Interacting Sequential Monte Carlo Samplers for Trans-dimensional Simulation

AJAY JASRA*, ARNAUD DOUCET†, DAVID A. STEPHENS* & CHRISTOPHER C. HOLMES⋆

*Department of Mathematics, Imperial College London, SW7 2AZ, UK

†Departments of Statistics & Computer Science, University of British Columbia, V6T 1Z2, Canada

⋆Department of Statistics, University of Oxford, OX1 3TG, UK

Abstract

In this article we present the methodology of interacting Sequential Monte Carlo (SMC) samplers. Sequential Monte Carlo samplers (Del Moral et al. 2005a) are methods for sampling from a sequence of densities on a common measurable space using Markov chain Monte Carlo (MCMC) (Metropolis et al. 1953; Hastings 1970) and sequential importance sampling/resampling (SIR) (Doucet et al. 2001; Liu 2001) methodology. One of the main problems with SMC samplers when simulating from trans-dimensional, multimodal static targets is that transition kernels do not mix which leads to low particle diversity. In such situations, sometimes under reasonable Markov kernels, poor Monte Carlo estimates may be derived. We present an interacting SMC approach for static inference, where SMC samplers are run in parallel on, initially, different regions of the state space and then moved onto the entire state space. Once the samplers reach a common space the samplers are combined and allowed to interact. The method is intended to increase the diversity of the population of samples. We demonstrate that interacting SMC has a Feynman-Kac (Del Moral 2004) representation and establish convergence results. We show how the methodology may be used for a trans-dimensional inference problem in Bayesian mixture modelling and also, using adaptive methods, a mixture modelling problem in population genetics.

[1]**AMS 2000 Subject Classification**: Primary 82C80; Secondary 60F99, 62F15

Key words: Sequential Monte Carlo, Markov chain Monte Carlo, Trans-dimensional simulation, Feynman-Kac formulae, Adaptive methods, Bayesian model selection

Short Title: Interacting SMC Samplers.

1 Introduction

One of the most challenging problems in modern statistical computation is that of simulating from multimodal and trans-dimensional target measures. Since the paper of Green (1995) (reversible jump MCMC), simulation from non-standard distributions which lie on state spaces of differing dimension is now possible. That is, π is the target probability measure on measurable space $(T, \mathcal{T})$ and $T = \bigcup_{k \in \mathcal{K}} \{k\} \times T_k$, $T_k \subseteq \mathbb{R}^k$, $\mathcal{K} \subseteq \mathbb{N}$. Such distributions arise quite often in many areas of statistics, for example in Bayesian mixture modelling (Richardson & Green 1997) and non-linear classification and regression (Denison et al. 2002).

In many applications it is often difficult to construct an appropriate reversible jump algorithm to correctly traverse the state-space; see for example Jasra et al. (2005a). This is often because the target density is multimodal and dimension jumping is difficult to achieve as information in a given dimension may be inadequate to provide a successful move into a different dimensional space.

A possible solution is to use population-based simulation; see Jasra et al. (2005b) for a review. This class of simulation method incorporates a collection of samples (often termed *particles* or population members) that are run in parallel and allowed to interact. The objective of population-based simulation is to improve the ability to explore the state-space, when compared to standard stochastic simulation approaches such as MCMC. One such approach is population-based MCMC (Liang & Wong 2001). A drawback is that, as for all MCMC methods, the population kernel needs to be ergodic. For example, if we wish to use adaptive methods (Andrieu & Robert 2001) it can be difficult to verify that the stochastic process is indeed ergodic; see Andrieu & Moulines (2003) for example. Another population-based simulation method, that does not rely on such properties of the transition kernel, is SMC samplers.

Sequential Monte Carlo methods have become an important tool in the analysis of sequential inference problems in statistics; see Doucet et al. (2001) and Liu (2001) for a review. However, it has recently been realized that such methodology may be applied for static inference problems; see Chopin (2002) and Del Moral & Doucet (2003) for

example. A particular SMC approach, SMC samplers, are methods designed to simulate from a set of probability measures $\{\pi_j : j = 0, \ldots, t, \pi_t \equiv \pi\}$ on common measurable space $(T, \mathcal{T})$ using SIR and MCMC methodology. They provide a flexible way to simulate from π, where, for the purposes of this article, the sequence $\pi_0, \ldots, \pi_{t-1}$ approaches π (in some sense).

A problem for SMC samplers is that of low particle diversity for multimodal, static targets (see the first example in Section 4). That is, particles may only represent a single mode of the target. In such situations, even under reasonable Markov kernels (they are able to sample the space in a reasonable amount of CPU time, but still take a long time), poor Monte Carlo estimates may be derived. Since the kernels do not move around the space *quickly*, many samples may be similar and estimates of expectations of functionals of interest can be incorrect. In this article we propose a method, which can be implemented with little extra computational cost (CPU, programming) to standard SMC methods, that attempts to ensure a diverse population (and hence all of the modes of the target are represented).

This article is structured as follows. In Section 2 we present a short summary of SMC samplers and our proposed methodology. In Section 3 we provide some theoretical analysis, in which we show that interacting SMC has a Feynman-Kac representation. From this viewpoint, our approach may be interpreted as a Feynman-Kac particle approximation with a change of approximation measure. We demonstrate, under minimal conditions, a bound on $\mathbb{L}_p$ distances (see Section 1.1 for a definition) for a class of functions of interest; this allows to provide a law of large numbers for Monte Carlo averages. We give a central limit theorem (CLT) for our particle approximation method; see also Del Moral & Guionnet (1999), Chopin (2004), Del Moral (2004) and Künsch (2005). Finally, we establish propagation of chaos properties (Del Moral & Miclo 2001). In Section 4 we give an example related to the Bayesian analysis of mixtures with an unknown number of components. Section 5 features an example related to population genetics; we combine SMC with adaptive methods which exploit certain features of our approach. In Section 6 we provide a discussion of our method as well as future work. The proofs of the results are given in the Appendix.

1.1 Notation

Throughout this article we adopt the following notation.

We denote a measurable space $(E, \mathcal{E})$ with the class of probability measures denoted $\mathcal{P}(E)$ and $\sigma-$finite measures $\mathcal{M}_\sigma(E)$. We assume that $\{x\} \in \mathcal{E}$, $x \in E$ (e.g. the Borel $\sigma-$algebra). Throughout the article we abuse the notation by denoting Radon-Nikodym derivatives of $\pi \in \mathcal{P}(E)$, $\lambda \in \mathcal{M}_\sigma(E)$, $\pi \ll \lambda$ as $d\pi/d\lambda$ (Section 3) or π (all other Sections). In all Sections other than 3 it is assumed that all probability measures have a common dominating ($\sigma-$finite) measure dx on the appropriate space and (in all Sections) that the Radon-Nikodym derivatives are bounded and strictly positive. We let $dx^{\otimes m} := dx \times \cdots \times dx$ (product m times).

We assume all functions $h \in \mathcal{B}_b(E)$, where $\mathcal{B}_b(E)$ is the class of bounded measurable functions, which, when associated with the supremum norm $||h|| := \sup_{x \in E} |h(x)|$ can be regarded as a Banach space. The $\mathbb{L}_p$ distance between two functions $f, h \in \mathcal{B}_b(E)$ is defined as $\mathbb{L}_p(f,h) := \left(\int_E |f(x) - h(x)|^p \mu(dx) \right)^{1/p}$, $p \geqslant 1$, $\mu \in \mathcal{P}(E)$. The oscillations of a function $f \in \mathcal{B}_b(E)$ are taken as $\mathrm{osc}(f) = \sup_{(x,y) \in E^2} |f(x) - f(y)|$. We denote $\pi(h) := \int_E h(x)\pi(dx)$, $\pi \in \mathcal{P}(E)$. The total variation distance between two probability measures μ, $\lambda \in \mathcal{P}(E)$ is taken as $||\mu - \lambda||_{TV} := \sup_{A \in \mathcal{E}} |\mu(A) - \lambda(A)|$. We also refer to the vector $x = (x_1, \ldots, x_n)$ as $x_{1:n}$ and the product space $E_0 \times \cdots \times E_n$ as $E_{0:n}$.

2 Proposed Methodology

2.1 SMC Samplers

The SMC method is essentially a sequential importance sampling/resampling approach where we sample a population or cloud of particles at time $n = 0, \ldots, t$, $\xi_n^{(N)} = (\xi_n^1, \ldots, \xi_n^N) \in T^N$ (assume the target is associated with state space $(T, \mathcal{T})$) from an importance function, and reweight in order to calculate expectations wrt π_n:

$$\mathbb{E}_{\pi_n}[h(X)] = \int_T h(x)\pi_n(x)dx. \tag{2.1}$$

SMC samplers begin by drawing $(\xi_0^1, \ldots, \xi_0^N)$ from $\eta_0 \in \mathcal{P}(T)$ and weight according

to $\pi_0 \in \mathcal{P}(T)$, i.e. $w_0(\xi_0^j) = \pi_0(\xi_0^j)/\eta_0(\xi_0^j)$, $j = 1, \ldots, N$ and estimate (2.1) by:

$$\widehat{\mathbb{E}_{\pi_0}[h(X)]} \quad = \quad \frac{\sum_{j=1}^N h(\xi_0^j) w_0(\xi_0^j)}{\sum_{j=1}^N w_0(\xi_0^j)}.$$

The method of SMC seeks to reweight the particles for π_1 and, in order to calculate the weights in practice, the method works on an extended state space $(T \times T, \mathcal{T} \times \mathcal{T})$. That is, $\xi_1^1, \ldots, \xi_1^N$ are drawn from a Markov kernel $K_1 : T \times \mathcal{T} \to [0, 1]$ and reweighting with respect to π_1, that is:

$$\begin{aligned} w_1(\xi_{0:1}^j) &\propto w_0(\xi_0^j) W_1(\xi_{0:1}^j) \\ W_1(\xi_{0:1}^j) &= \frac{\nu_1(\xi_{0:1}^j)}{\pi_0(\xi_0^j) K_1(\xi_0^j, \xi_1^j)} \end{aligned}$$

where ν_1 is a probability density wrt $dx^{\otimes 2}$ admitting π_1 as its marginal and W_1 is termed the incremental weight. The algorithm progresses in this manner until time t is reached ($\pi_t \equiv \pi$). It is well known that, for such importance sampling procedures, the weights will degenerate to zero, with the exception of a single particle which has weight 1. Resampling or selection is used to deal with this problem; see Doucet et al. (2001) or Liu & Chen (1998) for details. To see an SMC sampler in full, see Algorithm 2.1 (at the end of the paper) with $m = 1$ and no step 4.

For clarity of exposition we assume particles progress through a selection/mutation algorithm, that is, we sample $\xi_0^{(N)}$ (mutation) and then decide whether or not to perform a resampling step but always denoting the resulting particles $\widehat{\xi}_0^{(N)}$, i.e. continues as:

$$\xi_0^{(N)} \xrightarrow{\text{selection}} \widehat{\xi}_0^{(N)} \xrightarrow{\text{mutation}} \xi_1^{(N)}.$$

It should be noted that many other methods in the sequential Monte Carlo literature (e.g. resample move (Gilks & Berzuini 2001), the sequential particle filter of Chopin (2002) and population Monte Carlo (Cappé et al. (2004)) are special cases of SMC samplers.

2.2 Interacting SMC

Our method is based upon using m *parallel* SMC samplers, with an equal number of particles (as noted by Chopin (2004) this provides a way to monitor the degeneracy of

SMC simulation). For now, assume that we have m samplers with associated with target densities $\nu_{i,n}$ ($i \in \{1, \ldots, m\}$) which admit $\pi_{i,n}$ as the marginal (the $\pi_{i,n}$ are related to the π_n in some way to be described later) and importance functions $\eta_{i,0}$, $K_{i,1}, \ldots, K_{i,t}$.

The basic algorithm is given in Algorithm 2.1. We now discuss how this method is to be applied.

2.3 Methodology for Static Inference

Consider the case for which $\pi_0, \ldots, \pi_{t-1}$ are artificial densities used in order to allow movement around the state space (see Jasra et al. (2005b) for examples of such densities).

For each sampler i we target a sequence of probability measures $\pi_{i,n}$ (up to some terminal time $0 < n^* < t$) on measurable space $(T_{i,n}, \mathcal{T}_{i,n})$, with the assumption that $\forall i = 1, \ldots, n^*$ $T_{i,n} \subseteq T$, $T_{i,n} \in \mathcal{T}$, $\mathcal{T}_{i,n} \subseteq \mathcal{T}$ and $(T_{i,n}, \mathcal{T}_{i,n}) = (T, \mathcal{T})$ $\forall i, n^* + 1 \leqslant n \leqslant t$. We assume $\forall i$

$$T_{i,0} \subseteq T_{i,1} \subseteq \cdots \subseteq T_{i,n^*} = T$$

with corresponding filtration

$$\mathcal{T}_{i,0} \subseteq \mathcal{T}_{i,1} \subseteq \cdots \subseteq \mathcal{T}_{i,n^*} = \mathcal{T}.$$

Note that it is often a good idea to let

$$T \quad = \quad \bigcup_i T_{i,0} \quad T_{i,0} \cap T_{j,0} = \emptyset \ \forall i \neq j$$

so that all the samples (initially) lie on distinct parts of the state space T.

We define initial distributions $\eta_{i,0} \in \mathcal{P}(T_{i,0})$ and Markov kernels $K_{i,n} : T_{i,n-1} \times \mathcal{T}_{i,n} \to [0, 1]$ and take (as in Neal (2001) and Del Moral et al. (2005a)) the auxiliary densities:

$$\nu_{i,n}(x_{i,1:n}) \quad = \quad \pi_{i,n}(x_n) \prod_{q=0}^{n-1} L_{i,q}(x_{i,q+1}, x_{i,q})$$

with

$$\pi_{i,n}(x) \quad \propto \quad \pi_n(x) \mathbb{I}(x \in T_{i,n})$$

assuming

$$\int_{T_{i,n}} \pi_n(x) dx \quad > \quad 0$$

and $L_{i,n} : T_{i,n+1} \times \mathcal{T}_{i,n} \to [0,1]$ is a Markov kernel with time reversed index (backwards Markov kernel). It should be noted that $L_{i,n}$ is essentially arbitrary and, since the optimal kernel (in the sense of minimizing variance of importance weights) is unavailable, the sub optimal choice of Del Moral et al. (2005a) is:

$$L_{i,n}(x_{i,n+1}, x_{i,n}) = \frac{\pi_{i,n}(x_{i,n})K_{i,n+1}(x_{i,n}, x_{i,n+1})}{\pi_{i,n}(x_{i,n+1})} \tag{2.2}$$

where $K_{i,n+1}$ has invariant distribution $\pi_{i,n}$. Thus, the incremental weights become:

$$W_{i,n}(\widehat{\xi}^j_{i,n-1}, \xi^j_{i,n}) = \frac{\pi_{i,n}(\xi^j_{i,n})L_{i,n-1}(\xi^j_{i,n}, \widehat{\xi}^j_{i,n-1})}{\pi_{i,n-1}(\widehat{\xi}^j_{i,n-1})K_{i,n}(\widehat{\xi}^j_{i,n-1}, \xi^j_{i,n})}$$

where $\xi^j_{i,n}$ is the j^{th} particle for sampler i at time n.

At time n^* we sample all particles from the same Markov kernel K_{n^*} and then resample so that all samples are approximately distributed according to π_{n^*+1} (which is the same for each sampler), we then sample from K_{n^*+1} and form a new set of particles:

$$\xi^j_{n^*+1} = (\xi^j_{1,n^*+1}, \ldots, \xi^j_{m,n^*+1}) \quad j = 1, \ldots, N.$$

We then continue with a single sampler targeting ν^m_n (the product auxiliary density), $n \geqslant n^* + 1$ on $(\prod_{i=1}^m T, \bigvee_{i=1}^m \mathcal{T})$ where the density is

$$\nu^m_n(x_{1:m,n}) = \prod_{i=1}^m \nu_n(x_{i,n}).$$

The algorithm continues with Markov kernels K_n $(n^*+2 \leqslant n \leqslant t)$, incremental weights:

$$W_n(\widehat{\xi}^j_{1:m,n-1}, \xi^j_{1:m,n}) = \prod_{i=1}^m W_n(\widehat{\xi}^j_{i,n-1}, \xi^j_{i,n}).$$

2.4 Notes on the Algorithm

Before we present both a theoretical analysis of the method as well as numerical examples we discuss a couple of aspects of the algorithm.

Extending the Space. From the outset it appears that moving the particles onto larger state spaces will be problematic. We adopt the approach of not attempting to move the particles too much, by employing Markov kernels which propose the new dimensionality with high probability of being the same as the current one and using fixed dimensional

proposals which have small variance, taking the backwards kernel to be the reverse (we describe this in more detail in the examples). The point here is that we will use our method so that the problem of extending the space is side-stepped. For example, in the application of the algorithm in Section 4, samples will lie in the other regions (with respect to a given sampler). In our second example, in Section 5, kernels are constructed so that we may traverse the entire space easily and therefore efficiently extending the space is not of concern in practice.

Stratifying the Space. The initial stratification of the space is not as difficult for trans-dimensional problems as for fixed dimensional problems and we discuss this in an example specific way. It should be noted that the idea of partitioning the state space is not new in Monte Carlo methods, for example stratified sampling (see Robert & Casella (2004) and the references therein) or, more recently, the Wang-Landau algorithm (Wang & Landau 2001; Atachade & Liu 2004) operate on a stratified space. See Robert & Casella (2004) pp. 155-156 for further discussion.

Choice of n^*. The choice of when the samples are combined is also of importance. That is, we will combine particles from different dimensional spaces; thus it is likely that some SMC samplers will have higher variance (in terms of the importance weights) than others. In this article we combine particles instantly and, we have found that this does not adversely affect the algorithm, if this is not done too far from time from the target of interest (but not so far as the Markov kernels used do not have time to explore the space, far depending upon the problem at hand). That is, given that the initial regions have reasonable support under our target density, if the samples can adequately represent these parts of the space then the difference in variability of samples is not a substantial difficulty. We discuss better ways to do this in Section 6.

3 Theoretical Analysis

In the present Section we demonstrate that our algorithm admits a Feynman-Kac (Del Moral 2004) representation which allows us to appeal to several convergence results in this area.

3.1 Feynman-Kac Formulae

Feynman-Kac formulae may be described as follows. Consider a sequence of measurable spaces $(E_n, \mathcal{E}_n)$ with $\mathcal{E}_n-$measurable functions W_n, inhomogeneous Markov kernels M_n (from $(E_{n-1}, \mathcal{E}_{n-1})$ into $(E_n, \mathcal{E}_n)$) and initial distribution η_0. Associating the Markov chain with the probability space $(\Omega = \prod_{n\geqslant 0} E_n, \mathcal{F} = \bigvee_{n\geqslant 0} \mathcal{E}_n, (X_n)_{n\geqslant 0}, \mathbb{P}_{\eta_0})$ ($\mathbb{P}_{\eta_0}$ is the probability law of a Markov chain with initial distribution η_0 and transitions M_n).

Most of our discussion will be in terms of the $n-$time predicted and updated marginals, $\eta_n, \widehat{\eta}_n$ respectively:

$$\begin{aligned}
\eta_n(f_n) &= \frac{\gamma_n(f_n)}{\gamma_n(1)} = \int_{E_{0:n}} f_n(x_n)\mathbb{Q}_{\eta_0,n}(dx_{0:n}) \\
\widehat{\eta}_n(f_n) &= \frac{\widehat{\gamma}_n(f_n)}{\widehat{\gamma}_n(1)} = \int_{E_{0:n}} f_n(x_n)\widehat{\mathbb{Q}}_{\eta_0,n}(dx_{0:n}) \\
\gamma_n(f_n) &= \int_{E_{0:n}} f_n(x_n) \prod_{q=0}^{n-1} W_q(x_q)\mathbb{P}_{\eta_0,n}(dx_{0:n}) \\
\widehat{\gamma}_n(f_n) &= \int_{E_{0:n}} f_n(x_n) \prod_{q=0}^{n} W_q(x_q)\mathbb{P}_{\eta_0,n}(dx_{0:n})
\end{aligned}$$

where $\mathbb{Q}_{\eta_0,n}(dx_{0:n}) = \frac{1}{Z_n}\prod_{q=0}^{n-1} W_q(x_q)\mathbb{P}_{\eta_0,n}(dx_{0:n})$ is the predicted Feynman-Kac path measure, $\widehat{\mathbb{Q}}_{\eta_0,n}(dx_{0:n}) = \frac{1}{\widehat{Z}_n}\prod_{q=0}^{n} W_q(x_q)\mathbb{P}_{\eta_0,n}(dx_{0:n})$ is the updated Feynman-Kac path measure, $Z_n, \widehat{Z}_n$ are the normalizing constants, $\mathbb{P}_{\eta_0,n}(dx_{0:n})$ is the finite dimensional probability law of the inhomogeneous Markov chain, $f_n \in \mathcal{B}_b(E_n)$ and we adopt the convention $\prod_{\emptyset} = 1$.

3.2 Feynman-Kac Representation of Interacting SMC

We now demonstrate that the interacting SMC algorithm has a Feynman-Kac representation, assuming we perform (multinomial) resampling at every iteration. We adopt an interpretation of the algorithm identical to the interacting Metropolis model of Del Moral & Doucet (2003).

Let $E_n = \prod_{i=1}^m (T_{i,n} \times T_{i,n})$, $\mathcal{E}_n = \bigvee_{i=1}^m (\mathcal{T}_{i,n} \times \mathcal{T}_{i,n})$, $V_n = \prod_{i=1}^m T_{i,n}$, $\mathcal{V}_n = \bigvee_{i=1}^m \mathcal{T}_{i,n}$

and adopt initial distributions and Markov kernels:

$$\begin{aligned}
\eta_0(dx_{1:m,0}) &= \prod_{i=1}^{m} \delta_{y_i}(dy_{i,0})\eta_{i,0}(dy'_{i,0}) \\
M_n(x_{1:m,n-1}, dx_{1:m,n}) &= \begin{cases} \prod_{i=1}^{m} \delta_{y'_{i,n-1}}(dy_{i,n})K_{i,n}(y_{i,n}, dy'_{i,n}) & 1 \leqslant n \leqslant n^* - 1 \\ \prod_{i=1}^{m} \delta_{y'_{i,n-1}}(dy_{i,n})K_n(y_{i,n}, dy'_{i,n}) & n^* \leqslant n \leqslant t \end{cases}
\end{aligned}$$

where $x_{i,n} = (y_{i,n}, y'_{i,n})$ and $y_i \in T_{i,0}$.

Now, assume $\forall i = 1, \ldots, m$:

$$\begin{aligned}
\pi_{i,0} &\sim \eta_{i,0} \\
\left(\pi_{i,n} \times L_{i,n-1}\right)_2 &\sim \left(\pi_{i,n-1} \times K_{i,n}\right)_1 \\
\left(\pi_n \times L_{n-1}\right)_2 &\sim \left(\pi_{n-1} \times K_n\right)_1
\end{aligned}$$

where $\sim$ denotes mutual absolute continuity here and:

$$\begin{aligned}
\left(\pi_{i,n-1} \times K_{i,n}\right)_1(d(x, x')) &= \pi_{i,n-1}(dx)K_{i,n}(x, dx') \\
\left(\pi_{i,n} \times L_{i,n-1}\right)_2(d(x, x')) &= \pi_{i,n}(dx')L_{i,n-1}(x', dx).
\end{aligned}$$

Then define:

$$\begin{aligned}
W_0(x_{1:m,0}) &= \prod_{i=1}^{m} \frac{d\pi_{i,0}}{d\eta_{i,0}}(y'_{i,0}) \\
W_n(x_{1:m,n}) &= \prod_{i=1}^{m} \frac{d(\pi_{i,n} \times L_{i,n-1})_2}{d(\pi_{i,n-1} \times K_{i,n})_1}(y_{i,n}, y'_{i,n}) \quad 1 \leqslant n \leqslant n^* \\
W_n(x_{1:m,n}) &= \prod_{i=1}^{m} \frac{d(\pi_n \times L_{n-1})_2}{d(\pi_{n-1} \times K_n)_1}(y_{i,n}, y'_{i,n}) \quad n^* + 1 \leqslant n \leqslant t
\end{aligned}$$

note we have used the fact that $\pi_{i,n^*} \equiv \pi_{n^*}$.

The above discussion allows us to ascertain the following Proposition, which is essentially the time reversal formula of Del Moral (2004) (Lemma 5.5.1).

Proposition 3.1. *For any $h \in \mathcal{B}_b(V_t)$ we have that:*

$$\widehat{\eta}_t(h) = \int_{V_t} h(y_{1:m,t})\pi_t^{\otimes m}(dy_{1:m,t}).$$

Remark. The proof is straightforward and given the Appendix. The result shows that the above interpretation provides an alternative way to view the algorithm discussed in Section 2.3.

Whilst the above Feynman-Kac representation allows us to consider our algorithm from a purely importance sampling framework (that is, without resampling), a McKean interpretation (e.g. Del Moral (2004)) of the flow is easily provided, which allows us to incorporate selection steps.

3.3 Particle Approximation

In order to simulate from the model described above, we would need to sample from the distributions of interest in the selection steps: which we are unable to do. Therefore, we consider a particle approximation.

We simulate from the following distributions:

$$\begin{aligned}
\mathbb{P}(d\xi_0^{(N)}) &= \prod_{j=1}^{N} \eta_0(d\xi_0^j) \\
\mathbb{P}(d\xi_n^{(N)}|\xi_{n-1}^{(N)}) &= \begin{cases} \prod_{j=1}^{N} \Phi_n(r^m(\xi_{n-1}^{(N)}))(d\xi_n^j) & 1 \leqslant n \leqslant n^*+1 \\ \prod_{j=1}^{N} \Phi_n(r(\xi_{n-1}^{(N)}))(d\xi_n^j) & n^*+2 \leqslant n \leqslant t \end{cases}
\end{aligned}$$

where

$$\begin{aligned}
r^m(\xi_{n-1}^{(N)})(dx_{n-1}) &= \prod_{i=1}^{m} \left(\frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_{i,n-1}^j}(dx_{i,n-1}) \right) \\
r(\xi_{n-1}^{(N)})(dx_{n-1}) &= \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_{1:m,n-1}^j}(dx_{1:m,n-1}) \\
\Phi_n(\eta_{n-1})(dx_n) &= \int_{E_{n-1}} \Psi_{n-1}(\eta_{n-1})(dx_{n-1}) M_n(x_{n-1}, dx_n) \\
\Psi_n(\mu)(dx) &= \frac{W_n(x)\mu(dx)}{\mu(W_n)}
\end{aligned}$$

where $\mu \in \mathcal{P}(E_n)$, $\Psi_n : \mathcal{P}(E_n) \to \mathcal{P}(E_n)$ is the Boltzmann-Gibbs transformation and $\Phi_n : \mathcal{P}(E_{n-1}) \to \mathcal{P}(E_n)$. In the notation of Del Moral (2004) we have that $\eta_n^N = r^m$ for $1 \leqslant n \leqslant n^*$ and $\eta_n^N = r$ for $n^*+1 \leqslant n \leqslant t$. Note that the superscript N denotes the particle approximation. We leave our results in full generality as they apply for any

corresponding Feynman-Kac particle approximation with such a change of approximation measure.

For the theoretical analysis to follow, we now assume the algorithm has an extra time step with a Markov transition M_{t+1} which is Dirac measure, $(E_{t+1}, \mathcal{E}_{t+1}) = (E_t, \mathcal{E}_t)$ (hence $\eta_{t+1} = \widehat{\eta}_t$, $\eta_{t+1}^N = \widehat{\eta}_t^N$).

3.4 $\mathbb{L}_p$ Bounds

We now present a convergence result, related to a bound on $\mathbb{L}_p$ distances of $\widehat{\eta}_t^N(h) - \widehat{\eta}_t(h)$. In terms of statistical inference we are interested in functions in the following class:

$$\mathcal{S}_s(E_t) \quad = \quad \{h : h(x_{1:m,t}) = \frac{1}{m}\sum_{i=1}^{m} f(x_{i,t}) \cap f \in \mathcal{B}_b(T_{i,t} \times T_{i,t})\}.$$

The result we present, however, is given for a larger class of functions, in order to simplify the proof. This class is (for $1 \leqslant l < \infty$):

$$\mathcal{S}_p(E_t) \quad = \quad \{h : h(x_{1:m,t}) = \sum_{j=1}^{l} c_j\big(\prod_{i=1}^{m} h_{ij}(x_{i,t})\big) \cap -\infty < c_j < \infty \ \forall j \cap h_{ij} \in \mathcal{B}_b(T_{i,t} \times T_{i,t})\}.$$

The result is as follows and is a simple adaptation of Proposition 2.9 of Del Moral & Miclo (2000):

Lemma 3.2. *For the particle model defined above and for any $h \in \mathcal{S}_p(E_t)$, $p \geqslant 1$ there exists a finite $C_{m,t+1}^{(p)}$*

$$\sqrt{N}\mathbb{E}\Big[\big|[\widehat{\eta}_t^N - \widehat{\eta}_t](h)\big|^p\Big]^{1/p} \quad \leqslant \quad C_{m,t+1}^{(p)} \sum_{j=1}^{l} |c_j| \prod_{i=1}^{m} ||h_{ij}||.$$

This implies that $\{\widehat{\eta}_t^N(h) : N \geqslant 1\}$ converges almost surely to $\widehat{\eta}_t(h)$ as $N \to \infty$.

Remark. The proof may be found in the Appendix. Since $h \in \mathcal{S}_s(E_t) \subset \mathcal{S}_p(E_t)$, we have the bound $||f||C_{m,t+1}^{(p)}$ for functions of actual statistical interest, as well as a form of law of large numbers. This Lemma provides a theoretical justification of our approach.

3.5 Central Limit Theorem

In the present Section we present a CLT for our interacting SMC sampler.

To simplify the statement of the result we make the following definitions. Firstly, that $Q_{n+1}(f)(x_n) = W_n(x_n)M_{n+1}(f)(x_n)$ and $Q_{q,n} = Q_{q+1}Q_{q+2}\dots Q_n$ (we adopt the convention $Q_{n,n} = Id$ the identity operator). Secondly, the matrix $\Theta_{i,n}(h)$, $h = (h^1, \dots, h^d)$, $h^r \in \mathcal{B}_b(T_{i,n-1} \times T_{i,n})$, $r = 1, \dots, d$, $n > 0$ has $(j,r)^{th}$ element

$$\Theta_{i,n}^{jr}(h) \;=\; \sum_{q=0}^{n} \eta_{i,q}\big(Q_{i,(q,n)}[h_j - \eta_{i,n}(h_j)]Q_{i,(q,n)}[h_r - \eta_{i,n}(h_r)]\big)$$

where $Q_{i,(q,n)}$ is the marginal semigroup of $Q_{q,n}$. Thirdly, denote the product space of d-dimensional functions of interest as $\mathcal{S}_s(E_t)^d$, with $h^r(x_{1:m,t}) = 1/m\big(\sum_{i=1}^m f^r(x_{i,t})\big)$ and $f = (f^1, \dots, f^d)$. Finally, denote the $d-$dimensional normal distribution $\mathcal{N}_d(\mu, \Sigma)$ with mean μ and covariance Σ. Throughout we make the abuse of notation that for an operator Q, $Q(h) = (Q(h_1), \dots, Q(h_l))$.

Proposition 3.3. *Under the weak integrability conditions of Del Moral (2004), pp. 300-306 we have for any $h \in \mathcal{S}_s(E_t)^d$, $d \geqslant 1$*

$$\sqrt{N}\big(\widehat{\eta}_t^N(h) - \widehat{\eta}_t(h)\big) \;\Rightarrow\; \mathcal{N}_d(0, \Theta_{t+1}(h))$$

where

$$\Theta_{t+1}(h) \;=\; \widehat{\Theta}_{n^*}(Q_{n^*,t+1}(h,1)) + \sum_{q=n^*+1}^{t+1} \widetilde{\Theta}_q(Q_{q,t+1}[h - \widehat{\eta}_t(h)])$$

and

$$\begin{aligned}
\widetilde{\Theta}_q^{jr}(Q_{q,t+1}[h - \widehat{\eta}_t(h)]) &= \eta_q\big(Q_{q,t+1}[h^j - \widehat{\eta}_t(h^j)]Q_{q,t+1}[h^r - \widehat{\eta}_t(h^r)]\big) \\
\widehat{\Theta}_{n^*}^{jr}(Q_{n^*,t+1}(h,1)) &= \sum_{l=1}^{m}\Big[\alpha_l^2\Theta_{l,n^*}^{jr}(Q_{l,(n^*,t+1)}(f,1)) + \alpha_l\beta_l^j\Theta_{l,n^*}^{j(d+1)}(Q_{l,(n^*,t+1)}(f,1)) + \\
&\qquad \alpha_l\beta_l^r\Theta_{l,n^*}^{r(d+1)}(Q_{l,(n^*,t+1)}(f,1)) + \beta_l^j\beta_l^r\Theta_{l,n^*}^{(d+1)(d+1)}(Q_{l,(n^*,t+1)}(f,1))\Big]
\end{aligned}$$

with

$$\begin{aligned}
\alpha_l &= \frac{1}{m}\prod_{i=1,i\neq l}^{m}\eta_{i,n^*}(Q_{i,(n^*,t+1)}(1)) \\
\beta_l^j &= \frac{1}{m}\sum_{i=1,i\neq l}^{m}\eta_{i,n^*}(Q_{i,(n^*,t+1)}(f^j))\Big[\prod_{r=1,r\neq i,l}^{m}\eta_{r,n^*}(Q_{r,(n^*,t+1)}(1))\Big] - \\
&\qquad \widehat{\eta}_t(f^j)\prod_{i=1,i\neq l}^{m}\eta_{i,n^*}(Q_{i,(n^*,t+1)}(1)).
\end{aligned}$$

Remark. The proof may be found in the Appendix. Our result shows that sample path averages of functionals of interest converge weakly to a normal distribution with covariance $\Theta_{t+1}(h)$. The covariance expression is decomposed of the impact of the initial stratification ($\widehat{\Theta}_{n^*}$) and reverting to a single approximation measure (the $\widetilde{\Theta}_{n^*}$). It would be of interest to prove under what conditions the variance is inferior to an ordinary (parallel) SMC sampler. Contrary to our $\mathbb{L}_p$ result and the propagation of chaos properties below, the result does not hold for general Feynman-Kac particle approximations with a change of approximation measure.

3.6 Propagation of Chaos

Propagation of chaos is an important area in the theoretical analysis of sequential Monte Carlo methods; see Del Moral et al. (2005b) for example. Such properties establish that, for a fixed time horizon t and fixed block sizes of particles (q) that the particles become asymptotically independent with the correct (target) probability measure. That is, since the particles are actually statistically dependent, propagation of chaos properties demonstrate (as $N \to \infty$) that the particles behave as if they were independent.

For the following result we denote the $t-$time marginal distribution of the $1 \leqslant q \leqslant N$ particles as $\mathbb{P}^{(N,q)}_{\eta_0,[t]}$ (after mutation) and denote $\widehat{\mathbb{P}}^{(N,q)}_{\eta_0,[t]}$ after selection (note $\widehat{\mathbb{P}}^{(N,q)}_{\eta_0,[t]}(\cdot) = \mathbb{P}^{(N,q)}_{\eta_0,[t+1]}(\cdot)$). Denote the tensor product of functions $h \in \mathcal{B}_b(E_n)$ as $h \otimes \cdots \otimes h$ (product q times) as $h^{(q)}$ and let $Q^{(q)}_{n+1} = W^{(q)}_n M^{(q)}_{n+1}$ with $M^{(q)}_{n+1} = M_{n+1} \times \cdots \times M_{n+1}$ (product q times). We then have the following result, which relies strongly upon the theory of Del Moral (2004) (see also Theorem 1.1 of Del Moral & Miclo (2001)):

Theorem 3.4. *For any $1 \leqslant q \leqslant N$ we have:*

$$||\widehat{\mathbb{P}}^{(N,q)}_{\eta_0,[t]} - \widehat{\eta}^{\otimes q}_t||_{TV} \leqslant \frac{(q-1)^2}{N} + \frac{2q^2}{N} + \frac{1}{\eta^{\otimes q}_t(W^{(q)}_t)}\big[D^{(1)}_{1,t}(q,N,||W^{(q)}_t||) + \big(D^{(2)}_{1,t}(q,N,||W^{(q)}_t||)\big)^{1/2}\big]$$

where for $f \in \mathcal{B}_b(E_n^q)$, $1 \leqslant n \leqslant n^*$

$$\begin{aligned} D_{m,n}^{(1)}(q,N,||f||) &= \frac{2mq^2||f||}{N} + \frac{1}{\eta_{n-1}^{\otimes q}(W_{n-1}^{(q)})}\big[D_{m,n-1}^{(1)}(q,N,||Q_n^{(q)}(f)||) + \\ &\quad ||f||\big(D_{m,n-1}^{(2)}(q,N,||W_{n-1}^{(q)}||)\big)^{1/2}\big] \\ D_{m,n}^{(2)}(q,N,||f||) &= D_{m,n}^{(1)}(2q,N,||f||^2) + 2||f||D_{m,n}^{(1)}(q,N,||f||) \end{aligned}$$

with initial values $D_{m,0}^{(1)}(q,N,||f||) = 2mq^2||f||/N$, $D_{m,0}^{(2)}(q,N,||f||) = 6m(2q^2)||f||^2/N$, $f \in \mathcal{B}_b(E_0^q)$ *and*

$$\begin{aligned} D_{1,n^*+1}^{(1)}(q,N,||f||) &= \frac{2q^2}{N} + \frac{1}{\eta_{n^*}^{\otimes q}(W_{n^*}^{(q)}))}\big[D_{m,n^*}^{(1)}(q,N,||Q_{n^*}^{(q)}(f)||) + \\ &\quad ||f||\big(D_{m,n^*}^{(2)}(q,N,||W_{n^*}^{(q)}||)\big)^{1/2}\big] \\ D_{1,n^*+1}^{(2)}(q,N,||f||) &= D_{1,n^*+1}^{(1)}(2q,N,||f||^2) + 2||f||D_{1,n^*+1}^{(1)}(q,N,||f||) \end{aligned}$$

with similar recursions for $D_{1,n}^{(1)}$, $D_{1,n}^{(2)}$, $n^* + 2 \leqslant n \leqslant t+1$.

Remark. The proof may be found in the Appendix. We note the functions, $D_{1,t}^{(1)}$, $D_{1,t}^{(2)}$ for fixed q, m will tend to zero as $N \to \infty$. The result relaxes the assumption of Del Moral & Miclo (2001) and Del Moral (2004) that the potential functions need to be upper bounded. We note, however, that the penalty is that the rate at which the total variation distance decreases to zero is much slower. It would be of interest to prove that the rate could be increased.

4 Example 1: Bayesian Mixture Modelling

To demonstrate our methodology we consider the Bayesian analysis of mixture models with an unknown number of components.

Mixture models are typically used to model heterogeneous data, or as a simple means of density estimation, see McLachlan & Peel (2001) for an overview. Bayesian analysis using mixtures has been fairly recent e.g. Richardson & Green (1997) and there is often substantial difficulty in simulation from mixtures see Jasra et al. (2005c) for example.

4.1 Model

We use the model from Richardson & Green (1997). The model is as follows; data $y_1, \ldots, y_q$ are i.i.d with distribution

$$y_i|\theta_k \quad \sim \quad \sum_{r=1}^{k} \omega_r \mathcal{N}(\mu_r, \lambda_r^{-1})$$

where $\theta_k = (\mu_{1:k}, \lambda_{1:k}\ , \omega_{1:k-1})$ ($\omega_{1:0}$ is assumed to be null). We denote the parameter space as $\bigcup_{k\in\mathcal{K}}\{k\} \times \Theta_k$, with Θ_k the parameter space for the $k-$component mixture model (and β a hyperparameter below) and $\mathcal{K} \subset \mathbb{N}$. The priors, which are the same for each component $r = 1, \ldots, k$, are taken to be: $\mu_r \sim \mathcal{N}(\xi, \kappa^{-1})$, $\lambda_r|\beta \sim \mathcal{G}a(\alpha, \beta)$, $\beta \sim \mathcal{G}a(g, h)$, $\omega_{1:k-1}|k \sim \mathcal{D}(\delta)$, $k \sim \mathcal{U}_{\{1,\ldots,k_{\max}\}}$ where $\mathcal{D}(\delta)$ is the symmetric Dirichlet distribution with parameter δ, $\mathcal{G}a(\alpha, \beta)$ is the Gamma distribution, shape α, scale β and $\mathcal{U}_{\{1,\ldots,k_{\max}\}}$ is the Uniform distribution on the integers $1, \ldots, k_{\max}$ with $k_{\max}$ known.

4.2 Sequential Monte Carlo Sampler

For this example we take the auxiliary distributions to be:

$$\pi_n(\theta, \beta, k|y_{1:q}) \quad \propto \quad l(y_{1:q}; \theta, k)^{\gamma_n} p(\theta, k) p(\beta) \quad n \in \{0, \ldots, t\}$$

where $\gamma_n \in (0, 1)$ is an inverse temperature to be defined below, $l(\cdot)$ is the likelihood function, p denotes a generic probability density and the initial distribution (importance distribution) is the prior. To apply the SMC sampler we use the following Markov (MCMC) kernels (see Robert & Casella (2004) for a recent overview):

1. Update β via a Gibbs kernel.

2. Update $\mu_{1:k}$ via a Metropolis-Hastings (MH) kernel with additive normal random walk proposal.

3. Update $\lambda_{1:k}$ via a MH kernel with multiplicative log-normal random walk proposal.

4. Update $\omega_{1:k-1}$ via a MH kernel with additive normal random walk proposal on the logit scale.

5. Update (θ, k) via a birth/death reversible jump (RJ) kernel which is the same as in Richardson & Green (1997) except no latent variables are simulated.

Note that for the problem we will consider the MCMC kernels mix reasonably well. We also note, that for the trans-dimensional move, we are not constrained to use an RJ kernel. However, our intuition is that since trans-dimensional moves are notoriously difficult to construct, if we use a RJ kernel, a bad move will be rejected and we protect ourselves from losing good particles. This is at the cost of having to apply reversible moves. For example, placing kernels associated to a birth move (e.g. Green (1995)) in the importance weight can lead to variances which are typically very high. However, by 'filtering' such a proposal through an accept/reject mechanism, we may obtain a smaller variance of the importance weight and less weight degeneracy in practice.

We will apply the kernels in the following way. At odd time points we apply the Gibbs kernel 1 and even time points apply the cycle of kernels 2-5, the initial target density is the prior which we are able to sample from. We do not change the distribution at even time points. This ensures regular updates at the cost of increasing the variance of the importance weights. The backwards Markov kernels are taken to be the suboptimal choice in Del Moral et al. (2005a) (equation 2.2). Thus, using the invariance of the MCMC kernels, at odd time points (recall we do not change the distribution at even time points) we have (unnormalized) incremental weight:

$$W_n(x_{1:n}) \quad = \quad l(y_{1:q}; \theta_{n-1}, k_{n-1})^{\gamma_n - \gamma_{n-1}}$$

at even time points (and 0) the temperature parameters are equal with unit incremental weights.

4.3 Data

For this example we consider the Hidalgo stamp data. The data are 485 measurements (in cm) of stamp thickness for the printing of a stamp issue from different types of paper: see McLachlan & Peel (2001) for further details.

The priors were set as in Richardson & Green (1997) of which we refer the reader for further details. We note $k_{\max} = 30$.

4.4 Application of SMC Sampler

We ran the SMC sampler detailed above 20 times each with a population size of 500. We took $t = 750$ (since we do not change densities at even time points we run the algorithm for 1500 time points) and used the systematic resampling method with threshold 250 samples (applied after the Gibbs update).

The MH kernels had a decreasing value of the proposal variance and these were set so that the average acceptance rate was in the range $(0.15, 0.6)$ (as noted by Chopin (2002) this will not necessarily mean that the kernels mix well, but it is not clear how to construct a good global MCMC move (or Markov kernel) which is why we have resorted to population-based simulation in the first place).

The temperature sequence was taken to be piecewise linear with γ increasing uniformly by $1/15000$ (i.e. $\gamma_1 = 1/15000$, $\gamma_2 = 2/15000$ etc) for the first 150 distributions, $1/2500$ for the next 125, $1/750$ for the next 225 and finally $1/250$ for the last 250 distributions (see Figure 1 (a)). This choice was made to help ensure that the resampling occurred quite consistently; see Figure 1 (b) (the plot is for the first run of the algorithm).

In Figure 2 (a) we can observe the sampled k. The performance of the sampler appears to be poor. In Figure 2 (a) there are regions of points with sampled k around $3 - 5$ and $7 - 14$. However, there does not appear to be any runs of the algorithm that are producing a diverse set of samples in terms of k.

The problem in this example (we were unable to find algorithm settings that lead to satisfactory results) is that the Markov kernels employed are able to sample the state space, but do so quite slowly (see Jasra et al. (2005b) for the results of a RJ algorithm). As a result, samples that may be about to jump between modes (and hence in a low probability regions) are often lost at the resampling stage. We now demonstrate that our interacting SMC sampler is able to avoid these difficulties.

4.5 Application of Interacting SMC Sampler

For the application of our method we took $m = 3$ samplers each run with a population of 167 particles (thus storage requirements are similar to the SMC sampler above) 20 times.

We used the same MCMC kernels, and resampling procedure as for the SMC sampler. For the temperatures we had a uniform cooling schedule until time $t/2$ at which point we targeted our original density of interest and combined samplers (thus after resampling for the first time after $t/2$ we are effectively running (non-interacting) parallel MCMC algorithms). This was to allow enough time for the kernels to explore the full space.

For the sets we took a sequence of three sets for each sampler and constrained the densities to lie in sets of the form

$$T_{i,n} = \bigcup_{k \in \mathcal{K}_{i,n}} \{k\} \times \Theta_k$$

where we took $\mathcal{K}_{1,n} = \{1, \ldots, 5\}$, $\mathcal{K}_{2,n} = \{6, \ldots, 10\}$ $\mathcal{K}_{3,n} = \{8, \ldots, 20\}$ $(n \leqslant 145)$ (i.e. the three samplers are constrained to these sets for the first 145 time points) and $\mathcal{K}_{1,n} = \{1, \ldots, 10\}$, $\mathcal{K}_{2,n} = \{3, \ldots, 13\}$, $\mathcal{K}_{3,n} = \{5, \ldots, 25\}$ $(146 \leqslant n \leqslant 375)$ with $\mathcal{K}_{i,n} = \{1, \ldots, k_{\max}\}$ $i \in \{1, 2, 3\}$ for the rest of the algorithm. Our choice of sets is based upon the results of SMC sampler, i.e. the inability of the SMC sampler to adequately represent the modes $k \in \{3, \ldots, 5\}$ and $k \in \{7, \ldots, 14\}$. We took the first sets for the first 145 time points (corresponding to approximately 1/4 of the distributions), then the second for the next 130 (corresponding to 1/2 of the distributions).

To extend the space at time n we generate a new $k_{i,n}^j$ with large probability of retaining the same value and otherwise uniform over $\mathcal{K}_{i,n} \setminus \{k_{i,n-1}^j\}$ and then using the same random walk proposal densities above (and the Gibbs kernel on β_n^j) if $k_{i,n}^j = k_{i,n-1}^j$ otherwise drawing from the prior (for the parameters). The backwards kernel has distribution on k that is identical if the $k_{i,n}^j \in \mathcal{K}_{i,n-1}$ and the same random walk densities, otherwise a uniform distribution over $\mathcal{K}_{i,n-1}$ and the prior on the parameters (thus the incremental weight is similar to a Hastings ratio).

We note that, in this type of example, we see our method as a way to assist SMC sampling, therefore the usage of the information from the SMC sampler is used in our simulation design (see Section 6 for further discussion).

4.6 Performance of Interacting SMC Samplers

In Figure 2 (b) we can see the performance of the interacting SMC sampler.

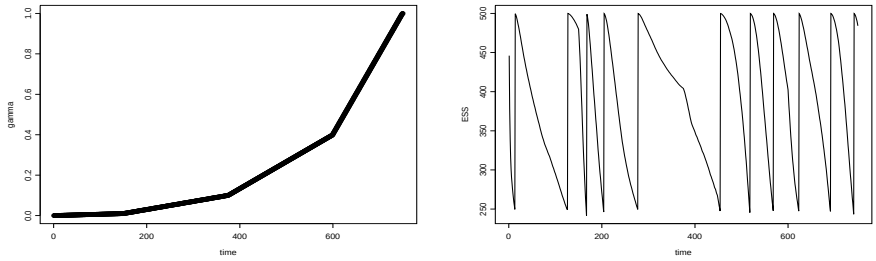
The sampled k (Figure 2 (b)) show that the interacting samplers have properly represented the full range of k which was missing under the standard SMC sampler. This may not mean the samples are correctly approximating the target, but the results are similar to a population-based RJMCMC algorithm that we ran (see Jasra et al. (2005b)).

Our experience with this method is that it is often best to allow a reasonable amount of time for the samplers to correctly approximate the target on the full space (i.e. linked to the fact that we are combining samplers with different variances). That is, when applying the method with the points are shifted onto the correct space close to time t, will mean poor estimates of quantities of interest. This must be counter-balanced with the cooling schedule. For example, if we allow unconstrained sampling when the inverse temperature is not close to 1 (this is dependent upon the problem) then it is likely that the sampling will revert to the original SMC sampler: we are unlikely to gain any advantage.

An important point is that if the Markov kernels used are unable to move around the space, then our method is unlikely to provide any improvement, unless we know the model probabilities *a priori*. That is, we will have samples that represent the entire space, but they may not be in correct proportion if the kernels do not mix. In the next example we provide a solution to this problem.

5 Example 2: Mixture Modelling in Population Genetics

For our second example we consider the analysis of multilocus genotype data using Bayesian mixture modelling. For the statistical model, we follow Pritchard et al. (2001) (model with no admixture), but in an attempt to make the example more realistic, we add some element of admixture. Note that, to our knowledge, **no** sampling method has ever been shown to work effectively for this class of models (that is, to move around the variable dimension support).



(a) Cooling Schedule. (b) Effective sample size.

Figure 1: Effective sample size plots (a) and temperature (b) from the SMC sampler; Hidalgo stamp data. We fitted the random beta model (Richardson & Green 1997) to the data, the output is for a single population of 500 samples from an SMC sampler using 750 tempered densities with a piecewise constant cooling schedule.

5.1 Model

In statistical terms, we may consider the data as a bilinear sequence of paired multinomial observations, that is $\mathbf{y} = (y_{111}, y_{112}, \ldots, y_{qL2})$ where y_{ilj} is observation $i \in \{1, \ldots, q\}$, at locus $l \in \{1, \ldots, L\}$ and site $j \in \{1, 2\}$. We have that $y_{ilj} \in \{1, \ldots, a_l\}$, i.e. there are a_l alleles at locus l. Given latent variable $z_{ilj} \in \{1, \ldots, k\}$, $k \in \{1, \ldots, k_{\max}\}$ and parameters $\boldsymbol{\theta}$ we take:

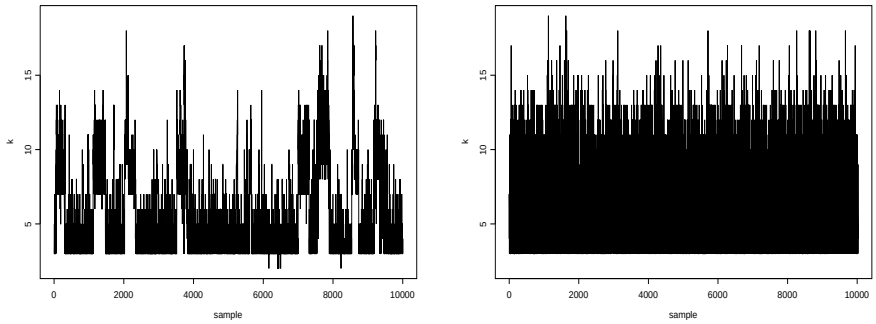
$$p(y_{ilj} = a|\boldsymbol{\theta}, z_{ilj} = r, k) \quad = \quad \theta_{rla}$$

where $\sum_{a=1}^{a_l} \theta_{rla} = 1 \ \forall r, l$ and the y_{ilj} are assumed independent given $\boldsymbol{\theta}, \mathbf{z}, k$.

For the priors k is taken to be uniform, the z_{ilj} are i.i.d given k and are uniform on the space. The $\boldsymbol{\theta}_{rl} = (\theta_{rl1}, \ldots, \theta_{rl(a_l-1)})$ are i.i.d given k and are symmetric Dirichlet $\mathcal{D}(\delta)$. Therefore we have posterior density for $(\mathbf{z}, \boldsymbol{\theta}, k)$, up to proportionality, is:

$$\pi(\mathbf{z}, \boldsymbol{\theta}, k|\mathbf{y}) \quad \propto \quad \left[\prod_{i,l,j} p(y_{ilj}|\boldsymbol{\theta}, z_{ilj}, k)\right] k! \frac{1}{k^{2Lq}} p(\boldsymbol{\theta}|k) p(k)$$

where the factorial is included due to the invariance of the priors and likelihoods to permutation of the labels of the parameters (see Jasra et al. (2005c) for example).



(a) SMC Sampler. (b) Interacting SMC Sampler.

Figure 2: Sampled k from the SMC samplers; Hidalgo stamp data. (a) is the standard SMC sampler, (b) is the interacting SMC sampler. Note that the SMC sampler consists of 20 runs with 500 particles each.

5.2 Interacting SMC Sampler

The SMC sampler will again employ a tempered densities strategy with MCMC updates. To assist a reversible jump move we describe below, the algorithm is performed after a reparameterization of the $\boldsymbol{\theta}$ on to the real line via a logit transformation (denote this $\boldsymbol{\phi}$).

For the sampler we will initially stratify the model space (i.e. there are $k_{\max}$ samplers) and then extend each sampler onto the entire space. This is done in a similar manner to the previous example except that the Gibbs kernels are employed.

The MCMC moves are used as follows:

1. Update $\mathbf{z}$ via a cycle of single site Gibbs updates.

2. Update $\boldsymbol{\phi}$ via a cycle of Gibbs updates.

3. Update $(\boldsymbol{\phi}, \mathbf{z}, k)$ via a mixture of reversible jump kernels (A) and (B).

The Markov kernel applied is thus a cycle of the above moves in the order given and are applied so that they leave the previous density in the sequence invariant. The backwards kernel is thus easily calculated using the form (2.2) and are thus the equation is omitted. For the Gibbs updates see Pritchard et al. (2001). Our moves are easily

applied even after reparameterization and model extension. The algorithm is initialized with a draw from the prior.

We note that the example will use adaptive methods, which can be justified in the following way. Suppose that we adapt some kernel K via some random variable ρ (e.g. the particle history), then the backwards kernel L_ρ may be also dependent on ρ, since (resetting the time counter):

$$\mathbb{E}\left[h(X')\frac{\pi_1(X')L_\rho(X',X)}{\eta_\rho(X)K_\rho(X,X')}\right] = \int h(x')\pi_1(x')L_\rho(x',x)\nu(\rho)dx'dxd\rho$$

where ν is the probability density for ρ. Application of Fubini's Theorem provides the appropriate unbiasedness.

Reversible Jump Move A

Move (A) is a simple (vanilla) RJ move employed to compare the move (B) with and is performed as follows. Propose to jump from state $(\boldsymbol{\phi}, \mathbf{z}, k)$ to $(\boldsymbol{\phi}', \mathbf{z}', k+1)$ with probability b_k and maintain the current $(\boldsymbol{\phi})$, generating new component parameters $\boldsymbol{\phi}_{k+1}$ say from the prior and then generating $\mathbf{z}'$ from the full conditional. In the reverse death (selected with probability d_{k+1}), select a component to die with uniform probability. We then invert the jump function and generate $\mathbf{z}$ from the full conditional.

The birth is accepted (when targeting density n) with probability $\min\{1, A\}$, with:

$$A = \frac{l(\mathbf{y};\boldsymbol{\phi}',\mathbf{z}',k+1)^{\gamma_n}\left(\frac{1}{k+1}\right)^{2Lq}(k+1)!}{l(\mathbf{y};\boldsymbol{\phi},\mathbf{z},k)^{\gamma_n}\left(\frac{1}{k}\right)^{2Lq}k!} \times \frac{d_{k+1}\pi_n(\mathbf{z}|\boldsymbol{\phi},k)}{(k+1)b_k\pi_n(\mathbf{z}'|\boldsymbol{\phi}',k+1)}$$

where $\pi_n(\mathbf{z}|\cdots)$ is the full conditional of the $\mathbf{z}$ given the time n.

Reversible Jump Move B

Move (B) is far more complicated and works assuming an initial stratification of the model space. We will assume that we have an approximation of the posterior in each dimension and use methods similar to those in Green (2003) and Hastie (2005).

More specifically, just before the time we extend the space we create a mixture approximation of the (current) target distribution in each dimension and use the methods

of Figuierdo & Jain (2002) to fit the mixture. In order to do this we select an identifiability constraint for the parameter space (recall that due to the invariance of the target density to relabelling the parameters, there are $k!$ symmetric modes (for a k component model) and thus for all the samples to represent a single mode we permute them) and fit a mixture of distributions to each model. We use the approximation in dimension k:

$$p_k(\boldsymbol{\phi}) \quad = \quad \sum_{j=1}^{g_k} v_j^k \left\{ \left[\prod_{l=1}^{L} \prod_{r=1}^{k} \mathcal{N}_{a_l-1}(\boldsymbol{\phi}_{rl}; \varrho_{rlj}^k, \varsigma_{rlj}^k) \right] k! \mathbb{I}(\phi_{1l^*j^*} < \cdots < \phi_{kl^*j^*}) \right\}$$

where (l^*, j^*) are the locus, allele pair we have chosen to identify the model and $\sum_{j=1}^{g_k} v_j^k = 1 \; \forall k$.

The move is as follows. We begin by permuting the sample to obey the identifiability constraint used in the approximation. In order to jump from $(\boldsymbol{\phi}, \mathbf{z}, k)$ to $(\boldsymbol{\phi}', \mathbf{z}', k+1)$, we select a birth with probability b_k and a component to add (c say) with uniform probability. We then select a component, p', of the approximation (in dimension $k+1$) to anchor on, with probability $v_{p'}^{k+1}$ with the reverse anchor chosen with probability:

$$r_p^k \quad \propto \quad v_p^k f_p(\boldsymbol{\phi})$$

where $f_p(\cdot)$ is the component (of the approximation) density. The jump function (assume for simplicity of notation that we add component $k+1$, but in general we my add any component), denoting the Cholesky decomposition of ς_{jlm} as φ_{jlm}, is:

$$\begin{aligned} \boldsymbol{\phi}'_{rl} &= \varrho_{rlp'}^{k+1} + \varphi_{rlp'}^{k+1} (\varphi_{rlp}^k)^{-1} [\boldsymbol{\phi}_{rl} - \varrho_{rlp}^k] \; r = 1, \ldots, k \\ \boldsymbol{\phi}'_{(k+1)l} &= \varrho_{(k+1)lp'}^{k+1} + \varphi_{(k+1)lp'}^{k+1} \mathbf{u}_l \end{aligned}$$

with $\mathbf{u}_l \sim \mathcal{N}_{a_l-1}(0, I_{(a_l-1)(a_l-1)})$. If the identifiability constraint is not satisfied we reject immediately. We propose $\mathbf{z}'$ from the full conditional. The reverse death is performed in much the same fashion, except we select a component to die with uniform probability and invert the appropriate jump function (i.e. conditional on the anchors and component to be removed).

We accept or reject the birth (when targeting density n) with probability $\min\{1, A\}$, with:

$$A \quad = \quad \frac{\pi_n(\mathbf{z}', \boldsymbol{\phi}', k+1|\mathbf{y})}{\pi_n(\mathbf{z}', \boldsymbol{\phi}, k|\mathbf{y})} \frac{\tau_{p'}^{k+1} v_p^k}{\tau_p^k v_{p'}^{k+1}} \frac{d_{k+1} \pi_n(\mathbf{z}|\boldsymbol{\phi}, k) |J|_{p,p',c}}{b_k \pi_n(\mathbf{z}'|\boldsymbol{\phi}', k+1) \prod_{l=1}^{L} \mathcal{N}_{a_l-1}(\mathbf{u}_l; 0, I_{(a_l-1)(a_l-1)})}$$

where $|J|_{p,p',c}$ is the Jacobian:

$$|J|_{p,p',k+1} = \left[\prod_{j=1}^{k}\prod_{l=1}^{L}\frac{|\varphi_{jlp'}^{k+1}|}{|\varphi_{jlp}^{k}|}\right]\prod_{l=1}^{L}|\varphi_{(k+1)lp'}^{k+1}|$$

and we use the notation $\mathcal{N}_k(\mathbf{x}; a, b)$ to denote the $k-$dimensional normal density evaluated at $\mathbf{x}$. We follow the accept/reject decision with a random permutation of the labels of the parameters, to allow invariance of the kernel.

5.3 The Data

For this example we use 50 (simulated) data points at 10 loci. The data were originally simulated from the package **POPGEN** and of size 200 data at 100 loci (and kindly simulated by Dr. J. Marchini, University of Oxford). The (original) data were generated so that there were 4 classes. We set $k_{\max} = 7$ (as in prior simulations we rarely sampled $k > 7$) and $\delta = 1$.

5.4 Interacting SMC Sampler

We ran 7 samplers for 600 time points. The temperatures increased uniformly from time 1 to 300 and then run on the target density. We stratified the space for 250 time points (at which time we constructed our mixture approximation). The final 300 time points were parallel MCMC sampling after the first time all samplers resampled (that is all the weights are uniform after this point and thus there is no reason to resample). We used systematic resampling with threshold half the sample size and ran each sampler with 1500 samples. For the approximation, we ordered on the first locus and allele.

Our choices are made for the following reasons (note that we do not claim any optimality for this sampler).

The samplers need to be run for a substantial time to allow appropriate movement around the state-space and to correct for the initial stratification. Another reason (for the time specification) is the kernels cannot be expected to mix quickly (even using the adaptive method). We ran parallel MCMC samplers, since for the adaptation procedure to be effective, the samples need to be close to the correct target. We also, as stated

above, want to correct for the stratification (which is more 'sudden' that for the previous example) and allow the samplers to interact.

The choice of identifiability constraint was arbitrary and we found that if we changed this, it seldom provided an improvement.

To improve the parallel MCMC run, we may use any of the strategies mentioned in Jasra et al. (2005a), but have not done so.

5.5 Performance of Interacting SMC Sampler

The algorithm was run on a Pentium 4 3 Ghz machine and took approximately 3.5 hours to run. The sampled k can be seen in Figure 3 and we can see that we have been able to successfully represent most of the models in the state space. For a vanilla reversible jump sampler (using jump A), we cannot correctly move around the space and even running the sampler for a long time will not be as reliable as the results presented here. This is because we have allowed interaction of the samples and parallel samplers.

To gauge the effectiveness of the adaptation method, we ran a reversible jump sampler for 50000 iterations (on the target) with the adaptive move and found that the acceptance rate for the birth of move A and B were 0.17% and 0.87% and the deaths were 0.21% and 1.3%, that is, the rate increases by about 8 times. Note this comparison used reversible jump, since the initial stratification and tempering can lead to distorted acceptance rates.

We note that our sampler has only returned a posterior which favours three classes, but due to the data reduction, there may not be enough signal in the data to suggest fitting four components.

6 Discussion

In this article we have presented an interacting sequential Monte Carlo method. We demonstrated that the method can significantly improve the performance of SMC samplers for mixtures of distributions with an unknown number of components.

Our method relies upon an initial stratification and tempering by extending the space. The choice of such stratification can be made based upon a pilot run of an SMC sampler.

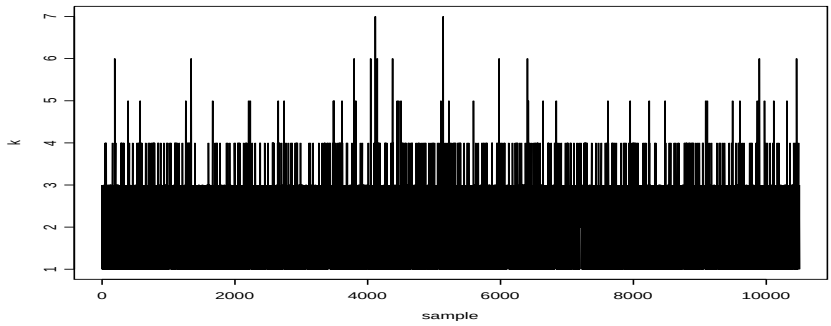


Figure 3: Sampled k from the SMC sampler; population genetics example. We ran 7 samplers, each with 1500 samples for 600 timepoints.

Indeed, we recommend that our method is best used in the situation where the original SMC sampler has not performed well (note that for our examples, without adaptation, the method requires little extra coding effort). Our second example ran samplers on each dimensionality, respectively, which solves the problem for small model spaces.

The example in Section 5 featured the use of adaptive methods. We believe that this is preferable to single chain adaptive MCMC, for the following reason. Our method effectively deals with the adaptation problem, that is, adaptive methods seek to find kernels which mix well over the space, but in order to this, the initial kernel must mix well over the space. Our approach uses, in the beginning, flat densities and population-based methods (i.e. resampling) to allow reasonable movement around the space and thus information to adapt the kernels.

We have noted that we combine particles from different state spaces, which are likely to have different variances in terms of the importance weights. This can mean (and we have observed this behaviour in simulations) that particles from certain regions are lost due to the fact that they do not have high target density when compared to other samplers. We recommended that the samplers are not combined until close to time t. A more satisfactory approach may be to use different tempering strategies (that is different for each sampler) so that samples have similar variance; it would be of interest to see if it is possible to derive an optimal set of temperatures so that samplers have approximately similar variance.

One method, that could deal with the above concerns, could be to only combine certain samplers. That is, to use instants $n_1^* < \cdots < n_r^*$ at which to combine the samples between samplers (at time n_r^* we have a single sampler). This approach would allow local diversity and provide an interesting way to allow the samplers to interact, via the selection step. From a theoretical point of view, it is clear that our convergence results may be extended to this case.

We have demonstrated our method for trans-dimensional problems only. The reasons for this are three-fold. Firstly, it is far easier to stratify the space for trans-dimensional problems than for most general simulation scenarios. Secondly, trans-dimensional problems are those for which it is most straightforward to detect poor coverage of the state-space. Lastly, trans-dimensional simulation is one of the most difficult problems in statistical computation and is vital for most areas of modelling. One way to use our methodology for fixed dimensional simulation would be to stratify the energy space (as in Kou et al. (2005)) and to extend this. This approach would provide an interesting area of future research.

Acknowledgement

The first author was supported by an Engineering and Physical Sciences Research Council Studentship. We would like to thank Raphael Gottardo and Adam Johansen for some of their comments on earlier versions and Jonathan Marchini for simulating the data for example 2.

A Proofs

Proof of Proposition 3.1. Consider $\widehat{\gamma}_t(h)$:

$$\widehat{\gamma}_t(h) \quad = \quad \int_{E_{0:t}} h(y'_{1:m,t}) \prod_{q=0}^{t} W_q(x_q) \mathbb{P}_{\eta_0}(dx_{0:t}).$$

Applying Fubini's Theorem and integrating over the Dirac measures we obtain:

$$\begin{aligned}\widehat{\gamma}_t(h) &= \int_{V_{0:t}} h(y'_{1:m,t}) \prod_{i=1}^{m} \Big\{ \frac{d\pi_{i,0}}{d\eta_{i,0}}(y'_{i,0}) \big(\prod_{q=1}^{n^*} \frac{d(\pi_{i,n} \times L_{i,n-1})_2}{d(\pi_{i,n-1} \times K_{i,n})_1}(y'_{i,n-1}, y'_{i,n}) \big) \times \\ &\quad \big(\prod_{q=n^*+1}^{t} \frac{d(\pi_n \times L_{n-1})_2}{d(\pi_{n-1} \times K_n)_1}(y'_{i,n-1}, y'_{i,n}) \big) \Big\} \mathbb{P}^K_{\eta_{i,0},t}(dy'_{i,0:t})\end{aligned}$$

where $\mathbb{P}^K_{\eta_{i,0},t}$ is the law of the inhomogeneous Markov chain:

$$\mathbb{P}^K_{\eta_{i,0},t}(dy_{i,0:t}) = \eta_{i,0}(dy_{i,0}) \big\{ \prod_{n=1}^{n^*-1} K_{i,n}(y_{i,n-1}, dy_{i,n}) \big\} \big\{ \prod_{n=n^*}^{t} K_n(y_{i,n-1}, dy_{i,n}) \big\}.$$

Now, we have that for all $i = 1, \ldots, m$:

$$\begin{aligned}\mathbb{P}^L_{\pi_t}(dy_{i,t:0}) &= \frac{d\pi_{i,0}}{d\eta_{i,0}}(y_{i,0}) \big\{ \prod_{q=1}^{n^*} \frac{d(\pi_{i,n} \times L_{i,n-1})_2}{d(\pi_{i,n-1} \times K_{i,n})_1}(y_{i,n-1}, y_{i,n}) \big\} \times \\ &\quad \big\{ \prod_{q=n^*+1}^{t} \frac{d(\pi_n \times L_{n-1})_2}{d(\pi_{n-1} \times K_n)_1}(y_{i,n-1}, y_{i,n}) \big\} \mathbb{P}^{K_i}_{\eta_{i,0}}(dy_{i,0:t})\end{aligned}$$

where

$$\mathbb{P}^L_{\pi_t}(dy_{i,t:0}) = \pi_t(dy_{i,t}) \big\{ \prod_{q=n^*}^{t-1} L_q(y_{i,q+1}, dy_{i,q}) \big\} \big\{ \prod_{q=0}^{n^*-1} L_{i,q}(y_{i,q+1}, dy_{i,q}) \big\}.$$

Therefore we can easily obtain:

$$\widehat{\gamma}_t(h) = \int_{V_t} h(y'_{1:m,t}) \pi_t^{\otimes m}(dy'_{1:m,t})$$

from which we can derive the result. □

Proof of Lemma 3.2. Recall that:

$$\begin{aligned}\eta^N_n(dx_{1:m}) &= \prod_{i=1}^{m} \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi^j_{i,n}}(dx_{i,n}) \quad 0 \leqslant n \leqslant n^* \\ \eta^N_n(dx_{1:m}) &= \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi^j_{1:m,n}}(dx_{1:m,n}) \quad n^* + 1 \leqslant n \leqslant t + 1.\end{aligned}$$

We first prove a result related to functions of the form $h(x_{1:m}) = \prod_{i=1}^{m} h_i(x_i)$ $h_i \in \mathcal{B}_b(T_{i,t} \times T_{i,t})$ for the $\mathbb{L}_p$ distances between $\eta^N_{n^*}$ and η_{n^*}, and then apply Minkowski's inequality so that the result applies for $h \in \mathcal{S}_p(E_{n^*})$. The result will allow a simple induction step in the proof of the Lemma.

Suppose $m = 1$, then by Proposition 2.9 of Del Moral & Miclo (2000) (note the fact that the measurable space changes in our context does not invalidate the application of the Proposition - it is valid when the measurable space changes with time index) we have:

$$\sqrt{N}\mathbb{E}\Big[\big|[\eta^N_{1,n^*} - \eta_{1,n^*}](h_1)\big|^p\Big]^{1/p} \leqslant ||h_1||C^{(p)}_{n^*}$$

where the measure $\eta^N_{i,n}$ refers to the marginal for the i^{th} Feynman-Kac particle approximation (resp. $\eta_{i,n}$ the $n-$time marginal). We thus conjecture that:

$$\sqrt{N}\mathbb{E}\Big[\big|\prod_{i=1}^m \eta^N_{i,n^*}(h_i) - \prod_{i=1}^m \eta_{1,n^*}(h_i)\big|^p\Big]^{1/p} \leqslant \big(\prod_{i=1}^m ||h_i||\big)C^{(p)}_{m,n^*} \tag{A.3}$$

for some finite $C^{(p)}_{m,n^*}$.

Assume the induction hypothesis (A.3) for $m = s$ and consider $m = s + 1$:

$$\begin{aligned}
\sqrt{N}\mathbb{E}\Big[|[\eta^N_{n^*} - \eta_{n^*}](h)|^p\Big]^{1/p} &= \sqrt{N}\mathbb{E}\Big[\big|\prod_{i=1}^{s+1} \eta^N_{i,n^*}(h_i) - \prod_{i=1}^{s+1} \eta_{i,n^*}(h_i)\big|^p\Big]^{1/p} \\
&= \sqrt{N}\mathbb{E}\Big[\big|\eta^N_{s+1,n^*}(h_{s+1})\big[\prod_{i=1}^{s} \eta^N_{i,n^*}(h_i) - \prod_{i=1}^{s} \eta_{i,n^*}(h_i)\big] + \\
&\quad \big(\prod_{i=1}^{s} \eta_{i,n^*}(h_i)\big)\big[\eta^N_{s+1,n^*}(h_{s+1}) - \eta_{s+1,n^*}(h_{s+1})\big]\big|^p\Big]^{1/p} \\
&\leqslant \big(\prod_{i=1}^{s+1} ||h_i||\big)C'^{(p)}_{s,n^*} + \big(\prod_{i=1}^{s+1} ||h_i||\big)C'^{(p)}_{n^*}
\end{aligned}$$

where we have used Minkowski's inequality, the mutual independence of the particle systems and Proposition 2.9 of Del Moral & Miclo (2000), which completes the induction proof with $C^{(p)}_{s+1,n^*} = C^{(p)}_{s,n^*} + C'^{(p)}_{n^*}$.

Now consider $h \in \mathcal{S}_p(E_{n^*})$, we have for a given m:

$$\begin{aligned}
\sqrt{N}\mathbb{E}\Big[\big|\sum_{j=1}^l c_j\big(\prod_{i=1}^m \eta^N_{i,n^*}(h_i) - \prod_{i=1}^m \eta_{i,n^*}(h_i)\big)\big|^p\Big]^{1/p} &\leqslant \sqrt{N}\sum_{j=1}^l |c_j|\mathbb{E}\Big[\big|\sum_{j=1}^l c_j \prod_{i=1}^m \eta^N_{i,n^*}(h_i) - \\
&\quad \prod_{i=1}^m \eta_{i,n^*}(h_i)\big|^p\Big]^{1/p} \\
&\leqslant C^{(p)}_{m,n^*}\sum_{j=1}^l |c_j| \prod_{i=1}^m ||h_{ij}||. \qquad \text{(A.4)}
\end{aligned}$$

To prove the Lemma, let $n^* + 1 \leqslant n \leqslant t + 1$ and assume that (A.4) holds for time $n - 1$. We conjecture that

$$\sqrt{N}\mathbb{E}\left[\left|[\eta_n^N - \eta_n](h)\right|^p\right]^{1/p} \leqslant C_{m,n}^{(p)} \sum_{j=1}^{l} |c_j| \prod_{i=1}^{m} ||h_{ij}||$$

for some finite $C_{m,n}^{(p)}$.

Now it is clear that by the Marcinkiewicz-Zygmund inequality (e.g. Shiryaev (1996)) that

$$\sqrt{N}\mathbb{E}\left[\left|[\eta_n^N - \Phi_n(\eta_{n-1}^N))](h)\right|^p\right]^{1/p} \leqslant ||h||B_p.$$

We may apply Minkowski's inequality to obtain:

$$\begin{aligned}\sqrt{N}\mathbb{E}\left[\left|[\eta_n^N - \eta_n](h)\right|^p\right]^{1/p} \leqslant\ & \sqrt{N}\mathbb{E}\left[\left|[\eta_n^N - \Phi_n(\eta_{n-1}^N))](h)\right|^p\right]^{1/p} + \\ & \sqrt{N}\mathbb{E}\left[\left|[\Phi_n(\eta_{n-1}^N) - \Phi_n(\eta_{n-1}))](h)\right|^p\right]^{1/p}. \end{aligned} \tag{A.5}$$

To deal with the second term on the RHS of (A.5), we use Lemma 2.2 of Del Moral & Miclo (2000), that is:

$$\begin{aligned}\Phi_n(\eta_{n-1}^N)(h) - \Phi_n(\eta_{n-1}))](h) =\ & \frac{1}{\eta_{n-1}(W_{n-1})}\Big(\big[\eta_{n-1}^N(Q_n(h)) - \eta_{n-1}(Q_n(h))\big] + \\ & \Phi_n(\eta_{n-1}^N)(h)\big[\eta_{n-1}(W_{n-1}) - \eta_{n-1}^N(W_{n-1})\big]\Big)\end{aligned}$$

(recall $Q_n(x_{n-1}, dx_n) = W_{n-1}(x_{n-1})M_n(x_{n-1}, dx_n)$) thus we have

$$\sqrt{N}\mathbb{E}\left[\left|[\Phi_n(\eta_{n-1}^N) - \Phi_n(\eta_{n-1}))](h)\right|^p\right]^{1/p} \leqslant \frac{2C_{m,n-1}^{(p)} \sum_{j=1}^{l} |c_j| \prod_{i=1}^{m} \left(||h_{ij}||||W_{i,n-1}||\right)}{\eta_{n-1}(W_{n-1})}$$

where we have used the fact that the potential functions are of product form (wrt particle systems) as are the semigroups Q_n and the induction hypothesis. As a result, we have:

$$\sqrt{N}\mathbb{E}\left[\left|[\eta_n^N - \eta_n](h)\right|^p\right]^{1/p} \leqslant C_{m,n}^{(p)} \sum_{j=1}^{l} |c_j| \prod_{i=1}^{m} ||h_{ij}||$$

which clearly ends the proof of the first part of the Lemma.

The second part of the Lemma is proved by considering the sets $A_N = \{\left|[\eta_{t+1}^N - \eta_{t+1}](h)\right|^p > (\frac{1}{N})^{p/2-1-\varepsilon}\}$ $(N \geqslant 1, \varepsilon > 0, p > 2(1+\varepsilon))$ and applying the first Borel-Cantelli Lemma. □

Proof of Proposition 3.3. We begin by proving that for any $n^* \leqslant n \leqslant t$ that $\eta_n(Q_{n,t+1}(\bar{h})) = 0$ with $\bar{h} = h - \eta_{t+1}(h)$; this will be useful later in the proof. Since $\gamma_n(1) = 1$ we note that

$$\begin{aligned}
\eta_n(Q_{n,t+1})(\bar{h})) &= \int_{E_{0:t+1}} \left[\prod_{q=0}^{t} W_q(x_q)\right](h(x_{t+1}) - \eta_{t+1}(h))\mathbb{P}_{\eta_0}(dx_{0:t+1}) \\
&= \eta_{t+1}(h)[1 - \int_{E_{0:t}} \left[\prod_{q=0}^{t} W_q(x_q)\right]\mathbb{P}_{\eta_0}(dx_{0:t})] \\
&= 0.
\end{aligned}$$

Our proof is constructed by proving a central limit theorem for the product measure $\sqrt{N}\eta^N_{n^*}(Q_{n^*,t+1}(\bar{h}))$ and then using this result as part of an inductive proof on the time parameter. We prove the result for $d = 1$ only as extension to larger d may be achieved via the Cramér-Wold device.

Let $n = n^*$ and arbitrary $f_i, g_i \in \mathcal{B}_b(T_{i,n-1} \times T_{i,n})$, $h_i = (f_i, g_i)$ as a simple corollory to Proposition 9.4.2 in Del Moral (2004) we have for each $i = 1, \ldots, m$:

$$\sqrt{N}\big(\eta^N_{i,n}(h_i) - \eta^N_{i,n}(h_i)\big) \;\Rightarrow\; \mathcal{N}_2(0, \Theta_{i,n}(h_i))$$

note that we have used the fact that $\gamma_{i,n}(1) = 1 \; \forall i, n$.

Since we are interested in a CLT for

$$\sqrt{N}\Big(\frac{1}{m}\sum_{i=1}^{m} \eta^N_{i,n}(Q_{i,(n,t+1)}(f)) \prod_{j=1, j\neq i}^{m} \eta^N_{j,n}(Q_{j,(n,t+1)}(1)) - \eta_{t+1}(h)\prod_{i=1}^{m}\eta_{i,n}(Q_{i,(n,t+1)}(1))\Big)$$

we consider the $\delta-$method with function:

$$F_{2m}(u_{11}, u_{12}, \ldots, u_{m1}, u_{m2}) \;=\; \frac{1}{m}\sum_{j=1}^{m} u_{j1} \prod_{i=1, i\neq j}^{m} u_{i2} - \eta_{t+1}(h)\prod_{i=1}^{m} u_{i2}.$$

As the particle systems are independent, application of the $\delta-$method results in

$$\sqrt{N}\eta^N_n(Q_{n,t+1}(\bar{h})) \;\Rightarrow\; \mathcal{N}\big(0, \sum_{i=1}^{m}\theta^2_{i,n}(h)\big) \tag{A.6}$$

where $\theta^2_{i,n}(h) = (\alpha_i, \beta_i)'\Theta_{i,n}(Q_{i,(n,t+1)}(f, 1))(\alpha_i, \beta_i)$ and we have dropped the superscripts for notational convenience.

Let $n = n^* + 1$, in order to prove a CLT for $\sqrt{N}\eta_n^N(Q_{n,t+1}(\bar{h}))$ we follow the approach of Chopin (2004) Lemma A.1 and consider the characteristic function:

$$\mathbb{E}\Big[\exp\{it\sqrt{N}[\eta_n^N(Q_{n,t+1}(\bar{h})]\Big] = \mathbb{E}\Big[\mathbb{E}\big[\exp\{it\sqrt{N}[\eta_n^N(Q_{n,t+1}(\bar{h})) - \Phi_n(\eta_{n-1}^N)(Q_{n,t+1}(\bar{h}))]\}\big|\mathcal{F}_{n-1}^N\big] \times \exp\{it\sqrt{N}[\Phi_n(\eta_{n-1}^N)(Q_{n,t+1}(\bar{h}))]\}\Big].$$

where we have denoted the $\sigma-$algebra generated by the particles at time $n-1$ as $\mathcal{F}_{n-1}^N$.

Consider

$$\sqrt{N}(\Phi_n(\eta_{n-1}^N)(Q_{n,t+1}(\bar{h}))) = \frac{\sqrt{N}\eta_{n-1}^N(W_{n-1}M_n(Q_{n,t+1}(\bar{h})))}{\eta_{n-1}^N(W_{n-1})}.$$

By assumption $W_{n-1}(x_{1:m,n}) > 0$ and since $g(y) = 1/y$ is continuous for $y \in \mathbb{R}^+$ we have by Lemma 3.2:

$$\frac{1}{\eta_{n-1}^N(W_{n-1})} \longrightarrow_p 1$$

where $\rightarrow_p$ denotes convergence in probability. By the result (A.6) and using the corollary to Slutsky's theorem we obtain:

$$\frac{\sqrt{N}\eta_{n-1}^N(W_{n-1}M_n(Q_{n,t+1}(\bar{h})))}{\eta_{n-1}^N(W_{n-1})} \Rightarrow \mathcal{N}\big(0, \sum_{i=1}^m \theta_{i,n-1}^2(h)\big)$$

where we have used the semigroup property of $Q_{n,t+1}$.

Additionally, conditional on $\mathcal{F}_{n-1}^N$, we have that

$$\frac{1}{\sqrt{N}}(Q_{n,t+1}(\bar{h})(\xi_{1:m,n}^j) - \Phi_n(\eta_{n-1}^N)(Q_{n,t+1}(\bar{h})))$$

forms a triangular array, with N i.i.d elements and satisfies the Lindeberg condition (see Del Moral (2004) p. 294) thus applying the CLT for triangular arrays we yield:

$$\sqrt{N}\big(\eta_n^N(Q_{n,t+1}(\bar{h})) - \Phi_n(\eta_{n-1}^N)(Q_{n,t+1}(\bar{h}))\big) \Rightarrow \mathcal{N}\big(0, \eta_n(Q_{n,t+1}(\bar{h})^2)\big)$$

where we have used

$$\lim_{N\to\infty} \mathbb{E}[(Q_{n,t+1}(\bar{h}) - \Phi_n(\eta_{n-1}^N)[Q_{n,t+1}(\bar{h})])^2|\mathcal{F}_{n-1}^N] \longrightarrow_{a.s} \eta_n\big((Q_{n,t+1}(\bar{h}))^2\big)$$

via Lemma 3.2 and clearly $Q_{n,t+1}(\bar{h}) \in \mathcal{S}_p(E_n)$ (where $\rightarrow_{a.s}$ denotes almost sure convergence). Application of Theorem 25.12 in Billingsley (1995) yields:

$$\lim_{N\to\infty} \mathbb{E}\Big[\exp\{it\sqrt{N}(\eta_n^N(h) - \eta_n(h))\}\Big] = \exp\big\{-t^2\big[\eta_n(Q_{n,t+1}(\bar{h})^2) + \sum_{i=1}^m \theta_{i,n-1}(h)\big]/2\big\}$$

and thus $\sqrt{N}\eta_n^N(Q_{n,t+1}(\bar{h})) \Rightarrow \mathcal{N}\big(0, \eta_n(Q_{n,t+1}(\bar{h})^2) + \sum_{i=1}^m \theta_{i,n-1}(h)\big)$.

We now propose the induction hypothesis:

$$\sqrt{N}\eta_n^N(Q_{n,t+1}(\bar{h})) \quad \Rightarrow \quad \mathcal{N}(0, \Theta_n(h)). \tag{A.7}$$

for $n^* + 2 \leqslant n \leqslant t+1$ and $f \in \mathcal{S}_s(E_n)$.

Following the above arguments for characteristic functions and applying the induction hypothesis (A.7) yields that the asymptotic variance is:

$$\Theta_n(h) \quad = \quad \Theta_{n-1}(h) + \eta_n(Q_{n,t+1}(\bar{h})^2)$$

which completes the proof. □

Proof of Theorem 3.4. We first consider $n = 0$ and assume $2 \leqslant q \leqslant N$ (the proof for $q = 1$ is trivial) with $f \in \mathcal{B}_b(E_0^q)$ and note that by Proposition 8.6.1 of Del Moral (2004):

$$|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f)\big]| \quad = \quad |\mathbb{E}\big[(\prod_{i=1}^m r^{\odot q}(\xi_i^{(N)}) R_{N_i}^{(q)})(f) - \eta_0^{\otimes q}(f)\big]|$$

where

$$\begin{aligned} r^{\odot q}(\xi_i^{(N)}) &= \frac{1}{(N)_q} \sum_{\alpha \in \langle q, N \rangle} \delta_{\xi_i^{\alpha(1)}, \ldots, \xi_i^{\alpha(q)}} \\ R_{N_i}^{(q)} &= \frac{(N)_q}{N^q} Id_i + \left(1 - \frac{(N)_q}{N^q}\right) \widetilde{R}_{N_i}^{(q)} \end{aligned}$$

where $(N)_q = N!/(N-q)!$, $\langle q, N \rangle$ is the set of all one-to-one mappings of $\{1, \ldots, q\}$ into $\{1, \ldots, N\}$ and α is such a mapping, $\widetilde{R}_{N_i}^{(q)} : T_{i,0}^{2q} \times \mathcal{T}_{i,0}^{2q} \to [0,1]$ is a Markov kernel (see Del Moral (2004) p. 289 for details) and Id_i is the identity operator on $T_{i,0}^{2q}$.

Now we have that for any $f \in \mathcal{B}_b(E_0^q)$:

$$\mathbb{E}[(\prod_{i=1}^m r^{\odot q}(\xi_i^{(N)}))(f)] \quad = \quad \eta_0^{\otimes q}(f)$$

since the product measure in the expectation operates on seperate parts of the function. Therefore:

$$|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f)\big]| \quad = \quad |\eta_0^{\otimes q}\big((\prod_{i=1}^m R_{N_i}^{(q)} - Id)(f)\big)|$$

where Id is the identity operator on E_0^q. Note that:

$$\begin{aligned}\prod_{i=1}^{m} R_{N_i}^{(q)} - Id &= \prod_{i=1}^{m}\left(\frac{(N)_q}{N^q}Id_i + \left(1-\frac{(N)_q}{N^q}\right)\widetilde{R}_{N_i}^{(q)}\right) - Id \\ &= \left[\prod_{i=1}^{m}\left(\frac{(N)_q}{N^q}Id_i + \left(1-\frac{(N)_q}{N^q}\right)\widetilde{R}_{N_i}^{(q)}\right) - \left(\frac{(N)_q}{N^q}\right)^m Id\right] - \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)Id \\ &= \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)\left[\widetilde{R}_N^{(q)} - Id\right]\end{aligned}$$

where $\widetilde{R}_N^{(q)}$ is a Markov kernel:

$$\widetilde{R}_N^{(q)} = \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)^{-1}\left[\prod_{i=1}^{m}\left(\frac{(N)_q}{N^q}Id_i + \left(1-\frac{(N)_q}{N^q}\right)\widetilde{R}_{N_i}^{(q)}\right) - \left(\frac{(N)_q}{N^q}\right)^m Id\right].$$

As a result:

$$\begin{aligned}|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f)\big]| &= \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)|\eta_0^{\otimes q}\big((\widetilde{R}_N^{(q)} - Id)(f)\big)| \\ &\leqslant \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)\eta_0^{\otimes q}\big(\widetilde{R}_N^{(q)}(|f - \eta_0^{\otimes q}(f)|)\big) \\ &\leqslant \left(1-\left(\frac{(N)_q}{N^q}\right)^m\right)\mathrm{osc}(f).\end{aligned}$$

Since $(1-(N)_q/N^q) \leqslant q^2/N$, elementary manipulations yield:

$$|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f)\big]| \leqslant \frac{2mq^2||f||}{N}. \tag{A.8}$$

We will also require a bound on $\mathbb{E}\big[((\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f))^2\big]$. This is obtained via:

$$\begin{aligned}\mathbb{E}\big[((\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f))^2\big] &= \mathbb{E}\big[(\eta_0^N)^{\otimes q}(f)^2 - \eta_0^{\otimes q}(f)^2\big] - \\ &\qquad 2\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f)\eta_0^{\otimes q}(f) - \eta_0^{\otimes q}(f)^2\big].\end{aligned}$$

Now

$$\begin{aligned}|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f)^2 - \eta_0^{\otimes q}(f)^2\big]| &= |\mathbb{E}\big[(\eta_0^N)^{\otimes (q,2)}(f^{(2)}) - \eta_0^{\otimes (q,2)}(f^{(2)})\big]| \\ &\leqslant \frac{2m(2q)^2||f||^2}{N}\end{aligned}$$

where $\eta^{\otimes(q,2)} = \eta^{\otimes q} \times \eta^{\otimes q}$, $f^{(2)} = f \otimes f$ and we have used the result (A.8) (note the inequality holds for any q as $|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f) - \eta_0^{\otimes q}(f)\big]| \leqslant \mathrm{osc}(f)$). Thus

$$|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(f)^2 - \eta_0^{\otimes q}(f)^2\big]| \leqslant \frac{6m(2q)^2||f||^2}{N}. \tag{A.9}$$

Now consider the time point $n = 1$, $f \in \mathcal{B}_b(E_1^q)$

$$\begin{aligned} |\mathbb{E}\big[(\eta_1^N)^{\otimes q}(f) - \eta_1^{\otimes q}(f)\big]| &= |I_1 + I_2| \\ I_1 &= \mathbb{E}\big[(\eta_1^N)^{\otimes q}(f) - \Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f)\big] \\ I_2 &= \mathbb{E}\big[\Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f) - \eta_1^{\otimes q}(f)\big] \end{aligned}$$

where $\Phi_n^{(q)} : \mathcal{P}(E_{n-1}^q) \to \mathcal{P}(E_n^q)$ is the semigroup:

$$\Phi_n^{(q)}(\mu)(f) = \frac{\mu(Q_n^{(q)}(f))}{\mu(Q_n^{(q)}(1))}$$

where $\mu \in \mathcal{P}(E_{n-1}^q)$.

Firstly, considering I_1:

$$\begin{aligned} |I_1| &\leqslant \mathbb{E}\big[|\mathbb{E}\big[(\eta_1^N)^{\otimes q}(f) - \Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f)|\mathcal{F}_0^N\big]|\big] \\ &\leqslant \frac{2mq^2||f||}{N} \end{aligned}$$

recall $\mathcal{F}_0^N$ is the $\sigma-$algebra generated by $\xi_0^{(N)}$ and we can argue in an analogous manner to (A.8).

Secondly, for I_2, applying Lemma 2.2 of Del Moral & Miclo (2000) we have:

$$\begin{aligned} |I_2| \leqslant \frac{1}{\eta_0^{\otimes q}(W_0^{(q)})}\Big[&|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(Q_1^{(q)}(f)) - \eta_0^{\otimes q}(Q_1^{(q)}(f))\big]| + \\ &|\mathbb{E}\big[\Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f)\big(\eta_0^{\otimes q}(W_0^{(q)}) - (\eta_0^N)^{\otimes q}(W_0^{(q)})\big)\big]|\Big]. \end{aligned}$$

By (A.8) we obtain:

$$|\mathbb{E}\big[(\eta_0^N)^{\otimes q}(Q_1^{(q)}(f)) - \eta_0^{\otimes q}(Q_1^{(q)}(f))\big]| \leqslant D_{m,0}^{(1)}(q, N, ||Q_1^{(q)}(f)||)$$

where we have denoted the bound on $|\mathbb{E}\big[\big((\eta_n^N)^{\otimes q}(f) - \eta_n^{\otimes q}(f)\big)^p\big]|$ as $D_{m,n}^{(p)}(q, N, ||f||)$. Applying Hölder's inequality we have:

$$\begin{aligned} |\mathbb{E}\big[\Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f)\big(\eta_0^{\otimes q}(W_0^{(q)}) - (\eta_0^N)^{\otimes q}(W_0^{(q)})\big)\big]| &\leqslant \mathbb{E}\big[\Phi_1^{(q)}((\eta_0^N)^{\otimes q})(f)^2\big]^{1/2} \times \\ &\quad \mathbb{E}\big[\big(\eta_0^{\otimes q}(W_0^{(q)}) - (\eta_0^N)^{\otimes q}(W_0^{(q)})\big)^2\big]^{1/2} \\ &\leqslant ||f||\mathbb{E}\big[\big(\eta_0^{\otimes q}(W_0^{(q)}) - (\eta_0^N)^{\otimes q}(W_0^{(q)})\big)^2\big]^{1/2}. \end{aligned}$$

We thus have:

$$D^{(1)}_{m,1}(q,N,||f||) = \frac{2mq^2||f||}{N} + \frac{1}{\eta^{\otimes q}(W_0^{(q)})}\big[D^{(1)}_{m,0}(q,N,||Q_1^{(q)}(f)||) + ||f||\big(D^{(2)}_{m,0}(q,N,||W_0^{(q)}||)\big)^{1/2}\big]$$

and arguing as for (A.9)

$$D^{(2)}_{m,1}(q,N,||f||) = D^{(1)}_{m,1}(2q,N,||f||^2) + 2||f||D^{(1)}_{m,1}(q,N,||f||).$$

Since the proof used for $n=1$ will apply for any $1 \leqslant n \leqslant n^*$ we thus define the recursion for $D^{(1)}_{m,n}$, $D^{(2)}_{m,n}$ in the statement of the proof (i.e. by induction).

To obtain the bound on the total variation distance, we note:

$$||\mathbb{P}^{(N,q)}_{\eta_0,[t+1]} - \eta^{\otimes q}_{t+1}||_{TV} = \sup_{f:E^q_{t+1}\to[0,1]} |\mathbb{P}^{(N,q)}_{\eta_0,[t+1]}(f) - \eta^{\otimes q}_{t+1}(f)|.$$

Now

$$\begin{aligned}|\mathbb{P}^{(N,q)}_{\eta_0,[t+1]}(f) - \eta^{\otimes q}_{t+1}(f)| &\leqslant |\mathbb{P}^{(N,q)}_{\eta_0,[t+1]}(f) - \mathbb{E}\big[(\eta^N_{t+1})^{\otimes q}(f)\big]| + |\mathbb{E}\big[(\eta^N_{t+1})^{\otimes q}(f)\big] - \eta^{\otimes q}_{t+1}(f)| \\ &\leqslant \frac{(q-1)^2}{N} + |\mathbb{E}\big[(\eta^N_{t+1})^{\otimes q}(f)\big] - \eta^{\otimes q}_{t+1}(f)|\end{aligned}$$

where we have used part of the proof of Theorem 8.3.3 of Del Moral (2004).

In order to complete the proof, note that since $m=1$ when $n = n^*+1$ (immediately after mutation) we have

$$\begin{aligned}D^{(1)}_{1,n^*+1}(q,N,||f||) &= \frac{2q^2}{N} + \frac{1}{\eta^{\otimes q}_{n^*}(W^{(q)}_{n^*}))}\big[D^{(1)}_{m,n^*}(q,N,||Q^{(q)}_{n^*}(f)||) + ||f||\big(D^{(2)}_{m,n^*}(q,N,||W^{(q)}_{n^*}||)\big)^{1/2}\big] \\ D^{(2)}_{1,n^*+1}(q,N,||f||) &= D^{(1)}_{1,n^*+1}(2q,N,||f||^2) + 2||f||D^{(1)}_{1,n^*+1}(q,N,||f||)\end{aligned}$$

and similar definitions for recursions (i.e. as in the statement of the proof) when $n \geqslant n^*+1$. Thus we have that:

$$\begin{aligned}|\mathbb{E}\big[(\eta^N_{t+1})^{\otimes q}(f)\big] - \eta^{\otimes q}_{t+1}(f)| &\leqslant D^{(1)}_{1,t+1}(q,N,||f||) \\ &\leqslant \frac{2q^2}{N} + \frac{1}{\eta^{\otimes q}_t(W^{(q)}_t)}\big[D^{(1)}_{1,t}(q,N,||W^{(q)}_t||) + \big(D^{(2)}_{1,t}(q,N,||W^{(q)}_t||)\big)^{1/2}\big]\end{aligned}$$

where we have used $||f|| \leqslant 1$ and that $D^{(1)}_{1,t+1}$ is non-decreasing function in terms of the norm.

Due to the above arguments:

$$|\mathbb{P}^{(N,q)}_{\eta_0,[t+1]}(f) - \eta^{\otimes q}_{t+1}(f)| \leqslant \frac{(q-1)^2}{N} + \frac{2q^2}{N} + \frac{1}{\eta^{\otimes q}_t(W^{(q)}_t)}\big[D^{(1)}_{1,t}(q,N,||W^{(q)}_t||) + \big(D^{(2)}_{1,t}(q,N,||W^{(q)}_t||)\big)^{1/2}\big]$$

since the construction is true for any $f \in \mathcal{B}_b(E^q_{t+1})$ with $||f|| \leqslant 1$ we have:

$$||\mathbb{P}^{(N,q)}_{\eta_0,[t+1]} - \eta^{\otimes q}_{t+1}||_{TV} \leqslant \frac{(q-1)^2}{N} + \frac{2q^2}{N} + \frac{1}{\eta^{\otimes q}_t(W^{(q)}_t)}\big[D^{(1)}_{1,t}(q,N,||W^{(q)}_t||) + \big(D^{(2)}_{1,t}(q,N,||W^{(q)}_t||)\big)^{1/2}\big]$$

as required. □

REFERENCES

Andrieu, C. & Moulines E. (2003). On the ergodicity properties of some adaptive MCMC algorithms, Technical Report, University of Bristol.

Andrieu, C. & Robert C. P. (2001). Controlled MCMC for optimal sampling, Technical Report, Universitié Paris Dauphine.

Atachade, Y. F. & Liu, J. S. (2004). The Wang-Landau algorithm for Monte Carlo computation in general state spaces, Technical Report, University of Ottawa.

Billingsley, P. (1995). *Probability and Measure*. Third edition. Wiley: New York.

Cappé, O., Gullin, A., Marin, J. M. & Robert, C. P. (2004). Population Monte Carlo. *J. Comp. Graph. Statist.*, **13**, 907-925.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, **89**, 539-552.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, **32**, 2385-2411.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

Del Moral, P. & Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. In *Séminaire de Probabilités XXXVII*, Ed. Azéma, J., Emery, M., Ledoux, M. and Yor, M., *Lecture Notes in Math.* **1832**, 415-446. Springer: Berlin.

Del Moral, P. & Guionnet, A. (1999). Central limit theorem for non-linear filtering and interacting particle systems. *Ann. Appl. Prob.*, **9**, 275-297.

Del Moral, P. & Miclo, L. (2000). Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. *Séminaire de Probabilitiés XXXIV*, Ed. Azéma, J., Emery, M., Ledoux, M. and Yor, M., *Lecture notes in Math.* **1729** 1–145. Springer: Berlin.

Del Moral, P. & Miclo, L. (2001). Genealogies and increasing propagation of chaos for Feynman-Kac and genetic models. *Ann. Appl. Prob.*, **11**, 1166-1198.

Del Moral, P., Doucet, A. & Jasra, A. (2005a). Sequential Monte Carlo samplers, Technical Report (under revision), University of Cambridge.

Del Moral, P., Doucet, A. & Peters, G. W. (2005b). Asymptotic and increasing propagation of chaos properties for genealogical particle models, Technical Report, Universitié Paul Sabatier.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Chichester: Wiley.

Doucet, A., De Freitas, J. F. G. & Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice.* Springer: New York.

Figuieredo, M. A. T. & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Patt. Analy. Mach. Intell.* **24**, 381–96.

Gilks, W. R. & Berzuini, C. (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. B* **63**, 127–46.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P. J. (2003b). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, (P.J. Green, N.L. Hjort & S. Richardson eds), 179-96 Oxford: Oxford University Press.

Hastie, D. (2005). *Towards Automatic Reversible Jump Markov chain Monte Carlo*, PhD thesis, University of Bristol.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Jasra, A., Stephens, D. A., & Holmes, C. C. (2005a). Population-based reversible jump Markov chain Monte Carlo, Technical Report, Imperial College London.

Jasra, A., Stephens, D. A., & Holmes, C. C. (2005b). On population-based simulation for static inference, Technical Report, Imperial College London.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005c). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling, *Statist. Sci.* **20**, 50–67.

Kou, S. C, Zhou, Q., & Wong, W. H. (2005). Equi-energy sampler with applications to statistical inference and statistical mechanics, *Ann. Statist.* (in press).

Künsch, H. R. (2005). Recursive Monte Carlo filters; algorithms and theoretical analysis, *Ann. Statist.* (in press).

Liang, F. & Wong, W. H. (2001). Real parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Am. Statist. Assoc.* **96**, 653–666.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer: New York.

Liu, J. S. & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems, *J. Amer. Statist. Assoc.*, **93**, 1032-1044.

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models.* Wiley: Chichester.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–92.

Neal, R. M. (2001). Annealed importance sampling. *Statist. and Comp.* **11**, 125–39.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2001). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixture models with an unknown number of components (with Discussion). *J. R. Statist. Soc. B* **59**, 731–792.

Robert, C. P. & Casella G. (2004). *Monte Carlo Statistical Methods.* Second edition. Springer: New York.

Shiryaev, A. N. (1996). *Probability.* Springer: New York.

Wang, F., & Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050–2053.

Algorithm 2.1 Interacting Sequential Monte Carlo Sampler.

1. (Initialization)

- Set $n = 0$.
- For $i = 1, \ldots, m$, $j = 1, \ldots, N$ draw $\xi_{i,0}^j \sim \eta_{i,0}$.
- Set

$$w_{i,0}(\xi_{i,0}^j) \propto \frac{\pi_{i,0}(\xi_{i,0}^j)}{\eta_{i,0}(\xi_{i,0}^j)}.$$

Iterate steps 2. and 3.

2.(Selection)

- For $i = 1, \ldots, m$.
- If $\left(\left[\sum_j w_{i,n}(\xi_{0:n}^j)^2 \right] / \left[\sum_l w_{i,n}(\xi_{0:n}^l) \right]^2 \right)^{-1} < L$ (for some threshold L), resample the particles for sampler i and set all weights equal to 1.

3.(Mutation)

- Set $n = n + 1$, if $n = n^*$ go to 4.
- For $i = 1, \ldots, m$, $j = 1, \ldots, N$ draw $\xi_{i,n}^j \sim K_{i,n}(\xi_{i,n-1}^j, \cdot)$.
- Reweight

$$\begin{aligned} W_{i,n}(\xi_{i,(0:n)}^j) &= \frac{\nu_{i,n}(\xi_{i,0:n}^j)}{\nu_{i,n-1}(\xi_{i,0:n-1}^j) K_{i,n}(\xi_{i,n-1}^j, \xi_{i,n}^j)} \\ w_{i,n}(\xi_{i,0:n}^j) &\propto w_{i,n-1}(\xi_{i,0:n-1}^j) W_{i,n}(\xi_{i,0:n}^j). \end{aligned}$$

4.(Single Sampler)

- Sample all particles with the same Markov kernel, reweight and then resample (for each sampler). Sample all particles from kernel K_{n^*+1}.
- Form new particles $\xi_{n^*+1}^j = (\xi_{1,n^*+1}^j, \ldots, \xi_{m,n^*+1}^j)$.
- Continue with a single SMC sampler targeting the appropriate densities.