(\theta)$ by the average $A_l^{b:I}(\theta) = (I - b + 1)^{-1} \sum_{i=b}^I G^l_\theta(X_{0:T}(i))$, for some fixed burn-in $0 \leq b \leq I$, which is consistent as the number of iterations $I \to \infty$. However, as the Markov chain is not started at stationarity, the MCMC average $A_l^{b:I}(\theta)$ is a biased estimator for any finite $I \in \mathbb{N}$. Although this burn-in bias can be reduced by increasing b, tuning it to control the bias is a difficult task in practice.

By building on the work of Glynn and Rhee (2014), Jacob et al. (2020a) showed how to obtain unbiased estimators of smoothing expectations by simulating a pair of coupled CPF chains $(X_{0:T}(i), \bar{X}_{0:T}(i))_{i=0}^\infty$ on the product space $\mathsf{Z}^l = \mathsf{X}^l \times \mathsf{X}^l$ with the same marginal law. This is achieved using a coupling of two CPFs as described in Algorithm 2, which we will refer to as the 2-CCPF. Writing $(X^\circ_{0:T}, \bar{X}^\circ_{0:T}) \sim \bar{M}^l_\theta(\cdot|X^\star_{0:T}, \bar{X}^\star_{0:T})$ to denote a pair of trajectories generated by the 2-CCPF kernel given $(X^\star_{0:T}, \bar{X}^\star_{0:T}) \in \mathsf{Z}^l$ as input, marginally we have $X^\circ_{0:T} \sim M^l_\theta(\cdot|X^\star_{0:T})$ and $\bar{X}^\circ_{0:T} \sim M^l_\theta(\cdot|\bar{X}^\star_{0:T})$. The two main ingredients of this coupling are the use of common Brownian increments (Steps 2a & 2b), and a coupling of the resampling distributions $\mathcal{R}(w_t^{1:N})$ and $\mathcal{R}(\bar{w}_t^{1:N})$ denoted as $\bar{\mathcal{R}}(w_t^{1:N}, \bar{w}_t^{1:N})$ (Steps 2e & 3a). We refer readers to the references in Jacob et al. (2020a) for various coupled resampling schemes. Our focus is the maximal coupling (Chopin and Singh, 2015; Jasra et al., 2017) that maximizes the probability of having identical ancestors at each step of the CPFs. Sampling from the maximal coupling can be done with the inverse transformation method; this is detailed in Algorithm 3, where we have suppressed the time dependence for notational simplicity. We could also consider an improvement of Step 5 that samples from the residual distributions with common uniform random variables. As the cost of implementing Algorithm 3 is $\mathcal{O}(N)$, the overall cost of Algorithm 2 is still $\mathcal{O}(NK_l)$.

We initialize by sampling $(X_{0:T}(0), \bar{X}_{0:T}(0))$ from a coupling $\bar{\nu}_\theta^l$ with ν_θ^l as its marginals. The choice of $\bar{\nu}_\theta^l$ could be explored but we will consider the independent coupling for simplicity. The pair of CPF chains is then generated as

$$X_{0:T}(1) \sim M_\theta^l(\cdot|X_{0:T}(0)), \quad (X_{0:T}(i+1), \bar{X}_{0:T}(i)) \sim \bar{M}_\theta^l(\cdot|X_{0:T}(i), \bar{X}_{0:T}(i-1)), \tag{21}$$

for $i \geq 1$. Marginally, $(X_{0:T}(i))_{i=0}^\infty$ and $(\bar{X}_{0:T}(i))_{i=0}^\infty$ have the same law as the Markov chain generated by (20). It can be shown that each application of 2-CCPF allows the chains to meet with some positive probability (Jacob et al., 2020a, Theorem 3.1), that depends on the number of trajectories N and observations T (Lee et al., 2020, Theorems 8 & 9). Moreover, the construction of 2-CCPF ensures that the chains are faithful, i.e. $X_{0:T}(i) = \bar{X}_{0:T}(i-1)$ for all $i \geq \tau_\theta^l$, where $\tau_\theta^l = \inf\{i \geq 1 : X_{0:T}(i) = \bar{X}_{0:T}(i-1)\}$ denotes the meeting time. Using the time-averaged estimator of Jacob et al. (2020a), an unbiased estimator of $S_l(\theta)$ is given by

$$\widehat{S}_l(\theta) = A_l^{b:I}(\theta) + \sum_{i=b+1}^{\tau_\theta^l - 1} \min\left(1, \frac{i-b}{I-b+1}\right)\left(G_\theta^l(X_{0:T}(i)) - G_\theta^l(\bar{X}_{0:T}(i-1))\right). \tag{22}$$

The second term corrects for the bias of the MCMC average $A_l^{b:I}(\theta)$ and is equal to zero if $b+1 > \tau_\theta^l - 1$. Under assumptions that will be stated in Section 4, $\widehat{S}_l(\theta)$ has finite variance and finite expected cost, for any choice of $N \geq 2$ and $0 \leq b \leq I$. Assuming that 2-CCPF costs twice as much as CPF, the cost of computing $\widehat{S}_l(\theta)$ is $\max\{2\tau_\theta^l - 1, I + \tau_\theta^l - 1\}$ applications of the CPF kernel M_θ^l.

3.3 Unbiased Estimation of Increments

We now consider unbiased estimation of the increment $I_l(\theta) = S_l(\theta) - S_{l-1}(\theta)$ at level $l \in \mathbb{N}$ to construct the term $\widehat{I}_l(\theta)$ in the independent sum estimator $\widehat{S}(\theta)$ in (19). A naive approach that employs the unbiased estimation framework in Section 3.2 to estimate the terms $S_{l-1}(\theta)$ and $S_l(\theta)$ independently, and take the difference to estimate the increment $I_l(\theta)$, will satisfy the unbiasedness requirement in (17). However, as the variance of the estimated increment will not decrease with the discretization level l under independent estimation of $S_{l-1}(\theta)$ and $S_l(\theta)$, one cannot choose a PMF $(P_l)_{l=0}^\infty$ such that the condition in (18) holds. This prompts an extension of the preceding framework that allows the terms in each increment to be estimated in a dependent manner, as illustrated in Figure 1. Our proposed methodology involves simulating a quadruple of coupled CPF chains, a pair $(X_{0:T}^{l-1}(i), \bar{X}_{0:T}^{l-1}(i))_{i=0}^\infty$

Algorithm 2 Two coupled CPF (2-CCPF) at parameter $\theta \in \Theta$ and discretization level $l \in \mathbb{N}_0$

Input: a pair of trajectories $(X^\star_{0:T}, \bar{X}^\star_{0:T}) = (X^\star_{s_k}, \bar{X}^\star_{s_k})_{k=0}^{K_l} \in \mathsf{Z}^l$. For time step $k = 0$

(1a) Set $X_0^n = x_\star$ and $\bar{X}_0^n = x_\star$ for $n \in \{1, \dots, N\}$.

(1b) Set $w_0^n = N^{-1}, \bar{w}_0^n = N^{-1}$ and $A_0^n = n, \bar{A}_0^n = n$ for $n \in \{1, \dots, N\}$.

For time step $k \in \{1, \dots, K_l\}$

(2a) Sample Brownian increment $V_{s_k}^n \sim \mathcal{N}_d(0_d, \Delta_l I_d)$ independently for $n \in \{1, \dots, N-1\}$.

(2b) Set $X_{s_k}^n = F_\theta^l(X_{s_{k-1}}^{A_{s_{k-1}}^n}, V_{s_k}^n)$ and $\bar{X}_{s_k}^n = F_\theta^l(\bar{X}_{s_{k-1}}^{\bar{A}_{s_{k-1}}^n}, V_{s_k}^n)$ for $n \in \{1, \dots, N-1\}$.

(2c) Set $X_{s_k}^N = X_{s_k}^\star$ and $\bar{X}_{s_k}^N = \bar{X}_{s_k}^\star$.

If there is an observation at time $t = s_k \in \{1, \dots, T\}$

(2d) Compute normalized weights $w_t^n \propto g_\theta(y_t|X_t^n)$ and $\bar{w}_t^n \propto g_\theta(y_t|\bar{X}_t^n)$ for $n \in \{1, \dots, N\}$.

(2e) If $t < T$, sample ancestors $(A_t^n, \bar{A}_t^n) \sim \bar{\mathcal{R}}(w_t^{1:N}, \bar{w}_t^{1:N})$ independently for $n \in \{1, \dots, N-1\}$ and set $A_t^N = N, \bar{A}_t^N = N$.

Else

(2f) Set $A_{s_k}^n = n$ and $\bar{A}_{s_k}^n = n$ for $n \in \{1, \dots, N\}$.

After the terminal step

(3a) Sample particle indexes $(B_T, \bar{B}_T) \sim \bar{\mathcal{R}}(w_T^{1:N}, \bar{w}_T^{1:N})$.

(3b) Set particle indexes $B_{s_k} = A_{s_k}^{B_{s_{k+1}}}$ and $\bar{B}_{s_k} = \bar{A}_{s_k}^{\bar{B}_{s_{k+1}}}$ for $k \in \{0, 1, \dots, K_l - 1\}$.

Output: a pair of trajectories $(X^\circ_{0:T}, \bar{X}^\circ_{0:T}) = (X_{s_k}^{B_{s_k}}, \bar{X}_{s_k}^{\bar{B}_{s_k}})_{k=0}^{K_l} \in \mathsf{Z}^l$.

Algorithm 3 Maximal coupling of two resampling distributions $\mathcal{R}(w^{1:N})$ and $\mathcal{R}(\bar{w}^{1:N})$ (2-Maximal)

Input: normalized weights $w^{1:N} = (w^n)_{n=1}^N$ and $\bar{w}^{1:N} = (\bar{w}^n)_{n=1}^N$.

(1) Compute the overlap $o^n = \min\{w^n, \bar{w}^n\}$ for $n \in \{1, \dots, N\}$.

(2) Compute the mass of the overlap $\mu = \sum_{n=1}^N o^n$ and normalize $O^n = o^n/\mu$ for $n \in \{1, \dots, N\}$.

(3) Compute the residuals $r^n = (w^n - o^n)/(1-\mu)$ and $\bar{r}^n = (\bar{w}^n - o^n)/(1-\mu)$ for $n \in \{1, \dots, N\}$.

With probability μ

(4) Sample $A \sim \mathcal{R}(O^{1:N})$ and set $\bar{A} = A$.

Otherwise

(5) Sample $A \sim \mathcal{R}(r^{1:N})$ and $\bar{A} \sim \mathcal{R}(\bar{r}^{1:N})$ independently.

Output: indexes $(A, \bar{A})$.

on Z^{l-1} for discretization level $l-1$, and another pair $(X_{0:T}^l(i), \bar{X}_{0:T}^l(i))_{i=0}^\infty$ on Z^l for discretization level l. This relies on the coupling of four CPFs detailed in Algorithm 4, which will be referred to as the 4-CCPF. In the algorithmic description, $(s_k)_{k=0}^{K_l}$ denotes a time discretization of $[0, T]$ at level l, and a trajectory at the coarser level $l-1$ is written as $X_{0:T}^{l-1} = (X_{s_{2k}}^{l-1})_{k=0}^{K_{l-1}}$. Given trajectories $(X_{0:T}^{l-1,\star}, \bar{X}_{0:T}^{l-1,\star}, X_{0:T}^{l,\star}, \bar{X}_{0:T}^{l,\star}) \in \mathsf{Z}^{l-1} \times \mathsf{Z}^l$ as input, we will write

$$(X_{0:T}^{l-1,\circ}, \bar{X}_{0:T}^{l-1,\circ}, X_{0:T}^{l,\circ}, \bar{X}_{0:T}^{l,\circ}) \sim \bar{M}_\theta^{l-1,l}(\cdot|X_{0:T}^{l-1,\star}, \bar{X}_{0:T}^{l-1,\star}, X_{0:T}^{l,\star}, \bar{X}_{0:T}^{l,\star}) \tag{23}$$

to denote the trajectories generated by 4-CCPF. The 4-CCPF kernel $\bar{M}_\theta^{l-1,l}$ is a four-marginal coupling of the CPF kernels M_θ^{l-1} and M_θ^l, in the sense that marginally, we have $X_{0:T}^{l-1,\circ} \sim M_\theta^{l-1}(\cdot|X_{0:T}^{l-1,\star})$ and $\bar{X}_{0:T}^{l-1,\circ} \sim M_\theta^{l-1}(\cdot|\bar{X}_{0:T}^{l-1,\star})$ at level $l-1$, and $X_{0:T}^{l,\circ} \sim M_\theta^l(\cdot|X_{0:T}^{l,\star})$

Algorithm 4 Four coupled CPF (4-CCPF) at parameter $\theta \in \Theta$ and discretization levels $l-1$ and $l \in \mathbb{N}$

Input: a pair of trajectories $(X_{0:T}^{l-1,\star}, \bar{X}_{0:T}^{l-1,\star}) = (X_{s_{2k}}^{l-1,\star}, \bar{X}_{s_{2k}}^{l-1,\star})_{k=0}^{K_{l-1}} \in \mathsf{Z}^{l-1}$ and a pair of trajectories $(X_{0:T}^{l,\star}, \bar{X}_{0:T}^{l,\star}) = (X_{s_k}^{l,\star}, \bar{X}_{s_k}^{l,\star})_{k=0}^{K_l} \in \mathsf{Z}^{l}$.

For time step $k = 0$

(1a) Set $X_0^{l-1,n} = x_\star, \bar{X}_0^{l-1,n} = x_\star$ and $X_0^{l,n} = x_\star, \bar{X}_0^{l,n} = x_\star$ for $n \in \{1, \ldots, N\}$.

(1b) Set $w_0^{l-1,n} = N^{-1}, \bar{w}_0^{l-1,n} = N^{-1}$ and $w_0^{l,n} = N^{-1}, \bar{w}_0^{l,n} = N^{-1}$ for $n \in \{1, \ldots, N\}$.

(1c) Set $A_0^{l-1,n} = n, \bar{A}_0^{l-1,n} = n$ and $A_0^{l,n} = n, \bar{A}_0^{l,n} = n$ for $n \in \{1, \ldots, N\}$.

For time step $k \in \{1, \ldots, K_l\}$

(2a) Sample Brownian increment $V_{s_k}^{l,n} \sim \mathcal{N}_d(0_d, \Delta_l I_d)$ at level l independently for $n \in \{1, \ldots, N-1\}$.

(2b) Set $X_{s_k}^{l,n} = F_\theta^l(X_{s_{k-1}}^{l,A_{s_{k-1}}^{l,n}}, V_{s_k}^{l,n})$ and $\bar{X}_{s_k}^{l,n} = F_\theta^l(\bar{X}_{s_{k-1}}^{l,\bar{A}_{s_{k-1}}^{l,n}}, V_{s_k}^{l,n})$ at level l for $n \in \{1, \ldots, N-1\}$.

(2c) Set $X_{s_k}^{l,N} = X_{s_k}^{l,\star}$ and $\bar{X}_{s_k}^{l,N} = \bar{X}_{s_k}^{l,\star}$.

If $k \in \{2, 4, \ldots, K_l\}$

(2d) Set Brownian increment $V_{s_k}^{l-1,n} = V_{s_{k-1}}^{l,n} + V_{s_k}^{l,n}$ at level $l-1$ for $n \in \{1, \ldots, N-1\}$.

(2e) Set $X_{s_k}^{l-1,n} = F_\theta^{l-1}(X_{s_{k-1}}^{l-1,A_{s_{k-1}}^{l-1,n}}, V_{s_k}^{l-1,n})$ and $\bar{X}_{s_k}^{l-1,n} = F_\theta^{l-1}(\bar{X}_{s_{k-1}}^{l-1,\bar{A}_{s_{k-1}}^{l-1,n}}, V_{s_k}^{l-1,n})$ at level $l-1$ for $n \in \{1, \ldots, N-1\}$.

(2f) Set $X_{s_k}^{l-1,N} = X_{s_k}^{l-1,\star}$ and $\bar{X}_{s_k}^{l-1,N} = \bar{X}_{s_k}^{l-1,\star}$.

If there is an observation at time $t = s_k \in \{1, \ldots, T\}$

(2g) Compute normalized weights $w_t^{l-1,n} \propto g_\theta(y_t | X_t^{l-1,n}), \bar{w}_t^{l-1,n} \propto g_\theta(y_t | \bar{X}_t^{l-1,n})$ and $w_t^{l,n} \propto g_\theta(y_t | X_t^{l,n}), \bar{w}_t^{l,n} \propto g_\theta(y_t | \bar{X}_t^{l,n})$ for $n \in \{1, \ldots, N\}$.

(2h) If $t < T$, sample ancestors $(A_t^{l-1,n}, \bar{A}_t^{l-1,n}, A_t^{l,n}, \bar{A}_t^{l,n}) \sim \bar{\mathcal{R}}(w_t^{l-1,1:N}, \bar{w}_t^{l-1,1:N}, w_t^{l,1:N}, \bar{w}_t^{l,1:N})$ independently for $n \in \{1, \ldots, N-1\}$ and set $A_t^{l-1,N} = N, \bar{A}_t^{l-1,N} = N, A_t^{l,N} = N, \bar{A}_t^{l,N} = N$.

Else

(2i) If $k \in \{2, 4, \ldots, K_l\}$, set $A_{s_k}^{l-1,n} = n$ and $\bar{A}_{s_k}^{l-1,n} = n$ at level $l-1$ for $n \in \{1, \ldots, N\}$.

(2j) Set $A_{s_k}^{l,n} = n$ and $\bar{A}_{s_k}^{l,n} = n$ at level l for $n \in \{1, \ldots, N\}$.

After the terminal step

(3a) Sample particle indexes $(B_T^{l-1}, \bar{B}_T^{l-1}, B_T^l, \bar{B}_T^l) \sim \bar{\mathcal{R}}(w_T^{l-1,1:N}, \bar{w}_T^{l-1,1:N}, w_T^{l,1:N}, \bar{w}_T^{l,1:N})$.

(3b) Set particle indexes $B_{s_{2k}}^{l-1} = A_{s_{2k}}^{l-1,B_{s_{2(k+1)}}^{l-1}}$ and $\bar{B}_{s_{2k}}^{l-1} = \bar{A}_{s_{2k}}^{l-1,\bar{B}_{s_{2(k+1)}}^{l-1}}$ at level $l-1$ for $k \in \{0, 1, \ldots, K_{l-1} - 1\}$.

(3c) Set particle indexes $B_{s_k}^l = A_{s_k}^{l,B_{s_{k+1}}^l}$ and $\bar{B}_{s_k}^l = \bar{A}_{s_k}^{l,\bar{B}_{s_{k+1}}^l}$ at level l for $k \in \{0, 1, \ldots, K_l - 1\}$.

Output: a pair of trajectories $(X_{0:T}^{l-1,\circ}, \bar{X}_{0:T}^{l-1,\circ}) = (X_{s_{2k}}^{l-1,B_{s_{2k}}^{l-1}}, \bar{X}_{s_{2k}}^{l-1,\bar{B}_{s_{2k}}^{l-1}})_{k=0}^{K_{l-1}} \in \mathsf{Z}^{l-1}$ and a pair of trajectories $(X_{0:T}^{l,\circ}, \bar{X}_{0:T}^{l,\circ}) = (X_{s_k}^{l,B_{s_k}^l}, \bar{X}_{s_k}^{l,\bar{B}_{s_k}^l})_{k=0}^{K_l} \in \mathsf{Z}^{l}$.

and $\bar{X}_{0:T}^{l,\circ} \sim M_\theta^l(\cdot | \bar{X}_{0:T}^{l,\star})$ at level l. The two main ingredients of 4-CCPF are the use of common Brownian increments within each level (Steps 2b & 2e) and across levels (Steps 2a & 2d), and an appropriate four-marginal coupling of the resampling distributions $\mathcal{R}(w_t^{l-1,1:N})$, $\mathcal{R}(\bar{w}_t^{l-1,1:N})$, $\mathcal{R}(w_t^{l,1:N})$, $\mathcal{R}(\bar{w}_t^{l,1:N})$ denoted by $\bar{\mathcal{R}}(w_t^{l-1,1:N}, \bar{w}_t^{l-1,1:N}, w_t^{l,1:N}, \bar{w}_t^{l,1:N})$ (Steps 2h & 3a). While the use of common Brownian increments is a standard choice in MLMC (Giles, 2008), constructing coupled resampling schemes that induce sufficient dependencies between the four CPF chains, for the variance of the estimated increment to decrease with the discretization level, requires new algorithmic design.

Algorithm 5 presents a coupled resampling scheme that will be the focus of our analysis in Section 4. Here we suppress the time dependence for notational simplicity. Given normalized weights $w^{l-1,1:N} = (w^{l-1,n})_{n=1}^N$, $\bar{w}^{l-1,1:N} = (\bar{w}^{l-1,n})_{n=1}^N$ at discretization level $l-1$ and $w^{l,1:N} = (w^{l,n})_{n=1}^N$, $\bar{w}^{l,1:N} = (\bar{w}^{l,n})_{n=1}^N$ at discretization level l, the algorithm samples ancestor indexes $(A^{l-1}, \bar{A}^{l-1}, A^l, \bar{A}^l)$ from the maximal coupling of the maximal couplings $\bar{\mathcal{R}}(w^{l-1,1:N}, w^{l,1:N})$ and $\bar{\mathcal{R}}(\bar{w}^{l-1,1:N}, \bar{w}^{l,1:N})$. That is, amongst all possible four-marginal couplings, this scheme maximizes the probabilities of having identical ancestors across levels, i.e. $A^{l-1} = A^l$ and $\bar{A}^{l-1} = \bar{A}^l$, and identical pair of ancestors within the levels, i.e. $(A^{l-1}, A^l) = (\bar{A}^{l-1}, \bar{A}^l)$. The cost of Algorithm 5 is random as it employs rejection samplers (Thorisson, 2000) in Steps 2b, 3b and 4, wherein $\mathcal{U}_{[0,1]}$ denotes the uniform distribution on $[0,1]$. As the expected cost is $\mathcal{O}(N)$, the overall cost of Algorithm 4 is $\mathcal{O}(NK_l)$ on average. Note that a naive approach to sample from the desired coupled resampling scheme based on the inverse transformation method in place of Algorithm 5 would involve a deterministic but prohibitive cost of $\mathcal{O}(N^2)$. In Step 4, we denote the respective PMF of $\bar{\mathcal{R}}(w^{l-1,1:N}, w^{l,1:N})$ and $\bar{\mathcal{R}}(\bar{w}^{l-1,1:N}, \bar{w}^{l,1:N})$ as

$$R^{l-1,l}(A,B) = \mathbb{I}_{\mathcal{D}}(A,B)o^A + \frac{(w^{l-1,A} - o^A)(w^{l,B} - o^B)}{1-\mu}, \tag{24}$$
$$\bar{R}^{l-1,l}(A,B) = \mathbb{I}_{\mathcal{D}}(A,B)\bar{o}^A + \frac{(\bar{w}^{l-1,A} - \bar{o}^A)(\bar{w}^{l,B} - \bar{o}^B)}{1-\bar{\mu}},$$

for $(A,B) \in \{1,\dots,N\}^2$, where $\mathbb{I}_{\mathcal{D}}(A,B)$ is the indicator function on the diagonal set $\mathcal{D} = \{(A,B) \in \{1,\dots,N\}^2 : A = B\}$, $o^n = \min\{w^{l-1,n}, w^{l,n}\}$ and $\bar{o}^n = \min\{\bar{w}^{l-1,n}, \bar{w}^{l,n}\}$ for $n \in \{1,\dots,N\}$ are the overlapping measures, and $\mu = \sum_{n=1}^N o^n$ and $\bar{\mu} = \sum_{n=1}^N \bar{o}^n$ are their corresponding mass. From the expressions in (24), one can check that the three cases considered in Algorithm 5 are necessary to ensure faithfulness of the pair of chains on each discretization level. More precisely, if the input trajectories satisfy $X_{0:T}^{l-1,\star} = \bar{X}_{0:T}^{l-1,\star}$ and/or $X_{0:T}^{l,\star} = \bar{X}_{0:T}^{l,\star}$, then under the 4-CCPF (23), the output trajectories satisfy $X_{0:T}^{l-1,\circ} = \bar{X}_{0:T}^{l-1,\circ}$ and/or $X_{0:T}^{l,\circ} = \bar{X}_{0:T}^{l,\circ}$ almost surely.

We note that the two-marginal couplings induced by the 4-CCPF kernel $\bar{M}_\theta^{l-1,l}$ on each discretization level are not the same as the 2-CCPF kernels $\bar{M}_\theta^{l-1}$ and $\bar{M}_\theta^{l}$. Although it is not a requirement of our methodology, this property would hold if we consider a modification of Algorithm 5 that samples from the maximal coupling of the maximal couplings $\bar{\mathcal{R}}(w^{l-1,1:N}, \bar{w}^{l-1,1:N})$ and $\bar{\mathcal{R}}(w^{l,1:N}, \bar{w}^{l,1:N})$. However, as with simple coupling strategies based on common random variables, such a coupled resampling scheme does not induce adequate dependencies between the CPF chains across discretization levels. This will be illustrated experimentally in Section 5.1. To understand the rationale behind Algorithm 5, we observe that the two-marginal couplings induced by the 4-CCPF kernel across discretization levels are given by the multilevel CPF (ML-CPF) described in Algorithm 6. That is, writing $M_\theta^{l-1,l}$ as the ML-CPF kernel, which is a coupling of the CPF kernels M_θ^{l-1} and M_θ^{l}, we have $(X_{0:T}^{l-1,\circ}, X_{0:T}^{l,\circ}) \sim M_\theta^{l-1,l}(\cdot | X_{0:T}^{l-1,\star}, X_{0:T}^{l,\star})$ and $(\bar{X}_{0:T}^{l-1,\circ}, \bar{X}_{0:T}^{l,\circ}) \sim M_\theta^{l-1,l}(\cdot | \bar{X}_{0:T}^{l-1,\star}, \bar{X}_{0:T}^{l,\star})$ under the 4-CCPF kernel in (23). The ML-CPF is similar to the multilevel particle filter of Jasra et al. (2017), who proposed multilevel estimators of filtering expectations that are non-asymptotically biased but consistent in the limit of our computational budget. Even though our objective is markedly different, as we seek non-asymptotically unbiased and

Algorithm 5 Maximal coupling of the maximal couplings $\bar{\mathcal{R}}(w^{l-1,1:N}, w^{l,1:N})$ and $\bar{\mathcal{R}}(\bar{w}^{l-1,1:N}, \bar{w}^{l,1:N})$ (4-Maximal)

Input: normalized weights $w^{l-1,1:N} = (w^{l-1,n})_{n=1}^N$, $\bar{w}^{l-1,1:N} = (\bar{w}^{l-1,n})_{n=1}^N$ at level $l-1$ and $w^{l,1:N} = (w^{l,n})_{n=1}^N$, $\bar{w}^{l,1:N} = (\bar{w}^{l,n})_{n=1}^N$ at level l.
(1) Sample $(A^{l-1}, A^l) \sim \bar{\mathcal{R}}(w^{l-1,1:N}, w^{l,1:N})$ using Algorithm 3.
If normalized weights at level $l-1$ are identical and normalized weights at level l are non-identical
(2a) Set $\bar{A}^{l-1} = A^{l-1}$.
(2b) With probability $\bar{w}^{l,\bar{A}^{l-1}}/\bar{w}^{l-1,\bar{A}^{l-1}}$, set $\bar{A}^l = \bar{A}^{l-1}$; otherwise sample $A \sim \mathcal{R}(\bar{w}^{l,1:N})$ and $U \sim \mathcal{U}_{[0,1]}$ until $U > \bar{w}^{l-1,A}/\bar{w}^{l,A}$, and set $\bar{A}^l = A$.
If normalized weights at level $l-1$ are non-identical and normalized weights at level l are identical
(3a) Set $\bar{A}^l = A^l$.
(3b) With probability $\bar{w}^{l-1,\bar{A}^l}/\bar{w}^{l,\bar{A}^l}$, set $\bar{A}^{l-1} = \bar{A}^l$; otherwise sample $A \sim \mathcal{R}(\bar{w}^{l-1,1:N})$ and $U \sim \mathcal{U}_{[0,1]}$ until $U > \bar{w}^{l,A}/\bar{w}^{l-1,A}$, and set $\bar{A}^{l-1} = A$.
Otherwise
(4) With probability $\bar{R}^{l-1,l}(A^{l-1}, A^l)/R^{l-1,l}(A^{l-1}, A^l)$, set $(\bar{A}^{l-1}, \bar{A}^l) = (A^{l-1}, A^l)$; otherwise sample $(A, B) \sim \bar{\mathcal{R}}(\bar{w}^{l-1,1:N}, \bar{w}^{l,1:N})$ and $U \sim \mathcal{U}_{[0,1]}$ until $U > R^{l-1,l}(A,B)/\bar{R}^{l-1,l}(A,B)$, and set $(\bar{A}^{l-1}, \bar{A}^l) = (A, B)$.
Output: indexes $(A^{l-1}, \bar{A}^{l-1}, A^l, \bar{A}^l)$.

(almost surely) finite cost estimators of $S(\theta)$ which is a smoothing expectation, the connection to MLMC estimation alludes to better convergence rates than Monte Carlo approaches based on the finest discretization level.

We now describe simulation of the quadruple of CPF chains $(X_{0:T}^{l-1}(i), \bar{X}_{0:T}^{l-1}(i), X_{0:T}^l(i), \bar{X}_{0:T}^l(i))_{i=0}^\infty$ using ML-CPF and 4-CCPF. We initialize the four chains $(X_{0:T}^{l-1}(0), \bar{X}_{0:T}^{l-1}(0), X_{0:T}^l(0), \bar{X}_{0:T}^l(0))$ from a four-marginal coupling $\bar{\nu}_\theta^{l-1,l}$ that satisfies $X_{0:T}^{l-1}(0), \bar{X}_{0:T}^{l-1}(0) \sim \nu_\theta^{l-1}$ and $X_{0:T}^l(0), \bar{X}_{0:T}^l(0) \sim \nu_\theta^l$. For simplicity, we assume $\bar{\nu}_\theta^{l-1,l}$ is such that each of the pairs across levels $(X_{0:T}^{l-1}(0), X_{0:T}^l(0))$ and $(\bar{X}_{0:T}^{l-1}(0), \bar{X}_{0:T}^l(0))$ independently follow a coupling of ν_θ^{l-1} and ν_θ^l, denoted as $\nu_\theta^{l-1,l}$. The choice of $\nu_\theta^{l-1,l}$ will be taken as the joint law of the time-discretized dynamics under common Brownian increments in our analysis. We then sample $(X_{0:T}^{l-1}(1), X_{0:T}^l(1)) \sim M_\theta^{l-1,l}(\cdot|X_{0:T}^{l-1}(0), X_{0:T}^l(0))$ with ML-CPF, and subsequently for $i \geq 1$, iteratively sample

$$(X_{0:T}^{l-1}(i+1), \bar{X}_{0:T}^{l-1}(i), X_{0:T}^l(i+1), \bar{X}_{0:T}^l(i)) \sim \bar{M}_\theta^{l-1,l}(\cdot|X_{0:T}^{l-1}(i), \bar{X}_{0:T}^{l-1}(i-1), X_{0:T}^l(i), \bar{X}_{0:T}^l(i-1)) \quad (25)$$

from 4-CCPF. Marginally, the single CPF chains have the same law as a Markov chain generated by (20) at discretization level $l-1$ or l. Since each application of 4-CCPF allows the pair of chains on each level to meet with some positive probability (see Lemma 29 in the supplementary material), and by construction remain faithful thereafter, we define the meeting time at level l as $\tau_\theta^l = \inf\{i \geq 1 : X_{0:T}^l(i) = \bar{X}_{0:T}^l(i-1)\}$ and the stopping time at level l as $\bar{\tau}_\theta^l = \max\{\tau_\theta^{l-1}, \tau_\theta^l\}$. Our construction has the desirable property that the 4-CCPF collapses to the ML-CPF after the stopping time, i.e. for $i > \bar{\tau}_\theta^l$, the transition in (25) is equivalent to sampling $(X_{0:T}^{l-1}(i+1), X_{0:T}^l(i+1)) \sim M_\theta^{l-1,l}(\cdot|X_{0:T}^{l-1}(i), X_{0:T}^l(i))$ and setting $\bar{X}_{0:T}^{l-1}(i) = X_{0:T}^{l-1}(i+1), \bar{X}_{0:T}^l(i) = X_{0:T}^l(i+1)$. For any choice of burn-in $b \in \mathbb{N}_0$ and number of iterations $I \geq b$, we can compute unbiased estimators $\widehat{S}_{l-1}(\theta)$ and $\widehat{S}_l(\theta)$

Algorithm 6 Multilevel CPF (ML-CPF) at parameter $\theta \in \Theta$ and discretization levels $l-1$ and $l \in \mathbb{N}$

Input: a trajectory $X_{0:T}^{l-1,\star} = (X_{s_{2k}}^{l-1,\star})_{k=0}^{K_{l-1}} \in \mathsf{X}^{l-1}$ and a trajectory $X_{0:T}^{l,\star} = (X_{s_k}^{l,\star})_{k=0}^{K_l} \in \mathsf{X}^l$.

For time step $k = 0$

(1a) Set $X_0^{l-1,n} = x_\star$ and $X_0^{l,n} = x_\star$ for $n \in \{1, \dots, N\}$.

(1b) Set $w_0^{l-1,n} = N^{-1}, w_0^{l,n} = N^{-1}$ and $A_0^{l-1,n} = n, A_0^{l,n} = n$ for $n \in \{1, \dots, N\}$.

For time step $k \in \{1, \dots, K_l\}$

(2a) Sample Brownian increment $V_{s_k}^{l,n} \sim \mathcal{N}_d(0_d, \Delta_l I_d)$ at level l independently for $n \in \{1, \dots, N-1\}$.

(2b) Set $X_{s_k}^{l,n} = F_\theta^l(X_{s_{k-1}}^{l,A_{s_{k-1}}^{l,n}}, V_{s_k}^{l,n})$ at level l for $n \in \{1, \dots, N-1\}$, and $X_{s_k}^{l,N} = X_{s_k}^{l,\star}$.

If $k \in \{2, 4, \dots, K_l\}$

(2c) Set Brownian increment $V_{s_k}^{l-1,n} = V_{s_{k-1}}^{l,n} + V_{s_k}^{l,n}$ at level $l-1$ for $n \in \{1, \dots, N-1\}$.

(2d) Set $X_{s_k}^{l-1,n} = F_\theta^{l-1}(X_{s_{k-1}}^{l-1,A_{s_{k-1}}^{l-1,n}}, V_{s_k}^{l-1,n})$ at level $l-1$ for $n \in \{1, \dots, N-1\}$, and $X_{s_k}^{l-1,N} = X_{s_k}^{l-1,\star}$.

If there is an observation at time $t = s_k \in \{1, \dots, T\}$

(2e) Compute normalized weights $w_t^{l-1,n} \propto g_\theta(y_t | X_t^{l-1,n})$ and $w_t^{l,n} \propto g_\theta(y_t | X_t^{l,n})$ for $n \in \{1, \dots, N\}$.

(2f) If $t < T$, sample ancestors $(A_t^{l-1,n}, A_t^{l,n}) \sim \bar{\mathcal{R}}(w_t^{l-1,1:N}, w_t^{l,1:N})$ independently for $n \in \{1, \dots, N-1\}$ and set $A_t^{l-1,N} = N, A_t^{l,N} = N$.

Else

(2i) If $k \in \{2, 4, \dots, K_l\}$, set $A_{s_k}^{l-1,n} = n$ at level $l-1$ for $n \in \{1, \dots, N\}$.

(2j) Set $A_{s_k}^{l,n} = n$ at level l for $n \in \{1, \dots, N\}$.

After the terminal step

(3a) Sample particle indexes $(B_T^{l-1}, B_T^l) \sim \bar{\mathcal{R}}(w_T^{l-1,1:N}, w_T^{l,1:N})$.

(3b) Set particle indexes $B_{s_{2k}}^{l-1} = A_{s_{2k}}^{l-1,B_{s_{2(k+1)}}^{l-1}}$ at level $l-1$ for $k \in \{0, 1, \dots, K_{l-1} - 1\}$.

(3c) Set particle indexes $B_{s_k}^l = A_{s_k}^{l,B_{s_{k+1}}^l}$ at level l for $k \in \{0, 1, \dots, K_l - 1\}$.

Output: a trajectory $X_{0:T}^{l-1,\circ} = (X_{s_{2k}}^{l-1,B_{s_{2k}}^{l-1}})_{k=0}^{K_{l-1}} \in \mathsf{X}^{l-1}$ and a trajectory $X_{0:T}^{l,\circ} = (X_{s_k}^{l,B_{s_k}^l})_{k=0}^{K_l} \in \mathsf{X}^l$.

of $S_{l-1}(\theta)$ and $S_l(\theta)$ using the time-averaged estimator in (22) based on the pair of CPF chains on level $l-1$ and l, respectively. We can then obtain an unbiased estimator of the increment $I_l(\theta)$ using the difference $\widehat{I}_l(\theta) = \widehat{S}_l(\theta) - \widehat{S}_{l-1}(\theta)$, which has finite variance and finite expected cost under the assumptions in Section 4. The cost of computing $\widehat{I}_l(\theta)$ is $\max\{2\tau_\theta^{l-1} - 1, I + \tau_\theta^{l-1} - 1\}/2 + \max\{2\tau_\theta^l - 1, I + \tau_\theta^l - 1\}$ applications of the CPF kernel M_θ^l (this assumes that $\mathrm{Cost}(M_\theta^{l-1,l}) = \mathrm{Cost}(M_\theta^{l-1}) + \mathrm{Cost}(M_\theta^l)$, $\mathrm{Cost}(\bar{M}_\theta^{l-1,l}) = 2\mathrm{Cost}(M_\theta^{l-1}) + 2\mathrm{Cost}(M_\theta^l)$ and $\mathrm{Cost}(M_\theta^l) = 2\mathrm{Cost}(M_\theta^{l-1})$).

3.4 Summary of Proposed Methodology and Choice of Tuning Parameters

We consolidate the algorithms presented in this section by summarizing our proposed methodology to unbiasedly estimate $S(\theta)$ below.

Input: PMF $(P_l)_{l=0}^\infty$, number of particles N, burn-in b, and number of iterations I.

(1) Sample highest discretization level L from $(P_l)_{l=0}^\infty$.

(2) Simulate a pair of coupled CPF chains $(X_{0:T}(i), \bar{X}_{0:T}(i))$ using CPF (Algorithm 1) and 2-CCPF (Algorithm 2) with N particles at discretization level 0 as described in (21) for iteration $i = 0, 1, \ldots, \max\{I, \tau_\theta^0\}$, and compute unbiased estimator $\widehat{I}_0(\theta)$ of $I_0(\theta)$ using time-averaged estimator (22) with burn-in b and I iterations.
(3) Independently for $l = 1, \ldots, L$, simulate a quadruple of coupled CPF chains $(X_{0:T}^{l-1}(i), \bar{X}_{0:T}^{l-1}(i), X_{0:T}^{l}(i), \bar{X}_{0:T}^{l}(i))$ using ML-CPF (Algorithm 6) and 4-CCPF (Algorithm 4) with N particles at discretization levels $l-1$ and l as described in (25) for iteration $i = 0, 1, \ldots, \max\{I, \bar{\tau}_\theta^l\}$, and compute unbiased estimators $\widehat{S}_{l-1}(\theta)$ and $\widehat{S}_l(\theta)$ of $S_{l-1}(\theta)$ and $S_l(\theta)$ using time-averaged estimator (22) with burn-in b and I iterations. Compute unbiased estimator of the increment $I_l(\theta)$ with $\widehat{I}_l(\theta) = \widehat{S}_l(\theta) - \widehat{S}_{l-1}(\theta)$.
(4) Compute unbiased estimator of $S(\theta)$ using independent sum estimator $\widehat{S}(\theta) = \sum_{l=0}^{L} \widehat{I}_l(\theta)/\mathcal{P}_l$, where $\mathcal{P}_l = \sum_{k=l}^{\infty} P_k$.

Output: unbiased estimator $\widehat{S}(\theta)$ of $S(\theta)$.

We will establish unbiasedness and finite variance properties of $\widehat{S}(\theta)$ in Section 4. The cost of the above procedure is $c(\theta) = \sum_{l=0}^{L} c_l(\theta)$, where $c_l(\theta)$ is the cost of computing $\widehat{I}_l(\theta)$. From Sections 3.2 and 3.3, we have $c_l(\theta) = a_\theta^l \times \mathrm{Cost}(M_\theta^l)$, with $a_\theta^l = \max(2\tau_\theta^0 - 1, I + \tau_\theta^0 - 1)$ for level $l = 0$, and $a_\theta^l = \max(2\tau_\theta^{l-1} - 1, I + \tau_\theta^{l-1} - 1)/2 + \max(2\tau_\theta^l - 1, I + \tau_\theta^l - 1)$ for level $l \in \mathbb{N}$. We take the view here that the cost per application of the CPF kernel M_θ^l is $\mathrm{Cost}(M_\theta^l) = NK_l = N2^lT$. Hence the expected cost of computing $\widehat{S}(\theta)$ is $\mathbb{E}[c(\theta)] = \sum_{l=0}^{\infty} \mathbb{E}[c_l(\theta)]\,\mathcal{P}_l = NT\sum_{l=0}^{\infty} \mathbb{E}\left[a_\theta^l\right] 2^l \mathcal{P}_l$. We will see that one cannot find a PMF $(P_l)_{l=0}^{\infty}$ so that the variance and expected cost of $\widehat{S}(\theta)$ are both finite. We defer further discussions and the selection of the distribution of L to Section 4, and assume for now that we have a given PMF $(P_l)_{l=0}^{\infty}$ that ensures finite variance but infinite expected cost. In this regime, we will refrain from discussions of asymptotic efficiency in the sense of Glynn and Whitt (1992).

We now discuss the choice of tuning parameters and various algorithmic considerations. In the above description, the choice of (N, b, I) could be level-dependent but optimizing these tuning parameters is outside the scope of this work. Following the discussion in Andrieu et al. (2010, Theorem 1) and the empirical findings in Jacob et al. (2020a), we will scale the number of particles N linearly with the number of observations T. Although the variance of $\widehat{I}_l(\theta)$ decreases as we increase the burn-in parameter $b \in \mathbb{N}_0$, setting b too large would be inefficient. Jacob et al. (2020a,b) proposed choosing b according to the distribution of the meeting time. In our context, as the stopping time $\bar{\tau}_\theta^l$ typically decreases as the level l increases, a conservative strategy is to select b based on the stopping time of a low discretization level, which can be simulated by running ML-CPF and 4-CCPF as in Step 3. We will illustrate this numerically in Section 5 and experiment with various choices of b. After selecting b, one can choose the number of iterations $I \geq b$ to further reduce the variance of $\widehat{I}_l(\theta)$, and hence that of $\widehat{S}(\theta)$, at a cost $\mathbb{E}[c(\theta)]$ that grows linearly with I. On the other hand, when employing unbiased estimators within an EM algorithm or a stochastic gradient method, taking large values of I to obtain low variance gradient estimators would be inefficient. Choosing the tuning parameters (N, b, I) to maximize the efficiency of the resulting parameter inference algorithm is a highly non-trivial problem, and could be the topic of future work.

Given a choice of (N, b, I), one can produce $R \in \mathbb{N}$ independent replicates $\widehat{S}(\theta)_r = \sum_{l=0}^{L_r} \widehat{I}_l(\theta)_r/\mathcal{P}_l, r \in \{1, \ldots, R\}$, of $\widehat{S}(\theta)$ in parallel, and compute the average $\bar{S}(\theta) = R^{-1} \sum_{r=1}^{R} \widehat{S}(\theta)_r$ to approximate $S(\theta)$. To see the connection to MLMC, we follow Vihola (2018, Example 3) by noting that $\bar{S}(\theta)$ has the same distribution as the random variable

$$\sum_{l=0}^{\infty} \frac{1}{\mathbb{E}\left[R_l\right]} \sum_{r=1}^{R_l} \widehat{I}_l(\theta)_r, \tag{26}$$

where $R_l = \sum_{r=1}^{R} \mathbb{I}(L_r \geq l)$ has expectation $\mathbb{E}[R_l] = R\mathcal{P}_l$. Vihola (2018) proposed new unbiased estimators with lower variance than $\bar{S}(\theta)$ by sampling the random variables $(L_r)_{r=1}^{R}$ that define $(R_l)_{l=0}^{\infty}$ in (26) using stratification. As the number of replicates $R \to \infty$, these improved estimators also attain the same limiting variance as the idealized MLMC estimator $\widetilde{S}(\theta) = \sum_{l=0}^{\infty} \tilde{R}_l^{-1} \sum_{r=1}^{\tilde{R}_l} \widehat{I}_l(\theta)_r$, where $\tilde{R}_l = \lfloor R\mathcal{P}_l \rfloor$ is allocated using the PMF $(P_l)_{l=0}^{\infty}$. From Vihola (2018, Thoerem 5), this asymptotic variance is given by $\lim_{R\to\infty} R\mathrm{Var}[\widetilde{S}(\theta)^j] = \sum_{l=0}^{\infty} \mathrm{Var}[\widehat{I}_l(\theta)^j]/\mathcal{P}_l < \mathrm{Var}[\widehat{S}(\theta)^j]$, for all $j \in \{1, \ldots, d_g\}$.

4. Analysis

This section is concerned with the theoretical validity of our approach. We first introduce some notation needed to state the assumptions which we will rely on. Let $(\mathsf{E}, \mathcal{E})$ be an arbitrary measurable space. We write $\mathcal{B}_b(\mathsf{E})$ as the collection of real-valued, bounded, and measurable functions on E. For real-valued $\varphi : \mathsf{E} \to \mathbb{R}$ (and vector-valued $\varphi : \mathsf{E} \to \mathbb{R}^d$), let $\mathcal{C}^j(\mathsf{E})$ (and $\mathcal{C}_d^j(\mathsf{E})$) denote the collection of $j \in \mathbb{N}$ times continuously differentiable functions, and $\mathcal{C}(\mathsf{E})$ (and $\mathcal{C}_d(\mathsf{E})$) for the collection of continuous functions. We write $\varphi \in \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$ if the real-valued function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz w.r.t. the $\mathbb{L}_2$-norm $\|\cdot\|_2$, i.e. if there exists a constant $C < \infty$ such that $|\varphi(x) - \varphi(y)| \leq C\|x - y\|_2$ for all $x, y \in \mathbb{R}^d$, and $\|\varphi\|_{\mathrm{Lip}}$ as the Lipschitz constant. We recall the definitions of $\Sigma(x) = \sigma(x)\sigma(x)^*$ and $b_\theta(x) = \Sigma(x)^{-1}\sigma(x)^* a_\theta(x)$ for $x \in \mathbb{R}^d$ as they appear in the following.

Assumption 1 *The drift function $a : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$ and diffusion coefficient $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfy:*

(i) (Smoothness) For any $\theta \in \Theta$, $a_\theta^j \in \mathcal{C}^2(\mathbb{R}^d)$ for all $j \in \{1, \ldots, d\}$ components of a_θ, and $\sigma^{j,k} \in \mathcal{C}^2(\mathbb{R}^d)$ for all $(j, k) \in \{1, \ldots, d\}$ components of σ. Also, for any $x \in \mathbb{R}^d$, $\theta \mapsto a_\theta^j(x) \in \mathcal{C}(\Theta)$ for all $j \in \{1, \ldots, d\}$.

(ii) (Uniform ellipticity) $\Sigma(x)$ is uniformly positive definite for all $x \in \mathbb{R}^d$.

(iii) (Globally Lipschitz) For any $\theta \in \Theta$, there exists a constant $C < \infty$ such that $|a_\theta^j(x) - a_\theta^j(x')| + |\sigma^{j,k}(x) - \sigma^{j,k}(x')| \leq C\|x - x'\|_2$ for all $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$ and $(j, k) \in \{1, \ldots, d\}^2$.

Assumption 2 *The drift function $a : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$, diffusion coefficient $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and conditional density $g : \Theta \times \mathbb{R}^d \times \mathbb{R}^{d_y} \to \mathbb{R}^+$ satisfy:*

(i) the inverse of $x \mapsto \Sigma(x)$ satisfies $[\Sigma^{-1}]^{j,k} \in \mathcal{B}_b(\mathbb{R}^d) \cap \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$ for all $(j, k) \in \{1, \ldots, d\}^2$.

(ii) *For any $\theta \in \Theta$, $a_\theta^j \in \mathcal{B}_b(\mathbb{R}^d)$ for all $j \in \{1, \ldots, d\}$, and $\sigma^{j,k} \in \mathcal{B}_b(\mathbb{R}^d)$ for all $(j,k) \in \{1, \ldots, d\}^2$.*

(iii) *For any $\theta \in \Theta$, there exists $0 < \underline{C} < \overline{C} < \infty$ such that $\underline{C} \leq g_\theta(y|x) \leq \overline{C}$ for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{d_y}$. In addition, for any $(\theta, y) \in \Theta \times \mathbb{R}^{d_y}$, we have $g_\theta(y|\cdot) \in \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$.*

(iv) *For any $(\theta, y) \in \Theta \times \mathbb{R}^{d_y}$, $[\nabla_\theta \log g_\theta(y|\cdot)]^j \in \mathcal{B}_b(\mathbb{R}^d) \cap \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$ for all $j \in \{1, \ldots, d_\theta\}$.*

(v) *For any $\theta \in \Theta$, $[\nabla_\theta [b_\theta]^j]^k, [\nabla_\theta (b_\theta^j)^2]^k \in \mathcal{B}_b(\mathbb{R}^d) \cap \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$ for all $(j,k) \in \{1, \ldots, d\} \times \{1, \ldots, d_\theta\}$.*

Assumptions 1 and 2 should be understood as sufficient conditions to verify the validity of our proposed methodology, and are not necessary for its implementation. Although some of these assumptions are strong, they have been adopted to simplify the exposition of our analysis, as is common in theoretical works on particle filtering. Some assumptions can be relaxed at the expense of more involved and lengthy technical arguments.

We first give an intermediate result on the convergence of the time-discretized approximation S_l defined in (6).

Theorem 1 *Under Assumptions 1 and 2, the choice of G_θ in (12) and G_θ^l in (15), for any $(T, \theta) \in \mathbb{N} \times \Theta$, there exists a constant $C < \infty$ such that for any $l \in \mathbb{N}_0$, $\|S_l(\theta) - S(\theta)\|_1 \leq C\Delta_l^{1/2}$.*

The proof of Theorem 1 which involves studying the time-discretization of diffusions can be found in Section B.2 of the supplementary material. The following result establishes the desired properties of our estimators.

Theorem 2 *Under Assumptions 1 and 2, the choice of G_θ in (12) and G_θ^l in (15), for any number of particles $N \geq 2$, burn-in $b \in \mathbb{N}_0$ and number of iterations $I \geq b$, there exists a choice of PMF $(P_l)_{l=0}^\infty$ such that for any $\theta \in \Theta$, the estimator $\widehat{S}(\theta)$ in (19) is unbiased and has finite variance.*

We remark that Theorem 2 can be extended to other functionals. For instance, in the context of the EM algorithm, the result follows under the same assumptions and only notational changes in the proof. The proof of Theorem 2 in Section B.3 of the supplementary material involves studying various aspects of the 4-CCPF chains and its initialization, followed by the properties of our estimators that are constructed using these coupled Markov chains. It follows from the proof that the left-hand side of (18) is upper-bounded by

$$C(\theta) \sum_{l=0}^{\infty} \mathcal{P}_l^{-1} \Delta_l^{2\phi}, \tag{27}$$

where $C(\theta) < \infty$ is a parameter-dependent constant and $\phi > 0$ is a constant determined in our analysis that does not depend on l; the exact value of ϕ is 'small' but positive. Hence any choice of PMF $(P_l)_{l=0}^\infty$ such that the sum (27) is finite would be valid; e.g. $P_l \propto \Delta_l^{2\phi\alpha}$ for any $\alpha \in (0, 1)$. Due to the technical complexity of the problem and algorithms under consideration, the rate in (27) is not sharp. We conjecture that the correct rate corresponds to having $\phi = 1/4$ and a better rate of $\phi = 1/2$ can be obtained in the case of constant diffusion coefficient σ.

Using Lemma 37 in the supplementary material, we upper-bound the expected cost $\mathbb{E}[c(\theta)]$ by

$$C(\theta, T, N, b, I) N T \sum_{l=0}^{\infty} 2^l \mathcal{P}_l, \tag{28}$$

where $C(\theta, T, N, b, I) < \infty$ is another constant that is independent of l. As we have $\phi \leq 1/2$, there is no choice of PMF $(P_l)_{l=0}^{\infty}$ that can keep both (27) and (28) finite. This is a consequence of employing the Euler–Maruyama discretization scheme (13) and the choice of coupled resampling scheme in Algorithm 5 despite their general applicability. Future work could consider the antithetic truncated Milstein scheme of Giles and Szpruch (2014) and improved coupled resampling schemes such as Ballesio et al. (2022). This issue with the Euler–Maruyama scheme is well-understood and studied by Rhee and Glynn (2015, Section 4) who suggested choosing the PMF $(P_l)_{l=0}^{\infty}$ to ensure unbiased estimators with finite variance but infinite expected cost. Our numerical implementations will follow their approach by choosing $P_l \propto \Delta_l^{1/2} l(\log_2(1+l))^2$ and $P_l \propto \Delta_l l(\log_2(1+l))^2$ in the case of non-constant and constant diffusion coefficients, respectively. Under these choices, the unbiased estimators achieve computational complexities that are similar to standard MLMC estimators (Giles, 2008). We stress that it is possible to achieve unbiased estimators with finite variance and finite expected cost using our computational framework in other settings (Heng et al., 2023). It is also worth noting that exact simulation algorithms for (unconditioned) diffusions that are applicable in similar generality as our work also have infinite expected cost (Blanchet and Zhang, 2020).

5. Applications

5.1 Ornstein–Uhlenbeck Process

We consider an Ornstein–Uhlenbeck process $X = (X_t)_{0 \leq t \leq T}$ in $\mathbb{R}$, defined by the SDE

$$dX_t = \theta_1(\theta_2 - X_t)dt + \sigma dW_t, \quad X_0 = 0. \tag{29}$$

The parameter $\theta_1 > 0$ can be interpreted as the speed of the mean reversion to the long-run equilibrium value $\theta_2 \in \mathbb{R}$. This corresponds to (1) with initial condition $x_\star = 0$, linear drift function $a_\theta(x) = \theta_1(\theta_2 - x)$ and constant diffusion coefficient $\sigma(x) = \sigma > 0$ for $x \in \mathbb{R}$. We assume that the process is observed at unit times with Gaussian measurement errors, i.e. $Y_t | X \sim g_\theta(\cdot | X_t) = \mathcal{N}(X_t, \theta_3)$ independently for $t \in \{1, \ldots, T\}$ and some $\theta_3 > 0$. We will generate observations $y_{1:T}$ by simulating from the model with parameter $\theta = (\theta_1, \theta_2, \theta_3) = (2, 7, 1)$. This setting is considered as it is possible to compute the score function $S(\theta)$ exactly using Kalman smoothing; see Section C.1 of the supplementary material for details and model-specific expressions to evaluate (15).

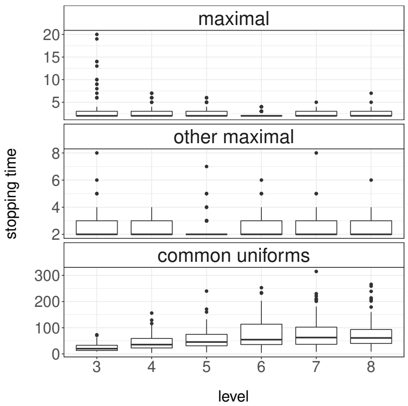
Figure 4a illustrates how the distribution of the stopping time $\bar{\tau}_\theta^l$ varies with the discretization level l on a simulated data set with $T = 25$ observations. We took $l = 3$ as the lowest discretization level as lower levels lead to numerically unstable trajectories.

As alluded earlier, the coupled resampling scheme proposed in Algorithm 5 (referred to as "maximal") leads to smaller and more stable stopping times for large enough levels. As alternatives, we consider a modification ("other maximal") that ensures 4-CCPF admits 2-CCPF as marginals on each level, and a scheme that uses common uniform random variables ("common uniforms"). While the schemes based on maximal couplings have comparable stopping times, the approach based on common uniform variables gives rise to significantly larger stopping times. Figure 4b reveals that these two alternative coupled resampling schemes do not induce sufficient dependencies between the four CPF chains. As the variance of the estimated increment does not decrease with the discretization level, this precludes their use within our unbiased estimation framework.

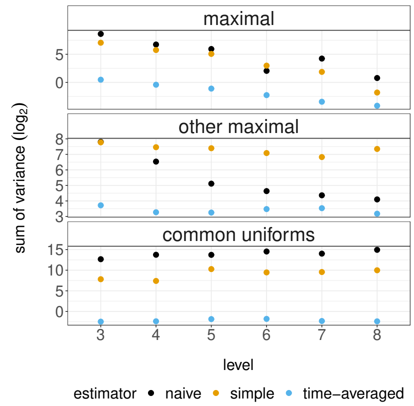
To show the impact of the choice of b and I, we considered three types of estimators corresponding to having $b = 0$ and $I = b$ ("naive"); $b =$ 90%-quantile($\bar{\tau}_\theta^l$) at level $l = 3$ and $I = b$ ("simple"); and $b =$ 90%-quantile($\bar{\tau}_\theta^l$) and $I = 10b$ ("time-averaged"), where 90%-quantile($\bar{\tau}_\theta^l$) denotes the 90% sample quantile of the stopping time at level l. The benefits of increasing b and I in terms of variance reduction are consistent with findings in Jacob et al. (2020a,b). Under our proposed coupling, all three choices yield estimators of increments whose variance decrease exponentially with the level, which agrees with our theoretical results (see Lemma 39 in the supplementary material). Hence we can employ any of these estimators within the estimation framework outlined in Section 3.4. Figure 4c displays the resulting squared error $\|\widehat{S}(\theta) - S(\theta)\|_2^2$ and cost of 100 independent replicates. This plot suggests having $N = 128$ particles is sufficient in the case of $T = 25$ observations. The choice of b and I also allows a tradeoff between error and cost. As we increase the number of observations T, Figure 4d shows it is important to scale the number of particles N linearly with T to obtain stable and non-exponentially increasing stopping times. Lastly, Figures 4e and 4f concern the averaging of independent replicates of the score estimator $(\widehat{S}(\theta)_r)_{r=1}^R$. Figure 4e shows that the average $\bar{S}(\theta) = R^{-1}\sum_{r=1}^R \widehat{S}(\theta)_r$ satisfies the standard Monte Carlo rate as $R \to \infty$, which is consistent with its properties in Theorem 2, at a linear cost in R as illustrated in Figure 4f.

5.2 Logistic Diffusion Model for Population Dynamics of Red Kangaroos

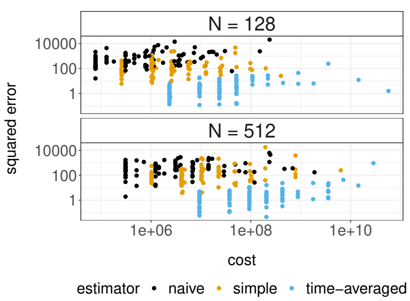
Next we consider an application from population ecology to model the dynamics of a population of red kangaroos (Macropus rufus) in New South Wales, Australia. Figure 5a displays data $y_{t_1}, \ldots, y_{t_P} \in \mathbb{N}_0^2$ from Caughley et al. (1987), which are double transect counts on $P = 41$ occasions at irregular times $(t_p)_{p=1}^P$ between 1973 to 1984. The latent population size $Z = (Z_t)_{t_1 \le t \le t_P}$ is assumed to follow a logistic diffusion process with environmental variance (Dennis and Costantino, 1988; Knape and De Valpine, 2012) defined by $dZ_t = (\theta_3^2/2 + \theta_1 - \theta_2 Z_t)Z_t dt + \theta_3 Z_t dW_t$, $Z_{t_1} \sim \mathcal{LN}(5, 10^2)$, where $\mathcal{LN}$ denotes the log-Normal distribution. The parameters $\theta_1 \in \mathbb{R}$ and $\theta_2 > 0$ can be seen as coefficients describing how the growth rate depends on the population size. As the parameter $\theta_3 > 0$ appears in the diffusion coefficient of the considered diffusion process we apply the Lamperti transformation $X_t = \Psi(Z_t) = \log(Z_t)/\theta_3$. By Itô's lemma, the transformed process $X = (X_t)_{t_1 \le t \le t_P}$ satisfies the SDE (1) with random initialization $X_{t_1} \sim \mu_\theta = \mathcal{N}(5/\theta_3, 10^2/\theta_3^2)$, drift function $a_\theta(x) = \theta_1/\theta_3 - (\theta_2/\theta_3)\exp(\theta_3 x)$ and unit diffusion coefficient $\sigma(x) = 1$ for $x \in \mathbb{R}$. The observations $(Y_{t_p})_{p=1}^P$ are modelled as conditionally



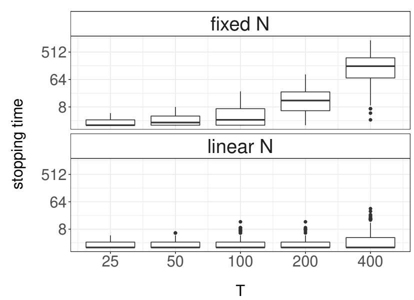
(a) Boxplots of stopping time $\bar{\tau}_\theta^l$ against discretization level l



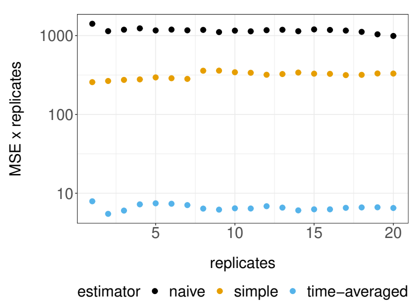
(b) Variance of score increment estimator summed over components $\sum_{j=1}^{d_\theta} \mathrm{Var}[\widehat{I}_l(\theta)^j]$ against discretization level l



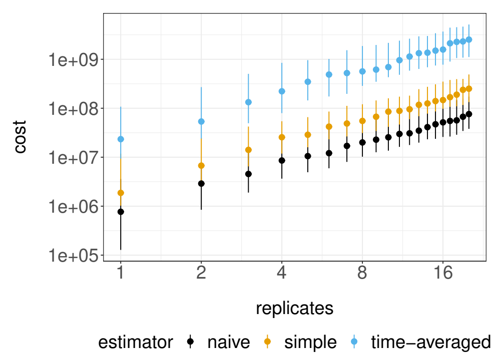
(c) Squared error $\|\widehat{S}(\theta) - S(\theta)\|_2^2$ against cost with $N \in \{128, 512\}$ particles



(d) Boxplots of stopping time $\bar{\tau}_\theta^l$ for discretization level $l = 6$ with $N = 128$ fixed or linearly increasing number of particles



(e) Mean squared error $\mathrm{MSE}(\theta) = \mathbb{E}[\|\bar{S}(\theta) - S(\theta)\|_2^2]$ against number of averaged replicates R



(f) Boxplots of cost against number of averaged replicates R

Figure 4: Behaviour of different coupling methods and score estimators at the data generating parameter $\theta = (2, 7, 1)$ of the Ornstein–Uhlenbeck model in Section 5.1. $T = 25$ observations and $N = 128$ particles were employed unless stated otherwise. These plots are based on 100 independent repetitions.

independent given X and negative Binomial distributed, i.e. the conditional density at time $t \in \{t_1, \dots, t_P\}$ is $g_\theta(y_t|x_t) = \mathcal{NB}(y_t^1; \theta_4, \exp(\theta_3 x_t))\mathcal{NB}(y_t^2; \theta_4, \exp(\theta_3 x_t))$, where $\theta_4 > 0$. We will use a parameterization of the negative Binomial distribution that is common in ecology, $\mathcal{NB}(y; r, \mu) = \frac{\Gamma(y+r)}{\Gamma(r)y!}(\frac{r}{r+\mu})^r(\frac{\mu}{r+\mu})^y$ for $y \in \mathbb{N}_0$, where $r > 0$ is the dispersion parameter and $\mu > 0$ is the mean parameter. The $d_\theta = 4$ unknown parameters to be inferred are $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta = \mathbb{R} \times (0, \infty)^3$.

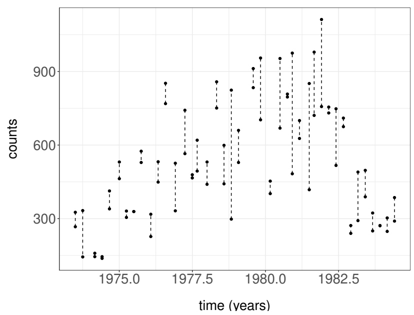
Application of our methodology for score estimation requires some minor modifications. As the initial distribution μ_θ depends on θ_3, the representation in (3) and (6) require adding $\nabla_\theta \log \mu_\theta(X_{t_1})$ to (12) and (15); see Section C.2 of the supplementary material for model-specific expressions. To deal with irregular observation times $(t_p)_{p=1}^P$, we set the step-size at discretization level zero as the size of the smallest time interval, i.e. $\Delta_0 = \min_{p=2,\dots,P} t_p - t_{p-1}$. Higher levels $l \in \mathbb{N}$ will employ $\Delta_l = \Delta_0 2^{-l}$. For level $l \in \mathbb{N}_0$, the first time interval $[t_1, t_2]$ is discretized using Δ_l sequentially, i.e. we set $s_k = t_1 + k\Delta_l$ for $k \in \{0, \dots, m_{l,1}\}$ with $m_{l,1} = \lfloor (t_2 - t_1)/\Delta_l \rfloor$, and $s_k = t_2$ for $k = m_{l,1} + 1$ if $(t_2 - t_1)/\Delta_l \notin \mathbb{N}$. The subsequent time intervals are then discretized in the same manner.

Figure 5b illustrates how the median and the 90% quantile of the stopping time $\bar{\tau}_\theta^l$ vary with the discretization level l, the impact of the number of particles N, and the benefits of employing adaptive resampling. As before, the coupled resampling scheme in Algorithm 5 results in stopping times that are smaller for higher discretization levels, with less variability over levels as the number of particles increases. Moreover, resampling only when the effective sample size is less than $N/2$ allows us to induce more dependencies between the multiple CPF chains at lower discretization levels. Using $N = 256$ particles and adaptive resampling, Figure 5c examines the rate at which the variance of the estimated increment decreases with the discretization level. Here we consider the "naive" and "simple" estimators described in Section 5.1, with a burn-in of $b = 90\%\text{-quantile}(\bar{\tau}_\theta^l)$ at level $l = 3$, and omit the more costly "time-averaged" estimator. From the plot, both type of estimators have similar rate of decay and are valid choices in our score estimation methodology. Using the "simple" estimator, Figure 5d verifies that the average of R independent replicates of the resulting score estimator $\bar{S}(\theta)$ satisfies the standard Monte Carlo rate as $R \to \infty$.

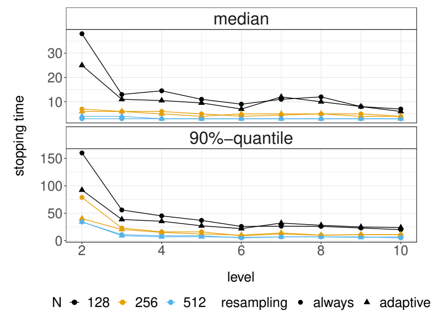
Lastly, we perform Bayesian parameter inference by employing our score estimators within the SGLD framework (Welling and Teh, 2011). We rely on logarithmic transformations to deal with positivity parameter constraints, and specify the prior distribution for the transformed parameters $(\theta_1, \log\theta_2, \log\theta_3, \log\theta_4)$ as $\mathcal{N}_{d_\theta}(\mu_0, \Sigma_0)$, with $\mu_0 = (0, -1, -1, -1)$ and $\Sigma_0 = \text{diag}(5^2, 2^2, 2^2, 2^2)$. As $\log\theta_3$ has a significantly different scale compared to the other parameters, we let the learning rate in (5) be component-dependent by taking $\varepsilon_m = \text{diag}((100 + m)^{-0.6}(10^{-2}, 10^{-2}, 10^{-4}, 10^{-2}))$ at iteration $m \geq 1$. The algorithmic settings used to produce score estimators are the same as in Figure 5d with $R = 1$ realization. Figure 5e shows the empirical average, weighted by the learning rates as in Welling and Teh (2011), over 7500 iterations of the resulting SGLD algorithm for each parameter.

5.3 Neural Network Model for Grid Cells in the Medial Entorhinal Cortex

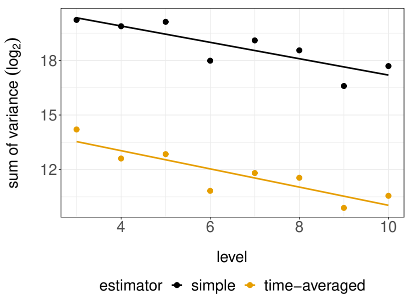
As our final application, we consider a neural network model for single neurons to analyze grid cells spike data (`https://www.ntnu.edu/kavli/research/grid-cell-data`) recorded in the medial entorhinal cortex of rats that were running on a linear track (Hafting



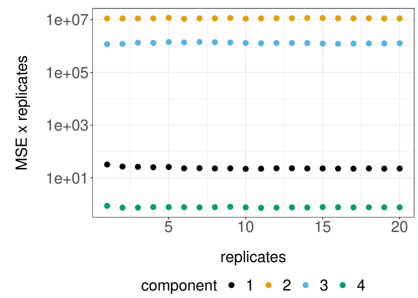
(a) Double transect counts



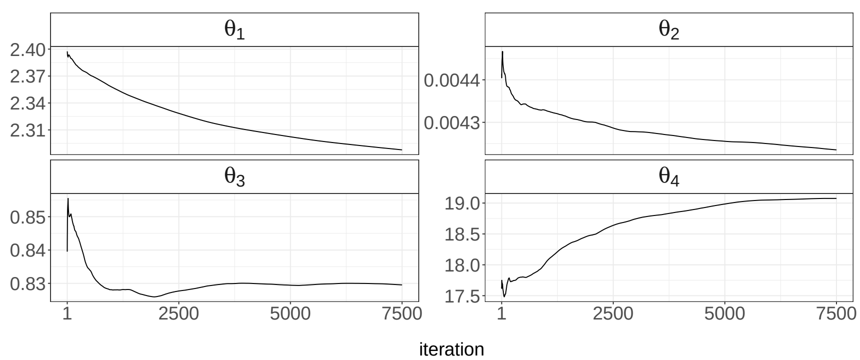
(b) Sample median and 90% quantile of stopping time $\bar{\tau}_\theta^l$ against discretization level l



(c) Variance of score increment estimator summed over components $\sum_{j=1}^{d_\theta} \mathrm{Var}[\widehat{I}_l(\theta)^j]$ against discretization level l



(d) Mean squared error $\mathrm{MSE}(\theta)^j = \mathbb{E}[|\bar{S}(\theta)^j - S(\theta)^j|^2]$ against number of averaged replicates R for component $j \in \{1, \dots, d_\theta\}$



(e) SGLD weighted empirical average for each component

Figure 5: Behaviour at parameter $\theta = (2.397, 4.429 \times 10^{-3}, 0.840, 17.631)$ of the logistic diffusion model in Section 5.2. $N = 256$ particles were employed unless stated otherwise. These plots are based on 100 independent repetitions.

et al., 2008). The neural states $Z_t = (Z_t^1, Z_t^2)$ of two grid cells that were simultaneously recorded is assumed to follow

$$dZ_t^1 = \left(\alpha_1 \tanh(\beta_1 Z_t^2 + \gamma_1) - \delta_1 Z_t^1\right) dt + \sigma_1 dW_t^1, \qquad (30)$$
$$dZ_t^2 = \left(\alpha_2 \tanh(\beta_2 Z_t^1 + \gamma_2) - \delta_2 Z_t^2\right) dt + \sigma_2 dW_t^2,$$

for $t \in [0,T]$, where $(\alpha_1, \alpha_2) \in \mathbb{R}^2$ controls the amplitude, $(\beta_1, \beta_2) \in \mathbb{R}^2$ describes the connectivity between the cells, $(\gamma_1, \gamma_2) \in \mathbb{R}^2$ are baseline levels, $(\delta_1, \delta_2) \in (0,\infty)^2$ determines the strength of the mean reversion towards the origin. We assume $Z_0 = (0,0)$ at the beginning of the experiment. This diffusion is motivated by an example in Kappen and Ruiz (2016), and modified for our purposes. To infer the unknown diffusivity parameters $(\sigma_1, \sigma_2) \in (0,\infty)^2$, we consider the transformation $X_t = (X_t^1, X_t^2) = \Psi(Z_t) = (Z_t^1/\sigma_1, Z_t^2/\sigma_2)$, which rescales each component of the diffusion. By Itô's formula, the transformed process $X = (X_t)_{0\le t\le T}$ satisfies the diffusion model (1) with initialization $x_\star = (0,0)$, drift function

$$a_\theta(x) = \begin{pmatrix} a_\theta^1(x) \\ a_\theta^2(x) \end{pmatrix} = \begin{pmatrix} \alpha_1 \tanh(\beta_1 \sigma_2 x^2 + \gamma_1)/\sigma_1 - \delta_1 x^1 \\ \alpha_2 \tanh(\beta_2 \sigma_1 x^1 + \gamma_2)/\sigma_2 - \delta_2 x^2 \end{pmatrix},$$

and diffusion coefficient $\sigma(x) = I_2$ for $x = (x^1, x^2) \in \mathbb{R}^2$.

The experimental data over a duration of $T = 20$ contains time stamps in $[0,T]$ when a spike at one of the two cells is recorded using tetrodes. Following Brown (2005), we adopt an inhomogenous Poisson point process to model these times. Let $t_p = pT2^{-6}$ for $p \in \{0,1,\dots,P\}$ denote a dyadic uniform discretization of $[0,T]$ into $P = 2^6$ time intervals. Given the latent process $X = (X_t)_{0\le t\le T}$, the number of spikes $Y_{t_p}^i$ occurring in each time interval $[t_{p-1}, t_p]$ at cell $i = 1,2$ is assumed to be conditionally independent of the other time intervals and the activity in the other cell, and follow a Poisson distribution with rate $\int_{t_{p-1}}^{t_p} \lambda_i(X_t^i)dt$. The intensity function for grid cell $i = 1,2$ is modelled as $\lambda_i(X_t^i) = \exp(\kappa_i + X_t^i)$, where $\kappa_i \in \mathbb{R}$ represents a baseline level. The observed counts $y_{t_p} = (y_{t_p}^1, y_{t_p}^2)$ for interval $p \in \{1,\dots,P\}$, computed from the experimental data, are displayed in Figure 6a. The conditional likelihood of the observation model is $p_\theta(y_{t_1},\dots,y_{t_P}|X) = \prod_{p=1}^P g_\theta(y_{t_p}|(X_t)_{t_{p-1}\le t\le t_p})$ with the intractable conditional density

$$g_\theta(y_{t_p}|(X_t)_{t_{p-1}\le t\le t_p}) = \prod_{i=1}^2 \mathcal{P}oi\left(y_{t_p}^i; \int_{t_{p-1}}^{t_p} \lambda_i(X_t^i)dt\right),$$

where $\mathcal{P}oi(y;\lambda) = \lambda^y \exp(-\lambda)/y!$ for $y \in \mathbb{N}_0$ denotes the PMF of a Poisson distribution with rate $\lambda > 0$. To approximate the conditional likelihood, at level $l \ge 6$, we discretize the time interval $[0,T]$ in a similar manner using $s_k = k\Delta_l$ for $k \in \{0,1,\dots,K_l\}$, where $\Delta_l = T2^{-l}$ is the step-size and $K_l = 2^l$ is the number of time steps. Under the time-discretized process $X_{0:T} = (X_{s_k})_{k=0}^{K_l}$, the resulting approximation of the conditional likelihood is $p_\theta^l(y_{t_1},\dots,y_{t_P}|X_{0:T}) = \prod_{p=1}^P g_\theta^l(y_{t_p}|(X_t)_{t_{p-1}\le t\le t_p})$ with the corresponding conditional density

$$g_\theta^l(y_{t_p}|(X_t)_{t_{p-1}\le t\le t_p}) = \prod_{i=1}^2 \mathcal{P}oi\left(y_{t_p}^i; \Delta_l \sum_{t:t_{p-1}\le t\le t_p} \lambda_i(X_t^i)\right). \qquad (31)$$

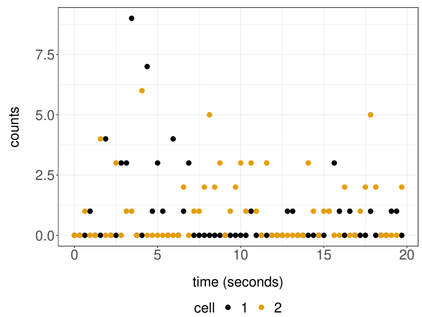
By using these level-dependent observation densities (31) in Section 2.2, our proposed methodology can then be applied. There are $d_\theta = 12$ parameters $\theta = (\theta_1, \theta_2)$ to be inferred, where $\theta_i = (\alpha_i, \beta_i, \gamma_i, \delta_i, \sigma_i, \kappa_i)$ denote the parameters associated to cell $i = 1, 2$. We refer the reader to Section C.3 of the supplementary material for model-specific expressions to implement score estimation.

We consider an extension of the proposed method based on the conditional ancestor sampling particle filter (CASPF) (Lindsten et al., 2014) as the basic algorithmic building block. In Figure 6b, we observe that the stopping times of CASPF are smaller than CPF for lower discretization levels, and similar for higher discretization levels. Although this is consistent with CASPF having better mixing properties than CPF (Lee et al., 2020), the use of CASPF is invalid in our setting as the algorithm is not well-defined as the discretization level goes to infinity. This issue stems from degeneracy of the transition kernel of the Euler-Maruyama discretization (13) as the step-size goes to zero; see also Beskos et al. (2021, Section 3) for a related discussion. Some pathological behaviour can be seen in Figure 6c which checks the validity of using both MCMC algorithms and "naive" and "simple" estimators (as described in Section 5.1) within our methodology. For "simple" estimators, the burn-in was taken as $b = 90\%$-quantile$(\bar{\tau}_\theta^l)$ at level $l = 11$. While the variance of the estimated increment decays with the discretization level for estimators based on CPF, it does not seem to be the case for estimators based on CASPF.

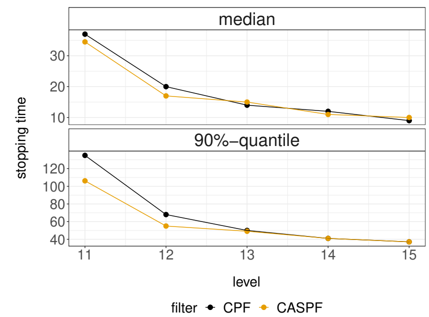
Lastly, we combine our score estimators and the SGA scheme in (4) to perform maximum likelihood estimation. The score estimation relies on the CPF algorithm and the "simple" estimator with a burn-in of $b = 100$. Positivity parameter constraints are dealt with using logarithmic transformations and a constant learning rate of $\varepsilon_m = 10^{-3}$ is employed. Figure 6d illustrates how the distribution of the Polyak–Ruppert average evolves over the iterations, estimated using independent runs of SGA. We note that only 86 out of 100 runs were considered, as there were 14 instances of the variance of the score estimator driving the SGA algorithm to regions of the parameter space where the stopping times are prohibitively large, causing the SGA to stall. This behaviour is due to poor mixing properties of the underlying CPF algorithm at very unlikely regions of the parameter space and the use of a constant learning rate. A discussion on how to improve the MCMC algorithm and adapt the learning rate is given in Section 6. The parameter estimates of β_1 and β_2 support the use of a joint model (30) for both grid cells, and indicates that these cells are positively dependent.

6. Discussion

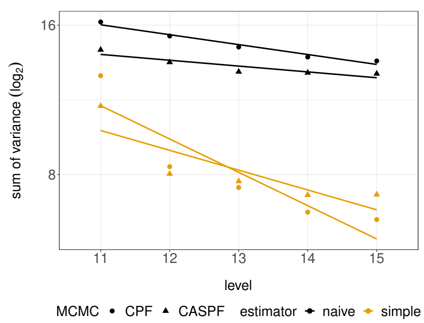
Although the proposed unbiased estimation methodology to remove both time-discretization error and MCMC burn-in bias is conceptually appealing and applicable to a large class of diffusion models, it is important to note that it has some computational limitations, as evidenced in Section 5.3. For our approach to be computationally feasible, the distribution of stopping times cannot be too heavy-tailed, i.e. the pair of CPF chains on each time-discretization level has to meet in reasonable computation time. Even if our proposed coupling is adequate, it follows from the coupling inequality that this is impractical if the marginal CPF kernel has poor mixing properties. Hence the underlying assumption here is



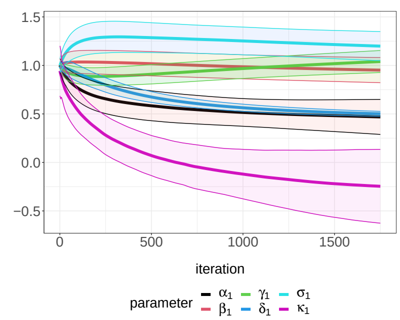
(a) Counts $y_{t_p} = (y^1_{t_p}, y^2_{t_p})$ on time intervals of duration $T2^{-6} = 0.3125$



(b) Sample median and 90% quantile of stopping time $\bar{\tau}^l_\theta$ against discretization level l



(c) Variance of score increment estimator summed over components $\sum_{j=1}^{d_\theta} \mathrm{Var}[\widehat{I}_l(\theta)^j]$ against discretization level l



(d) Mean ($\pm$ one standard deviation) of Polyak–Ruppert average for parameter estimates of θ_1 over 86 runs of SGA

Figure 6: Behaviour at parameter $\theta = (1, \dots, 1)$ of the neural network model in Section 5.3. The algorithmic settings involve $N = 256$ particles and adaptive resampling. These plots are based on 1000 independent repetitions unless stated otherwise.

that the corresponding BPF is performing sufficiently well (Lindsten et al., 2015; Andrieu et al., 2018).

Common regimes where this is not the case include high-dimensional state spaces (Snyder et al., 2008) and highly informative observations (Del Moral and Murray, 2015) with low probability under the law of the diffusion in (1) and the observation model. For some class of models, the curse of dimensionality can be tackled using particle filters that are modified to exploit certain properties of the model (Rebeschini and van Handel, 2015; Beskos et al., 2017). Further investigation is required to understand if these ideas can be used with our approach. Highly informative observations require carefully designed particle filters that simulate particle dynamics in a manner that incorporates information from the entire observation sequence (Richard and Zhang, 2007; Guarniero et al., 2017; Heng et al., 2020). In our setting of diffusion models, the optimal particle dynamics follows a diffusion process that is constructed using a Doob's h-transform (Rogers and Williams, 2000, p. 83). The numerical approximation of Doob's h-transform and its use within particle filtering has been explored in works such as Ruiz and Kappen (2017); Park and Ionides (2020); Mider et al. (2021); Chopin et al. (2023). As approximations of the optimal particle dynamics follow the diffusion process (1) with a change in the drift function, we anticipate the use of these methods to be quite straightforward within our unbiased estimation framework.

Even if our unbiased estimation methodology is performant, its application within parameter inference schemes may still be challenging when the parameter space is high-dimensional and the marginal likelihood function is highly non-convex with many local maxima. In such settings, we anticipate that the use of adaptive learning rate methods to improve the performance of stochastic gradient ascent (Zeiler, 2012; Duchi et al., 2011; Kingma and Ba, 2014), and momentum algorithms to accelerate its convergence (Nesterov, 1983; Qian, 1999; Sutskever et al., 2013).

Acknowledgments

J. Heng was funded by CY Initiative of Excellence ("Investissements d'Avenir" ANR-16-IDEX-0008. A. Jasra was supported by CUHK Shenzhen.

Appendix A. Parameter Dependence in Diffusion Coefficient

We consider how one can extend the ideas in this article to accommodate the case where the diffusion coefficient also depends on the parameter $\theta \in \Theta$. In this case, we have the SDE

$$dX_t = a_\theta(X_t)dt + \sigma_\theta(X_t)dW_t, \quad X_0 = x_\star \in \mathbb{R}^d,$$

where $\sigma : \Theta \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is assumed to be such that $\theta \mapsto \sigma_\theta$ is invertible and satisfies the conditions in Assumption 1 uniformly in $\theta \in \Theta$. Moreover, we shall suppose that all the conditions in Schauer et al. (2017) hold. For $k \in \mathbb{N}_0$ and $t \in [k, k+1]$, consider the diffusion bridge

$$dX_t = a^\circ_\theta(X_t)dt + \sigma_\theta(X_t)dW_t, \quad X_k = x_k, \; X_{k+1} = x_{k+1}, \tag{32}$$

where the drift function a°_θ is described in Schauer et al. (2017). Given a Brownian path $\boldsymbol{W}_k = (W_t)_{k\le t\le k+1}$, we denote the path-wise solution of the diffusion bridge as $F_{\theta,k}(\boldsymbol{W}_k, x_k, x_{k+1})$. Furthermore, for $G : \Theta\times\mathbb{R}^d \to \mathbb{R}$ given in Schauer et al. (2017, Equation 2.3) and a process $\boldsymbol{Z}_k = (Z_t)_{k\le t\le k+1}$, we define the functional $H_{\theta,k}(\boldsymbol{Z}_k) = \int_k^{k+1} G_\theta(Z_t)dt$. Using the change of measure in Schauer et al. (2017) along with the approach in Beskos et al. (2021) and Yonekura and Beskos (2022), one can write the marginal likelihood of observations $y_{1:T} = (y_t)_{t=1}^T$ as

$$p_\theta(y_{1:T}) = \widetilde{\mathbb{E}}_\theta\left[\prod_{t=1}^T g_\theta(y_t|X_t)\exp\left\{\sum_{t=0}^{T-1} H_{\theta,t}(F_{\theta,t}(\boldsymbol{W}_t, X_t, X_{t+1}))\right\}\right]. \tag{33}$$

In the above, $\widetilde{\mathbb{E}}_\theta$ denotes expectation w.r.t. the probability measure $\widetilde{\mathbb{P}}_\theta$ defined as

$$\widetilde{\mathbb{P}}_\theta(d(x_1,\ldots,x_T,\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{T-1})) = \prod_{t=1}^{T}\left\{\widetilde{p}_\theta(x_{t-1},x_t)dx_t\right\}\widetilde{\mathbb{W}}(d(\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{T-1})),$$

where $\widetilde{p}_\theta(x_{t-1},x_t)$ is the transition density of an auxiliary process on a unit time interval as constructed in Schauer et al. (2017) that is known and can be sampled, and $\widetilde{\mathbb{W}}(d(\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{T-1})) = \bigotimes_{k=0}^{T-1}\mathbb{W}(d\boldsymbol{W}_k)$ is given by the Wiener measure $\mathbb{W}$.

EM algorithm. The expectation step of an EM algorithm will involve computing

$$S(\theta,\theta_\star) = \breve{\mathbb{E}}_{\theta_\star}\left[\sum_{t=1}^{T}\log g_\theta(y_t|X_t) + \sum_{t=1}^{T}\log\widetilde{p}_\theta(X_{t-1},X_t) + \sum_{t=1}^{T-1}H_{\theta,t}(F_{\theta,t}(\mathbf{W}_t,X_t,X_{t+1}))\right], \tag{34}$$

where $\breve{\mathbb{E}}_\theta$ denotes expectation w.r.t. the probability measure

$$\breve{\mathbb{P}}_\theta(d(x_1,\ldots,x_T,\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{T-1})) = p_\theta(y_{1:T})^{-1}\times \prod_{t=1}^{T}g_\theta(y_t|x_t)\exp\Big\{\sum_{t=0}^{T-1}H_{\theta,t}(F_{\theta,t}(\mathbf{W}_t,x_t,x_{t+1}))\Big\}\widetilde{\mathbb{P}}_\theta(d(x_1,\ldots,x_T,\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{T-1})). \tag{35}$$

Gradient-based methods. Under regularity conditions, one can differentiate (33) and represent the score function $S(\theta) = \nabla_\theta\log p(y_{1:T})$ as

$$\breve{\mathbb{E}}_\theta\left[\sum_{t=1}^{T}\nabla_\theta\log g_\theta(y_t|X_t) + \sum_{t=1}^{T}\nabla_\theta\log\widetilde{p}_\theta(X_{t-1},X_t) + \sum_{t=1}^{T-1}\nabla_\theta H_{\theta,t}(F_{\theta,t}(\mathbf{W}_t,X_t,X_{t+1}))\right], \tag{36}$$

Practical implementation will require a discretized approximation of $H_{\theta,t}(F_{\theta,t}(\cdot))$ and $\nabla_\theta H_{\theta,t}(F_{\theta,t}(\cdot))$ which involves a gradient w.r.t. θ of a path-wise solution of the diffusion bridge (32). Although Euler-type approximations can be obtained, the resulting bias in the sense of Theorem 1 is significantly more complicated to analyze and is thus left as future work. We stress that only small modifications to our proposed methodology is necessary to obtain unbiased estimators of (34) and (36) in this case. A similar approach is considered in Beskos et al. (2021) for a class of continuous-time models. Alternatively, one could also consider using Malliavin techniques (Fournié et al., 1999), instead of the ideas described here.

Appendix B. Theoretical Analysis

B.1 Introduction and Preliminaries

Section B.2 provides some results on time-discretization of diffusions, which are needed for the proofs associated to Theorem 1 as well as the 4-CCPF (Algorithm 4). Our main technical arguments associated to Theorem 2 are given in Section B.3, followed by several remarks about the proofs and discussions of alternative strategies. This section of the appendix is intended to be read in the order in which it is presented. Some familiarity with the approach in Jasra et al. (2017) is also useful.

Note that our results concerning $\mathbb{L}_r$-norms are stated for $r \in [1, \infty)$; and can be extended to the case $r \in (0, 1)$ by Hölder's inequality. We will use this fact without further elaboration. Throughout our arguments, C will represent a finite constant whose value may change from line to line, but does not depend upon the discretization level. Any other dependencies in the various parameters considered will be made explicit in the statement of our results.

B.2 Results on Time-Discretized Diffusion Processes

In this section, we consider two diffusion process $X = (X_t)_{t\geq 0}$ and $X^\star = (X_t^\star)_{t\geq 0}$ on the filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P}_\theta)$ following (1), with the respective initial conditions $X_0 = x \in \mathbb{R}^d$ and $X_0^\star = x_\star \in \mathbb{R}^d$, and driven by the same Brownian motion. We will consider Euler discretizations (13) of $(X_t)_{t\geq 0}$ and $(X_t^\star)_{t\geq 0}$ at some given level l, denoted as $\widetilde{X}_{0:T}$ and $\widetilde{X}^\star_{0:T}$, driven by the same Brownian motion and with the initial conditions $\widetilde{X}_0 = x$ and $\widetilde{X}_0^\star = x_\star$. The expectation operator for the described processes is written as $\mathbb{E}_\theta$.

In addition to the previously defined terms $b_\theta(x) = \Sigma(x)^{-1}\sigma(x)^* a_\theta(x)$ and $\Sigma(x) = \sigma(x)\sigma(x)^*$, we introduce the function $\rho_\theta(x) = b_\theta(x)^*\Sigma(x)^{-1}\sigma(x)^*$ which allows us to rewrite (12) as

$$G_\theta(X) = -\frac{1}{2}\int_0^T \nabla_\theta \|b_\theta(X_t)\|_2^2 dt + \int_0^T \nabla_\theta \rho_\theta(X_t) dX_t + \sum_{t=1}^T \nabla_\theta \log g_\theta(y_t|X_t),$$

and (15) as

$$G_\theta^l(X_{0:T}) = \\ -\frac{1}{2}\sum_{k=1}^{K_l} \nabla_\theta \|b_\theta(X_{s_{k-1}})\|_2^2 \Delta_l + \sum_{k=1}^{K_l} \nabla_\theta \rho_\theta(X_{s_{k-1}})(X_{s_k} - X_{s_{k-1}}) + \sum_{t=1}^T \nabla_\theta \log g_\theta(y_t|X_t). \quad (37)$$

For notational convenience, we define the $d \times 1$ vector of derivatives of ρ_θ as

$$\kappa_{\theta,i}(x)^* = \Big(\frac{\partial}{\partial\theta_i}[\rho_\theta(x)]^1, \ldots, \frac{\partial}{\partial\theta_i}[\rho_\theta(x)]^d\Big),$$

for any $(i, x) \in \{1, \ldots, d_\theta\} \times \mathbb{R}^d$, and the conditional likelihood given states $x_1, \ldots, x_T \in \mathbb{R}^d$ as $\varphi_\theta(x_1, \ldots, x_T) = \prod_{t=1}^T g_\theta(y_t|x_t)$. We now give the proof of Theorem 1 followed by several technical lemmata that are required to establish the theorem.

Proof [Proof of Theorem 1] We consider the proof for any given component $i \in \{1, \ldots, d_\theta\}$ and decompose the error of the score function (6) at level $l \in \mathbb{N}_0$ as

$$[S_l(\theta) - S(\theta)]^i = T_1 + T_2 \quad (38)$$

where

$$T_1 = \frac{\mathbb{E}_\theta[\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star)[G_\theta^l(\widetilde{X}_{0:T}^\star)]^i]}{\mathbb{E}_\theta[\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star)]\mathbb{E}_\theta[\varphi_\theta(X_1^\star, \dots, X_T^\star)]}\Big(\mathbb{E}_\theta[\varphi_\theta(X_1^\star, \dots, X_T^\star)] - \mathbb{E}_\theta[\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star)]\Big),$$

$$T_2 = \frac{1}{\mathbb{E}_\theta[\varphi_\theta(X_1^\star, \dots, X_T^\star)]}\Big(\mathbb{E}_\theta[\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star)[G_\theta^l(\widetilde{X}_{0:T}^\star)]^i] - \mathbb{E}_\theta[\varphi_\theta(X_1^\star, \dots, X_T^\star)[G_\theta(X^\star)]^i]\Big).$$

Thus our objective is to provide bounds on the quantities T_1 and T_2 to conclude the proof.

For T_1, using Assumption 2, one has the upper-bound

$$T_1 \leq C \sum_{t=1}^{T} \mathbb{E}_\theta[\|\widetilde{X}_t^\star - X_t^\star\|_2],$$

then by using results on the convergence of Euler approximations (Kloeden and Platen, 2013), for $r > 0$

$$\mathbb{E}_\theta\big[\|\widetilde{X}_t^\star - X_t^\star\|_2^r\big]^{1/r} \leq C\Delta_l^{1/2} \tag{39}$$

one has

$$T_1 \leq C\Delta_l^{1/2}. \tag{40}$$

Note that using standard results on weak errors for diffusions one can improve this upper-bound to $T_1 \leq C\Delta_l$.

For T_2, using Assumption 2, we have $T_2 \leq C(T_3 + T_4)$ where

$$T_3 = \mathbb{E}_\theta\big[\{\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star) - \varphi_\theta(X_1^\star, \dots, X_T^\star)\}[G_\theta^l(\widetilde{X}_{0:T}^\star)]^i\big],$$
$$T_4 = \mathbb{E}_\theta\big[\varphi_\theta(X_1^\star, \dots, X_T^\star)\{[G_\theta^l(\widetilde{X}_{0:T}^\star)]^i - [G_\theta(X^\star)]^i\}\big].$$

For T_3, using Cauchy-Schwarz, we have the upper-bound

$$T_3 \leq \mathbb{E}_\theta\big[\{\varphi_\theta(\widetilde{X}_1^\star, \dots, \widetilde{X}_T^\star) - \varphi_\theta(X_1^\star, \dots, X_T^\star)\}^2\big]^{1/2}\mathbb{E}_\theta\big[\{[G_\theta^l(\widetilde{X}_{0:T}^\star)]^i\}^2\big]^{1/2}.$$

As the second term is bounded by C, we consider only the first. We have the upper-bound

$$T_3 \leq C \sum_{t=1}^{T} \mathbb{E}_\theta\big[\|\widetilde{X}_t^\star - X_t^\star\|_2^2\big]^{1/2} \leq C\Delta_l^{1/2}. \tag{41}$$

For T_4, noting that φ_θ is a bounded function under Assumption 2, applying Lemma 4 allows one to conclude that $T_4 \leq C\Delta_l^{1/2}$. Therefore, using $T_2 \leq C(T_3 + T_4)$ along with (41), we have

$$T_2 \leq C\Delta_l^{1/2}. \tag{42}$$

Combining (40) and (42) with (38) allows one to conclude the proof. ■

Remark 3 *We have adopted a strong error approach in our analysis to simplify the arguments involved. At the expense of more involved and lengthy arguments, we note that the upper-bound of Theorem 1 can be sharpened to $\mathcal{O}(\Delta_l)$ if one takes a weak error approach.*

Lemma 4 *Under Assumptions 1 and 2, for any* $(T, r, \theta, i) \in \mathbb{N} \times [1, \infty) \times \Theta \times \{1, \ldots, d_\theta\}$, *there exists a constant* $C < \infty$ *such that for any* $(l, x) \in \mathbb{N}_0 \times \mathbb{R}^d$

$$\mathbb{E}_\theta\Big[\big|[G_\theta^l(\widetilde{X}_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big]^{1/r} \leq C\Delta_l^{1/2},$$

with $\widetilde{X}_0 = X_0 = x$.

Proof We have that

$$\mathbb{E}_\theta\Big[\big|[G_\theta^l(\widetilde{X}_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big] \leq C(T_1 + T_2) \tag{43}$$

where

$$T_1 = \mathbb{E}_\theta\Big[\big|[G_\theta^l(\widetilde{X}_{0:T})]^i - [G_\theta^l(X_{0:T})]^i\big|^r\Big],$$
$$T_2 = \mathbb{E}_\theta\Big[\big|[G_\theta^l(X_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big],$$

where $X_{0:T} = (X_{s_k})_{k=0}^{K_l}$ are the states of the process $(X_t)_{t\geq 0}$ at the discretization times of the process $\widetilde{X}_{0:T}$. From (37), we have that $T_1 \leq C\sum_{j=3}^{6} T_j$, where

$$T_3 = \mathbb{E}_\theta\Big[\Big|\sum_{t=1}^{T}\Big\{\big[\nabla_\theta \log g_\theta(y_t|\widetilde{X}_t)\big]^i - \big[\nabla_\theta \log g_\theta(y_t|X_t)\big]^i\Big\}\Big|^r\Big],$$

$$T_4 = \Delta_l^r \mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l}\Big\{\big[\nabla_\theta\|b_\theta(\widetilde{X}_{s_{k-1}})\|_2^2\big]^i - \big[\nabla_\theta\|b_\theta(X_{s_{k-1}})\|_2^2\big]^i\Big\}\Big|^r\Big],$$

$$T_5 = \mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}[\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}]\Big|^r\Big],$$

$$T_6 = \mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l}\kappa_{\theta,i}(X_{s_{k-1}})^*[(\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}) - (X_{s_k} - X_{s_{k-1}})]\Big|^r\Big].$$

The term T_3 can be treated in almost the same manner as T_1 in the proof of Theorem 1, i.e. using a similar argument to the proof of the bound on T_1 in Theorem 1, one can deduce that

$$T_3 \leq C\Delta_l^{r/2}. \tag{44}$$

For T_4, using the fact that $\partial/\partial\theta_i[b_\theta^2]^j \in \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$ for any $(i,j) \in \{1,\ldots,d_\theta\} \times \{1,\ldots,d\}$, we have by first applying Minkowski's inequality

$$T_4 \leq C\Delta_l^r\Big(\sum_{k=1}^{K_l}\mathbb{E}_\theta\big[\|\widetilde{X}_{s_{k-1}} - X_{s_{k-1}}\|_2^r\big]^{1/r}\Big)^r.$$

Then using (39), it follows that

$$T_4 \leq C\Delta_l^{r/2}. \tag{45}$$

The terms T_5 and T_6 are bounded in Lemmata 6-7, so combining (44), (45) and the aforementioned lemmata with $T_1 \leq C\sum_{j=3}^{6} T_j$ yields

$$T_1 \leq C\Delta_l^{r/2}.$$

By Lemma 8, $T_2 \leq C\Delta_l^{r/2}$ and thus by (43) the proof is concluded. ■

Corollary 5 *Under Assumptions 1 and 2, for any $(T, r, \theta) \in \mathbb{N} \times [1, \infty) \times \Theta$, there exists a constant $C < \infty$ such that for any $(l, x) \in \mathbb{N}_0 \times \mathbb{R}^d$*

$$\mathbb{E}_\theta\Big[\big\|G_\theta^l(\widetilde{X}_{0:T}) - G_\theta(X)\big\|_2^r\Big]^{1/r} \leq C\Delta_l^{1/2}.$$

Proof By Minkowski's inequality

$$\mathbb{E}_\theta\Big[\big\|G_\theta^l(\widetilde{X}_{0:T}) - G_\theta(X)\big\|_2^r\Big]^{1/r} \leq \left(\sum_{i=1}^{d_\theta} \mathbb{E}_\theta\Big[\big|[G_\theta^l(\widetilde{X}_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big]^{2/r}\right)^{1/2}$$

so the proof follows by Lemma 4. ■

Lemma 6 *Under Assumptions 1 and 2, for any $(T, r, \theta, i) \in \mathbb{N} \times [1, \infty) \times \Theta \times \{1, \ldots, d_\theta\}$, there exists a constant $C < \infty$ such that for any $(l, x) \in \mathbb{N}_0 \times \mathbb{R}^d$*

$$\mathbb{E}_\theta\Bigg[\Big|\sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}[\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}]\Big|^r\Bigg] \leq C\Delta_l^{r/2},$$

with $\widetilde{X}_0 = X_0 = x$.

Proof We have the decomposition

$$\sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}[\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}] = M_{K_l} + R_{K_l},$$

where

$$M_{K_l} = \sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}\sigma(\widetilde{X}_{s_{k-1}})[W_{s_k} - W_{s_{k-1}}],$$

$$R_{K_l} = \Delta_l\sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}a_\theta(\widetilde{X}_{s_{k-1}}).$$

Thus by the C_r-inequality,

$$\mathbb{E}_\theta\Bigg[\Big|\sum_{k=1}^{K_l}\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}[\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}]\Big|^r\Bigg] \leq C\big(\mathbb{E}_\theta[|M_{K_l}|^r] + \mathbb{E}_\theta[|R_{K_l}|^r]\big). \quad (46)$$

We will bound the two terms on the R.H.S. of (46) individually. *Bound for* $\mathbb{E}_\theta[|M_{K_l}|^r]$. If we define

$$M_u = \sum_{k=0}^{u-1} \{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}\sigma(\widetilde{X}_{s_{k-1}})[W_{s_k} - W_{s_{k-1}}]$$

for any $u \in \mathbb{N}_0$, then $(M_u, \mathscr{F}_{u\Delta_l})_{u\in\mathbb{N}_0}$ is a martingale. It follows from this fact and from an application of the Burkholder-Gundy-Davis (BGD) inequality that

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l}(\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}\sigma(\widetilde{X}_{s_{k-1}})[W_{s_k} - W_{s_{k-1}}])^2\Big|^{r/2}\Big],$$

from which Minkowski's inequality yields

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\bigg(\sum_{k=1}^{K_l}\mathbb{E}_\theta\Big[\Big|\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\}\sigma(\widetilde{X}_{s_{k-1}})[W_{s_k} - W_{s_{k-1}}]\Big|^r\Big]^{2/r}\bigg)^{r/2}.$$

Using the C_r-inequality d^2 times, we obtain the bound

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\Bigg(\sum_{k=1}^{K_l}\bigg(\sum_{(m,j)\in\{1,\dots,d\}^2}\mathbb{E}_\theta\Big[\Big|\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}}) - \kappa_{\theta,i}(X_{s_{k-1}})\}^m \times \sigma(\widetilde{X}_{s_{k-1}})^{m,j}[W_{s_k} - W_{s_{k-1}}]^j\Big|^r\Big]\bigg)^{2/r}\Bigg)^{r/2}.$$

Using the fact that $\sigma^{m,j} \in \mathcal{B}_b(\mathbb{R}^d)$ along with the Cauchy-Schwarz inequality yields

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\Bigg(\sum_{k=1}^{K_l}\bigg(\sum_{(m,j)\in\{1,\dots,d\}^2}\mathbb{E}_\theta\Big[\Big|\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}}) - \kappa_{\theta,i}(X_{s_{k-1}})\}^m\Big|^{2r}\Big]^{1/2} \times \mathbb{E}_\theta\Big[\Big|[W_{s_k} - W_{s_{k-1}}]^j\Big|^{2r}\Big]^{1/2}\bigg)^{2/r}\Bigg)^{r/2}. \quad (47)$$

Since it holds that $[\kappa_{\theta,i}]^m \in \mathrm{Lip}_{\|\cdot\|}(\mathbb{R}^d)$, it follows from the same type of inequality as (39) that

$$\mathbb{E}_\theta\Big[\Big|\{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}}) - \kappa_{\theta,i}(X_{s_{k-1}})\}^m\Big|^{2r}\Big]^{1/2} \leq C\Delta_l^{r/2}, \quad (48)$$

and by standard properties of Brownian motion, we obtain

$$\mathbb{E}_\theta\Big[\Big|[W_{s_k} - W_{s_{k-1}}]^j\Big|^{2r}\Big]^{1/2} \leq C\Delta_l^{r/2}. \quad (49)$$

Combining (47) with (48) and (49) yields the upper-bound

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\Delta_l^{r/2}. \quad (50)$$

Bound for $\mathbb{E}_\theta[|R_{K_l}|^r]$. We have the upper-bound by Minkowski's inequality

$$\mathbb{E}_\theta[|R_{K_l}|^r] \leq \Delta_l^r \bigg(\sum_{k=1}^{K_l} \mathbb{E}_\theta\Big[\big| \{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}})^* - \kappa_{\theta,i}(X_{s_{k-1}})^*\} a_\theta(\widetilde{X}_{s_{k-1}}) \big|^r \Big]^{1/r} \bigg)^r.$$

Then applying the C_r-inequality d times and using the assumption that $a_\theta^j \in \mathcal{B}_b(\mathbb{R}^d)$ we have the upper-bound

$$\mathbb{E}_\theta[|R_{K_l}|^r] \leq C\Delta_l^r \bigg(\sum_{k=1}^{K_l} \sum_{j=1}^{d} \mathbb{E}_\theta\Big[\big| \{\kappa_{\theta,i}(\widetilde{X}_{s_{k-1}}) - \kappa_{\theta,i}(X_{s_{k-1}})\}^j \big|^r \Big]^{1/r} \bigg)^r.$$

Using the same argument to obtain (48), we have

$$\mathbb{E}_\theta[|R_{K_l}|^r] \leq C\Delta_l^r \Delta_l^{-r/2} = C\Delta_l^{r/2}. \tag{51}$$

Combining (50) and (51) with (46) allows us to conclude. ■

Lemma 7 *Under Assumptions 1 and 2, for any* $(T, r, \theta, i) \in \mathbb{N} \times [1, \infty) \times \Theta \times \{1, \ldots, d_\theta\}$, *there exists a constant* $C < \infty$ *such that for any* $(l, x) \in \mathbb{N}_0 \times \mathbb{R}^d$

$$\mathbb{E}_\theta\bigg[\Big| \sum_{k=1}^{K_l} \kappa_{\theta,i}(X_{s_{k-1}})^* \big[(\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}) - (X_{s_k} - X_{s_{k-1}}) \big] \Big|^r \bigg] \leq C\Delta_l^{r/2},$$

with $\widetilde{X}_0 = X_0 = x$.

Proof We have the decomposition

$$\sum_{k=1}^{K_l} \kappa_{\theta,i}(X_{s_{k-1}})^* \big[(\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}) - (X_{s_k} - X_{s_{k-1}}) \big] = M_{K_l} + R_{K_l},$$

where

$$M_{K_l} = \sum_{k=1}^{K_l} \sum_{(m,j) \in \{1,\ldots,d\}^2} \kappa_{\theta,i}(X_{s_{k-1}})^m \int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j,$$

$$R_{K_l} = \sum_{k=1}^{K_l} \sum_{j=1}^{d} \kappa_{\theta,i}(X_{s_{k-1}})^j \int_{s_{k-1}}^{s_k} \big[a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_s) \big]^j ds.$$

Thus by the C_r-inequality,

$$\mathbb{E}_\theta\bigg[\Big| \sum_{k=1}^{K_l} \kappa_{\theta,i}(X_{s_{k-1}})^* \big[(\widetilde{X}_{s_k} - \widetilde{X}_{s_{k-1}}) - (X_{s_k} - X_{s_{k-1}}) \big] \Big|^r \bigg] \leq C\big(\mathbb{E}_\theta[|M_{K_l}|^r] + \mathbb{E}_\theta[|R_{K_l}|^r] \big). \tag{52}$$

We will bound the two terms on the R.H.S. of (52) individually.

Bound for $\mathbb{E}_\theta[|M_{K_l}|^r]$. By applying the C_r-inequality d^2 times we have the upper-bound

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C \sum_{(m,j)\in\{1,\ldots,d\}^2} \mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l} \kappa_{\theta,i}(X_{s_{k-1}})^m \int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big|^r\Big].$$

For any $(m,j)\in\{1,\ldots,d\}^2$, we define

$$M_u^{m,j} = \sum_{k=0}^{u-1} \kappa_{\theta,i}(X_{s_{k-1}})^m \int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j$$

for $u\in\mathbb{N}_0$. As $(M_u^{m,j}, \mathscr{F}_{u\Delta_l})_{u\in\mathbb{N}_0}$ is a martingale, applying the BGD inequality yields

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C \sum_{(m,j)\in\{1,\ldots,d\}^2} \mathbb{E}_\theta\Big[\Big|\sum_{k=1}^{K_l} \Big(\kappa_{\theta,i}(X_{s_{k-1}})^m \int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big)^2\Big|^{r/2}\Big].$$

Applying Minkowski's inequality and using the fact that $[\kappa_{\theta,i}]^m \in \mathcal{B}_b(\mathbb{R}^d)$, we obtain

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C \sum_{(m,j)\in\{1,\ldots,d\}^2} \left(\sum_{k=1}^{K_l} \mathbb{E}_\theta\Big[\Big|\int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big|^r\Big]^{2/r}\right)^{r/2}. \quad (53)$$

Now we deal with the expectation on the R.H.S. of (53). Using the martingale property of the stochastic integral, it follows from applying the BGD inequality again that

$$\begin{aligned}\mathbb{E}_\theta\Big[\Big|\int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big|^r\Big] &\leq C\mathbb{E}_\theta\Big[\Big|\int_{s_{k-1}}^{s_k} \big\{[\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j}\big\}^2 ds\Big|^{r/2}\Big] \\ &\leq C\Delta_l^{r/2-1}\mathbb{E}_\theta\Big[\int_{s_{k-1}}^{s_k} \big|[\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j}\big|^r ds\Big],\end{aligned}$$

where we have used Jensen's inequality in the second line. Using the C_r-inequality, we have

$$\mathbb{E}_\theta\Big[\Big|\int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big|^r\Big] \leq$$
$$C\Delta_l^{r/2-1}\int_{s_{k-1}}^{s_k} \Big\{\mathbb{E}_\theta\Big[\big|[\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_{s_{k-1}})]^{m,j}\big|^r\Big] + \mathbb{E}_\theta\Big[\big|[\sigma(X_{s_{k-1}}) - \sigma(X_s)]^{m,j}\big|^r\Big]\Big\} ds.$$

Using the fact that $\sigma^{m,j}\in \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$, we then obtain

$$\mathbb{E}_\theta\Big[\Big|\int_{s_{k-1}}^{s_k} [\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)]^{m,j} dW_s^j\Big|^r\Big] \leq$$
$$C\Delta_l^{r/2-1}\int_{s_{k-1}}^{s_k} \Big\{\mathbb{E}_\theta\big[\|\widetilde{X}_{s_{k-1}} - X_{s_{k-1}}\|_2^r\big] + \mathbb{E}_\theta\big[\|X_{s_{k-1}} - X_s\|_2^r\big]\Big\} ds.$$

By the property (39) for $r > 0$ (Ikeda and Watanabe, 2014), it follows that

$$\sup_{(t,s)\in[0,T]^2} \mathbb{E}_\theta\left[\|X_t - X_s\|_2^r\right] \leq C|t-s|^{r/2}, \tag{54}$$

hence we obtain the upper-bound

$$\mathbb{E}_\theta\left[\Big|\int_{s_{k-1}}^{s_k} \left[\sigma(\widetilde{X}_{s_{k-1}}) - \sigma(X_s)\right]^{m,j} dW_s^j\Big|^r\right] \leq C\Delta_l^{r/2-1}\Delta_l^{1+r/2} = C\Delta_l^r. \tag{55}$$

Combining (55) with (53) gives

$$\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\Delta_l^{r/2}. \tag{56}$$

Bound for $\mathbb{E}_\theta[|R_{K_l}|^r]$. Using Minkowski's inequality followed by Jensen's inequality, we obtain the following bounds

$$\begin{aligned}\mathbb{E}_\theta[|R_{K_l}|^r] &\leq \left(\sum_{k=1}^{K_l}\sum_{j=1}^{d}\mathbb{E}_\theta\left[\Big|\kappa_{\theta,i}(X_{s_{k-1}})^j\int_{s_{k-1}}^{s_k}[a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_s)]^j ds\Big|^r\right]^{1/r}\right)^r \\ &\leq \left(\sum_{k=1}^{K_l}\sum_{j=1}^{d}\Delta_l^{1-1/r}\mathbb{E}_\theta\left[\big|\kappa_{\theta,i}(X_{s_{k-1}})^j\big|^r\int_{s_{k-1}}^{s_k}\big|[a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_s)]^j\big|^r ds\right]^{1/r}\right)^r.\end{aligned}$$

Using the assumption $[\kappa_{\theta,i}]^j \in \mathcal{B}_b(\mathbb{R}^d)$, the decomposition

$$[a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_s)]^j = [a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_{s_{k-1}})]^j + [a_\theta(X_{s_{k-1}}) - a_\theta(X_s)]^j,$$

and the C_r-inequality, we have

$$\begin{aligned}\mathbb{E}_\theta[|R_{K_l}|^r] \leq \Bigg(\sum_{k=1}^{K_l}\sum_{j=1}^{d}C\Delta_l^{1-1/r}\bigg(&\int_{s_{k-1}}^{s_k}\mathbb{E}_\theta\left[\big|[a_\theta(\widetilde{X}_{s_{k-1}}) - a_\theta(X_{s_{k-1}})]^j\big|^r\right]ds + \\ &\int_{s_{k-1}}^{s_k}\mathbb{E}_\theta\left[\big|[a_\theta(X_{s_{k-1}}) - a_\theta(X_s)]^j\big|^r\right]ds\bigg)^{1/r}\Bigg)^r.\end{aligned}$$

Since we have assumed that $a_\theta^j \in \mathrm{Lip}_{\|\cdot\|_2}(\mathbb{R}^d)$, one has

$$\begin{aligned}\mathbb{E}_\theta[|R_{K_l}|^r] \leq \Bigg(\sum_{k=1}^{K_l}\sum_{j=1}^{d}C\Delta_l^{1-1/r}\bigg(&\int_{s_{k-1}}^{s_k}\mathbb{E}_\theta\left[\|\widetilde{X}_{s_{k-1}} - X_{s_{k-1}}\|_2^r\right]ds + \\ &\int_{s_{k-1}}^{s_k}\mathbb{E}_\theta\left[\|X_{s_{k-1}} - X_s\|_2^r\right]ds\bigg)^{1/r}\Bigg)^r.\end{aligned}$$

Using (39) and (54), we therefore obtain

$$\mathbb{E}_\theta[|R_{K_l}|^r] \leq C\Delta_l^{r/2}. \tag{57}$$

Combining (56) and (57) with (52) allows us to conclude. ■

Lemma 8 *Under Assumptions 1 and 2, for any* $(T, r, \theta, i) \in \mathbb{N} \times [1, \infty) \times \Theta \times \{1, \ldots, d_\theta\}$, *there exists a constant* $C < \infty$ *such that for any* $(l, x) \in \mathbb{N}_0 \times \mathbb{R}^d$

$$\mathbb{E}_\theta\Big[\big|[G_\theta^l(X_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big] \leq C\Delta_l^{r/2},$$

with $\widetilde{X}_0 = X_0 = x$.

Proof We have the decomposition

$$\big|[G_\theta^l(X_{0:T})]^i - [G_\theta(X)]^i\big| = R_{K_l}^{(1)} + M_{K_l} + R_{K_l}^{(2)},$$

where

$$\begin{aligned}
R_{K_l}^{(1)} &= \sum_{k=1}^{K_l} \int_{s_{k-1}}^{s_k} \Big([\nabla_\theta \|b_\theta(X_{s_{k-1}})\|_2^2]^i - [\nabla_\theta \|b_\theta(X_s)\|_2^2]^i\Big) ds, \\
M_{K_l} &= \sum_{k=1}^{K_l} \sum_{(m,j)\in\{1,\ldots,d\}^2} \int_{s_{k-1}}^{s_k} [\kappa_{\theta,i}(X_{s_{k-1}})^m \sigma(\widetilde{X}_{s_{k-1}}) - \kappa_{\theta,i}(X_s)^m \sigma(X_s)]^{m,j} dW_s^j, \\
R_{K_l}^{(2)} &= \sum_{k=1}^{K_l} \sum_{j=1}^{d} \int_{s_{k-1}}^{s_k} \Big(\kappa_{\theta,i}(X_{s_{k-1}})^j a_\theta(X_{s_{k-1}})^j - \kappa_{\theta,i}(X_s)^j a_\theta(X_s)^j\Big) ds.
\end{aligned}$$

Thus, by the C_r-inequality, we have

$$\mathbb{E}_\theta\Big[\big|[G_\theta^l(X_{0:T})]^i - [G_\theta(X)]^i\big|^r\Big] \leq C\Big(\mathbb{E}_\theta\big[|M_{K_l}|^r\big] + \mathbb{E}_\theta\big[\big|R_{K_l}^{(1)}\big|^r\big] + \mathbb{E}_\theta\big[\big|R_{K_l}^{(2)}\big|^r\big]\Big).$$

In order to prove that $\mathbb{E}_\theta[|M_{K_l}|^r] \leq C\Delta_l^{r/2}$, one can rely on very similar arguments to (56) in the proof of Lemma 7. Likewise, for $m \in \{1, 2\}$ one can prove that $\mathbb{E}_\theta[|R_{K_l}^{(m)}|^r] \leq C\Delta_l^{r/2}$ using similar arguments to (57) in the proof of Lemma 7. ■

Lemma 9 *Under Assumptions 1 and 2, for any* $(T, r, \theta) \in \mathbb{N} \times [1, \infty) \times \Theta$, *there exists a constant* $C < \infty$ *such that for any* $(x, x_\star) \in \mathbb{R}^d \times \mathbb{R}^d$

$$\mathbb{E}_\theta\big[\|G_\theta(X) - G_\theta(X^\star)\|_2^r\big]^{1/r} \leq C\|x - x_\star\|_2,$$

with $X_0 = x$ *and* $X_0^\star = x_\star$.

Proof This proof follows a similar type of arguments to those considered in the proofs of Lemmata 7-8. The main difference is that one must use the following result which can be deduced from (Rogers and Williams, 2000, Corollary v.11.7) and Grönwall's inequality)

$$\sup_{t\in[0,T]} \mathbb{E}_\theta[\|X_t - X_t^\star\|_2^{2r}]^{\frac{1}{2r}} \leq C\|x - x_\star\|_2$$

instead of using Euler convergence. Given that the proofs of Lemmata 7-8 are repetitive, these arguments are omitted. ■

Lemma 10 *Under Assumptions 1 and 2, for any $(T, r, \theta) \in \mathbb{N} \times [1, \infty) \times \Theta$, there exists a constant $C < \infty$ such that for any $(l, x, x_\star) \in \mathbb{N} \times \mathbb{R}^d \times \mathbb{R}^d$*

$$\mathbb{E}_\theta\big[\|G_\theta^l(\widetilde{X}_{0:T}) - G_\theta^{l-1}(\widehat{X}_{0:T}^\star)\|_2^r\big]^{1/r} \leq C\big(\Delta_l^{1/2} + \|x - x_\star\|_2\big),$$

with $\widetilde{X}_0 = x$ and $\widetilde{X}_0^\star = x_\star$, where $\widehat{X}_{0:T}^\star = (\widetilde{X}_{s_{2k}}^\star)_{k=0}^{K_{l-1}}$ are the sequence of states in $\widetilde{X}_{0:T}^\star = (\widetilde{X}_{s_k}^\star)_{k=0}^{K_l}$ at the time steps corresponding to level $l-1$.

Proof The expectation in the statement of the lemma is upper-bounded by $\sum_{j=1}^3 T_j$, where

$$\begin{aligned} T_1 &= \mathbb{E}_\theta\big[\|G_\theta^l(\widetilde{X}_{0:T}) - G_\theta(X)\|_2^r\big]^{1/r}, \\ T_2 &= \mathbb{E}_\theta\big[\|G_\theta(X) - G_\theta(X^\star)\|_2^r\big]^{1/r}, \\ T_3 &= \mathbb{E}_\theta\big[\|G_\theta^{l-1}(\widehat{X}_{0:T}^\star) - G_\theta(X^\star)\|_2^r\big]^{1/r}. \end{aligned}$$

For T_1 and T_3, one can apply Corollary 5; for T_2, we use Lemma 9. This allows us to conclude the result. ■

We now introduce two additional functions which will be useful in the following section. For $(t, l) \in \{1, \ldots, T\} \times \mathbb{N}_0$, we define $G_{\cdot,t}^l : \Theta \times (\mathbb{R}^d)^{2^l t+1} \to \mathbb{R}^{d_\theta}$ as

$$\begin{aligned} G_{\theta,t}^l(X_{0:t}) = &-\frac{1}{2}\sum_{k=1}^{2^l t} \nabla_\theta \|b_\theta(X_{s_{k-1}})\|_2^2 \Delta_l + \sum_{k=1}^{2^l t} \nabla_\theta \rho_\theta(X_{s_{k-1}})(X_{s_k} - X_{s_{k-1}}) + \\ &\sum_{p=1}^{t} \nabla_\theta \log g_\theta(y_p|X_p), \end{aligned}$$

and $G_{\cdot,t-1:t}^l : \Theta \times \mathbb{R}^d \times (\mathbb{R}^d)^{2^l} \to \mathbb{R}^{d_\theta}$ as

$$\begin{aligned} G_{\theta,t-1:t}^l(X_{t-1}, X_{t-1+\Delta_l:t}) = &-\frac{1}{2}\sum_{k=0}^{2^l-1} \nabla_\theta \|b_\theta(X_{t-1+k\Delta_l})\|_2^2 \Delta_l + \\ &\sum_{k=0}^{2^l-1} \nabla_\theta \rho_\theta(X_{t-1+k\Delta_l})(X_{t-1+(k+1)\Delta_l} - X_{t-1+k\Delta_l}) + \nabla_\theta \log g_\theta(y_t|X_t). \end{aligned}$$

From (37), we note that $G_{\theta,T}^l(X_{0:T}) = G_\theta^l(X_{0:T})$. The following remarks will facilitate our proofs in the next section.

Remark 11 *One can easily extend Lemma 10 as follows. Under Assumptions 1 and 2, for any $(t, r, \theta) \in \{1, \ldots, T\} \times [1, \infty) \times \Theta$, there exists a constant $C < \infty$ such that for any $(l, x, x_\star) \in \mathbb{N} \times \mathbb{R}^d \times \mathbb{R}^d$*

$$\mathbb{E}_\theta\Big[\big\|G_{\theta,t}^l(\widetilde{X}_{0:t}) - G_{\theta,t}^{l-1}(\widehat{X}_{0:t}^\star)\big\|_2^r\Big]^{1/r} \leq C\big(\Delta_l^{1/2} + \|x - x_\star\|_2\big),$$

with $\widetilde{X}_0 = x$ and $\widetilde{X}_0^\star = x_\star$, where $\widehat{X}_{0:t}^\star = (\widetilde{X}_{s_{2k}}^\star)_{k=0}^{2^{l-1}t}$.

Remark 12 *One can also deduce the following result. Under Assumptions 1 and 2, for any* $(t,r,\theta)\in\{1,\dots,T\}\times[1,\infty)\times\Theta$*, there exists a constant* $C<\infty$ *such that for any* $(l,x,x_\star)\in\mathbb{N}\times\mathbb{R}^d\times\mathbb{R}^d$

$$\mathbb{E}_\theta\Big[\big\|G^l_{\theta,t-1:t}(x,\widetilde{X}_{t-1+\Delta_l:t})-G^{l-1}_{\theta,t-1:t}(x_\star,\widehat{X}^\star_{t-1+\Delta_{l-1}:t})\big\|_2^r\Big|\widetilde{X}_{t-1}=x,\widetilde{X}^\star_{t-1}=x_\star\Big]^{1/r}$$
$$\leq C\big(\Delta_l^{1/2}+\|x-x_\star\|_2\big),$$

where $\widehat{X}^\star_{t-1+\Delta_{l-1}:t}=(\widetilde{X}^\star_{t-1+\Delta_{l-1}})_{k=1}^{2^{l-1}}$.

B.3 Results on Coupled Conditional Particle Filters

We begin with some definitions associated to Algorithm 4. For any $(i,t,s)\in\{1,\dots,N\}\times\mathbb{N}_0\times\{l-1,l\}$, we will write $A_t^s(i)=A_t^{s,i}$ and $\bar{A}_t^s(i)=\bar{A}_t^{s,i}$. Using this notation, for any $(t,l)\in\{0,\dots,T-1\}\times\mathbb{N}$, we define

$$\begin{aligned}\mathsf{S}_t^l=\{i\in\{1,\dots,N\}:A_t^l(i)=A_t^{l-1}(i),\\ A^l_{t-1}\circ A^l_t(i)=A^{l-1}_{t-1}\circ A^{l-1}_t(i),\dots,A^l_0\circ\cdots\circ A^l_t(i)=A^{l-1}_0\circ\cdots\circ A^{l-1}_t(i)\},\end{aligned}$$

and

$$\begin{aligned}\bar{\mathsf{S}}_t^l=\{i\in\{1,\dots,N\}:\bar{A}_t^l(i)=\bar{A}_t^{l-1}(i),\\ \bar{A}^l_{t-1}\circ \bar{A}^l_t(i)=\bar{A}^{l-1}_{t-1}\circ \bar{A}^{l-1}_t(i),\dots,\bar{A}^l_0\circ\cdots\circ \bar{A}^l_t(i)=\bar{A}^{l-1}_0\circ\cdots\circ \bar{A}^{l-1}_t(i)\}.\end{aligned}$$

For $(l,\beta,C)\in\mathbb{N}\times\mathbb{R}^+\times\mathbb{R}^+$, we also introduce the following sets

$$\begin{aligned}\mathsf{B}^l_{\beta,C}&=\{(x_{0:T},\bar{x}_{0:T})\in\mathsf{X}^l\times\mathsf{X}^{l-1}:\|x_t-\bar{x}_t\|_2\leq C\Delta_l^\beta,t\in\{1,\dots,T\}\},\\ \mathsf{G}^l_{\beta,C}&=\{(x_{0:T},\bar{x}_{0:T})\in\mathsf{X}^l\times\mathsf{X}^{l-1}:\|G^l_{\theta,t}(x_{0:t})-G^{l-1}_{\theta,t}(\bar{x}_{0:t})\|_2\leq C\Delta_l^\beta,t\in\{1,\dots,T\}\}.\end{aligned}$$

B.3.1 Results associated to Steps (1) and (2) of Algorithm 4

We consider Steps (1) and (2) of Algorithm 4 where the input pairs of trajectories are taken as $(X^{l-1,\star}_{0:T},\bar{X}^{l-1,\star}_{0:T})=(x^{l-1}_{0:T},\bar{x}^{l-1}_{0:T})\in\mathsf{Z}^{l-1}$ and $(X^{l,\star}_{0:T},\bar{X}^{l,\star}_{0:T})=(x^{l}_{0:T},\bar{x}^{l}_{0:T})\in\mathsf{Z}^{l}$. In order to analyze the algorithm, it is useful to define the simulated trajectories recursively at time steps $t\in\{1,\dots,T\}$, for any $i\in\{1,\dots,N\}$, as

$$\begin{aligned}(X^{l-1,i}_{0:t},\bar{X}^{l-1,i}_{0:t})&=\Big((X^{l-1,A^{l-1,i}_{t-1}}_{0:t-1},X^{l-1,i}_{t-1+\Delta_{l-1}:t}),(\bar{X}^{l-1,\bar{A}^{l-1,i}_{t-1}}_{0:t-1},\bar{X}^{l-1,i}_{t-1+\Delta_{l-1}:t})\Big),\\ (X^{l,i}_{0:t},\bar{X}^{l,i}_{0:t})&=\Big((X^{l,A^{l,i}_{t-1}}_{0:t-1},X^{l,i}_{t-1+\Delta_{l}:t}),(\bar{X}^{l,\bar{A}^{l,i}_{t-1}}_{0:t-1},\bar{X}^{l,i}_{t-1+\Delta_{l}:t})\Big).\end{aligned}$$

After the completion of Step (2), we consider the output given by the two collections of pairs of trajectories $(X^{l-1,i}_{0:T},\bar{X}^{l-1,i}_{0:T})_{i=1}^N$ and $(X^{l,i}_{0:T},\bar{X}^{l,i}_{0:T})_{i=1}^N$. Under these conditions, the expectation associated to the law of Steps (1) and (2) of Algorithm 4 is denoted as $\check{\mathbb{E}}^{l-1,l}_\theta$.

Lemma 13 *Under Assumptions 1 and 2, for any* $(t, r, \theta, C') \in \{1, \ldots, T\} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, *there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \ldots\}$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$, *it holds that*

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{1}{N}\sum_{i\in \mathsf{S}_{t-1}^l} \|X_t^{l,i} - X_t^{l-1,i}\|_2^r\right]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$ *also holds, then*

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{1}{N}\sum_{i\in \bar{\mathsf{S}}_{t-1}^l} \|\bar{X}_t^{l,i} - \bar{X}_t^{l-1,i}\|_2^r\right]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. The proof is almost identical to Jasra et al. (2017, Lemma D.3.). The only difference is if $A_p^{l,i} = N$ for some $(p, i) \in \{1, \ldots, t\} \times \{1, \ldots, N\}$ with $i \in \mathsf{S}_p^l$, but in such a case, one can use that $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$. ■

Remark 14 *Implicit in the proof of Lemma 13 is the following result. Under Assumptions 1 and 2, for any* $(t, r, \theta, C') \in \{1, \ldots, T\} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, *there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \ldots\}$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$, *it holds that*

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{1}{N}\sum_{i\in \mathsf{S}_{t-1}^l} \left\|X_{t-1}^{l,A_{t-1}^{l,i}} - X_{t-1}^{l-1,A_{t-1}^{l-1,i}}\right\|_2^r\right]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$ *also holds, then*

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{1}{N}\sum_{i\in \bar{\mathsf{S}}_{t-1}^l} \left\|\bar{X}_{t-1}^{l,\bar{A}_{t-1}^{l,i}} - \bar{X}_{t-1}^{l-1,\bar{A}_{t-1}^{l-1,i}}\right\|_2^r\right]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

Lemma 15 *Under Assumptions 1 and 2, for any* $(t, \theta, C') \in \{1, \ldots, T\} \times \Theta \times \mathbb{R}^+$, *there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \ldots\}$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$, *it holds that*

$$1 - \check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{\mathrm{Card}(\mathsf{S}_{t-1}^l)}{N}\right] \leq C\Delta_l^{\frac{1}{2}\wedge\beta},$$

where Card$(\cdot)$ *denotes the cardinality of a set. If* $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$ *also holds, then*

$$1 - \check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{\mathrm{Card}(\bar{\mathsf{S}}_{t-1}^l)}{N}\right] \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. The proof is almost identical to Jasra et al. (2017, Lemma D.4.). The only difference is if $A_p^{l,i} = N$ for some $(p, i) \in \{1, \ldots, t\} \times \{1, \ldots, N\}$ with $i \in \mathsf{S}_p^l$, but in such a case, one can use that $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l$ and Lemma 13. ■

Lemma 16 *Under Assumptions 1 and 2, for any $(t, r, \theta, C') \in \{1, \ldots, T\} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, there exists a constant $C < \infty$ such that for any $(l, \beta, N) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \ldots\}$ and any $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$, it holds that*

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$ also holds, then

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\bar{\mathsf{S}}_{t-1}^l}\big\|G_{\theta,t}^l(\bar{X}_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(\bar{X}_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} \leq C\Delta_l^{\frac{1}{2}\wedge\beta}.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. The proof is by induction on t. The initialization holds for $i \in \{1, \ldots, N-1\}$ by the result stated in Remark 11. The case $i = N$ is trivial as $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$.

We now consider

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} \leq C(T_1 + T_2), \tag{58}$$

where

$$T_1 = \check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\big\|G_{\theta,t-1:t}^l(X_{t-1}^{l,i}, X_{t-1+\Delta_l:t}^{l,i}) - G_{\theta,t-1:t}^{l-1}(X_{t-1}^{l-1,i}, X_{t-1+\Delta_{l-1}:t}^{l-1,i})\big\|_2^r\Big]^{1/r},$$

$$T_2 = \check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\Big\|G_{\theta,t-1}^l(X_{0:t-1}^{l,A_{t-1}^{l,i}}) - G_{\theta,t-1}^{l-1}(X_{0:t-1}^{l-1,A_{t-1}^{l-1,i}})\Big\|_2^r\Big]^{1/r}.$$

So we consider bounds on T_1 and T_2.

For T_1, by applying the result in Remark 12, we have

$$T_1 \leq C\Big(\Delta_l^{1/2} + \check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\Big\|X_{t-1}^{l,A_{t-1}^{l,i}} - X_{t-1}^{l-1,A_{t-1}^{l-1,i}}\Big\|_2^r\Big]^{1/r}\Big).$$

Then applying the result in Remark 14 gives

$$T_1 \leq C\Delta_l^{\frac{1}{2}\wedge\beta}. \tag{59}$$

For T_2, one can use the same approach as in the proof of Jasra et al. (2017, Lemma D.3.) from the bottom of page 3092, along with the induction hypothesis, to obtain

$$T_2 \leq C\Delta_l^{\frac{1}{2}\wedge\beta}. \tag{60}$$

Combining (58) with (59) and (60) concludes the proof. ■

Lemma 17 *Under Assumptions 1 and 2, for any $(t,r,\theta) \in \{1,\dots,T\}\times[1,\infty)\times\Theta$, there exists a constant $C<\infty$ such that for any $(l,s,N,i) \in \mathbb{N}\times\{l-1,1\}\times\{2,3,\dots\}\times\{1,\dots,N\}$ and any $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{X}^l \times \mathsf{X}^{l-1}$, it holds that*

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\left\|G_{\theta,t}^s(X_{0:t}^{s,i})\right\|_2^r\right] \leq C\Big(1+\sum_{p=1}^t \left\|G_{\theta,p-1:p}^s(x_{p-1}^{s,N}, x_{p-1+\Delta_s:p}^{s,N})\right\|_2^r\Big).$$

Also for any $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{X}^l \times \mathsf{X}^{l-1}$

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\left\|G_{\theta,t}^s(\bar{X}_{0:t}^{s,i})\right\|_2^r\right] \leq C\Big(1+\sum_{p=1}^t \left\|G_{\theta,p-1:p}^s(\bar{x}_{p-1}^{s,N}, \bar{x}_{p-1+\Delta_s:p}^{s,N})\right\|_2^r\Big).$$

Proof We only consider the first inequality as it is the same proof for the second inequality. We consider a proof by induction. In the case of $t=1$, one can use the boundedness properties of the appropriate terms along with the martingale-remainder methods in the proof of Lemma 6 to deduce the given bound, except for the case $i=N$, which is exhibited in the bound.

Now consider the case of $t>1$, we have the upper-bound

$$\check{\mathbb{E}}_\theta^{l-1,l}\left[\left\|G_{\theta,t}^s(X_{0:t}^{s,i})\right\|_2^r\right] \leq C(T_1+T_2), \tag{61}$$

where

$$T_1 = \check{\mathbb{E}}_\theta^{l-1,l}\left[\left\|G_{\theta,t-1:t}^s(X_{t-1}^{s,i}, X_{t-1+\Delta_s:t}^{s,i})\right\|_2^r\right],$$
$$T_2 = \check{\mathbb{E}}_\theta^{l-1,l}\left[\left\|G_{\theta,t-1}^s(X_{0:t-1}^{s,A_{t-1}^{s,i}})\right\|_2^r\right].$$

So we consider bounds on T_1 and T_2.

For T_1, by the same argument as for the initialization

$$T_1 \leq C\Big(1+\left\|G_{\theta,t-1:t}^s(x_{t-1}^{s,N}, x_{t-1+\Delta_s:t}^{s,N})\right\|_2^r\Big). \tag{62}$$

For T_2, we first note that for the resampling probabilities associated to $A_{t-1}^{s,i}$, we can deduce the following upper-bound using Assumption 2

$$\frac{g_\theta(y_{t-1}|x_{t-1}^i)}{\sum_{j=1}^N g_\theta(y_{t-1}|x_{t-1}^j)} \leq \frac{C}{N}. \tag{63}$$

Averaging over the resampling indexes, using (63) and the induction hypothesis one has

$$T_2 \leq C\Big(1 + \sum_{p=1}^{t-1} \|G^s_{\theta,p-1:p}(x^{s,N}_{p-1}, x^{s,N}_{p-1+\Delta_s:p})\|_2^r\Big). \tag{64}$$

Combining (61) with (62) and (64) concludes the proof. ■

Corollary 18 *Under Assumptions 1 and 2, for any* $(t, r, \theta) \in \{1, \dots, T\} \times [1, \infty) \times \Theta$*, there exists a constant* $C < \infty$ *such that for any* $(l, s, N, i) \in \mathbb{N} \times \{l-1, 1\} \times \{2, 3, \dots\} \times \{1, \dots, N\}$ *and any* $(x^l_{0:T}, x^{l-1}_{0:T}) \in \mathsf{X}^l \times \mathsf{X}^{l-1}$*, it holds that*

$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\|X^{s,i}_t\|_2^r\Big] \leq C\Big(1 + \sum_{p=1}^{t} \|x^{s,N}_p\|_2^r\Big).$$

Also for any $(\bar{x}^l_{0:T}, \bar{x}^{l-1}_{0:T}) \in \mathsf{X}^l \times \mathsf{X}^{l-1}$

$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\|\bar{X}^{s,i}_t\|_2^r\Big] \leq C\Big(1 + \sum_{p=1}^{t} \|\bar{x}^{s,N}_p\|_2^r\Big).$$

Proof We only consider the first inequality as it is the same proof for the second inequality. We consider a proof by induction. In the case of $t = 1$, one can use the boundedness properties of the appropriate terms along with the martingale-remainder methods in the proof of Lemma 6 to deduce the given bound, except for the case $i = N$, which is exhibited in the bound.

For $t > 1$, when $i \in \{1, \dots, N-1\}$, repeating the argument of the initialization, one has

$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\|X^{s,i}_t\|_2^r\Big] \leq C\Big(1 + \check{\mathbb{E}}^{l-1,l}_\theta\Big[\Big\|X^{s,A^{s,i}_{t-1}}_{t-1}\Big\|_2^r\Big]\Big).$$

Then one can repeat the argument that leads to (64). The case $i = N$ is trivially true. ■

Remark 19 *The results in Lemmata 16 and 17 can be extended to the case where one considers*

$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\frac{1}{N}\sum_{i\in\mathsf{S}^l_{t-1}} \Big\|G^l_{\theta,t}(X^{l,A^{l,i}_t}_{0:t}) - G^{l-1}_{\theta,t}(X^{l-1,A^{l-1,i}_t}_{0:t})\Big\|_2^r\Big]$$
$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\frac{1}{N}\sum_{i\in\bar{\mathsf{S}}^l_{t-1}} \|G^l_{\theta,t}(\bar{X}^{l,\bar{A}^{l,i}_t}_{0:t}) - G^{l-1}_{\theta,t}(\bar{X}^{l-1,\bar{A}^{l-1,i}_t}_{0:t})\|_2^r\Big]^{1/r},$$

and

$$\check{\mathbb{E}}^{l-1,l}_\theta\Big[\Big\|G^s_{\theta,t}(X^{s,A^{s,i}_t}_{0:t})\Big\|_2^r\Big], \quad \check{\mathbb{E}}^{l-1,l}_\theta\Big[\Big\|G^s_{\theta,t}(\bar{X}^{s,\bar{A}^{s,i}_t}_{0:t})\Big\|_2^r\Big],$$

by using very similar arguments to the proof of those lemmata.

Lemma 20 *Under Assumptions 1 and 2, for any* $(t, r, \theta, C') \in \{1, \dots, T\} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, *there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N, i, \delta) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \dots\} \times \{1, \dots, N\} \times \mathbb{R}^+$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$, *it holds that*

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^s(x_{p-1}^{s,N}, x_{p-1+\Delta_s:p}^{s,N})\big\|_2^r\Big).$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$ *also holds, then*

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(\bar{X}_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(\bar{X}_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^s(\bar{x}_{p-1}^{s,N}, \bar{x}_{p-1+\Delta_s:p}^{s,N})\big\|_2^r\Big).$$

Proof We only consider the first inequality as it is the same proof for the second inequality. We have

$$\check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r} = T_1 + T_2, \tag{65}$$

where

$$T_1 = \check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r \mathbb{I}_{\mathsf{S}_{t-1}^l}(i)\Big]^{1/r},$$
$$T_2 = \check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r \mathbb{I}_{(\mathsf{S}_{t-1}^l)^c}(i)\Big]^{1/r}.$$

So we consider bounds on T_1 and T_2.

For T_1, we have the upper-bound

$$T_1 \leq C\check{\mathbb{E}}_\theta^{l-1,l}\Big[\frac{1}{N}\sum_{i\in\mathsf{S}_{t-1}^l}\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^r\Big]^{1/r}$$

then applying Lemma 16 gives

$$T_1 \leq C\Delta_l^{\frac{1}{2}\wedge\beta}. \tag{66}$$

For T_2, applying Hölder's inequality gives

$$T_2 \leq \check{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l(X_{0:t}^{l,i}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^{\frac{r(1+\delta)}{\delta}}\Big]^{\frac{\delta}{(1+\delta)}} \check{\mathbb{E}}_\theta^{l-1,l}\big[\mathbb{I}_{(\mathsf{S}_{t-1}^l)^c}(i)\big]^{\frac{1}{r(1+\delta)}}. \tag{67}$$

Note that

$$\check{\mathbb{E}}_\theta^{l-1,l}\big[\mathbb{I}_{(\mathsf{S}_{t-1}^l)^c}(i)\big] = 1 - \check{\mathbb{E}}_\theta^{l-1,l}\left[\frac{\mathrm{Card}(\mathsf{S}_{t-1}^l)}{N}\right] \leq C\Delta_l^{\frac{1}{2}\wedge\beta}, \tag{68}$$

where we have used Lemma 15, and

$$\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l,i})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^{\frac{r(1+\delta)}{\delta}}\Big]^{\frac{\delta}{(1+\delta)}} \leq C\Big(\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l,i})\big\|_2^{\frac{r(1+\delta)}{\delta}}\Big]^{\frac{\delta}{(1+\delta)}}+\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l-1}(X_{0:t}^{l-1,i})\big\|_2^{\frac{r(1+\delta)}{\delta}}\Big]^{\frac{\delta}{(1+\delta)}}\Big).$$

Then applying Lemma 17 and combining with (67) and (68), one can deduce that

$$T_2 \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1+\sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^{s}(x_{p-1}^{s,N},x_{p-1+\Delta_s:p}^{s,N})\big\|_2^{r}\Big). \tag{69}$$

Combining (65), (66) and (69) completes the proof. ■

Corollary 21 *Under Assumptions 1 and 2, for any* $(t,r,\theta,C')\in\{1,\dots,T\}\times[1,\infty)\times\Theta\times\mathbb{R}^+$*, there exists a constant* $C<\infty$ *such that for any* $(l,\beta,N,i,\delta)\in\mathbb{N}\times\mathbb{R}^+\times\{2,3,\dots\}\times\{1,\dots,N\}\times\mathbb{R}^+$ *and any* $(x_{0:T}^{l},x_{0:T}^{l-1})\in\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}$*, it holds that*

$$\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|X_t^{l,i}-X_t^{l-1,i}\big\|_2^{r}\Big]^{1/r}\leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1+\sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|x_p^{s,N}\big\|_2^{r}\Big).$$

If $(\bar{x}_{0:T}^{l},\bar{x}_{0:T}^{l-1})\in\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}$ *also holds, then*

$$\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|\bar{X}_t^{l,i}-\bar{X}_t^{l-1,i}\big\|_2^{r}\Big]^{1/r}\leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1+\sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|\bar{x}_p^{s,N}\big\|_2^{r}\Big).$$

Remark 22 *Using Remark 19, one can extend Lemma 20 using a similar argument to its proof to obtain the same bound on*

$$\check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\Big\|G_{\theta,t}^{l}(X_{0:t}^{l,A_t^{l,i}})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1,A_t^{l-1,i}})\Big\|_2^{r}\Big]^{1/r}.$$

A similar statement applies to Corollary 21.

We introduce the following sets, which will be of use later on in our proofs. For $(t,l)\in\{0,\dots,T-1\}\times\mathbb{N}$, we define

$$\check{\mathsf{S}}_t^{l}=\{i\in\{1,\dots,N-1\}:\ A_t^{l}(i)=\bar{A}_t^{l}(i)\neq N, A_{t-1}^{l}\circ A_t^{l}(i)=\bar{A}_{t-1}^{l}\circ\bar{A}_t^{l}(i)\neq N,\dots, A_0^{l}\circ\cdots\circ A_t^{l}(i)=\bar{A}_0^{l}\circ\cdots\circ\bar{A}_t^{l}(i)\neq N\} \tag{70}$$

and

$$\check{\mathsf{S}}_t^{l-1}=\{i\in\{1,\dots,N-1\}:\ A_t^{l-1}(i)=\bar{A}_t^{l-1}(i)\neq N, A_{t-1}^{l-1}\circ A_t^{l-1}(i)=\bar{A}_{t-1}^{l}\circ\bar{A}_t^{l}(i)\neq N,\dots, A_0^{l-1}\circ\cdots\circ A_t^{l-1}(i)=\bar{A}_0^{l-1}\circ\cdots\circ\bar{A}_t^{l-1}(i)\neq N\}. \tag{71}$$

Lemma 23 *Under Assumptions 1 and 2, for any* $(t,\theta,N)\in\{0,\dots,T-1\}\times\Theta\times\{2,3,\dots\}$, *there exists a constant* $\varepsilon\in(0,1)$ *such that for any* $(l,i)\in\mathbb{N}\times\{1,\dots,N-1\}$ *and any* $((x_{0:T}^{l-1},\bar{x}_{0:T}^{l-1}),(x_{0:T}^{l},\bar{x}_{0:T}^{l}))\in\mathsf{Z}^{l-1}\times\mathsf{Z}^{l}$, *it holds that*

$$\check{\mathbb{E}}_{\theta}^{l-1,l}\big[\mathbb{I}_{\check{\mathsf{S}}_t^l}(i)\big]\wedge\check{\mathbb{E}}_{\theta}^{l-1,l}\big[\mathbb{I}_{\check{\mathsf{S}}_t^{l-1}}(i)\big]\geq\varepsilon.$$

Proof To proceed we first introduce some notation. For $(t,l)\in\{1,\dots,T\}\times\mathbb{N}$, we define

$$\begin{aligned}\check{\vartheta}_t^l(i,j) \quad = \quad & \left\{\frac{g_\theta(y_t|x_t^{l,i})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l,j_1})}\wedge\frac{g_\theta(y_t|x_t^{l-1,j})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l-1,j_1})}\right\}\mathbb{I}_{\{i\}}(j)+\\ & \left(\frac{g_\theta(y_t|x_t^{l,i})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l,j_1})}-\left\{\frac{g_\theta(y_t|x_t^{l,i})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l,j_1})}\wedge\frac{g_\theta(y_t|x_t^{l-1,i})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l-1,j_1})}\right\}\right)\times\\ & \left(\frac{g_\theta(y_t|x_t^{l-1,j})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l-1,j_1})}-\left\{\frac{g_\theta(y_t|x_t^{l,j})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l,j_1})}\wedge\frac{g_\theta(y_t|x_t^{l-1,j})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l-1,j_1})}\right\}\right)\times\\ & \left(1-\sum_{j_2=1}^N\left\{\frac{g_\theta(y_t|x_t^{l,j_2})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l,j_1})}\wedge\frac{g_\theta(y_t|x_t^{l-1,j_2})}{\sum_{j_1=1}^N g_\theta(y_t|x_t^{l-1,j_1})}\right\}\right)^{-1}\end{aligned}$$

which is the maximal coupling of the resampling distributions across levels. We write $\check{\bar{\vartheta}}_t^l(i,j)$ when one replaces $(x_t^{l,1:N},x_t^{l-1,1:N})$ with $(\bar{x}_t^{l,1:N},\bar{x}_t^{l-1,1:N})$. We will write the maximal coupling (in the above sense with independent residuals) of $\check{\vartheta}_t^l(j_1,j_2)$ and $\check{\bar{\vartheta}}_t^l(j_3,j_4)$ for $(j_1,\dots,j_4)\in\{1,\dots,N\}^4$, as $\overline{\vartheta}_t^l(j_1,\dots,j_4)$. We also define $\mathsf{D}^l=\{(x_{0:T},\bar{x}_{0:T})\in\mathsf{Z}^l : x_{0:T}=\bar{x}_{0:T}\}$ and

$$\begin{aligned}&\overline{\vartheta}_t^{(l)}(j_1,j_3)=\\ &\mathbb{I}_{(\mathsf{D}^l)^c}(x_{0:T}^l,\bar{x}_{0:T}^l)\sum_{(j_2,j_4)\in\{1,\dots,N\}^2}\overline{\vartheta}_t^{(l)}(j_1,\dots,j_4)+\mathbb{I}_{\mathsf{D}^l}(x_{0:T}^l,\bar{x}_{0:T}^l)\mathbb{I}_{\{j_1\}}(j_3)\frac{g_\theta(y_t|x_t^{l,j_1})}{\sum_{j=1}^N g_\theta(y_t|x_t^{l,j})}.\end{aligned}\tag{72}$$

$\overline{\vartheta}_t^{(l)}(j_1,j_3)$ is the distribution of the resampled indexes within a level under Algorithm 5. One can make a similar definition for $\overline{\vartheta}_t^{(l-1)}(j_2,j_4)$.

We give the proof in the case of l only as the proof for $l-1$ is similar. The proof is by induction on t and the initial case $t=0$ is trivial by definition. For $t\geq 1$, we have

$$\begin{aligned}\check{\mathbb{E}}_{\theta}^{l-1,l}\big[\mathbb{I}_{\check{\mathsf{S}}_t^l}(i)\big] &= \check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\sum_{j=1}^{N-1}\mathbb{I}_{\check{\mathsf{S}}_{t-1}^l}(i)\overline{\vartheta}_t^{(l)}(j,j)\Big]\\ &\geq \check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{\check{\mathsf{S}}_{t-1}^l}(1)\overline{\vartheta}_t^{(l)}(1,1)\Big]\\ &\geq \check{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{\check{\mathsf{S}}_{t-1}^l}(1)\Big]\frac{C}{N}\\ &\geq \varepsilon\end{aligned}$$

where we have used Assumption 2 to establish that

$$\overline{\vartheta}_t^{(l)}(1,1) \geq \frac{C}{N} \tag{73}$$

on the third line, and the induction hypothesis in the final line. This completes our proof. ■

B.3.2 Results associated to the Entirety of Algorithm 4

We now consider Algorithm 4 in its entirety. We will denote expectation and probability w.r.t. a single step of the corresponding 4-CCPF kernel $\bar{M}_\theta^{l-1,l}$ by $\bar{\mathbb{P}}_\theta^{l-1,l}$ and $\bar{\mathbb{E}}_\theta^{l-1,l}$, respectively.

Corollary 24 *Under Assumptions 1 and 2, for any $(T, r, \theta, C') \in \mathbb{N} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, there exists a constant $C < \infty$ such that for any $(l, \beta, N, \delta) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \dots\} \times \mathbb{R}^+$ and any $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$, it holds that*

$$\bar{\mathbb{E}}_\theta^{l-1,1}\Big[\big\|G_\theta^l\big(X_{0:T}^{l,B_T^l}\big) - G_\theta^{l-1}\big(X_{0:T}^{l-1,B_T^{l-1}}\big)\big\|_2^r\Big]^{1/r} \leq$$
$$C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^T \sum_{s=l-1}^l \big\|G_{\theta,p-1:p}^s(x_{p-1}^s, x_{p-1+\Delta_s:p}^s)\big\|_2^r\Big).$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$ also holds, then

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l\big(\bar{X}_{0:T}^{l,\bar{B}_T^l}\big) - G_\theta^{l-1}\big(\bar{X}_{0:T}^{l-1,\bar{B}_T^{l-1}}\big)\big\|_2^r\Big]^{1/r} \leq$$
$$C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^T \sum_{s=l-1}^l \big\|G_{\theta,p-1:p}^s(\bar{x}_{p-1}^s, \bar{x}_{p-1+\Delta_s:p}^s)\big\|_2^r\Big).$$

Proof This follows from the discussion in Remark 22. ■

Remark 25 *By following the discussion in Remark 22, one can also extend Corollary 24 to a bound of the type*

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l\big(X_{0:t}^{l,B_T^l}\big) - G_{\theta,t}^{l-1}\big(X_{0:t}^{l-1,B_T^{l-1}}\big)\big\|_2^r\Big]^{1/r} \leq$$
$$C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^t \sum_{s=l-1}^l \big\|G_{\theta,p-1:p}^s(x_{p-1}^s, x_{p-1+\Delta_s:p}^s)\big\|_2^r\Big),$$

for $t \in \{1, \dots, T\}$.

Remark 26 *One also use the discussion of Remark 22 to extend Lemma 17 to*

$$\bar{\mathbb{E}}_\theta^{l-1,1}\Big[\big\|G_{\theta,t}^l\big(X_{0:t}^{l,B_T^l}\big)\big\|_2^r\Big] \leq C\Big(1 + \sum_{p=1}^t \big\|G_{\theta,p-1:p}^l(x_{p-1}^l, x_{p-1+\Delta_l:p}^l)\big\|_2^r\Big),$$

and similarly for $\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_{\theta,t}^l\big(\bar{X}_{0:t}^{l,\bar{B}_T^l}\big)\big\|_2^r\Big]$.

Corollary 27 *Under Assumptions 1 and 2, for any* $(t, r, \theta, C') \in \{1, \dots, T\} \times [1, \infty) \times \Theta \times \mathbb{R}^+$*, there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N, i, \delta) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \dots\} \times \{1, \dots, N\} \times \mathbb{R}^+$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$*, it holds that*

$$\bar{\mathbb{E}}_\theta^{l-1,1}\Big[\big\|X_t^{l,B_T^l} - X_t^{l-1,B_T^{l-1}}\big\|_2^r\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^t \sum_{s=l-1}^l \|x_p^s\|_2^r\Big).$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$ *also holds, then*

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|\bar{X}_t^{l,\bar{B}_T^l} - \bar{X}_t^{l-1,\bar{B}_T^{l-1}}\big\|_2^r\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\Big(1 + \sum_{p=1}^t \sum_{s=l-1}^l \|\bar{x}_p^s\|_2^r\Big).$$

Proof This follows from Corollary 21 and the discussion in Remark 26. ■

Lemma 28 *Under Assumptions 1 and 2, for any* $(T, \theta, C') \in \mathbb{N} \times \Theta \times \mathbb{R}^+$*, there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N, \delta, \gamma) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \dots\} \times \mathbb{R}^+ \times (0, \frac{\frac{1}{2}\wedge\beta}{\beta(1+\delta)})$ *and any* $(x_{0:T}^l, x_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$*, it holds that*

$$\bar{\mathbb{E}}_\theta^{l-1,1}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l)^c}\big(X_{0:T}^{l,B_T^l}, X_{0:T}^{l-1,B_T^{l-1}}\big)\Big] \leq \\ C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)} - \gamma\beta}\Big(1 + \sum_{p=1}^T \sum_{s=l-1}^l \big\{\|x_p^s\|_2^\gamma + \|G_{\theta,p-1:p}^s(x_{p-1}^s, x_{p-1+\Delta_s:p}^s)\|_2^\gamma\big\}\Big).$$

If $(\bar{x}_{0:T}^l, \bar{x}_{0:T}^{l-1}) \in \mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l$ *also holds, then*

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l)^c}\big(\bar{X}_{0:T}^{l,\bar{B}_T^l}, \bar{X}_{0:T}^{l-1,\bar{B}_T^{l-1}}\big)\Big] \leq \\ C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)} - \gamma\beta}\Big(1 + \sum_{p=1}^T \sum_{s=l-1}^l \big\{\|\bar{x}_p^s\|_2^\gamma + \|G_{\theta,p-1:p}^s(\bar{x}_{p-1}^s, \bar{x}_{p-1+\Delta_s:p}^s)\|_2^\gamma\big\}\Big).$$

Proof We only consider the first inequality as it is the same proof for the second inequality. For any $t \in \{1, \dots, T\}$, by Markov's inequality and Corollary 27, we have

$$\bar{\mathbb{P}}_\theta^{l-1,l}\Big(\big\|X_t^{l,B_T^l} - X_t^{l-1,B_T^{l-1}}\big\|_2 > C'\Delta_l^\beta\Big) \leq C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)} - \gamma\beta}\Big(1 + \sum_{p=1}^t \sum_{s=l-1}^l \|x_p^s\|_2^\gamma\Big).$$

Similarly, for any $t \in \{1, \dots, T\}$, by Markov's inequality and the results discussed in Remark 25

$$\bar{\mathbb{P}}_\theta^{l-1,l}\Big(\big\|G_{\theta,t}^l(X_{0:t}^{l,B_T^l}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,B_T^{l-1}})\big\|_2 > C'\Delta_l^\beta\Big) \\ \leq C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)} - \gamma\beta}\Big(1 + \sum_{p=1}^t \sum_{s=l-1}^l \|G_{\theta,p-1:p}^s(x_{p-1}^s, x_{p-1+\Delta_s:p}^s)\|_2^\gamma\Big).$$

Hence there exists a constant $C < \infty$ which depends on T but not l such that the result holds. ■

We recall the definition of $\mathsf{D}^l = \{(x_{0:T}, \bar{x}_{0:T}) \in \mathsf{Z}^l : x_{0:T} = \bar{x}_{0:T}\}$.

Lemma 29 *Under Assumptions 1 and 2, for any* $(T, \theta, N) \in \mathbb{N} \times \Theta \times \{2, 3, \dots\}$*, there exists a constant* $\varepsilon \in (0,1)$ *such that for any* $l \in \mathbb{N}$ *and any* $((x^l_{0:T}, \bar{x}^l_{0:T}), (x^{l-1}_{0:T}, \bar{x}^{l-1}_{0:T})) \in \mathsf{Z}^l \times \mathsf{Z}^{l-1}$*, it holds that*

$$\bar{\mathbb{E}}^{l-1,l}_\theta\big[\mathbb{I}_{\mathsf{D}^l}\big(X^{l,B^l_T}_{0:T}, \bar{X}^{l,\bar{B}^l_T}_{0:T}\big)\big] \wedge \bar{\mathbb{E}}^{l-1,l}_\theta\big[\mathbb{I}_{\mathsf{D}^{l-1}}\big(X^{l-1,B^{l-1}_T}_{0:T}, \bar{X}^{l-1,\bar{B}^{l-1}_T}_{0:T}\big)\big] \geq \varepsilon.$$

Proof Recall the definition (72) of $\overline{\vartheta}^{(l)}_t(j_1, j_3)$ in the proof of Lemma 23. This can be extended to time T using the same construction for both level l and $l-1$ and will correspond to the marginal distributions of $(B^l_T, \bar{B}^l_T)$ and $(B^{l-1}_T, \bar{B}^{l-1}_T)$. We denote these two probability distributions as $\overline{\vartheta}^{(l)}_t(B^l_T, \bar{B}^l_T)$ and $\overline{\vartheta}^{(l-1)}_t(B^{l-1}_T, \bar{B}^{l-1}_T)$. Also recall the definitions of $\check{\mathsf{S}}^l_t$ and $\check{\mathsf{S}}^{l-1}_t$ in (70)-(71).

We give the proof for level l only as the case of level $l-1$ is almost identical. We have the following inequalities

$$\begin{aligned}\bar{\mathbb{E}}^{l-1,l}_\theta\big[\mathbb{I}_{\mathsf{D}^l}\big(X^{l,B^l_T}_{0:T}, \bar{X}^{l,\bar{B}^l_T}_{0:T}\big)\big] &\geq \check{\mathbb{E}}^{l-1,l}_\theta\Big[\sum_{j=1}^{N-1} \mathbb{I}_{\check{\mathsf{S}}_{T-1}}(j)\overline{\vartheta}^{(l)}_T(j,j)\Big] \\ &\geq \check{\mathbb{E}}^{l-1,l}_\theta\Big[\mathbb{I}_{\check{\mathsf{S}}_{T-1}}(1)\overline{\vartheta}^{(l)}_T(1,1)\Big] \\ &\geq \varepsilon.\end{aligned}$$

In the first line, we have noted that for $(x^{l,B^l_T}_{0:T}, \bar{x}^{l,\bar{B}^l_T}_{0:T}) \in \mathsf{D}^l$ to occur, one must at least pick two equal indexes of pairs of particles at level l which were equal at time step $T-1$ of Algorithm 4. In the final line, we have used (73) and Lemma 23. This concludes the proof. ■

B.3.3 Results associated to the Initialization

Recall from Section 3.3 that the two pairs of CPF chains on $\mathsf{Z}^{l-1} \times \mathsf{Z}^l$ are initialized by sampling pairs $(X^{l-1,\star}_{0:T}, X^{l,\star}_{0:T})$ and $(\bar{X}^{l-1}_{0:T}, \bar{X}^l_{0:T})$ independently from $\nu^{l-1,l}_\theta$, and sampling $(X^{l-1}_{0:T}, X^l_{0:T}) \sim M^{l-1,l}_\theta(\cdot|X^{l-1,\star}_{0:T}, X^{l,\star}_{0:T})$ using the ML-CPF in Algorithm 6. We will denote the law of the tuple $(X^{l-1}_{0:T}, \bar{X}^{l-1}_{0:T}, X^l_{0:T}, \bar{X}^l_{0:T})$ under this initialization by $\check{\nu}^{l-1,l}_\theta$. Expectations w.r.t. $\nu^{l-1,l}_\theta$, $\check{\nu}^{l-1,l}_\theta$ and the ML-CPF kernel $M^{l-1,l}_\theta$ will be written as $\mathbb{E}^{l-1,l}_{\theta,\nu}$, $\check{\mathbb{E}}^{l-1,l}_{\theta,\nu}$ and $\mathbb{E}^{l-1,l}_\theta$, respectively.

Lemma 30 *Under Assumptions 1 and 2, for any* $(t, r, \theta) \in \{1, \dots, T\} \times [1, \infty) \times \Theta$*, there exists a constant* $C < \infty$ *such that for any* $(l, \beta, N, \delta) \in \mathbb{N} \times (0, \frac{1}{2}) \times \{2, 3, \dots\} \times \mathbb{R}^+$*, we have*

$$\check{\mathbb{E}}^{l-1,l}_{\theta,\nu}\Big[\big\|G^l_{\theta,t}(X^l_{0:t}) - G^{l-1}_{\theta,t}(X^{l-1}_{0:t})\big\|^r_2\Big]^{1/r} \leq C\Delta_l^{\frac{\beta}{r(1+\delta)}},$$

and

$$\check{\mathbb{E}}^{l-1,l}_{\theta,\nu}\Big[\big\|G^l_{\theta,t}(\bar{X}^l_{0:t}) - G^{l-1}_{\theta,t}(\bar{X}^{l-1}_{0:t})\big\|^r_2\Big]^{1/r} \leq C\Delta_l^{\frac{1}{2}}.$$

Proof The second inequality simply follows from Remark 11, so we only consider the first. We have

$$\check{\mathbb{E}}_{\theta,\nu}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]^{1/r}=\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\mathbb{E}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]\Big]^{1/r}.$$

We note that by construction

$$\mathbb{E}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]=\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big].$$

We consider the decomposition

$$\check{\mathbb{E}}_{\theta,\nu}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]=T_1+T_2, \tag{74}$$

where

$$T_1=\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}}(X_{0:T}^{l,\star},X_{0:T}^{l-1,\star})\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]\Big],$$
$$T_2=\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l,\star},X_{0:T}^{l-1,\star})\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]\Big],$$

for any $0<C'<\infty$. We will deal with both terms separately.

For T_1, applying the result in Remark 25 gives

$$T_1\le C\Delta_l^{\frac{\beta}{(1+\delta)}}\Big(1+\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s,\star},X_{p-1+\Delta_s:p}^{s,\star})\big\|_2^r\Big]\Big). \tag{75}$$

We can use boundedness properties of the appropriate terms along with the martingale-remainder methods in the proof of Lemma 6 to deduce that

$$T_1\le C\Delta_l^{\frac{\beta}{(1+\delta)}}.$$

For T_2, applying Hölder's inequality for any $\varrho>0$ gives

$$T_2\le T_3T_4,$$

where

$$T_3=\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l,\star},X_{0:T}^{l-1,\star})\Big]^{\frac{1}{(1+\varrho)}},$$
$$T_4=\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta,t}^{l}(X_{0:t}^{l})-G_{\theta,t}^{l-1}(X_{0:t}^{l-1})\big\|_2^r\Big]^{\frac{(1+\varrho)}{\varrho}}\Big]^{\frac{\varrho}{(1+\varrho)}}.$$

We now bound T_3 and T_4. For any $t\in\{1,\ldots,T\}$, using properties of the coupled Euler–Maruyama discretization and Markov's inequality, we have

$$\mathbb{P}_{\theta,\nu}^{l-1,l}\Big(\big\|X_t^{l,\star}-X_t^{l-1,\star}\big\|_2>C'\Delta_l^{\beta}\Big)\le C\Delta_l^{\alpha(\frac{1}{2}-\beta)}$$

for any $\alpha > 0$, where $\mathbb{P}_{\theta,\nu}^{l-1,l}$ denotes probability under $\nu_{\theta}^{l-1,l}$. Similarly, for any $t \in \{1,\dots,T\}$, it follows from Remark 11 that

$$\mathbb{P}_{\theta,\nu}^{l-1,l}\Big(\big\|G_{\theta,t}^{l}(X_{0:t}^{l,\star}) - G_{\theta,t}^{l-1}(X_{0:t}^{l-1,\star})\big\|_2 > C'\Delta_l^{\beta}\Big) \leq C\Delta_l^{\alpha(\frac{1}{2}-\beta)}$$

for any $\alpha > 0$. Hence there exists a constant $C < \infty$, that depends on T but not l, such that if $\alpha = (1+\varrho)$

$$T_3 \leq C\Delta_l^{(\frac{1}{2}-\beta)}. \tag{76}$$

For T_4, one can use the results discussed in Remark 26, along with the above argument (below (75)) to control terms such as $\mathbb{E}_{\theta,\nu}^{l-1,l}\big[\sum_{p=1}^{t}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s,\star},\dots,X_{p}^{s,\star})\big\|_2^r\big]$ to deduce that $T_4 \leq C$. Thus we have shown that

$$T_2 \leq C\Delta_l^{(\frac{1}{2}-\beta)}. \tag{77}$$

Combining (74)-(77) completes the proof. ■

Lemma 31 *Under Assumptions 1 and 2, for any $(T,\theta,C',\delta,\gamma) \in \mathbb{N}\times\Theta\times\mathbb{R}^+\times\mathbb{R}^+\times(0,\frac{1}{(1+\delta)})$, there exists a constant $C < \infty$ such that for any $(l,\beta,N) \in \mathbb{N}\times(0,\frac{1}{2})\times\{2,3,\dots\}$*

$$\check{\mathbb{E}}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l},X_{0:T}^{l-1})\Big] \leq C(\Delta_l)^{\{\beta(\frac{1}{(1+\delta)}-\gamma)\}\wedge(\frac{1}{2}-\beta)},$$

and

$$\check{\mathbb{E}}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(\bar{X}_{0:T}^{l},\bar{X}_{0:T}^{l-1})\Big] \leq C\Delta_l^{\frac{1}{2}-\beta}.$$

Proof As the proof of the second inequality is contained within the calculations to obtain (76), we will only consider the first inequality. We have

$$\check{\mathbb{E}}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l},X_{0:T}^{l-1})\Big] \leq T_1 + T_2, \tag{78}$$

where

$$T_1 = \mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l},X_{0:T}^{l-1})\Big]\mathbb{I}_{\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}}(X_{0:T}^{l,\star},X_{0:T}^{l-1,\star})\Big],$$
$$T_2 = \mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l,\star},X_{0:T}^{l-1,\star})\Big].$$

By Lemma 28, we have

$$T_1 \leq C(\Delta_l)^{\beta(\frac{1}{(1+\delta)}-\gamma)}\mathbb{E}_{\theta,\nu}^{l-1,l}\Big[\Big(1+\sum_{p=1}^{t}\sum_{s=l-1}^{l}\{\|X_p^{s,\star}\|_2^{\gamma} + \big\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s,\star},X_{p-1+\Delta_s:p}^{s,\star})\big\|_2^{\gamma}\}\Big)\Big].$$

The expectation can be controlled using the argument below (75), so we have

$$T_1 \leq C(\Delta_l)^{\beta(\frac{1}{(1+\delta)}-\gamma)}. \tag{79}$$

For T_2, using the second inequality in the statement of the lemma

$$T_2 \leq C\Delta_l^{\frac{1}{2}-\beta}. \tag{80}$$

Combining (78) with (79) and (80) concludes the proof. ■

B.3.4 Results associated to Score Estimation Methodology

We now study the two pairs of CPF chains $(X_{0:T}^{l-1}(i), \bar{X}_{0:T}^{l-1}(i))_{i=0}^{\infty}$ and $(X_{0:T}^{l}(i), \bar{X}_{0:T}^{l}(i))_{i=0}^{\infty}$ that are assumed to be run indefinitely even if both pairs of chains have met. We will denote the corresponding expectations by $\bar{\mathbb{E}}_{\theta}^{l-1,l}$. For any level $l \in \mathbb{N}_0$ and any probability measure π defined on X^l, we denote by $\mathbb{L}_2(\pi)$ the set of all measurable functions $\psi : \mathsf{X}^l \to \mathbb{R}$ such that $\pi(\psi^2) = \int_{\mathsf{X}^l} \psi(x)^2 \pi(dx) \in (0, \infty)$. The following results will involve the smoothing distribution π_{θ}^{l} defined in (14).

Lemma 32 *Under Assumptions 1 and 2, for any $(T, \theta) \in \mathbb{N} \times \Theta$, there exists $(\varepsilon, C) \in (0,1) \times \mathbb{R}^+$ such that for any $(l, N, i, \psi) \in \mathbb{N} \times \{2, 3, \dots\} \times \mathbb{N} \times \mathbb{L}_2(\pi_{\theta}^{l})$*

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(X_{0:T}^{l}(i))] \vee \bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(\bar{X}_{0:T}^{l}(i-1))] \leq C\varepsilon^i \pi_{\theta}^{l}(\psi^2)^{1/2} + \pi_{\theta}^{l}(|\psi|).$$

Also if $\psi \in \mathbb{L}_2(\pi_{\theta}^{l-1})$ then

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(X_{0:T}^{l-1}(i))] \vee \bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(\bar{X}_{0:T}^{l-1}(i-1))] \leq C\varepsilon^i \pi_{\theta}^{l-1}(\psi^2)^{1/2} + \pi_{\theta}^{l-1}(|\psi|).$$

Proof We will prove the result for $\bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(X_{0:T}^{l}(i))]$ only. The other results can be obtained in a similar way.

Marginally, the sequence $(X_{0:T}^{l}(i))_{i=1}^{\infty}$ is a Markov chain that has the initial distribution

$$\check{\nu}_{\theta}^{l}(dx_{0:T}) = \int_{\mathsf{X}^l} \nu_{\theta}^{l}(dx_{0:T}^{\star}) M_{\theta}^{l}(dx_{0:T} | x_{0:T}^{\star}) dx_{0:T}^{\star}$$

and Markov transition kernel M_{θ}^{l} as described in Algorithm 1. By Andrieu et al. (2018, Theorem 1b), one has

$$\left|\bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(X_{0:T}^{l}(i))] - \pi_{\theta}^{l}(\psi)\right| \leq \Big(\int_{\mathsf{X}^l} \frac{\nu_{\theta}^{l}(x_{0:T}^{\star})}{\pi_{\theta}^{l}(x_{0:T}^{\star})} \nu_{\theta}^{l}(dx_{0:T}^{\star}) \Big) \varepsilon^{i+1} \pi_{\theta}^{l}(\psi^2)^{1/2},$$

where we note that the extra power in ε follows as $X_{0:T}^{l}(0) \sim \nu_{\theta}^{l}$. Using Assumptions 2, it follows that

$$\left|\bar{\mathbb{E}}_{\theta}^{l-1,l}[\psi(X_{0:T}^{l}(i))] - \pi_{\theta}^{l}(\psi)\right| \leq C\varepsilon^i \pi_{\theta}^{l}(\psi^2)^{1/2},$$

and from here the proof is easily completed. ■

Lemma 33 *Under Assumptions 1 and 2, for any $(T, r, \theta, C') \in \mathbb{N} \times [1, \infty) \times \Theta \times \mathbb{R}^+$, there exists a constant $C < \infty$ such that for any $(l, \beta, N, \delta, i) \in \mathbb{N} \times \mathbb{R}^+ \times \{2, 3, \dots\} \times \mathbb{R}^+ \times \mathbb{N}$*

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta}^{l}(X_{0:T}^{l}(i)) - G_{\theta}^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r \mathbb{I}_{\mathsf{B}_{\beta,C'}^{l} \cap \mathsf{G}_{\beta,C'}^{l}}(X_{0:T}^{l}(i-1), X_{0:T}^{l-1}(i-1))\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}},$$

and

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta}^{l}(\bar{X}_{0:T}^{l}(i)) - G_{\theta}^{l-1}(\bar{X}_{0:T}^{l-1}(i))\big\|_2^r \mathbb{I}_{\mathsf{B}_{\beta,C'}^{l} \cap \mathsf{G}_{\beta,C'}^{l}}(\bar{X}_{0:T}^{l}(i-1), \bar{X}_{0:T}^{l-1}(i-1))\Big]^{1/r} \leq C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. By Corollary 24, we have the upper-bound

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\big\|G_{\theta}^{l}(X_{0:T}^{l}(i)) - G_{\theta}^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r \mathbb{I}_{\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}}(X_{0:T}^{l}(i-1), X_{0:T}^{l-1}(i-1))\Big]^{1/r} \leq$$
$$C(\Delta_l^{\frac{1}{2}\wedge\beta})^{\frac{1}{r(1+\delta)}}\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\Big(1+\sum_{p=1}^{T}\sum_{s=l-1}^{l}\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s}(i-1), X_{p-1+\Delta_s:p}^{s}(i-1))\|_2^r\Big)\Big].$$

Note that $\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s}, X_{p-1+\Delta_s:p}^{s})\|_2^r \in \mathbb{L}_2(\pi_\theta^s)$ (see the argument below (75)). In addition the expectation of the square of this function w.r.t. π_θ^s is bounded uniformly in s (one can use Assumptions 2 to upper-bound expectations w.r.t. π_θ^s by expectations w.r.t. ν_θ^s). Hence using Lemma 32, we obtain

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\Big(1+\sum_{p=1}^{T}\sum_{s=l-1}^{l}\big\|G_{\theta,p-1:p}^{s}(X_{p-1}^{s}(i-1), X_{p-1:\Delta_s:p}^{s}(i-1))\big\|_2^r\Big)\Big] \leq C, \tag{81}$$

which allows us to conclude the proof. ■

Lemma 34 *Under Assumptions 1 and 2, for any* $(T,\theta,C',\beta,\delta,\gamma) \in \mathbb{N}\times\Theta\times(\mathbb{R}^+)^3\times(0,\frac{\frac{1}{2}\wedge\beta}{\beta(1+\delta)})$*, there exists a constant* $C<\infty$ *such that for any* $(l,N,i)\in\mathbb{N}\times\{2,3,\dots\}\times\mathbb{N}$

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l}(i), X_{0:T}^{l-1}(i))\mathbb{I}_{\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}}(X_{0:T}^{l}(i-1), X_{0:T}^{l-1}(i-1))\Big] \leq C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)}-\gamma\beta},$$

and

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(\bar{X}_{0:T}^{l}(i), \bar{X}_{0:T}^{l-1}(i))\mathbb{I}_{\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l}}(\bar{X}_{0:T}^{l}(i-1), \bar{X}_{0:T}^{l-1}(i-1))\Big] \leq C(\Delta_l)^{\frac{\frac{1}{2}\wedge\beta}{(1+\delta)}-\gamma\beta}.$$

Proof The proof is essentially identical to that of Lemma 33, except one must use Lemma 28 instead of Corollary 24. ■

Lemma 35 *Under Assumptions 1 and 2, for any* $(T,\theta,C',\beta,\delta,\gamma)\in\mathbb{N}\times\Theta\times\mathbb{R}^+\times(0,\frac{1}{2})\times\mathbb{R}^+\times(0,\frac{1}{(1+\delta)})$*, there exists a constant* $C<\infty$ *such that for any* $(l,N,i)\in\mathbb{N}\times\{2,3,\dots\}\times\mathbb{N}$

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(X_{0:T}^{l}(i), X_{0:T}^{l-1}(i))\Big] \leq C(i+1)(\Delta_l)^{\{\beta(\frac{1}{(1+\delta)}-\gamma)\}\wedge(\frac{1}{2}-\beta)},$$

and

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^{l}\cap\mathsf{G}_{\beta,C'}^{l})^c}(\bar{X}_{0:T}^{l}(i), \bar{X}_{0:T}^{l-1}(i))\Big] \leq C(i+1)(\Delta_l)^{\{\beta(\frac{1}{(1+\delta)}-\gamma)\}\wedge(\frac{1}{2}-\beta)}.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. The proof is by induction on i. The initialization follows by Lemma 31. For the induction step, one can easily conclude by using Lemma 34 and the induction hypothesis. ■

Lemma 36 *Under Assumptions 1 and 2, for any* $(T, r, \theta, \beta, \delta, \gamma) \in \mathbb{N} \times [1, \infty) \times \Theta \times (0, \frac{1}{2}) \times \mathbb{R}^+ \times (0, \frac{1}{(1+\delta)})$*, there exists a constant* $C < \infty$ *such that for any* $(l, N, i) \in \mathbb{N} \times \{2, 3, \dots\} \times \mathbb{N}$

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r\Big]^{1/r} \leq C(i+1)\Delta_l^\phi,$$

where $\phi = \frac{\beta}{r(1+\delta)} \wedge \frac{1}{(1+\delta)}(\{\beta(\frac{1}{(1+\delta)} - \gamma)\} \wedge (\frac{1}{2} - \beta))$*, and*

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(\bar{X}_{0:T}^l(i)) - G_\theta^{l-1}(\bar{X}_{0:T}^{l-1}(i))\big\|_2^r\Big]^{1/r} \leq C(i+1)\Delta_l^\phi.$$

Proof We only consider the first inequality as it is the same proof for the second inequality. The proof is by induction on i. The initialization follows from Lemma 30. For the induction step, one has

$$\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r\Big]^{1/r} \leq C(T_1 + T_2),$$

where

$$T_1 = \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r \mathbb{I}_{\mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l}(X_{0:T}^l(i-1), X_{0:T}^{l-1}(i-1))\Big]^{1/r},$$
$$T_2 = \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^r \mathbb{I}_{(\mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l)^c}(X_{0:T}^l(i-1), X_{0:T}^{l-1}(i-1))\Big]^{1/r},$$

for any $0 < C' < \infty$. For T_1, one can apply Lemma 33 to obtain $T_1 \leq C\Delta_l^\phi$. For T_2, one can use Hölder's inequality to get the bound

$$T_2 \leq \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^{r\frac{1+\delta}{\delta}}\Big]^{\frac{\delta}{r(1+\delta)}} \times \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\mathbb{I}_{(\mathsf{B}_{\beta,C'}^l \cap \mathsf{G}_{\beta,C'}^l)^c}(X_{0:T}^l(i), X_{0:T}^{l-1}(i))\Big]^{\frac{1}{(1+\delta)}}.$$

To deal with the leftmost expectation on the R.H.S. one can rely on the same argument that led to (81) and for the other expectation one can use Lemma 35. This allows us to obtain

$$T_2 \leq C(i+1)\Delta_l^\phi,$$

and conclude the proof. ∎

In the following, we will employ the notation $\mathsf{A}_i = \{j \in \mathbb{N} : j > i\}$ for $i \in \mathbb{N}$.

Lemma 37 *Under Assumptions 1 and 2, for any* $(T, \theta, N) \in \mathbb{N} \times \Theta \times \{2, 3, \dots\}$*, there exists* $(\varepsilon, C) \in (0, 1) \times \mathbb{R}^+$ *such that for any* $(l, i) \in \mathbb{N}^2$

$$\bar{\mathbb{E}}_\theta^{l-1,l}\big[\mathbb{I}_{\mathsf{A}_i}(\bar{\tau}_\theta^l)\big] \leq C\varepsilon^i.$$

Proof We have

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}[\mathbb{I}_{\mathsf{A}_i}(\bar{\tau}_{\theta}^{l})] \leq \bar{\mathbb{E}}_{\theta}^{l-1,l}[\mathbb{I}_{\mathsf{A}_i\times\mathsf{A}_i}(\tau_{\theta}^{l-1},\tau_{\theta}^{l})] + \bar{\mathbb{E}}_{\theta}^{l-1,l}[\mathbb{I}_{\mathsf{A}_i^c\times\mathsf{A}_i}(\tau_{\theta}^{l-1},\tau_{\theta}^{l})] + \bar{\mathbb{E}}_{\theta}^{l-1,l}[\mathbb{I}_{\mathsf{A}_i\times\mathsf{A}_i^c}(\tau_{\theta}^{l-1},\tau_{\theta}^{l})].$$

By Lemma 29, we have $\bar{\mathbb{E}}_{\theta}^{l-1,l}[\mathbb{I}_{\mathsf{A}_i}(\tau_{\theta}^{s})] \leq C\varepsilon^i$ for $s\in\{l-1,l\}$, and so the proof is now easily completed. ■

Remark 38 *It is worth noting that Lemmata 23, 29 and 37 are the only cases where our bounds have constants that depend upon N.*

Lemma 39 *Under Assumptions 1 and 2, for any $(T,\theta,\beta,\delta,\gamma,N,b)\in\mathbb{N}\times\Theta\times(0,\frac{1}{2})\times\mathbb{R}^+\times(0,\frac{1}{(1+\delta)})\times\{2,3,\ldots,\}\times\mathbb{N}_0$, there exists a constant $C<\infty$ such that for any $l\in\mathbb{N}$*

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\big[\|\widehat{I}_l(\theta)\|_2^2\big] \leq C\Delta_l^{2\phi},$$

where $\phi = \frac{\beta}{2(1+\delta)} \wedge \frac{1}{(1+\delta)}(\{\beta(\frac{1}{(1+\delta)}-\gamma)\}\wedge(\frac{1}{2}-\beta))$.

Proof We recall that $\widehat{I}_l(\theta) = \widehat{S}_l(\theta) - \widehat{S}_{l-1}(\theta)$, where $\widehat{S}_{l-1}(\theta)$ and $\widehat{S}_l(\theta)$ are time-averaged estimators of the form in (22). For level $s\in\{l-1,l\}$, note that we can rewrite the time-averaged estimator as $\widehat{S}_s(\theta) = (I-b+1)^{-1}\sum_{k=b}^{I}\widehat{S}_s^k(\theta)$ with

$$\widehat{S}_s^k(\theta) = G_\theta^s(X_{0:T}^s(k)) + \sum_{i=b+1}^{\tau_\theta^s-1}\left(G_\theta^s(X_{0:T}^s(i)) - G_\theta^s(\bar{X}_{0:T}^s(i-1))\right).$$

Hence we can rewrite

$$\widehat{I}_l(\theta) = \frac{1}{I-b+1}\sum_{k=b}^{I}\left(\widehat{S}_l^k(\theta) - \widehat{S}_{l-1}^k(\theta)\right). \tag{82}$$

Since we have

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\big[\|\widehat{I}_l(\theta)\|_2^2\big] \leq C\sum_{k=b}^{I}\bar{\mathbb{E}}_{\theta}^{l-1,l}\big[\|\widehat{S}_l^k(\theta) - \widehat{S}_{l-1}^k(\theta)\|_2^2\big] \tag{83}$$

using the representation in (82), it suffices to establish $\bar{\mathbb{E}}_{\theta}^{l-1,l}\big[\|\widehat{S}_l^k(\theta) - \widehat{S}_{l-1}^k(\theta)\|_2^2\big] \leq C\Delta_l^{2\phi}$.

We consider the decomposition

$$\bar{\mathbb{E}}_{\theta}^{l-1,l}\big[\|\widehat{S}_l^k(\theta) - \widehat{S}_{l-1}^k(\theta)\|_2^2\big] \leq C\sum_{j=1}^{2}T_j,$$

where

$$T_1 = \bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\|G_\theta^l(X_{0:T}^l(k)) - G_\theta^{l-1}(X_{0:T}^{l-1}(k))\|_2^2\Big],$$

$$T_2 = \bar{\mathbb{E}}_{\theta}^{l-1,l}\Big[\|\sum_{i=k+1}^{\bar{\tau}_\theta^l}\{G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i)) + G_\theta^l(\bar{X}_{0:T}^l(i)) - G_\theta^{l-1}(\bar{X}_{0:T}^{l-1}(i))\}\|_2^2\Big].$$

For T_1, one can apply Lemma 36 to obtain

$$T_1 \leq C\Delta_l^{2\phi}.$$

For T_2, we have

$$T_2 = \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\| \sum_{i=k+1}^{\infty} \mathbb{I}_{\mathsf{A}_i}(\bar{\tau}_\theta^l)\{G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i)) + G_\theta^l(\bar{X}_{0:T}^l(i)) - G_\theta^{l-1}(\bar{X}_{0:T}^{l-1}(i))\}\|_2^2\Big].$$

To shorten the notations, set

$$\upsilon_i = G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i)) + G_\theta^l(\bar{X}_{0:T}^l(i)) - G_\theta^{l-1}(\bar{X}_{0:T}^{l-1}(i))$$

where we denote the j^{th}-component of υ_i as $[\upsilon_i]^j$, $j \in \{1, \ldots, d\}$. Then by application of Minkowski's inequality, we have

$$T_2 \leq \sum_{j=1}^{d_\theta}\left(\sum_{i=k+1}^{\infty} \bar{\mathbb{E}}_\theta^{l-1,l}[\mathbb{I}_{\mathsf{A}_i}(\bar{\tau}_\theta^l)\{[\upsilon_i]^j\}^2]^{1/2}\right)^2.$$

Then applying the Cauchy-Schwarz inequality

$$T_2 \leq \sum_{j=1}^{d_\theta}\left(\sum_{i=k+1}^{\infty} \bar{\mathbb{E}}_\theta^{l-1,l}[\mathbb{I}_{\mathsf{A}_i}(\bar{\tau}_\theta^l)]^{1/4}\bar{\mathbb{E}}_\theta^{l-1,l}[\{[\upsilon_i]^j\}^4]^{1/4}\right)^2.$$

Now, using standard properties of the $\mathbb{L}_2$-norm along with Lemma 37

$$T_2 \leq C\Big(\sum_{i=k+1}^{\infty} (\varepsilon^{1/4})^i \bar{\mathbb{E}}_\theta^{l-1,l}[\|\upsilon_i\|_2^4]^{1/4}\Big)^2.$$

It is simple to ascertain that:

$$\bar{\mathbb{E}}_\theta^{l-1,l}[\|\upsilon_i\|_2^4]^{1/4} \leq C\Big(\bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(X_{0:T}^l(i)) - G_\theta^{l-1}(X_{0:T}^{l-1}(i))\big\|_2^4\Big] + \bar{\mathbb{E}}_\theta^{l-1,l}\Big[\big\|G_\theta^l(\bar{X}_{0:T}^l(i)) - G_\theta^{l-1}(\bar{X}_{0:T}^{l-1}(i))\big\|_2^4\Big]\Big)^{1/4}.$$

Therefore, applying Lemma 36 gives

$$T_2 \leq C\Delta_l^{2\phi}\Big(\sum_{i=k+1}^{\infty} (\varepsilon^{1/4})^i (i+1)\Big)^2 \leq C\Delta_l^{2\phi}. \tag{84}$$

Combining (83)-(84) concludes the proof.

■

Remark 40 *The strategy in the proof of Lemma 39 can be improved by using martingale methods and Wald's equality for Markov chains as considered in Heng et al. (2023). This strategy was not adopted as it would require more complicated arguments given the technical complexity of the problem and algorithms in this article.*

Remark 41 *A better rate of ϕ can be obtained in Lemma 39 if we consider the case of constant diffusion coefficient σ.*

Remark 42 *One can employ the approaches in Lemmata 32, 37 and 39 to establish that the expected value of $\widehat{I}_0(\theta)$ is upper-bounded by a finite constant.*

Proof [Proof of Theorem 2] We have to establish that (17) and (18) hold for some choice of PMF $(P_l)_{l=0}^{\infty}$. The unbiasedness property in (17) can be established using the same approach as in Jacob et al. (2020a, Theorem 3.1). Assumption 1 of Jacob et al. (2020a) is implied by Assumption 2(*iii*); Assumption 2 of Jacob et al. (2020a) can be verified by inspecting the construction in the proof of Lemma 29; and Assumption 3 of Jacob et al. (2020a) follows from the fact that $\|G_\theta^l\|_2^r \in \mathbb{L}_2(\pi_\theta^l)$ for any $(l,r) \in \mathbb{N}_0 \times [1,\infty)$ and (Andrieu et al., 2018, Theorem 1b). For the condition in (18), we apply Theorem 1, Lemma 39 and Remark 42 to obtain

$$\sum_{l=0}^{\infty} \mathcal{P}_l^{-1} \left\{ \mathrm{Var}\left[\widehat{I}_l(\theta)^j\right] + \left(S_l(\theta)^j - S(\theta)^j\right)^2 \right\} < C \sum_{l=0}^{\infty} \frac{\Delta_l^{2\phi}}{\mathcal{P}_l}$$

for all $j \in \{1, \ldots, d_\theta\}$, where Var denotes variance under $\bar{\mathbb{E}}_\theta^{l-1,l}$ for all $l \in \mathbb{N}$. We can conclude the proof by selecting for instance $P_l \propto \Delta_l^{2\phi\alpha}$ for any $\alpha \in (0,1)$. ■

Remark 43 *The approach in Lemma 32 also suggests an alternative method of proof. If one could identify the invariant distribution of the ML-CPF kernel $M_\theta^{l-1,l}$ (Algorithm 6) and establish an ergodic theorem as in Andrieu et al. (2018, Theorem 1b), one could then study the expectation of differences of the type $\|G_\theta^l - G_\theta^{l-1}\|_2^r$ under the invariant distribution. Characterizing the invariant distribution could follow the ideas in Jasra and Yu (2020). This potentially interesting strategy is left as a topic for future work.*

Appendix C. Model-Specific Expressions

C.1 Ornstein–Uhlenbeck Process

For this example, we have $\Sigma(x) = \sigma^2$, $b_\theta(x) = \sigma^{-1} a_\theta(x)$ for $x \in \mathbb{R}$ and $\theta \in \Theta = (0,\infty) \times \mathbb{R} \times (0,\infty)$. To evaluate (12) and (15), the gradients required are given by

$$\nabla_\theta a_\theta(x) = ((\theta_2 - x), \theta_1, 0), \quad \nabla_\theta \log g_\theta(y|x) = \left(0, 0, -\frac{1}{2\theta_3} + \frac{(y-x)^2}{2\theta_3^2}\right),$$

for $x \in \mathbb{R}$ and $\theta = (\theta_1, \theta_2, \theta_3) \in \Theta$. In this example, the score function $S(\theta)$ can be computed using

$$\nabla_\theta \log p_\theta(y_{1:T}) = \sum_{t=1}^{T} \int_{\mathbb{R}^2} \left\{\nabla_\theta \log p_\theta(dx_t|x_{t-1}) + \nabla_\theta \log g_\theta(y_t|x_t)\right\} p_\theta(dx_{t-1}, dx_t|y_{1:T}),$$

where the transition kernel of the SDE (29) on a unit interval is

$$p_\theta(dx_t|x_{t-1}) = \mathcal{N}\left(x_t; \theta_2 + (x_{t-1} - \theta_2)\exp(-\theta_1), \frac{\sigma^2(1 - \exp(-2\theta_1))}{2\theta_1}\right) dx_t,$$

and the marginal of the smoothing distribution $p_\theta(dx_{t-1}, dx_t|y_{1:T})$ is a Gaussian distribution whose mean and covariance can be obtained using a Kalman smoother.

C.2 Logistic Diffusion Model for Population Dynamics of Red Kangaroos

In this application, we have $\Sigma(x) = 1$ and $b_\theta(x) = a_\theta(x)$ for $x \in \mathbb{R}$ and $\theta \in \Theta = \mathbb{R} \times (0, \infty)^3$. Evaluation of (12) and (15) require the following expressions. Firstly, we have

$$\begin{aligned}\nabla_\theta \log \mu_\theta(x) &= (0, 0, \partial_{\theta_3} \log \mu_\theta(x), 0)\\ \partial_{\theta_3} \log \mu_\theta(x) &= \frac{1}{\theta_3} - \frac{\theta_3}{10^2}(x - 5/\theta_3)^2 - \frac{5}{10^2}(x - 5/\theta_3),\end{aligned}$$

and

$$\nabla_\theta a_\theta(x) = \left(\frac{1}{\theta_3}, -\frac{1}{\theta_3}\exp(\theta_3 x), -\frac{\theta_1}{\theta_3^2} - \frac{\theta_2}{\theta_3^2}\exp(\theta_3 x)(\theta_3 x - 1), 0\right),$$

for $x \in \mathbb{R}$ and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta$. The conditional density can be written as

$$\begin{aligned}g_\theta(y|x) &= \mathcal{NB}(y^1; \theta_4, \exp(\theta_3 x))\mathcal{NB}(y^2; \theta_4, \exp(\theta_3 x))\\ &= \frac{\Gamma(y^1 + \theta_4)\Gamma(y^2 + \theta_4)}{\Gamma(\theta_4)^2 (y^1)!(y^2)!}\left(\frac{\theta_4}{\theta_4 + \exp(\theta_3 x)}\right)^{2\theta_4}\left(\frac{\exp(\theta_3 x)}{\theta_4 + \exp(\theta_3 x)}\right)^{y^1+y^2}\end{aligned}$$

for $y = (y^1, y^2)$ and $x \in \mathbb{R}$. Hence

$$\nabla_\theta \log g_\theta(y|x) = (0, 0, \partial_{\theta_3} \log g_\theta(y|x), \partial_{\theta_4} \log g_\theta(y|x)),$$

with

$$\partial_{\theta_3} \log g_\theta(y|x) = -\frac{2\theta_4 x \exp(\theta_3 x)}{\theta_4 + \exp(\theta_3 x)} + (y^1 + y^2)x\left(1 - \frac{\exp(\theta_3 x)}{\theta_4 + \exp(\theta_3 x)}\right),$$

and

$$\begin{aligned}\partial_{\theta_4} \log g_\theta(y|x) = \psi(y^1 + \theta_4) + \psi(y^2 + \theta_4) - 2\psi(\theta_4) + 2\left\{\log(\theta_4) - \log(\theta_4 + \exp(\theta_3 x))\right\}\\ + 2\left(1 - \frac{\theta_4}{\theta_4 + \exp(\theta_3 x)}\right) - \frac{(y^1 + y^2)}{(\theta_4 + \exp(\theta_3 x))},\end{aligned}$$

where $x \mapsto \psi(x) = (d/dx)\log\Gamma(x)$ denotes the digamma function.

C.3 Neural Network Model for Grid Cells in the Medial Entorhinal Cortex

In this application, we have $\Sigma(x) = I_2$ and $b_\theta(x) = a_\theta(x)$ for $x = (x^1, x^2) \in \mathbb{R}^2$ and $\theta \in \Theta$. The following expressions are needed to evaluate (12) and (15). The non-zero entries of the Jacobian matrix $\nabla_\theta a_\theta(x) \in \mathbb{R}^{d \times d_\theta}$ are given by

$$\begin{aligned}
\partial_{\alpha_1} a_\theta^1(x) &= \tanh(\beta_1\sigma_2 x^2 + \gamma_1)/\sigma_1, \\
\partial_{\beta_1} a_\theta^1(x) &= \alpha_1\sigma_2 x^2 \left(1 - \tanh^2(\beta_1\sigma_2 x^2 + \gamma_1)\right)/\sigma_1, \\
\partial_{\gamma_1} a_\theta^1(x) &= \alpha_1 \left(1 - \tanh^2(\beta_1\sigma_2 x^2 + \gamma_1)\right)/\sigma_1, \\
\partial_{\delta_1} a_\theta^1(x) &= -x^1, \\
\partial_{\sigma_1} a_\theta^1(x) &= -\alpha_1 \tanh(\beta_1\sigma_2 x^2 + \gamma_1)/\sigma_1^2, \\
\partial_{\sigma_2} a_\theta^1(x) &= \alpha_1\beta_1 x^2 \left(1 - \tanh^2(\beta_1\sigma_2 x^2 + \gamma_1)\right)/\sigma_1, \\
\partial_{\alpha_2} a_\theta^2(x) &= \tanh(\beta_2\sigma_1 x^1 + \gamma_2)/\sigma_2, \\
\partial_{\beta_2} a_\theta^2(x) &= \alpha_2\sigma_1 x^1 \left(1 - \tanh^2(\beta_2\sigma_1 x^1 + \gamma_2)\right)/\sigma_2, \\
\partial_{\gamma_2} a_\theta^2(x) &= \alpha_2 \left(1 - \tanh^2(\beta_2\sigma_1 x^1 + \gamma_2)\right)/\sigma_2, \\
\partial_{\delta_2} a_\theta^2(x) &= -x^2, \\
\partial_{\sigma_1} a_\theta^2(x) &= \alpha_2\beta_2 x^1 \left(1 - \tanh^2(\beta_2\sigma_1 x^1 + \gamma_2)\right)/\sigma_2, \\
\partial_{\sigma_2} a_\theta^2(x) &= -\alpha_2 \tanh(\beta_2\sigma_1 x^1 + \gamma_2)/\sigma_2^2.
\end{aligned}$$

The partial derivatives of the log-conditional density are all zero except the ones w.r.t. κ_1 and κ_2, which can be expressed as

$$\partial_{\kappa_i} \log g_\theta^l(y_{t_p}|(X_t)_{t_{p-1}\leq t\leq t_p}) = y_{t_p}^i - \Delta_l \sum_{t:t_{p-1}\leq t\leq t_p} \lambda_i(X_t^i),$$

for $i = 1, 2$.

References

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Christophe Andrieu, Anthony Lee, and Matti Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2): 842–872, 2018.

Marco Ballesio, Ajay Jasra, Erik von Schwerin, and Raul Tempone. A Wasserstein coupled particle filter for multilevel estimation. *Stochastic Analysis and Applications*, pages 1–40, 2022.

Alexandros Beskos and Gareth O Roberts. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005.

Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6):1077–1098, 2006a.

Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382, 2006b.

Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth Roberts. Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics*, 37(1):223–245, 2009.

Alexandros Beskos, Dan Crisan, Ajay Jasra, Kengo Kamatani, and Yan Zhou. A stable particle filter for a class of high-dimensional state-space models. *Advances in Applied Probability*, 49(1):24–48, 2017.

Alexandros Beskos, Dan Crisan, Ajay Jasra, Nikolas Kantas, and Hamza Ruzayqat. Score-based parameter estimation for a class of continuous-time state space models. *SIAM Journal on Scientific Computing*, 43(4):A2555–A2580, 2021.

Jose Blanchet and Fan Zhang. Exact simulation for multivariate Itô diffusions. *Advances in Applied Probability*, 52(4):1003–1034, 2020.

Mark Briers, Arnaud Doucet, and Simon Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61, 2010.

Emery N. Brown. The theory of point processes for neural systems. In C.C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, editors, *Methods and Models in Neurophysics*, pages 691–726. Elsevier, 2005.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

Graeme Caughley, Neil Shepherd, and Jeff Short. *Kangaroos: their ecology and management in the sheep rangelands of Australia*. Cambridge University Press, 1987.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

Nicolas Chopin and Sumeetpal S Singh. On particle Gibbs sampling. *Bernoulli*, 21(3): 1855–1883, 2015.

Nicolas Chopin, Andras Fulop, Jeremy Heng, and Alexandre H Thiery. Computational Doob h-transforms for online filtering of discretely observed diffusions. In *International Conference on Machine Learning*, pages 5904–5923. PMLR, 2023.

Pierre Del Moral and Lawrence M Murray. Sequential Monte Carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):969–997, 2015.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22, 1977.

Brian Dennis and Robert F Costantino. Analysis of steady-state populations with the gamma abundance model: application to Tribolium. *Ecology*, 69(4):1200–1213, 1988.

Randal Douc, Aurélien Garivier, Eric Moulines, and Jimmy Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777, 2008.

Paul Fearnhead, David Wyncoll, and Jonathan Tawn. A sequential smoothing algorithm with linear computational cost. *Biometrika*, 97(2):447–464, 2010.

Eric Fournié, Jean-Michel Lasry, Jérôme Lebuchoux, Pierre-Louis Lions, and Nizar Touzi. Applications of Malliavin calculus to Monte Carlo methods in finance. *Finance and Stochastics*, 3(4):391–412, 1999.

Michael B Giles. Multilevel Monte Carlo path simulation. *Operations research*, 56(3): 607–617, 2008.

Michael B Giles and Lukasz Szpruch. Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *The Annals of Applied Probability*, 24 (4):1585–1620, 2014.

Peter W Glynn and Chang-han Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

Peter W Glynn and Ward Whitt. The asymptotic efficiency of simulation estimators. *Operations research*, 40(3):505–520, 1992.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE proceedings F (radar and signal processing)*, 140 (2):107–113, 1993.

Pieralberto Guarniero, Adam M Johansen, and Anthony Lee. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112(520):1636–1647, 2017.

Torkel Hafting, Marianne Fyhn, Tora Bonnevie, May-Britt Moser, and Edvard I Moser. Hippocampus-independent phase precession in entorhinal grid cells. *Nature*, 453(7199): 1248–1252, 2008.

Jeremy Heng, Adrian N. Bishop, George Deligiannidis, and Arnaud Doucet. Controlled sequential Monte Carlo. *The Annals of Statistics*, 48(5):2904–2929, 2020.

Jeremy Heng, Ajay Jasra, Kody JH Law, and Alexander Tarakanov. On unbiased estimation for discretized models. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):616–645, 2023.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

Pierre E Jacob, Fredrik Lindsten, and Thomas B Schön. Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530): 721–729, 2020a.

Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020b.

Ajay Jasra and Fangyuan Yu. Central limit theorems for coupled particle filters. *Advances in Applied Probability*, 52(3):942–1001, 2020.

Ajay Jasra, Kengo Kamatani, Kody JH Law, and Yan Zhou. Multilevel particle filters. *SIAM Journal on Numerical Analysis*, 55(6):3068–3096, 2017.

Hilbert Johan Kappen and Hans Christian Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 162(5):1244–1266, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Peter E Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.

Jonas Knape and Perry De Valpine. Fitting complex population models by combining particle filters with Markov chain Monte Carlo. *Ecology*, 93(2):256–263, 2012.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Anthony Lee, Sumeetpal S Singh, and Matti Vihola. Coupled conditional backward sampling particle filter. *Annals of Statistics*, 48(5):3066–3089, 2020.

Fredrik Lindsten, Michael I Jordan, and Thomas B Schon. Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15:2145–2184, 2014.

Fredrik Lindsten, Randal Douc, and Eric Moulines. Uniform ergodicity of the particle Gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797, 2015.

Don Mcleish. A general method for debiasing a Monte Carlo. *Monte Carlo Methods and Applications*, 17:301–315, 2011.

Marcin Mider, Moritz Schauer, and Frank Van der Meulen. Continuous-discrete smoothing of diffusions. *Electronic Journal of Statistics*, 15(2):4295–4342, 2021.

Yu E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. akad. nauk Sssr*, 269(3):543–547, 1983.

Joonha Park and Edward L Ionides. Inference on high-dimensional implicit dynamic models using a guided intermediate resampling filter. *Statistics and Computing*, 30(5):1497–1522, 2020.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

Patrick Rebeschini and Ramon van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

Chang-Han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

Jean-Francois Richard and Wei Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411, 2007.

L Chris G Rogers and David Williams. *Diffusions, Markov processes and Martingales: Volume 2: Itô Calculus*, volume 2. Cambridge university press, 2000.

Hans-Christian Ruiz and Hilbert J Kappen. Particle smoothing for hidden diffusion processes: Adaptive path integral smoother. *IEEE Transactions on Signal Processing*, 65 (12):3191–3203, 2017.

Moritz Schauer, Frank Van Der Meulen, and Harry Van Zanten. Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli*, 23(4A):2917–2950, 2017.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

Helle Sørensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354, 2004.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147. PMLR, 2013.

Vladislav B Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304, 2017.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17, 2016.

Hermann Thorisson. *Coupling, stationarity, and regeneration*. Springer New York, 2000.

Matti Vihola. Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66 (2):448–462, 2018.

Tianze Wang and Guanyang Wang. Unbiased Multilevel Monte Carlo methods for intractable distributions: MLMC meets MCMC. *Journal of Machine Learning Research*, 24(249):1–40, 2023.

Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Shouto Yonekura and Alexandros Beskos. Online smoothing for diffusion processes observed with noise. *Journal of Computational and Graphical Statistics*, 2022.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.