On particle Gibbs sampling

Nicolas Chopin, Sumeetpal Singh

Bernoulli **21**(3), 2015, 1855–1883
DOI: 10.3150/14-BEJ629

On particle Gibbs sampling

NICOLAS CHOPIN[1] and SUMEETPAL S. SINGH[2]

[1]*CREST-ENSAE and HEC Paris, 3 Avenue Pierre Larousse, 92235 Malakoff, France. E-mail: nicolas.chopin@ensae.fr*
[2]*Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK. E-mail: sss40@eng.cam.ac.uk*

The particle Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm to sample from the full posterior distribution of a state-space model. It does so by executing Gibbs sampling steps on an extended target distribution defined on the space of the auxiliary variables generated by an interacting particle system. This paper makes the following contributions to the theoretical study of this algorithm. Firstly, we present a coupling construction between two particle Gibbs updates from different starting points and we show that the coupling probability may be made arbitrarily close to one by increasing the number of particles. We obtain as a direct corollary that the particle Gibbs kernel is uniformly ergodic. Secondly, we show how the inclusion of an additional Gibbs sampling step that reselects the ancestors of the particle Gibbs' extended target distribution, which is a popular approach in practice to improve mixing, does indeed yield a theoretically more efficient algorithm as measured by the asymptotic variance. Thirdly, we extend particle Gibbs to work with lower variance resampling schemes. A detailed numerical study is provided to demonstrate the efficiency of particle Gibbs and the proposed variants.

Keywords: Feynman–Kac formulae; Gibbs sampling; particle filtering; particle Markov chain Monte Carlo; sequential Monte Carlo

1. Introduction

PMCMC (particle Markov chain Monte Carlo [1]) is a new set of MCMC algorithms devised for inference in state-space models which has attracted considerable attention in statistics. It has in a short time triggered intense scientific activity spanning methodological [5, 15, 23, 27] and applied work, the latter in domains as diverse as ecology [20], electricity forecasting [14], finance [21], systems biology [11], social networks [10] and hydrology [25]. One appeal of PMCMC is that it makes it possible to perform "plug-and-play" inference for complex hidden Markov models, that is, the only requirement is that one needs to be able to sample from the Markov transition of the hidden chain, which is in most cases non-demanding, in contrast to previous approaches based on standard MCMC.

1350-7265

Each PMCMC step generates an interacting particle system; see [8, 9] and [3] for general references on particle algorithms (also known as Sequential Monte Carlo algorithms). Several instances of PMCMC may be analysed as *exact Monte Carlo approximations* of an ideal algorithm, that is, as a noisy version of an ideal algorithm where some intractable quantity is replaced by an unbiased Monte Carlo estimate (computed from the interacting particle system). Such algorithms are analysed in detail in [2]. The term 'exact' in the phrase 'exact Monte Carlo' highlights the fact that, despite being an approximation of an ideal algorithm, PMCMC samples exactly from the distribution of interest.

However, this interpretation does not seem applicable to variants of PMCMC involving a particle Gibbs step. While particle Gibbs also generates a complete interacting particle system at each iteration, it does so conditionally on the trajectory for one particle being fixed, and it does not replace an intractable quantity of an ideal algorithm with an unbiased estimator.

The objective of this paper is to undertake a theoretical study of particle Gibbs to try to support its very favourable performance observed in practice. For this, we design a coupling construction between two particle Gibbs updates that start from different trajectories and establish that the coupling probability may be made arbitrarily large by increasing the number of particles N. As a direct corollary, we conclude that the transition kernel of particle Gibbs is uniformly ergodic (under suitable conditions). This strong result supports why particle Gibbs can be expected, and does indeed, perform so well in practice. Our coupling construction is maximal for some special cases and appears unique in the literature on particle systems.

Secondly, we show how the inclusion of an additional backward sampling step that reselects the ancestors of the particle Gibbs' extended target distribution, first proposed by [26] and now a popular approach in practice to improve mixing [16], does indeed yield a theoretically more efficient algorithm as measured by the asymptotic variance of the central limit theorem. Thirdly, and as another way to enhance mixing, we extend the original particle Gibbs sampler (which is based on the multinomial resampling scheme as presented in the original paper of [1]) to work with lower variance residual or systematic resampling schemes. This variety of implementation of particle Gibbs raises an obvious question: which variant performs best in practice? We present numerical comparisons in a particular example, which suggests that the backward sampling strongly improves the mixing of particle Gibbs, and, when it cannot be implemented, then residual and systematic resampling leads to significantly better mixing than multinomial resampling.

The plan of the paper is the following. Section 2 sets up the notation and defines the particle Gibbs algorithm. This section reviews the original particle Gibbs algorithm of [1] and presents a reinterpretation of particle Gibbs as a Markov kernel to facilitate the analysis to follow in the later sections. Some supporting technical results are also presented. Section 3 proves that the particle Gibbs kernel is uniformly ergodic. To that effect, a coupling construction is obtained such that the coupling probability between two particle Gibbs updates may be made arbitrarily large for N large enough. Section 4 discusses the backward sampling step proposed by [26], and establishes dominance of particle Gibbs with this backward sampling step over the version without. Section 5 discusses how to extend particle Gibbs to alternative resampling schemes. Section 6

presents a numerical comparison of the variants of particle Gibbs discussed in the previous sections. Section 7 concludes.

2. Definition of the particle Gibbs sampler

2.1. Notation

For $m \leq n$, we denote by $m{:}n$ the range of integers $\{m, \dots, n\}$, and we use extensively the semicolon short-hand for collections of random variables, for example, $X_{0:T} = (X_0, \dots, X_T)$, $X_t^{1:N} = (X_t^1, \dots, X_t^N)$, and even in an nested form, $X_{0:T}^{1:N} = (X_0^{1:N}, \dots, X_T^{1:N})$; more generally X_t^v, where v is a vector in $\mathbb{N}^+$ will refer to the collection $(X_t^n)_{n \in v}$. These short-hands are also used for realisations of these random variables, which are in lower case, for example, $x_{0:t}$ or $x_t^{1:N}$. The sub-vector containing the t first components of some vector Z_T is denoted by $[Z_T]_t$.

For a vector $r^{1:N}$ of probabilities, $r^n \in [0,1]$ and $\sum_{n=1}^N r^n = 1$, we denote by $\mathcal{M}(r^{1:N})$ the multinomial distribution which produces outcome n with probability r^n, $n \in 1{:}N$. For reals x, y, let $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$. The integer part of x is $\lfloor x \rfloor$, and the positive part is $x^+ = x \vee 0$. The cardinal of a finite set $\mathcal{C}$ is denoted as $|\mathcal{C}|$.

For a complete separable metric space $\mathcal{X}$, we denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$. For a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, a kernel $K : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ and a measurable function f defined on $\mathcal{X}$, we use the following standard notation: $\mu(f) = \int_{\mathcal{X}} \mathrm{d}\mu f$, Kf is the application $x \to \int_{\mathcal{X}} K(x, \mathrm{d}x') f(x')$, and μK is the probability measure $(\mu K)(A) = \int_{\mathcal{X}} \mu(\mathrm{d}x) K(x, A)$. The atomic measure at $a \in \mathcal{X}$ is denoted by $\delta_a(\mathrm{d}x)$. We denote by $\mu \otimes K$ the measure $\mu(\mathrm{d}x)K(x, \mathrm{d}x')$ on the product space $\mathcal{X} \times \mathcal{X}$. Finally, we shall often use the same symbol for distributions and densities; for example, $m_0(\mathrm{d}x_0) = m_0(x_0)\,\mathrm{d}x_0$ means that the distribution $m_0(\mathrm{d}x_0)$ admits the function $x_0 \to m_0(x_0)$ as a probability density relative to some sigma-finite dominating measure $\mathrm{d}x_0$.

2.2. The target distribution

Let $\mathcal{X}$ be a complete separable metric space, and $(X_t)_{t \geq 0}$ a discrete-time $\mathcal{X}$-valued Markov chain, with initial law $m_0(\mathrm{d}x_0) = m_0(x_0)\,\mathrm{d}x_0$, and transition law $m_t(x_{t-1}, \mathrm{d}x_t) = m_t(x_{t-1}, x_t)\,\mathrm{d}x_t$, where $\mathrm{d}x_0$, $\mathrm{d}x_t$ are appropriately chosen (possibly identical) sigma-finite dominating measures. Let $(G_t)_{t \geq 0}$ be a sequence of $\mathcal{X} \to \mathbb{R}^+$ potential functions. In the context of hidden Markov models, typically $G_t(x_t) = g(x_t, y_t)$, the density (with respect to some dominating measure $\mathrm{d}y$) of observation y_t of the $\mathcal{Y}$-valued random variable Y_t, conditional on state $X_t = x_t$.

It is convenient to work directly with the path model, that is, we define $Z_t = X_{0:t}$ (and $z_t = x_{0:t}$) taking values in $\mathcal{X}^{t+1}$, and slightly abusing notation, we extend the domain of G_t from $\mathcal{X}$ to $\mathcal{X}^{t+1}$ as follows: $G_t(z_t) = G_t(x_t)$. The Z_t's form a time inhomogeneous Markov kernel, with initial law $q_0(\mathrm{d}z_0) = m_0(\mathrm{d}x_0)$, and transition

$$q_t(z_{t-1}, \mathrm{d}z_t') = \delta_{z_{t-1}}(\mathrm{d}x_{0:t-1}') m_t(x_{t-1}, x_t')\,\mathrm{d}x_t'$$

that is, keep all of z_{t-1} and append new state x_t, from Markov transition $m_t(x_{t-1}, \mathrm{d}x_t)$. The associated (Feynman–Kac) path measures are

$$\mathbb{Q}_t(\mathrm{d}z_t) = \mathbb{Q}_t(\mathrm{d}x_{0:t}) = \frac{1}{\mathcal{Z}_t} G_0(x_0) m_0(\mathrm{d}x_0) \prod_{s=1}^{t} \{G_s(x_s) m_s(x_{s-1}, \mathrm{d}x_s)\}, \tag{1}$$

where $\mathcal{Z}_t$ is defined as

$$\mathcal{Z}_t = \int_{\mathcal{X}^{t+1}} G_0(x_0) m_0(\mathrm{d}x_0) \prod_{s=1}^{t} \{G_s(x_s) m_s(x_{s-1}, \mathrm{d}x_s)\}$$

assuming from now on that $0 < \mathcal{Z}_t < +\infty$. The target distribution to be sampled from is $\mathbb{Q}_T(\mathrm{d}z_T)$ for some fixed T, which can also be interpreted as the full posterior of a state-space model.

The fact that we work directly with the path Z_t, and path-valued potential functions $G_t(z_t)$, reveals that our results could be extended easily to the situation where in the original formulation for X_t, the potential function depended on past values, for example, $G_t(x_{t-1}, x_t)$. In that way, one may consider, for instance, more general algorithms where particles are mutated according to a proposal kernel that may differ from the Markov kernel of the considered model. However, in the only part of the paper (Section 4) where we shall revert to the original formulation based on X_t, we will stick to the standard case where G_t depends only on x_t for the sake of clarity.

Andrieu *et al.* [1] introduced an MCMC algorithm that samples from (1) by defining an extended target distribution (which admits (1) as its marginal) and then constructing a Gibbs sampler for this extended target. In the next section, we review this construction of theirs.

2.3. The extended target and the particle Gibbs sampler

The starting point in the definition of [1]'s extended target distribution that admits (1) as its marginal is the joint distribution of all the random variables generated in the course of the execution of an (interacting) particle algorithm that targets the path measures given in (1). We refer the reader to [3, 8] for a review of particle algorithms that target Feynman–Kac path measures.

The particle representation $\mathbb{Q}_t^N(\mathrm{d}z_t)$ is the empirical measure defined as, for $t \geq 0$,

$$\mathbb{Q}_t^N(\mathrm{d}z_t) = \frac{1}{N} \sum_{n=1}^{N} \delta_{Z_t^n}(\mathrm{d}z_t),$$

where the particles $Z_t^{1:N} = (Z_t^1, \ldots, Z_t^N)$ are defined recursively as follows. First, $Z_0^{1:N}$ is obtained by sampling N times independently from $m_0(x_0)\,\mathrm{d}x_0$. To progress from time t

to time $t+1$, $t \geq 0$, the pair $(A_t^{1:N}, Z_{t+1}^{1:N})$ is generated jointly from

$$\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N}) \prod_{n=1}^{N} q_{t+1}(z_t^{a_t^n}, \mathrm{d}z_{t+1}^n),$$

conditionally on $Z_t^{1:N} = z_t^{1:N}$, where the A_t^n's, $n \in 1:N$, jointly sampled from the resampling distribution $\varrho_t(z_t^{1:N}, \mathrm{d}A_t^{1:N})$ are the ancestor variables, that is, A_t^n is the label of the particle at time t which generated particle Z_{t+1}^n at time $t+1$. (Since the A_t^n's are integer-valued, the dominating measure of kernel $\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N})$ is simply the counting measure.)

The law of the collection of random variables $(Z_{0:T}^{1:N}, A_{0:T-1}^{1:N})$ generated from time 0 to some final time $T \geq 1$ is therefore

$$\vartheta_T^N(\mathrm{d}z_{0:T}^{1:N}, \mathrm{d}a_{0:T-1}^{1:N}) = m_0^{\otimes N}(\mathrm{d}z_0^{1:N}) \prod_{t=1}^{T} \left\{ \varrho_{t-1}(z_{t-1}^{1:N}, \mathrm{d}a_{t-1}^{1:N}) \prod_{n=1}^{N} [q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n)] \right\}.$$

The simplest choice for ϱ_t is what is usually referred to as the multinomial resampling scheme, namely the A_t^n's are drawn independently from the multinomial distribution $\mathcal{M}(W_t^{1:N}(z_t^{1:N}))$, where the W_t^n's are the normalised weights

$$W_t^n(z_t^{1:N}) \triangleq \frac{G_t(z_t^n)}{\sum_{m=1}^{N} G_t(z_t^m)}, \qquad n \in 1:N. \tag{2}$$

Then $\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N})$, $t \geq 0$ equals

$$\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N}) = \left\{ \prod_{n=1}^{N} W_t^{a_t^n}(z_t^{1:N}) \right\} \mathrm{d}a_t^{1:N}. \tag{3}$$

For now, we assume this particular choice for ϱ_t, and our main results will therefore be specific to multinomial resampling. Note, however, that we will discuss alternative resampling schemes at the end of the paper; see Section 5.

We now state an intermediate result which is needed to ensure the validity of the extended target (5) below.

Proposition 1. *One has*

$$\mathbb{E}_{\vartheta_T^N} \left[\prod_{t=0}^{T} \left\{ \frac{1}{N} \sum_{n=1}^{N} G_t(Z_t^n) \right\} \right] = \mathcal{Z}_T. \tag{4}$$

See, for example, Lemma 3 in [7]. In order to state the particle Gibbs sampler and prove it leaves $\mathbb{Q}_T(\mathrm{d}z_T)$ invariant, we commence first with the definition of the following extended distribution π_T^N of [1] whose sampling space is the sampling space of the measure

ϑ_T^N augmented to include a discrete random variable $N^\star \in 1:N$,

$$\begin{aligned}
&\pi_T^N(\mathrm{d}z_{0:T}^{1:N}, \mathrm{d}a_{0:T-1}^{1:N}, \mathrm{d}n^\star) \\
&\quad = \frac{1}{\mathcal{Z}_T}\vartheta_T^N(\mathrm{d}z_{0:T}^{1:N}, \mathrm{d}a_{0:T-1}^{1:N})\left[\prod_{t=0}^{T-1}\left\{\frac{1}{N}\sum_{n=1}^{N} G_t(Z_t^n)\right\}\right]\frac{1}{N}G_T(z_T^{n^\star}) \\
&\quad = \frac{1}{\mathcal{Z}_T} m_0^{\otimes N}(\mathrm{d}z_0^{1:N}) \qquad (5) \\
&\qquad \times \prod_{t=1}^{T}\left[\left\{\frac{1}{N}\sum_{n=1}^{N} G_{t-1}(z_{t-1}^n)\right\}\prod_{n=1}^{N}\{W_{t-1}^{a_{t-1}^n}(z_{t-1}^{1:N})\,\mathrm{d}a_{t-1}^n q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n)\}\right] \\
&\qquad \times \frac{1}{N}G_T(z_T^{n^\star}),
\end{aligned}$$

again assuming (3). The fact that the expression above does define a correct probability law (with a density that integrates to one) is an immediate consequence of the unbiasedness property given in (4).

Proposition 2. *The distribution π_T^N is such that the marginal distribution of the random variable $Z_T^\star \triangleq Z_T^{N^\star}$ is $\mathbb{Q}_T$.*

This proposition is proved in [1]. To verify this result, the expectation of functions of $Z_T^\star$ may be computed by integrating out the variables in the reverse order $n^\star$, $x_T^{1:N}$, $a_T^{1:N}, \ldots, x_1^{1:N}$, $a_0^{1:N}$, $x_0^{1:N}$. We now proceed to state the Gibbs algorithm of [1].

Given a sample from π_T^N, we can trace the ancestry of the variable $Z_T^\star = Z_T^{N^\star}$ as follows. Let $B_t^\star$ for $t \in 0:T$ be the index of the time t ancestor particle of trajectory $Z_T^\star$, which is defined recursively backward as $B_T^\star = N^\star$, then $B_t^\star = A_t^{B_{t+1}^\star}$, for $t = T-1, \ldots, 0$. Finally, let $Z_t^\star = Z_t^{B_t^\star}$ for $t \in 0:T$, so that $Z_t^\star$ is precisely the first $t+1$ components of $Z_T^\star$, that is, $Z_t^\star = [Z_T^\star]_{t+1}$.

Let $Z_t^{1:N\setminus\star}$ be the ordered collection of the $N-1$ trajectories Z_t^n such that $n \neq B_t^\star$ (i.e., $n \neq N^\star$ when $t = T$), and $Z_{0:T}^{1:N\setminus\star} = (Z_0^{1:N\setminus\star}, \ldots, Z_T^{1:N\setminus\star})$. Define similarly $A_{0:T-1}^{1:N\setminus\star} = (A_0^{1:N\setminus\star}, \ldots, A_{T-1}^{1:N\setminus\star})$, where $A_t^{1:N\setminus\star}$ is $A_t^{1:N}$ excluding $A_t^{B_{t+1}^\star}$. It is convenient to apply the following one-to-one transformation to the argument of π_T^N:

$$(z_{0:T}^{1:N}, a_{0:T-1}^{1:N}, n^\star) \leftrightarrow (z_{0:T}^{1:N\setminus\star}, a_{0:T-1}^{1:N\setminus\star}, z_{0:T}^\star, b_{0:T-1}^\star, n^\star).$$

With a slight abuse of notation, we identify the law induced by this transformation (going to the representation with the $b_t^\star$ variables) as π_T^N as well:

$$\begin{aligned}
&\pi_T^N(\mathrm{d}z_{0:T}^{1:N\setminus\star}, \mathrm{d}a_{0:T-1}^{1:N\setminus\star}, \mathrm{d}z_{0:T}^\star, \mathrm{d}b_{0:T-1}^\star, \mathrm{d}n^\star) \\
&\quad = \frac{1}{N^{T+1}}(\mathrm{d}b_{0:T-1}^\star\,\mathrm{d}n^\star)\mathbb{Q}_T(\mathrm{d}z_T^\star)\prod_{t=0}^{T-1}\delta_{([z_T^\star]_{t+1})}(\mathrm{d}z_t^\star) \qquad (6)
\end{aligned}$$

$$\times \prod_{n \neq b_0^\star} m_0(\mathrm{d}z_0^n) \left[\prod_{t=1}^{T} \prod_{n \neq b_t^\star} W_{t-1}^{a_{t-1}^n}(z_{t-1}^{1:N})\, \mathrm{d}a_{t-1}^n q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n) \right].$$

Passage from (5) to (6) is straightforward. It is worth noting that the marginal law of $(B_{0:T-1}^\star, N^\star)$ is the uniform law on the product space $(1:N)^{T+1}$.

Given a sample $Z_T = z_T$ from $\mathbb{Q}_T$, consider the following three step sampling procedure that transports $Z_T = z_T$ to define a new random variable $Z_T' \in \mathcal{X}^{T+1}$. Step 1 is to sample the ancestors $(B_{0:T-1}^\star, N^\star)$ of the random variable $Z_{0:T}^\star = z_{0:T}$ from $\pi_T^N(\mathrm{d}b_{0:T-1}^\star, \mathrm{d}n^\star)$; step 2 is to generate the $N-1$ remaining trajectories $(Z_{0:T}^{1:N\setminus\star}, A_{0:T-1}^{1:N\setminus\star})$ conditional on the trajectory $(Z_{0:T}^\star, B_{0:T-1}^\star, N^\star)$ from

$$\pi_T^N(\mathrm{d}z_{0:T}^{1:N\setminus\star}, \mathrm{d}a_{0:T-1}^{1:N\setminus\star} | z_{0:T}^\star, b_{0:T-1}^\star, n^\star).$$

(There is no specific difficulty in performing step 2, details to follow, which is pretty much equivalent to the problem of generating a particle filter for T time steps.) Note that steps 1 and 2 are both Gibbs step with respect to (6). Step 3 is to resample the index $N^\star$ from (5); hence $N^\star$ is sampled from $\pi_T^N(\mathrm{d}n^\star | z_{0:T}^{1:N}, a_{0:T-1}^{1:N}) = \mathcal{M}(W_T^{1:N}(z_T^{1:N}))$, which is also a Gibbs step, but this time with respect to (5); recall that $W_T^n(z_T^{1:N}) = G_T(z_T^n)/\sum_m G_T(z_T^m)$. It follows from Proposition 2 that the law of $Z_T' = Z_T^{N^\star}$ is also $\mathbb{Q}_T$.

Steps 1 to 3 therefore define a Markov kernel P_T^N that maps $\mathcal{X}^{T+1} \to \mathcal{P}(\mathcal{X}^{T+1})$ and has $\mathbb{Q}_T(\mathrm{d}z_T)$ as its invariant measure. In practice, however, step 1 is redundant and we may as well set $(b_{0:T-1}^\star, n^\star)$ to the (arbitrary) value $(1, \ldots, 1)$ before applying steps 2 and 3, as per the following remark.

Remark 1. The image of $z_T^\star \in \mathcal{X}^{T+1}$ under P_T^N is unchanged by the choice of $(b_{0:T-1}^\star, n^\star)$ for the realization of $(B_{0:T-1}^\star, N^\star)$ in the initialization of the CPF kernel.

This remark follows from the fact that the joint distribution of $Z_T^{1:N}$ in (5) is exchangeable. On the other hand, we shall see in Section 4 that the equivalent representation of the particle Gibbs kernel as an update that involves a step that re-simulates $(B_{0:T-1}^\star, N^\star)$ will be useful to establish certain properties.

To conclude, and following [1], the CPF kernel may be defined as the succession of the following two steps, from current value $z_T^\star \in \mathcal{X}^{T+1}$.

CPF-1 Generate the $N-1$ remaining trajectories of the particle system by sampling from the conditional distribution (deduced from (6)):

$$\begin{aligned} &\pi_T^N(\mathrm{d}z_{0:T}^{2:N}, \mathrm{d}a_{0:T-1}^{2:N} | Z_{0:T}^1 = z_{0:T}^\star, A_{0:T-1}^1 = (1, \ldots, 1), N^\star = 1) \\ &\quad = m_0^{\otimes(N-1)}(\mathrm{d}z_0^{2:N}) \prod_{t=1}^{T} \left[\prod_{n=2}^{N} W_{t-1}^{a_{t-1}^n}(z_{t-1}^{1:N})\, \mathrm{d}a_{t-1}^n q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n) \right] \end{aligned} \tag{7}$$

sequentially, that is, sample independently $Z_0^n \sim m_0(\mathrm{d}z_0)$ for all $n \in 2:N$, then sample independently $A_0^n \sim \mathcal{M}(W_0^{1:N}(z_0^{1:N}))$ for all $n \in 2:N$, and so on. (This

is equivalent to running a particle algorithm, except that the trajectory with labels $(1,\dots,1)$ is kept fixed.)

CPF-2 Sample $N^\star$ from $\mathcal{M}(W_T^{1:N}(z_T^{1:N}))$, that is, perform a Gibbs update of $N^\star$ conditional on all the other variables, relative to (6), and return trajectory $Z^{N^\star}$.

With all these considerations, one sees that the CPF algorithm defines the following kernel $P_T^N : \mathcal{X}^{T+1} \to \mathcal{P}(\mathcal{X}^{T+1})$: for $z_T^\star \in \mathcal{X}^{T+1}$,

$$\begin{aligned}&(P_T^N\varphi)(z_T^\star)\\ &\quad= \int P_T^N(z_T^\star, \mathrm{d}z_T')\varphi(z_T') \qquad (8)\\ &\quad= \mathbb{E}_{\pi_T^N}\left\{\frac{G_T(z_T^\star)}{G_T(z_T^\star)+\sum_{m=2}^N G_T(Z_T^m)}\varphi(z_T^\star) + \sum_{n=2}^N \frac{G_T(Z_T^n)}{G_T(z_T^\star)+\sum_{m=2}^N G_T(Z_T^m)}\varphi(Z_T^n)\right\}.\end{aligned}$$

3. A coupling of the particle Gibbs Markov kernel

This section is dedicated to establishing Theorem 3 below. We first make the following assumption, which is a common assumption to establish the stability of a Feynman–Kac system (e.g., [8]).

Assumption (G). *There exists a sequence of finite positive numbers $\{g_t\}_{t\geq 0}$ such that $0 < G_t(x_t) \leq g_t$ for all $x_t \in \mathcal{X}$, $t \geq 0$. Moreover,*

$$\int m_0(\mathrm{d}x_0)G_0(x_0) \geq \frac{1}{g_0}, \qquad \inf_{x_{t-1}\in\mathcal{X}} \int m_t(x_{t-1}, \mathrm{d}x_t)G_t(x_t) \geq \frac{1}{g_t}, \qquad t > 0.$$

Loosely speaking this assumption prevents the reference trajectory of the particle Gibbs kernel from dominating the other particles during resampling.

Theorem 3. *Under Assumption (G), for any $\varepsilon \in (0,1)$ and $T \in \mathbb{N}^+$, there exists $N_0 \in \mathbb{N}^+$, such that, for all $N \geq N_0$, $x_{0:T}$, $\check{x}_{0:T} \in \mathcal{X}^{T+1}$, and $\varphi : \mathcal{X}^{T+1} \to [-1,1]$, one has*

$$|P_T^N(\varphi)(x_{0:T}) - P_T^N(\varphi)(\check{x}_{0:T})| \leq \varepsilon.$$

The supremum with respect to φ of the bounded quantity is the total variation between the two corresponding distributions (defined by kernel P_T^N and the two starting points $x_{0:T}$, $\check{x}_{0:T}$). A direct corollary of Theorem 3 is that, for N large enough, the kernel P_T^N is arbitrarily close to the independent kernel that samples from $\mathbb{Q}_T$. This means that, again for N large enough, the kernel P_T^N is uniformly ergodic (see, e.g., [22] for a definition), with an arbitrarily small ergodicity coefficient.

The proof of Theorem 3 is based on coupling: let $\bar{\pi}(\mathrm{d}z_t^\star, \mathrm{d}\check{z}_t^\star)$ be a joint distribution for the couple $(Z_t^\star, \check{Z}_t^\star)$, such that the marginal distribution of $Z_t^\star$, respectively. $\check{Z}_t^\star$, is

$P_T^N(x_{0:T}, \mathrm{d}z_T^\star)$, respectively, $P_T^N(\check{x}_{0:T}, \mathrm{d}\check{z}_T^\star)$. Then

$$\begin{aligned} P_T^N(\varphi)(x_{0:T}) - P_T^N(\varphi)(\check{x}_{0:T}) &= \mathbb{E}_{\bar{\pi}}\{\varphi(Z_T^\star) - \varphi(\check{Z}_T^\star)\} \\ &= \mathbb{E}_{\bar{\pi}}\{(\varphi(Z_T^\star) - \varphi(\check{Z}_T^\star))\mathbb{I}_{\{Z_T \neq \check{Z}_T\}}\} \\ &\leq 2\mathbb{P}_{\bar{\pi}}(Z_T^\star \neq \check{Z}_T^\star). \end{aligned}$$

The following section describes the particular coupling construction we are using. Section 3.2 then establishes that this particular coupling ensures that

$$\mathbb{P}_{\bar{\pi}}(Z_T^\star \neq \check{Z}_T^\star) \leq \varepsilon/2 \tag{9}$$

for N large enough, which concludes the proof.

3.1. Coupling construction

The coupling operates on the extended space corresponding to the support of the conditional distribution (7). The idea is to construct two conditional particle systems generated marginally from (7), that is, two systems of $N-1$ trajectories, denoted, respectively, $(Z_{0:T}^{2:N}, A_{0:T-1}^{2:N})$ and $(\check{Z}_{0:T}^{2:N}, \check{A}_{0:T-1}^{2:N})$, that complement, respectively, the trajectory $x_{0:T}$ (first system) and $\check{x}_{0:T}$ (second system), in such a way that these trajectories coincide as much as possible. We will denote by $\mathcal{C}_t \subset 1\!:\!N$ the set which contains the particle labels n such that Z_t^n and $\check{Z}_t^n$ are coupled. Let $\mathcal{C}_t^c = (1\!:\!N) \setminus \mathcal{C}_t$; by construction, $\mathcal{C}_t^c$ always contains 1, since the frozen trajectory is relabelled as trajectory 1 in (7). Before we define recursively $\mathcal{C}_t$, $(Z_{0:T}^{2:N}, A_{0:T-1}^{2:N})$ and $(\check{Z}_{0:T}^{2:N}, \check{A}_{0:T-1}^{2:N})$, we need to introduce several quantities, such as the following empirical measures, for $t \geq 0$,

$$\xi_{\mathcal{C}_t} = \sum_{n \in \mathcal{C}_t} \delta_{Z_t^n}(\mathrm{d}z_t), \qquad \xi_{\mathcal{C}_t^c} = \sum_{n \in \mathcal{C}_t^c} \delta_{Z_t^n}(\mathrm{d}z_t), \qquad \check{\xi}_{\mathcal{C}_t^c} = \sum_{n \in \mathcal{C}_t^c} \delta_{\check{Z}_t^n}(\mathrm{d}z_t),$$

the following probability measures, $\mu_0(\mathrm{d}z_0) = m_0(\mathrm{d}z_0)$, and for $t \geq 1$,

$$\begin{aligned} \mu_t(\mathrm{d}z_t) &= \int_{\mathcal{X}^t} \Psi_{G_{t-1}}(\xi_{\mathcal{C}_{t-1}})(\mathrm{d}z_{t-1}) q_t(z_{t-1}, \mathrm{d}z_t), \\ \mu_t^c(\mathrm{d}z_t) &= \int_{\mathcal{X}^t} \Psi_{G_{t-1}}(\xi_{\mathcal{C}_{t-1}^c})(\mathrm{d}z_{t-1}) q_t(z_{t-1}, \mathrm{d}z_t), \\ \check{\mu}_t^c(\mathrm{d}\check{z}_t) &= \int_{\mathcal{X}^t} \Psi_{G_{t-1}}(\check{\xi}_{\mathcal{C}_{t-1}^c})(\mathrm{d}\check{z}_{t-1}) q_t(z_{t-1}, \mathrm{d}z_t), \end{aligned}$$

where

$$\Psi_{G_{t-1}}(\xi_{\mathcal{C}_{t-1}})(\mathrm{d}z_{t-1}) = \frac{\xi_{\mathcal{C}_{t-1}}(\mathrm{d}z_{t-1}) G_{t-1}(z_{t-1})}{\int_{\mathcal{X}^t} \xi_{\mathcal{C}_{t-1}}(\mathrm{d}z_{t-1}) G_{t-1}(z_{t-1})},$$

the measures $\Psi_{G_{t-1}}(\xi_{\mathcal{C}_{t-1}^c})(\mathrm{d}z_{t-1})$ and $\Psi_{G_{t-1}}(\check{\xi}_{\mathcal{C}_{t-1}^c})(\mathrm{d}z_{t-1})$ being defined similarly, and finally the constants

$$\lambda_{t-1} = \frac{\xi_{\mathcal{C}_{t-1}}(G_{t-1})}{\xi_{\mathcal{C}_{t-1}}(G_{t-1}) + \xi_{\mathcal{C}_{t-1}^c}(G_{t-1})}, \qquad \check{\lambda}_{t-1} = \frac{\xi_{\mathcal{C}_{t-1}}(G_{t-1})}{\xi_{\mathcal{C}_{t-1}}(G_{t-1}) + \check{\xi}_{\mathcal{C}_{t-1}^c}(G_{t-1})},$$

and the measures

$$\begin{aligned}
\nu_t &= \frac{|\lambda_{t-1} - \check{\lambda}_{t-1}|}{1 - \lambda_{t-1} \wedge \check{\lambda}_{t-1}} \mu_t + \frac{1 - \lambda_{t-1} \vee \check{\lambda}_{t-1}}{1 - \lambda_{t-1} \wedge \check{\lambda}_{t-1}} \mu_t^c, \\
\check{\nu}_t &= \frac{|\lambda_{t-1} - \check{\lambda}_{t-1}|}{1 - \lambda_{t-1} \wedge \check{\lambda}_{t-1}} \mu_t + \frac{1 - \lambda_{t-1} \vee \check{\lambda}_{t-1}}{1 - \lambda_{t-1} \wedge \check{\lambda}_{t-1}} \check{\mu}_t^c, \\
\kappa_t(\mathrm{d}z_t, \mathrm{d}\check{z}_t) &= \nu_t(\mathrm{d}z_t)\check{\mu}_t^c(\mathrm{d}\check{z}_t)\mathbb{I}_{\{\lambda_{t-1} > \check{\lambda}_{t-1}\}} + \mu_t^c(\mathrm{d}z_t)\check{\nu}_t(\mathrm{d}\check{z}_t)\mathbb{I}_{\{\check{\lambda}_{t-1} \geq \lambda_{t-1}\}}.
\end{aligned}$$

We now construct $\mathcal{C}_t$ and the two particle systems as follows. First, set $\mathcal{C}_0 = 2\!:\!N$, hence $\mathcal{C}_0^c = \{1\}$, draw Z_0^n independently from m_0, and set $\check{Z}_0^n = Z_0^n$, for all $n \in 2\!:\!N$. Recall that $Z_0^1 = x_0$ and $\check{Z}_0^1 = \check{x}_0$.

To progress from time $t-1 \geq 0$ to time t, we note that there is a λ_{t-1} (resp., $\check{\lambda}_{t-1}$) probability that A_{t-1}^n (resp., $\check{A}_{t-1}^n$) is drawn from $\mathcal{C}_{t-1}$, for any $n \in 2\!:\!N$. Hence, the maximum coupling probability for $(A_{t-1}^n, \check{A}_{t-1}^n)$ is $\lambda_{t-1} \wedge \check{\lambda}_{t-1}$. Thus, with probability $\lambda_{t-1} \wedge \check{\lambda}_{t-1}$, we sample A_{t-1}^n from $\mathcal{C}_{t-1}$ (with probability proportional to $G_{t-1}(Z_{t-1}^m)$, for $m \in \mathcal{C}_{t-1}$), $Z_t^n \sim q_t(Z_{t-1}^{A_{t-1}^n}, \mathrm{d}Z_t)$, and take $(\check{A}_{t-1}^n, \check{Z}_t^n) = (A_{t-1}^n, Z_t^n)$. Marginally, $Z_t^n = \check{Z}_t^n$ is drawn from μ_t, and we set $n \in \mathcal{C}_t$.

Conditional on not being coupled (hence, we set $n \in \mathcal{C}_t^c$), (A_{t-1}^n, Z_t^n) and $(\check{A}_{t-1}^n, \check{Z}_{t-1}^n)$ may be sampled independently using the same ideas. Assume $\lambda_{t-1} \leq \check{\lambda}_{t-1}$. With probability $(\check{\lambda}_{t-1} - \lambda_{t-1})$, one should sample A_{t-1}^n from $\mathcal{C}_{t-1}^c$ and $\check{A}_{t-1}^n$ from $\mathcal{C}_{t-1}$. And with probability $1 - \lambda_{t-1} \vee \check{\lambda}_{t-1}$, both A_t^n and $\check{A}_t^n$ may be sampled from $\mathcal{C}_t^c$. Either way, $Z_t^n \sim q_t(Z_{t-1}^{A_{t-1}^n}, \mathrm{d}Z_t)$, $\check{Z}_t^n \sim q_t(\check{Z}_{t-1}^{\check{A}_{t-1}^n}, \mathrm{d}Z_t)$, independently. By symmetry, the case $\lambda_{t-1} \geq \check{\lambda}_{t-1}$ works along the same lines. Marginally (when integrating out A_{t-1}^n and $\check{A}_{t-1}^n$), and conditional on not being coupled, the pair $(Z_t^n, \check{Z}_t^n)$ is drawn from κ_t. Clearly, this construction maintains the correct marginal distribution for the two particle systems.

At the final time T, the trajectories $Z_T^\star$, $\check{Z}_T^\star$ that are eventually selected, that is, the output of Markov kernels $P_T^N(x_{0:T}, \mathrm{d}z_T^\star)$ and $P_T^N(\check{x}_{0:T}, \mathrm{d}\check{z}_T^\star)$ may be coupled exactly in the same way: with probability $\lambda_T \wedge \check{\lambda}_T$, they are taken to be equal, and $Z_T = Z_T^\star$ is sampled from μ_T; and with probability $(1 - \lambda_T \wedge \check{\lambda}_T)$, $(Z_T^\star, \check{Z}_T^\star)$ is sampled from κ_T.

The motivation for this coupling construction is that it is the *maximal coupling* (see [17] for a definition) for quantifying the total variation norm between CPF kernels $P_T^N(x_{0:T}, \cdot)$ and $P_T^N(\check{x}_{0:T}, \cdot)$ when either $T = 0$ or $T > 0$ and m_t is a Dirac measure for all t. Details of proof of this fact can be obtained from the authors.

3.2. Proof of inequality (9)

We now prove that the coupling construction described in the previous section is such that inequality (9) holds for N large enough.

By construction, one has that $\mathbb{P}(Z_T^\star = \check{Z}_T^\star) \geq \mathbb{E}(\lambda_T \wedge \check{\lambda}_T)$. Consider the event

$$\mathcal{A} = \left\{ \frac{\xi_{\mathcal{C}_T}(G_T)}{|\mathcal{C}_T|} \geq \frac{\mu_T(G_T)}{2} \right\}.$$

Given Assumption (G), and the definition of λ_T, one has

$$1 - \lambda_T \leq \frac{g_T |\mathcal{C}_T^c|}{\xi_{\mathcal{C}_T}(G_T)}$$

and the same inequality holds for $1 - \check{\lambda}_T$, which leads to

$$(1 - \lambda_T \wedge \check{\lambda}_T) \times \mathbb{I}_{\mathcal{A}} \leq 2 \frac{|\mathcal{C}_T^c| g_T}{|\mathcal{C}_T| \mu_T(G_T)} \times \mathbb{I}_{\mathcal{A}} \leq 2 \frac{|\mathcal{C}_T^c| g_T^2}{|\mathcal{C}_T|} \times \mathbb{I}_{\mathcal{A}},$$

where the second inequality is due to Assumption (G). Therefore, for $k_1, \ldots, k_T \in 1:N$,

$$\begin{aligned}
&\mathbb{E}\{(\lambda_T \wedge \check{\lambda}_T) \times \mathbb{I}_{\mathcal{A}} || \mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T\} \\
&\quad \geq \left(1 - 2 \frac{k_T}{N - k_T} g_T^2\right)^+ \mathbb{E}\{\mathbb{I}_{\mathcal{A}} || \mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T\}.
\end{aligned}$$

Conditional on $Z_{T-1}^{1:N}$, and $n \in \mathcal{C}_T$, Z_T^n is an independent draw from $\mu_T(\mathrm{d}z_T)$. Thus, in order to lower bound the probability of event $\mathcal{A}$, we may apply Hoeffding's inequality [12] to the empirical mean $\xi_{\mathcal{C}_T}(G_T)/|\mathcal{C}_T|$ as follows. Again per Assumption (G), noting that $0 < G_T(z_T^n) \leq g_T$ and the one-step predicted potential is bounded below uniformly by g_T^{-1},

$$\begin{aligned}
&\mathbb{E}\{(1 - \mathbb{I}_{\mathcal{A}}) || \mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T, Z_{T-1}^{1:N} = z_{T-1}^{1:N}\} \\
&\quad = \mathbb{P}\left(\frac{\xi_{\mathcal{C}_T}(G_T)}{|\mathcal{C}_T|} < \frac{\mu_T(G_T)}{2} || \mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T, Z_{T-1}^{1:N} = z_{T-1}^{1:N} \right) \\
&\quad \leq \exp\left(-2(N - k_T) \frac{\mu_T(G_T)^2}{4 g_T^2} \right) \\
&\quad \leq \exp\left(-\frac{(N - k_T)}{2 g_T^4} \right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\{(\lambda_T \wedge \check{\lambda}_T) \times \mathbb{I}_{\mathcal{A}} || C_1| = N - k_1, \ldots, |C_T| = N - k_T\} \\
&\quad \geq \left(1 - 2 \frac{k_T}{N - k_T} g_T^2\right)^+ \left\{ 1 - \exp\left(-\frac{(N - k_T)}{2 g_T^4} \right) \right\}.
\end{aligned}$$

Finally, for any sequence of integers $L_{1:T} \in (1:N)^T$,

$$\begin{aligned}\mathbb{E}(\lambda_T \wedge \check{\lambda}_T) &\geq \sum_{k_1=1}^{L_1} \cdots \sum_{k_T=1}^{L_T} \mathbb{E}\{(\lambda_T \wedge \check{\lambda}_T) \times \mathbb{I}_{\mathcal{A}} || C_1| = N - k_1, \ldots, |C_T| = N - k_T\} \\ &\quad \times \mathbb{P}(|\mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T) \\ &\geq \left(1 - 2\frac{L_T}{N - L_T} g_T^2\right)\left\{1 - \exp\left(-\frac{(N - L_T)}{2g_T^4}\right)\right\} \\ &\quad \times \sum_{k_1=1}^{L_1} \cdots \sum_{k_T=1}^{L_T} \mathbb{P}(|\mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T)\end{aligned}$$

provided N is large enough. To conclude, we resort to a technical lemma, proven in the following section, that states it is possible to choose $L_1, \ldots, L_T$ large enough so as to make the sum of probabilities in the last line above as large as needed. In addition, for L_T fixed and N large enough, the two factors in front of that sum are arbitrarily close to one.

3.3. Technical lemma

Lemma 4. *Under Assumption (G), and for any $\delta \in (0,1)$, $T \in \mathbb{N}^+$, there exist positive integers N_0, $L_1, \ldots, L_T$ such that for any $N \geq N_0$ and $x_{0:T}$, $\check{x}_{0:T} \in \mathcal{X}^{T+1}$,*

$$\sum_{k_1=1}^{L_1} \cdots \sum_{k_T=1}^{L_T} \mathbb{P}(|\mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T) \geq (1 - \delta)^{3T}.$$

Proof. Let $\omega_t = \lambda_t \wedge \check{\lambda}_t$ and recall that ω_t is the probability (conditional on $Z_t^{1:N}$) that $n \in \mathcal{C}_{t+1}$, that is, that particles Z_{t+1}^n and $\check{Z}_{t+1}^n$ are coupled. Thus, and using the fact that the particle system is exchangeable, one has

$$\begin{aligned}&\mathbb{P}(|\mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T) \\ &\quad = \left\{\prod_{t=1}^{T} \binom{N-1}{N-k_t}\right\} \mathbb{P}(\mathcal{C}_1^c = (1:k_1), \ldots, \mathcal{C}_T^c = (1:k_T)) \\ &\quad = \int \prod_{t=0}^{T-1} \binom{N-1}{N-k_{t+1}} \left\{\prod_{n=2}^{k_t} \kappa_t(\mathrm{d}z_t^n, \mathrm{d}\check{z}_t^n) \prod_{n=k_t+1}^{N} \mu_t(\mathrm{d}z_t^n)\right\} (1-\omega_t)^{(k_{t+1}-1)} \omega_t^{(N-k_{t+1})} \\ &\quad \geq \int \prod_{t=0}^{T-1} \left\{\prod_{n=2}^{k_t} \kappa_t(\mathrm{d}z_t^n, \mathrm{d}\check{z}_t^n) \prod_{n=k_t+1}^{N} \mu_t(\mathrm{d}z_t^n)\right\} \left\{\frac{(N-1)!(1-\omega_t)^{(k_{t+1}-1)}}{(N-k_{t+1})!}\right\} \left\{\frac{\omega_t^{(N-k_{t+1})}}{(k_{t+1}-1)!}\right\} \mathbb{I}_{\mathcal{A}_t}\end{aligned} \tag{10}$$

with the convention that $k_0 = 1$, that $\mu_0(\mathrm{d}z_0) = m_0(\mathrm{d}z_0)$, and that empty products equal one, and defining the event $\mathcal{A}_t$ as

$$\mathcal{A}_t = \left\{ \frac{\xi_{\mathcal{C}_t}(G_t)}{|\mathcal{C}_t|} \geq \frac{\mu_t(G_t)}{2} \right\}. \tag{11}$$

Note that the two integrals above are with respect to a joint distribution which corresponds to a chain rule decomposition that works forward in time: κ_1 and μ_1 are distributions conditional on $Z_0^{1:N}$ and so on. In addition, these chained conditional distributions are such that $|\mathcal{C}_t| = N - k_t$ with probability one.

For the sake of transparency, we complete the proof for $T = 2$, but we note that exactly the same steps employed may be extended to the general case where $T > 2$.

The key idea is to replace the two factors in the integrand of (10) with their large N values which we now define. Let

$$\Lambda_t = |\mathcal{C}_t| \times \frac{\xi_{\mathcal{C}_t^c}(G_t) \vee \check{\xi}_{\mathcal{C}_t^c}(G_t)}{\xi_{\mathcal{C}_t}(G_t)} \qquad \text{for } |\mathcal{C}_t| > 0 \tag{12}$$

and set $\Lambda_t = 0$ if $|\mathcal{C}_t| = 0$. Using Lemma 5, stated and proved at the end of this section, one has, for fixed k_1, k_2, δ, and N large enough, that the integral in (10) is larger than

$$\begin{aligned} &(1-\delta)^2 \int_{\mathcal{A}_0} \left\{ \prod_{n=2}^{N} \mu_0(\mathrm{d}z_0^n) \right\} \frac{\mathrm{e}^{-\Lambda_0} \Lambda_0^{k_1-1}}{(k_1-1)!} \\ &\quad \times \int_{\mathcal{A}_1} \left\{ \prod_{n=2}^{k_1} \kappa_1(\mathrm{d}z_1^n, \mathrm{d}\check{z}_1^n) \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\} \frac{\mathrm{e}^{-\Lambda_1} \Lambda_1^{k_2-1}}{(k_2-1)!}. \end{aligned} \tag{13}$$

We now explain how to choose L_1, L_2 such that, for N large enough,

$$\begin{aligned} &\int_{\mathcal{A}_0} \left\{ \prod_{n=2}^{N} \mu_0(\mathrm{d}z_0^n) \right\} \sum_{k_1=1}^{L_1} \frac{\mathrm{e}^{-\Lambda_0} \Lambda_0^{k_1-1}}{(k_1-1)!} \\ &\quad \times \int_{\mathcal{A}_1} \left\{ \prod_{n=2}^{k_1} \kappa_1(\mathrm{d}z_1^n, \mathrm{d}\check{z}_1^n) \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\} \sum_{k_2=1}^{L_2} \frac{\mathrm{e}^{-\Lambda_1} \Lambda_1^{k_2-1}}{(k_2-1)!} \geq (1-\delta)^4. \end{aligned} \tag{14}$$

First note that, given Assumption (G), and since $|\mathcal{C}_0| = N - 1$, $|\mathcal{C}_1| = N - k_1$ (with probability one under the conditional distribution that appears in (10), for $t = 1$, as explained above), and since $\xi_{\mathcal{C}_t}(G_t) \geq |\mathcal{C}_t| \mu_t(G_t)/2$ (by event $\mathcal{A}_t$), one has (again with probability one under the same conditional distribution):

$$0 \leq \Lambda_0 \times \mathbb{I}_{\mathcal{A}_0} \leq 2g_0^2 \times \mathbb{I}_{\mathcal{A}_0}, \qquad 0 \leq \Lambda_1 \times \mathbb{I}_{\mathcal{A}_1} \leq 2g_1^2 k_1 \times \mathbb{I}_{\mathcal{A}_1} \tag{15}$$

for $N > k_1$ (otherwise the probability that $|\mathcal{C}_1| = N - k_1$ would be zero). Choose L_1, then L_2, such that

$$\sum_{k_1=1}^{L_1} \frac{\mathrm{e}^{-2g_0^2}(2g_0^2)^{k_1-1}}{(k_1-1)!} \geq 1-\delta,$$
$$\sum_{k_2=1}^{L_2} \frac{\mathrm{e}^{-(2g_1^2 L_1)}(2g_1^2 L_1)^{k_2-1}}{(k_2-1)!} \geq 1-\delta.$$

Since $x \to \mathrm{e}^{-x}\sum_{k=0}^{L} x^k/k!$ is a decreasing function for $x > 0$, this choice of L_2 ensures that

$$\int_{\mathcal{A}_1} \left\{ \prod_{n=2}^{k_1} \kappa_1(\mathrm{d}z_1^n, \mathrm{d}\check{z}_1^n) \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\} \sum_{k_2=1}^{L_2} \frac{\mathrm{e}^{-\Lambda_1}\Lambda_1^{k_2-1}}{(k_2-1)!}$$
$$\geq (1-\delta) \int_{\mathcal{A}_1} \left\{ \prod_{n=2}^{k_1} \kappa_1(\mathrm{d}z_1^n, \mathrm{d}\check{z}_1^n) \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\}.$$

By Hoeffding's inequality (in the same way as in the previous section),

$$\int_{\mathcal{A}_1^{\mathrm{c}}} \left\{ \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\} \leq \exp\left\{ -2(N-k_1)\frac{\mu_1(G_1)^2}{4g_1^2} \right\}$$
$$\leq \exp\left\{ -(N-k_1)\frac{g_1^{-4}}{2} \right\},$$

where the last inequality follows from Assumption (G). Using the same calculations for the first integral in (14) (i.e., applying Hoeffding's inequality to $\mathcal{A}_0$ again in the same way), one obtains eventually

$$\int_{\mathcal{A}_0} \left\{ \prod_{n=2}^{N} \mu_0(\mathrm{d}z_0^n) \right\} \sum_{k_1=1}^{L_1} \frac{\mathrm{e}^{-\Lambda_0}\Lambda_0^{k_1-1}}{(k_1-1)!}$$
$$\times \int_{\mathcal{A}_1} \left\{ \prod_{n=2}^{k_1} \kappa_1(\mathrm{d}z_1^n, \mathrm{d}\check{z}_1^n) \prod_{n=k_1+1}^{N} \mu_1(\mathrm{d}z_1^n) \right\} \sum_{k_2=1}^{L_2} \frac{\mathrm{e}^{-\Lambda_1}\Lambda_1^{k_2-1}}{(k_2-1)!}$$
$$\geq (1-\delta)^2 \left[1 - \exp\left\{ -(N-L_1)\frac{g_1^{-4}}{2} \right\} \right] \left[1 - \exp\left\{ -(N-1)\frac{g_0^{-4}}{2} \right\} \right]$$
$$\geq (1-\delta)^4$$

for N large enough and, therefore, combining this with (13), one may conclude that

$$\sum_{k_1=1}^{L_1}\sum_{k_2=1}^{L_2}\mathbb{P}(|\mathcal{C}_1|=N-k_1,|\mathcal{C}_2|=N-k_2)\geq(1-\delta)^6$$

provided N is taken to be large enough. □

To conclude the proof, we state and prove the following lemma, which we used in the proof above in order to replace the two last factors in (10) by their large-N values.

Lemma 5. *Assume (G). For any given $\delta>0$ and positive integers $k_1,\ldots,k_T$, there exists a positive integer N_0 such that the following inequalities hold for all $N\geq N_0$ and $x_{0:T}$, $\check{x}_{0:T}\in\mathcal{X}^{T+1}$:*

$$\omega_t^{(N-k_{t+1})}\exp(\Lambda_t)\times\mathbb{I}_{\mathcal{A}_t}\geq(1-\delta)\times\mathbb{I}_{\mathcal{A}_t},$$
$$\frac{(N-1)!}{(N-k_{t+1})!}(1-\omega_t)^{(k_{t+1}-1)}\frac{1}{\Lambda_t^{k_{t+1}-1}}\times\mathbb{I}_{\mathcal{A}_t}\geq(1-\delta)\times\mathbb{I}_{\mathcal{A}_t}.$$

Proof. Given that $|\mathcal{C}_t|=N-k_t$ and the respective definitions of ω_t and Λ_t, one has

$$\frac{1-\omega_t}{\omega_t}=\frac{\Lambda_t}{N-k_t}$$

and, therefore, for N large enough and k_t fixed, and conditional on $\mathbb{I}_{\mathcal{A}_t}=1$, the probability $1-\omega_t$ may be made arbitrarily small, given that $\Lambda_t\mathbb{I}_{\mathcal{A}_t}$ is a bounded quantity; see (15). Since $\log(1+x)\geq x-x^2$ for $x\geq-1/2$, one has, for N large enough (so that $x=\omega_t-1\geq-1/2$), and conditional on $\mathbb{I}_{\mathcal{A}_t}=1$,

$$\Lambda_t+(N-k_{t+1})\log\omega_t\geq\Lambda_t\biggl\{1-\frac{N-k_{t+1}}{N-k_t}\omega_t-\frac{N-k_{t+1}}{(N-k_t)^2}\Lambda_t\omega_t^2\biggr\},$$

which can be clearly made arbitrarily small (in absolute value) by taking N large enough, since both ω_t and Λ_t are bounded quantities. The second inequality may be proved along the same lines. □

4. Backward sampling

This section discusses the backward sampling (BS) step proposed by [26] so as to improve the mixing of particle Gibbs. It is convenient in this section to revert to standard notation based on the initial process X_t, rather than on notation based on trajectories $Z_t=X_{0:t}$. Thus, we now consider the following (extended) invariant distribution for the CPF kernel

$$\pi_T^N(\mathrm{d}x_{0:T}^{1:N},\mathrm{d}a_{0:T-1}^{1:N},\mathrm{d}n^\star)$$

$$
\begin{aligned}
&= \frac{1}{\mathcal{Z}_T} m_0^{\otimes N}(\mathrm{d}x_0^{1:N}) \\
&\quad \times \prod_{t=1}^{T} \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} G_{t-1}(x_{t-1}^n) \right] \prod_{n=1}^{N} [W_{t-1}^{a_{t-1}^n}(x_{t-1}^{1:N})\, \mathrm{d}a_{t-1}^n m_t(x_{t-1}^{a_{t-1}^n}, \mathrm{d}x_t^n)] \right\} \\
&\quad \times \frac{1}{N} G_T(x_T^{n^\star}),
\end{aligned}
\tag{16}
$$

where $W_t^n(x_t^{1:N}) = G_t(x_t^n)/\sum_{m=1}^N G_t(x_t^m)$; compared to (5), this equation represents a simple change of variables.

In this new set of notation, the nth trajectory Z_T^n becomes a deterministic function of the particle system $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N})$, that may be defined as follows: $Z_T^n = (X_0^{B_0^n}, \ldots, X_T^{B_T^n})$, where the indexes B_T^n's are defined recursively as: $B_T^n = n$, $B_t^n = A_t^{B_{t+1}^n}$, for $t \in 0:(T-1)$. Similarly, as noted before, $Z_T^\star = Z_T^{N^\star} = (X_0^{B_0^\star}, \ldots, X_T^{B_T^\star})$, with $B_T^\star = N^\star$, $B_t^\star = A_t^{B_{t+1}^\star}$, that is, $Z_T^\star$ is a deterministic function of $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N}, N^\star)$.

Whiteley [26], in his discussion of [1] (see also [16]), suggested to add the following BS (backward sampling) step to a particle Gibbs update.

CPF-3 Let $B_T^\star = N^\star$, then, recursively for $t = T-1$ to $t = 0$, sample index $B_t^\star = A_t^{B_{t+1}^\star} \in 1:N$, conditionally on $B_{t+1}^\star = b$, from the distribution

$$
\begin{aligned}
&\pi_T^N(A_t^b = a_t^b | X_{0:T}^{1:N} = x_{0:T}^{1:N}, A_t^{-b} = a_t^{-b}, A_{0:t-1}^{1:N} = a_{0:t-1}^{1:N}, A_{t+1:T}^{1:N} = a_{t+1:T}^{1:N}, N^\star = n^\star) \\
&\quad \propto W_t^{a_t^b} m_{t+1}(x_t^{a_t^b}, x_{t+1}^b)
\end{aligned}
\tag{17}
$$

and set $X_t^\star = x_t^{b_t^\star}$. (Recall that $Z_T^\star = (X_0^{B_0^\star}, \ldots, X_T^{B_T^\star})$.)

In (17), A_t^{-b} is $A_t^{1:N}$ minus A_t^b, a_t^{-b} is defined similarly, and $m_{t+1}(x_t^{a_t^b}, x_{t+1}^b)$ is the probability density (relative to measure $\mathrm{d}x_{t+1}$) of conditional distribution $m_{t+1}(x_t^{a_t^b}, \mathrm{d}x_{t+1})$ evaluated at point x_{t+1}^b.

This extra step amounts to update the ancestral lineage of the selected trajectory up to time t, recursively in time, from $t = T-1$ to time $t = 0$. It is straightforward to show that (17) is the conditional distribution of random variable $B_t^\star = A_t^{B_{t+1}^\star}$, conditional on $B_{t+1}^\star = b$ and the other auxiliary variables of the particle system, relative to the joint distribution (16). As such, this extra step leaves $\mathbb{Q}_T(\mathrm{d}x_{0:T})$ invariant.

It should be noted that the BS step may be implemented only when the density $m_{t+1}(x_t, x_{t+1})$ admits an explicit expression, which is unfortunately not the case for several models of practical interest. Finally, Remark 1 also applies to the CPF-BS kernel.

Remark 2. Let $P_T^{N,B}$ denote the CPF-BS (CPF with backward sampling) Markov kernel. Then the image of $z_T^\star \in \mathcal{X}^{T+1}$ under $P_T^{N,B}$ is unchanged by the choice of $(b_{0:T-1}^\star, n^\star)$ for the realization of $(B_{0:T-1}^\star, N^\star)$ in the initialization, that is, step CPF-1.

4.1. Reversibility, covariance ordering and asymptotic efficiency

To compare PG with backward sampling with PG without backward sampling, one might be tempted to use Peskun ordering. The following counter-example shows that unfortunately the former does not dominate the latter in the Peskun sense.

Example 6. Let $\mathcal{X} = \mathbb{R}$, $(X_t)_{t\geq 0}$ be an i.i.d. sequence with marginal law $m(\mathrm{d}x)$, and let the potentials be unit valued, that is, $G_t(x_t) = 1$ for all t. Then one can show that the CPF kernel with backward sampling does not dominate the CPF kernel in Peskun sense. For example, let $\mathcal{B} = \mathcal{B}_0 \times \cdots \times \mathcal{B}_T \subset \mathcal{X}^{T+1}$, where $m(\mathcal{B}_t) = \varepsilon$ for all t. If we choose a reference trajectory $x_{0:T} \notin \mathcal{B}$ but $x_t \in \mathcal{B}_t$ for $t \neq T$, then it is easy to show that (e.g., when $T = 2$ and $N = 2$) that the probability of hitting $\mathcal{B}$ when starting from $x_{0:T}$, that is, $P_T^N(x_{0:T}, \mathcal{B})$, is higher without backward sampling than with it. In this example, a chosen trajectory that coalesces with the reference trajectory has more chance of hitting set $\mathcal{B}$.

One does observe in practice that Backward sampling (BS) brings improvement to the decay of the autocorrelation function of successive samples of $X_{0:T}$ generated by the particle Gibbs sampler, that is, more rapid decay compared to not implementing BS; see [16] and our numerical experiments in (6). However, how much improvement depends on the transition kernel $m_t(x_{t-1}, \mathrm{d}x_t)$ of the hidden state process $(X_t)_{t\geq 0}$. If only $X_0 \sim m_0$ is random while $m_t(x_{t-1}, \mathrm{d}x_t)$ is a point mass at x_{t-1} for $t \geq 1$, then it is clear that BS will bring no improvement. We can however prove that, regardless of m_t, the empirical average of the successive samples from a CPF kernel with BS will have an asymptotic variance no larger than the asymptotic variance of the empirical average of successive samples from the corresponding CPF kernel without BS. The asymptotic variance here is the variance of the limiting Gaussian distribution characterised by the usual $\sqrt{n}$-central limit theorem (CLT) for Markov chains; see, for example, [22].

The following result due to [24] (see also [19]), formalises this comparison, or ordering, of two Markov transition kernels having the same stationary measure via the asymptotic variance given by the CLT. We call this efficiency ordering.

Theorem 7 ([24]). *For $\xi_0, \xi_1, \ldots$ successive samples from a reversible Markov transition kernel H (on some general state space) with stationary measure π, where $\xi_0 \sim \pi$, and for $f \in L^2(\pi) = \{f : \int \pi(\mathrm{d}x) f(x)^2 < \infty\}$ let*

$$v(f, H) = \lim_{n\to\infty} \frac{1}{n} \mathrm{Variance}\left(\sum_{i=0}^{n-1} f(\xi_i)\right).$$

Let P_1 and P_2 be two reversible Markov kernels with stationary measure π such that $\mathbb{E}_{\pi\otimes P_1}\{g(\xi_0)g(\xi_1)\} \leq \mathbb{E}_{\pi\otimes P_2}\{g(\xi_0)g(\xi_1)\}$ for all $g \in L^2(\pi)$. Then $v(f, P_1) \leq v(f, P_2)$ for all $f \in L^2(\pi)$ and P_1 is said to dominate P_2 in efficiency ordering.

Note that in the original version of this theorem by [24] the requirement on P_1 and P_2 for $v(f,P_1) \leq v(f,P_2)$ for all $f \in L^2(\pi)$ is that P_1 dominates P_2 in Peskun ordering. However, Tierney's proof actually makes use of the weaker Peskun implied property of lag-1 domination instead, as also noted in Theorem 4.2 of [19].

To prove efficiency ordering, we must prove first that the CPF kernel, with or without BS, is reversible. Following reversibility, we then need to show that the CPF kernel with BS has smaller lag 1 autocorrelation compared to the CPF kernel without BS; this property if holds is called *lag-one domination*.

Proposition 8. *The CPF kernel is reversible.*

Proof. This result is based on the equivalent representation of the CPF kernel described in Section 2 (see Remark 1) which regenerates both the labels $(B^\star_{0:T-1}, N^\star)$ of the frozen trajectory, and the $N-1$ remaining trajectories $(X^{1:N\setminus\star}_{0:T}, A^{1:N\setminus\star}_{0:T-1})$. Consider a measurable function $h : \mathcal{X}^{T+1} \times \mathcal{X}^{T+1} \to \mathbb{R}$ and let $(Z^\star_T, \check{Z}^\star_T) = (X^\star_{0:T}, \check{X}^\star_{0:T}) \sim \mathbb{Q}_T \otimes P^N_T$ then

$$\begin{aligned}
\mathbb{E}\{h(Z^\star_T, \check{Z}^\star_T)\} &= \int \mathbb{Q}_T(\mathrm{d}x^\star_{0:T}) \int \pi^N_T(\mathrm{d}x^{1:N\setminus\star}_{0:T}, \mathrm{d}a^{1:N\setminus\star}_{0:T-1}, \mathrm{d}b^\star_{0:T-1}, \mathrm{d}n^\star | x^\star_{0:T}) \\
&\quad \times \int \pi^N_T(\mathrm{d}\check{n}^\star | x^{1:N}_{0:T}, a^{1:N}_{0:T-1}) h(z^{n^\star}_T, z^{\check{n}^\star}_T) \\
&= \int \pi^N_T(\mathrm{d}x^{1:N}_{0:T}, \mathrm{d}a^{1:N}_{0:T-1}, \mathrm{d}n^\star) \int \pi^N_T(\mathrm{d}\check{n}^\star | x^{1:N}_{0:T}, a^{1:N}_{0:T-1}) h(z^{n^\star}_T, z^{\check{n}^\star}_T) \\
&= \int \pi^N_T(\mathrm{d}x^{1:N}_{0:T}, \mathrm{d}a^{1:N}_{0:T-1}, \mathrm{d}\check{n}^\star) \int \pi^N_T(\mathrm{d}n^\star | x^{1:N}_{0:T}, a^{1:N}_{0:T-1}) h(z^{n^\star}_T, z^{\check{n}^\star}_T) \\
&= \mathbb{E}\{h(\check{Z}^\star_T, Z^\star_T)\},
\end{aligned}$$

where the second equality uses Remark 1. We have also used a change of variables, that is, $z^{n^\star}_T$, respectively, $z^{\check{n}^\star}_T$, must be understood as a certain deterministic function of $(x^{1:N}_{0:T}, a^{1:N}_{0:T-1}, n^\star)$, respectively, $(x^{1:N}_{0:T}, a^{1:N}_{0:T-1}, \check{n}^\star)$, in the equations above; see notation at the beginning of this section, specifically in the paragraph before step CPF-3. □

We now prove a similar result for CPF-BS, the CPF kernel with backward sampling.

Proposition 9. *The CPF-BS kernel (CPF with backward sampling) is reversible.*

Proof. Consider a $\mathcal{X}^{T+1} \times \mathcal{X}^{T+1} \to \mathbb{R}$ measurable function h and let $(Z^\star_T, \check{Z}^\star_T) = (X^\star_{0:T}, \check{X}^\star_{0:T}) \sim \mathbb{Q}_T \otimes P^{N,B}_T$. To evaluate $\mathbb{E}\{h(Z^\star_T, \check{Z}^\star_T)\}$, we first invoke Remark 2 and then the following observation: the image of CPF-BS would be unchanged if step CPF-3 would be replaced by a Gibbs step that would update the complete genealogy, that is, replace $A^{1:N}_{0:T-1}$ by $\check{A}^{1:N}_{0:T-1}$, a sample from $\pi^N_T(\mathrm{d}a^{1:N}_{0:T-1} | x^{1:N}_{0:T}, n^\star)$. This is because the A^n_t's

are independent conditionally on $(X_{0:T}^{1:N}, N^\star)$. Thus,

$$\begin{aligned}
\mathbb{E}\{h(Z_T^\star, \check{Z}_T^\star)\} &= \int \pi_T^N(\mathrm{d}x_{0:T}^{1:N}, \mathrm{d}a_{0:T-1}^{1:N}, \mathrm{d}n^\star) \\
&\quad \times \int \pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(\mathrm{d}x_{0:T}^{1:N}, \mathrm{d}\overline{a}_{0:T-1}^{1:N}) \int \pi_T^N(\mathrm{d}n^\star | x_{0:T}^{1:N}) \pi_T^N(\mathrm{d}a_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star) \\
&\quad \times \int \pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \mathbb{E}\{h(\check{Z}_T^\star, Z_T^\star)\},
\end{aligned}$$

where the second equality is based on the fact that one may generate $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N}, N^\star) \sim \pi_T^N$ as: $(X_{0:T}^{1:N}, \overline{A}_{0:T-1}^{1:N}, N^\star) \sim \pi_T^N$, then update $\overline{A}_{0:T-1}^{1:N}$ as $A_{0:T-1}^{1:N}$ through the Gibbs step $\pi_T^N(\mathrm{d}a_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$, and the third equality is a simple change of variables. The simplification of $\pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N})$ into $\pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N})$ (first equality onward) reflects the fact that step CPF-2 does not depend on $a_{0:T-1}^{1:N}$. □

The final result shows that the CPF-BS kernel dominates the CPF kernel in lag-one autocorrelation.

Theorem 10. *The CPF-BS kernel, denoted $P_T^{N,B}$, dominates the CPF kernel in lag one autocorrelation, that is, let h be square integrable function* then

$$0 \leq \mathbb{E}_{\mathbb{Q}_T \otimes P_T^{N,B}}\{h(Z_T^\star) h(\check{Z}_T^\star)\} \leq \mathbb{E}_{\mathbb{Q}_T \otimes P_T^N}\{h(Z_T^\star) h(\check{Z}_T^\star)\}.$$

Proof. We use again the facts that $\pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N})$ reduces into $\pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N})$, and that, under multinomial resampling, step CPF-3 may be replaced by a Gibbs step that updates the complete genealogy as in the proof of Proposition 9.

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_T \otimes P_T^{N,B}}\{h(Z_T^\star) h(\check{Z}_T^\star)\} \\
&= \int \pi_T^N(\mathrm{d}x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}n^\star | x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}a_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star) \\
&\quad \times \int \pi_T^N(\mathrm{d}\check{n}^\star | x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}) h(z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(\mathrm{d}x_{0:T}^{1:N}) \left(\int \pi_T^N(\mathrm{d}n^\star | x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}a_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star) h(z_T^{n^\star}) \right)^2 \\
&\leq \int \pi_T^N(\mathrm{d}x_{0:T}^{1:N}) \int \pi_T^N(\mathrm{d}a_{0:T-1}^{1:N} | \, \mathrm{d}x_{0:T}^{1:N}) \left(\int \pi_T^N(\mathrm{d}n^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) h(z_T^{n^\star}) \right)^2 \\
&= \mathbb{E}_{\mathbb{Q}_T \otimes P_T^N}\{h(Z_T^\star) h(\check{Z}_T^\star)\}.
\end{aligned}$$

The penultimate line uses Jensen inequality. The last line is indeed the same expectation but under $\mathbb{Q}_T \otimes P_T^N$ (no BS step). □

We are now in position to state the main result of this section.

Theorem 11. *The CPF-BS kernel dominates the CPF kernel in efficiency ordering.*

Proof. This is a direct consequence of Theorem 7 and Propositions 8 and 9. □

5. Alternative resampling schemes

We mentioned in the previous section that the backward sampling step is not always applicable, as it relies on the probability density of the Markov kernel m_t being tractable. In this section, we discuss another way to improve the performance of particle Gibbs through the introduction of resampling schemes that are less noisy than multinomial resampling. The intuition is that such resampling schemes tend to reduce path degeneracy in particle systems, and thus should lead to better mixing for particle Gibbs; see, for example, [13] for some results on path degeneracy.

We no longer assume that the resampling distribution ϱ_t is (3), and we rewrite π_T^N under the more general expression (using the same notation as in Section 2)

$$\begin{aligned}
&\pi_T^N(\mathrm{d}z_{0:T}^{1:N}, \mathrm{d}a_{0:T-1}^{1:N}, \mathrm{d}n^\star) \\
&\quad = \frac{1}{\mathcal{Z}_T} m_0^{\otimes N}(\mathrm{d}z_0^{1:N}) \\
&\qquad \times \prod_{t=1}^{T} \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} G_{t-1}(z_{t-1}^n) \right] \varrho_{t-1}(z_{t-1}^{1:N}, \mathrm{d}a_{t-1}^{1:N}) \prod_{n=1}^{N} [q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n)] \right\} \frac{1}{N} G_T(z_T^{n^\star}).
\end{aligned} \tag{18}$$

Recall that to establish validity of particle Gibbs, we applied the following change of variables:

$$(z_{0:T}^{1:N}, a_{0:T-1}^{1:N}, n^\star) \leftrightarrow (z_{0:T}^{1:N\setminus\star}, a_{0:T-1}^{1:N\setminus\star}, z_{0:T}^\star, b_{0:T-1}^\star, n^\star),$$

to π_T^N, which led to distribution (6), which is such that $Z_T^\star = Z_T^{N^\star}$ has marginal distribution $\mathbb{Q}_T(\mathrm{d}z_T)$. To generalise (6) to resampling schemes other than multinomial resampling, we assume that the resampling distribution is *marginally unbiased*: the joint distribution $\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N})$ (for fixed $z_t^{1:N}$) is such that the marginal distribution of a single component A_t^n is the discrete distribution which assigns probability W_t^m to outcome $m \in 1:N$. (We shall see that, up to a trivial modification, standard resampling schemes fulfil this condition.)

Under marginal unbiasedness, $\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N})$ may be decomposed as follows, for any $n \in 1:N$:

$$\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N}) = \{W_t^{a_t^n}(z_t^{1:N})\, \mathrm{d}a_t^n\} \varrho_t^c(z_t^{1:N}, \{\mathrm{d}a_t^{1:N\setminus n} | A_t^n = a_t^n\}),$$

where the second factor above denotes the distribution of the $(N-1)$ labels $A_t^{1:N\backslash n}$ conditional on $A_t^n = a_t^n$ which corresponds to the joint distribution $\varrho_t(z_t^{1:N}, \mathrm{d}a_t^{1:N})$. Thus, applying the change of variables above to π_T^N gives

$$\begin{aligned}
&\pi_T^N(\mathrm{d}z_{0:T}^{1:N\backslash\star}, \mathrm{d}a_{0:T-1}^{1:N\backslash\star}, \mathrm{d}z_{0:T}^{\star}, \mathrm{d}b_{0:T-1}^{\star}, \mathrm{d}n^{\star}) \\
&\quad = \frac{1}{N^{T+1}}(\mathrm{d}b_{0:T-1}^{\star}\,\mathrm{d}n^{\star})\mathbb{Q}_T(\mathrm{d}z_T^{\star}) \prod_{t=0}^{T-1} \delta_{([z_T^{\star}]_{t+1})}(\mathrm{d}z_t^{\star}) \qquad (19) \\
&\qquad \times \prod_{n\neq b_0^{\star}} m_0(\mathrm{d}z_0^n) \left[\prod_{t=1}^{T} \varrho_{t-1}^c(z_{t-1}^{1:N}, \{\mathrm{d}a_{t-1}^{1:N\backslash b_t^{\star}} | A_{t-1}^{b_t^{\star}} = b_{t-1}^{\star}\}) \prod_{n\neq b_t^{\star}} q_t(z_{t-1}^{a_{t-1}^n}, \mathrm{d}z_t^n)\right].
\end{aligned}$$

Inspection of the distribution above reveals that step CPF-1 (as defined in Section 2), that is, the Gibbs step that regenerates the complete particle system conditional on one "frozen" trajectory $z_{0:T}^{\star}$, now requires to sample at each iteration t from the conditional resampling distribution ϱ_{t-1}^c. The two next sections explains how to do so for two popular resampling schemes, namely residual resampling and systematic resampling.

To simplify notation in the next sections, we will remove any dependency in t, and consider the generic problem of deriving, from a certain distribution of N labels $A^{1:N}$ based on normalised weights $W^{1:N}$, the conditional distribution of $A^{1:N}$ given that one component is fixed.

5.1. Conditional residual resampling

The standard definition of residual resampling [18] is recalled as Algorithm 1. It is clear that this resampling scheme is such that the number of off-springs of particles n is a random variable with expectation NW^n (assuming $W^{1:N}$ is the vector of the N normalised weights used as input). To obtain a resampling distribution that is marginally unbiased (as defined in the previous section), we propose the following simple modification: we run Algorithm 1, and then we permute randomly the output: $A^{1:N} = \bar{A}^{\sigma(1:N)}$ where σ is chosen uniformly among the $N!$ permutations on the set $1:N$.

Another advantage of randomly permuting the labels obtained by residual resampling is that it makes the particle system exchangeable, as with multinomial resampling (as-

Algorithm 1 Residual resampling

Input: normalised weights $W^{1:N}$

Output: a vector of N random labels $\bar{A}^{1:N} \in 1:N$

(a) Compute $r^n = NW^n - \lfloor NW^n \rfloor$ (for each $n \in 1:N$) and $R = \sum_{n=1}^N r^n$.

(b) Construct $\bar{A}^{1:(N-R)}$ as the ordered vector of size $(N-R)$ that contains $\lfloor NW^n \rfloor$ copies of value n for each $n \in 1:N$.

(c) For each $n \in (N-R+1):N$, sample $\bar{A}^n \sim \mathcal{M}(r^{1:N}/R)$.

Algorithm 2 Conditional residual resampling

Input: normalised weights $W^{1:N}$

Output: a vector of N random labels $A^{1:N} \in 1:N$ such that $A^1 = 1$

(a) Compute $r^n = NW^n - \lfloor NW^n \rfloor$ (for each $n \in 1:N$) and $R = \sum_{n=1}^N r^n$.
(b) Generate $U \sim \mathcal{U}[0,1]$.
(c) **If** $U < \lfloor NW^1 \rfloor / NW^1$, **then** generate $\bar{A}^{1:N}$ using Algorithm 1;
Else generate $\bar{A}^{2:N}$ exactly as in Algorithm 1, except that the number of multinomial draws in step (c) is $R-1$ instead of R. (Thus $\bar{A}^{2:N}$ contains $\lfloor NW^n \rfloor$ deterministic copies of value n, for each $n \in 1:N$, and $(R-1)$ random copies.)
(d) Let $A^1 = 1$, and $A^{2:N} = \bar{A}^{\sigma(2:N)}$, where σ is a random $2:N \to 2:N$ permutation.

suming that residual resampling is applied at every iteration t of the particle algorithm). Thus, using the same line of reasoning as in Section 2, we see that one may arbitrarily relabel the frozen trajectory $z_T^\star$ as $(1,\ldots,1)$ before applying step CPF-1. Therefore, it is sufficient to derive an algorithm to sample from the distribution of labels $A^{2:N}$, conditional on $A^1 = 1$.

We observe that, under residual resampling (with randomly permuted output), the probability that A^1 is set to one of the $\lfloor NW^1 \rfloor$ "deterministic" copies of label 1 is $\lfloor NW^1 \rfloor / NW^1$. This remark leads to Algorithm 2, which generates a vector $A^{1:N}$ of N labels such that $A^1 = 1$.

In practice, assuming conditional residual resampling is applied at every iteration of the particle algorithm (i.e., when generating $(X_{0:T}^{2:N}, A_{0:T-1}^{2:N})$ conditional on $X_{0:T}^1$), step (d) of Algorithm 2 may be omitted, as the actual order of particles with labels $2:N$ do not play any role in the following iterations (and, therefore, has no bearing on the image of the CPF kernel).

5.2. Conditional systematic resampling

The systematic resampling algorithm of [4] consists in creating N off-springs, based on the normalised weights $W^{1:N}$, as follows. Let U a uniform variate in $[0,1]$, $v^0 = 0$, $v^n = \sum_{m=1}^n NW^m$, and set $\bar{A}^n = m$ for the N pairs (n,m) such that $v^{m-1} \leq U + n - 1 < v^m$. The standard algorithm to perform systematic resampling (for a given U, sampled from $\mathcal{U}([0,1])$ beforehand) is recalled as Algorithm 3.

To obtain a resampling distribution that is marginally unbiased, we propose to randomly cycle the output: $A^{1:N} = \bar{A}^{c(1:N)}$, where $c:(1:N) \to (1:N)$ is drawn uniformly among the N cycles of length N. Recall that a cycle c is a permutation such that for a certain $c_0 \in 1:N$ and for all $n \in 1:N$, $c(n) = c_0 + n$ if $c_0 + n \leq N$, $c(n) = c_0 + n - N$ otherwise.

Cycle randomisation is slightly more convenient that permutation randomisation when it comes to deriving the conditional systematic resampling algorithm. It is also slightly cheaper in computational terms. Under cycle randomisation, the particle system is no

Algorithm 3 Systematic resampling (for a given U)

Input: normalised weights $W^{1:N}$, and $U \in [0,1]$
Output: a vector of N random labels $A^{1:N} \in 1:N$

(a) Compute the cumulative weights as $v^n = \sum_{m=1}^{n} NW^m$ for $n \in 1:N$.
(b) Set $s \leftarrow U$, $m \leftarrow 1$.
(c) **For** $n = 1:N$
  While $v^m < s$ **do** $m \leftarrow m+1$.
  $\bar{A}^n \leftarrow m$, and $s \leftarrow s+1$.
  End For

longer exchangeable (assuming systematic resampling is carried out at each iteration), but it remains true that one has the liberty to relabel arbitrarily the frozen trajectory, say with labels $(1,\ldots,1)$, without changing the image of the PG kernel. (A proof may be obtained from the corresponding author.) Thus, as in the previous section, it is sufficient to derive the algorithm to simulate $A^{2:N}$ conditional on $A^1 = 1$.

A distinctive property of systematic resampling is that the number of off-springs of particle n is either $\lfloor NW^n \rfloor$ or $\lfloor NW^n \rfloor + 1$. In particular, the algorithm starts by creating $\lfloor NW^1 \rfloor$ "deterministic" copies of particle 1, then adds one extra "random copy," with probability $r^1 = NW^1 - \lfloor NW^1 \rfloor$, and so on. When N off-springs have been obtained, the output is randomly cycled. Thus, conditional on $A^1 = 1$, the probability that a deterministic copy of 1 was moved to position 1 is $\lfloor NW^1 \rfloor / NW^1$. This observation leads to the Algorithm 4 for generating from $A^{2:N}$ conditional on $A^1 = 1$.

5.3. Note on backward sampling for alternative resampling schemes

It is possible to adapt the backward sampling step (see Section 4) to residual or systematic resampling. Unfortunately, the corresponding algorithmic details are quite involved, and the results are not very satisfactory (in the sense of not improving strongly the mixing

Algorithm 4 Conditional systematic resampling

Input: normalised weights $W^{1:N}$
Output: a vector of N random labels $A^{1:N} \in 1:N$ such that $A^1 = 1$

(a) **If** $NW^1 \leq 1$, sample $U \sim \mathcal{U}[0, NW^1]$.
  Else Set $r^1 = NW^1 - \lfloor NW^1 \rfloor$. With probability $\frac{r^1(\lfloor NW^1 \rfloor + 1)}{NW^1}$, sample $U \sim \mathcal{U}[0, r^1]$, otherwise sample $U \sim \mathcal{U}[r^1, 1]$.
(b) Run Algorithm 3 with inputs $W^{1:N}$ and U; call $\bar{A}^{1:N}$ the output.
(c) Choose C uniformly from the set of cycles such that $\bar{A}^{C(1)} = 1$ and set $A^{1:N} = \bar{A}^{C(1:N)}$.

of particle Gibbs relative to the version without a backward sampling step); see [6] for details. This seems related to the strong dependence between the labels $A_t^{1:N}$ that is induced by residual resampling and particularly systematic resampling, which therefore makes it more difficult to update one single component of this vector.

As pointed out by a referee, it is straightforward to adapt *ancestor sampling* [15], which is an alternative approach to backward sampling for rejuvenating the ancestry of the frozen trajectory, to the alternative resampling schemes. The mixing gains of doing so deserves further investigation.

6. Numerical experiments

The focus of our numerical experiments is on comparing the four variants of particle Gibbs discussed in this paper, corresponding to the three different resampling schemes (multinomial, residual, systematic), and whether the extra backward sampling is performed or not (assuming multinomial resampling).

We consider the following state-space model:

$$X_0 \sim N(\mu, \sigma^2), \qquad X_{t+1}|X_t = x_t \sim N(\mu + \rho(x_t - \mu), \sigma^2), \qquad Y_t|X_t = x_t \sim \text{Poisson}(\mathrm{e}^{x_t})$$

for $t \in 0:T$, hence one may take $G_t(x_t) = \exp\{-\mathrm{e}^{x_t} + y_t x_t\}$, where y_t is the observed value of Y_t. This model is motivated by [28] who consider a similar model for photon counts in X-ray astrophysics. The parameters μ, ρ, σ are assumed to be unknown, and are assigned the following (independent) prior distributions: $\rho \sim \text{Uniform}[-1, 1]$, $\mu \sim N(m_\mu, s_\mu^2)$ and $1/\sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma)$; let $\theta = (\mu, \rho, \sigma^2)$. (We took $m_\mu = 0$, $s_\mu = 10$, $a_\sigma = b_\sigma = 1$ in our simulations.) We run a Gibbs sampler that targets the posterior distribution of $(\theta, X_{0:T})$, conditional on $Y_{0:T} = y_{0:T}$, by iterating (a) the Gibbs step that samples from $\theta|X_{0:T}, Y_{0:T}$, described below; and (b) the particle Gibbs step discussed in this paper, which samples from $X_{0:T}|\theta, Y_{0:T}$. Direct calculations show that step (a) may be decomposed into the following successive three operations, which sample from the full conditional distribution of each component of θ, conditional on the other components of θ and $X_{0:T}$:

$$1/\sigma^2|X_{0:T} = x_{0:T}, Y_{0:T}, \mu, \rho \sim \text{Gamma}\left(a_\sigma + \frac{T+1}{2}, b_\sigma + \frac{1}{2}\tilde{x}_0^2 + \frac{1}{2}\sum_{t=0}^{T-1}(\tilde{x}_{t+1} - \rho\tilde{x}_t)^2\right),$$

$$\rho|X_{0:T} = x_{0:T}, Y_{0:T}, \mu, \sigma \sim N_{[-1,1]}\left(\frac{\sum_{t=0}^{T-1}\tilde{x}_t\tilde{x}_{t+1}}{\sum_{t=0}^{T-1}\tilde{x}_t^2}, \frac{\sigma^2}{\sum_{t=0}^{T-1}\tilde{x}_t^2}\right),$$

$$\mu|X_{0:T} = x_{0:T}, Y_{0:T}, \rho, \sigma \sim N\left(\frac{1}{\lambda_\mu}\left\{\frac{m_\mu}{s_\mu^2} + \frac{x_0 + (1-\rho)\sum_{t=0}^{T-1}(x_{t+1} - \rho x_t)}{\sigma^2}\right\}, \frac{1}{\lambda_\mu}\right),$$

where we have used the short-hand notation $\tilde{x}_t = x_t - \mu$,

$$\lambda_\mu = \frac{1}{s_\mu^2} + \frac{1 + T(1-\rho)^2}{\sigma^2},$$

and where $N_{[-1,1]}(m, s^2)$ denotes the Gaussian distribution truncated to the interval $[-1, 1]$.

In each case, we run our Gibbs sampler for 10^5 iterations, and discard the first 10^4 iterations as a burn-in period. Apart from the resampling scheme, and whether or not backward sampling is used, the only tuning parameter for the algorithm is the number of particles N in the particle Gibbs step.

6.1. First dataset

The first dataset we consider is simulated from the model, with $T + 1 = 400$, $\mu = 0$, $\rho = 0.9$, $\sigma = 0.5$.

Figure 1 reports the ACF (Autocorrelation function) of certain components of $(\theta, X_{0:T})$ for the four considered variants of our algorithm, for $N = 200$.

Clearly, the version which includes a backward sampling step performs best. The version based on systematic resampling comes second. This suggests that, in situations where backward sampling cannot be implemented, one may expect that using systematic resampling should be beneficial.

It is also worthwhile to look at the update rates of X_t with respect to t which is defined as the proportion of iterations where X_t changes value; see left panel of Figure 2. This figure reveals that backward sampling increases very significantly the probability of updating X_t to a new value, especially at small t values, to a point where this proportion is close to one. This also suggests that good performance for backward sampling might be obtained with a smaller value of N.

To test this idea, we ran the four variants of our Gibbs sampler, but with $N = 20$. The right side of Figure 2 reveals the three non-backward sampling algorithms provide useless results because components of $X_{0:300}$ hardly ever change values. For the same reasons, the ACFs of these variants do not decay at reasonable rate (which are not shown here).

To summarise, in this particular exercise, implementing backward sampling is very beneficial, as it leads to good mixing even if N is small. If backward sampling could not be implemented, then using systematic resampling may also improve performance, but not to same extent as backward sampling, as it may still require to take N to a larger value to obtain reasonable performance.

6.2. Second dataset

We consider a second dataset, simulated from the model with $T + 1 = 200$, $\mu = \log(5000)$, $\rho = 0.5$, $\sigma = 0.1$. (These values are close to the posterior expectation for the real dataset of [28].)

The interest of this example relative to the first one is twofold. Firstly, the positive impact of backward sampling is even bigger in this case. We have to increase N to $N = 1000$ to obtain non-zero update rates for the three variants that do not use backward sampling, whereas good update rates may be obtained for $N = 20$ for either multinomial or residual resampling, when backward sampling is used; see Figure 3.

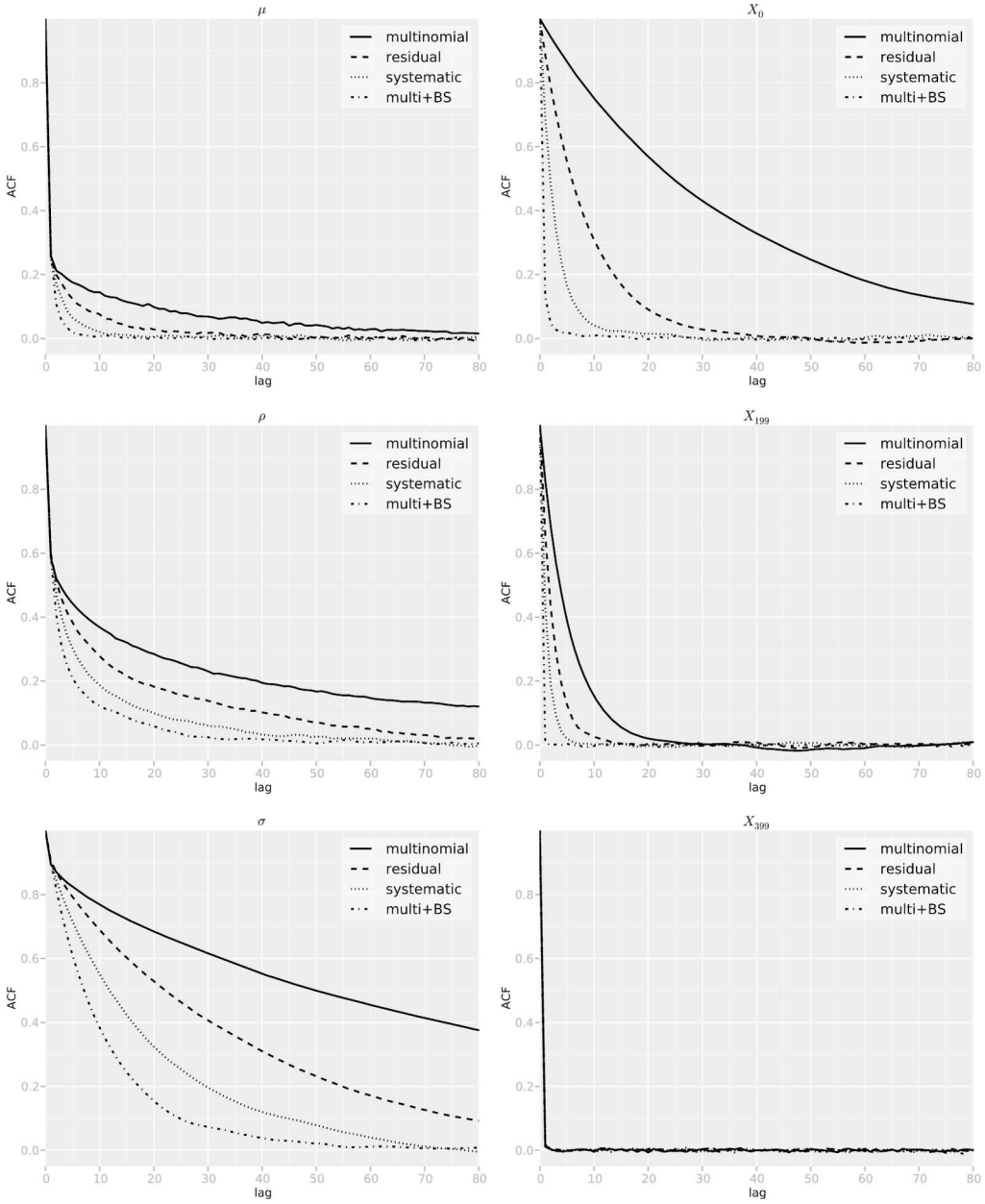

Figure 1. First dataset: ACF for different components of $(\theta, X_{0:T})$ and the four considered variants of particle Gibbs ($N = 200$).

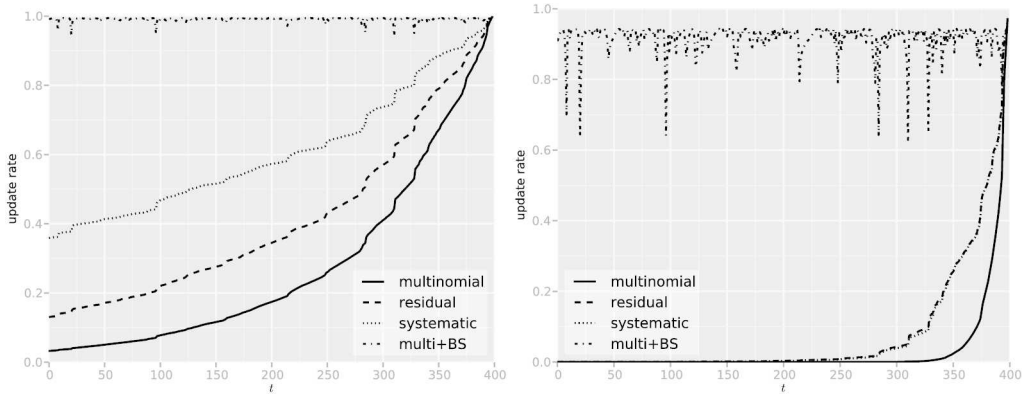

Figure 2. First dataset and resulting update rates of X_t versus $t \in 0:399$. Left plot is for $N = 200$ and right is for $N = 20$. For $N = 20$, forward only versions of systematic and residual perform similarly.

Secondly, we observe that backward sampling leads to excellent performance even when N is small, see also the ACF in Figure 4, which are close to the ACF of an independent process. Thus, the performance of that variant of particle Gibbs seems to on par with the algorithm of [28], which is specialised to this particular model (whereas particle Gibbs may be used in a more general class of models).

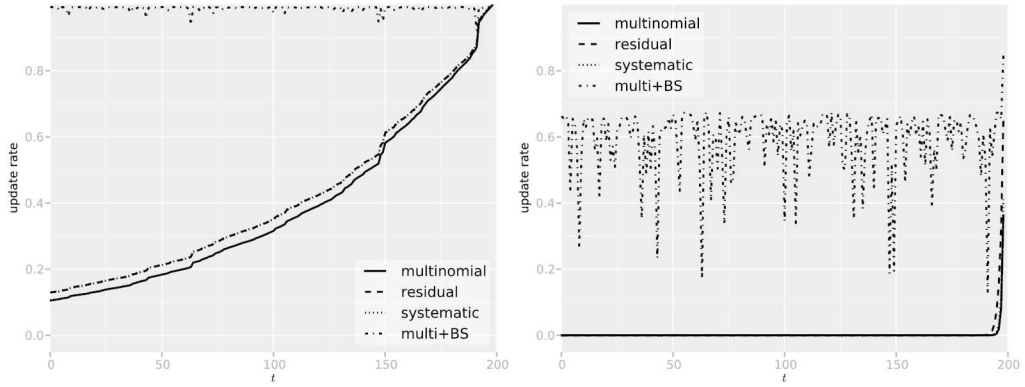

Figure 3. Second dataset, same plots as Figure 2, with $N = 1000$ (left panel) and $N = 20$ (right panel). Same legend as Figure 1. In the left plot, residual and systematic are largely indistinguishable. In the right plot, the three forward only schemes indistinguishable before $t \approx 190$.

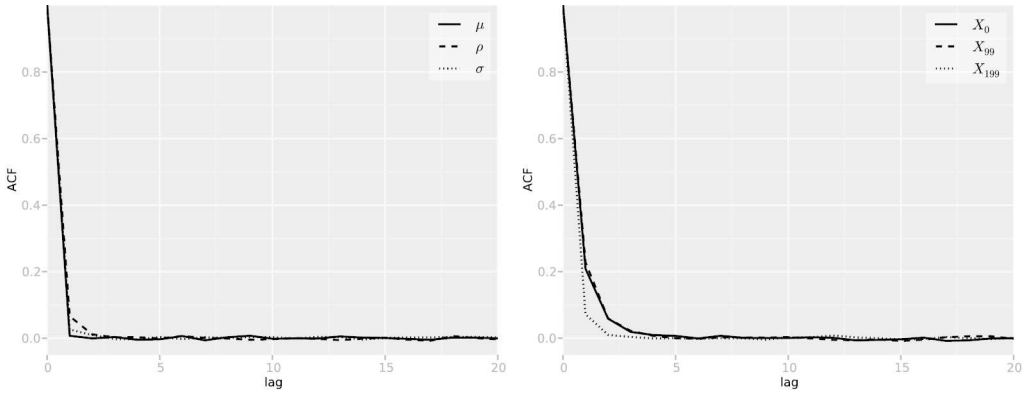

Figure 4. Second dataset: ACF for certain components of $(\theta, X_{0:T})$ for particle Gibbs with backward sampling ($N = 20$).

7. Discussion and conclusions

We now discuss the main practical conclusions that one can draw from our numerical studies.

First, they are many situations where backward resampling cannot be implemented, in particular when the probability density of the Markov transition is not tractable. In that case, our simulations suggests that one should run particle Gibbs with systematic resampling, as this leads to better mixing. A possible explanation is that, when only a forward pass is performed, the lower variability of systematic resampling makes it less likely that the proposed trajectories in the particle system coalesce with the fixed trajectory during the resampling steps. Therefore, the particle Gibbs step is more likely to output a trajectory which is different than the previous one.

Second, when backward sampling can be implemented, it should be used, as this makes it possible to set N to a significantly smaller value while maintaining good mixing; see also [15, 16] for similar findings.

In all cases, we recommend inspecting (on top of ACF plots) the same type of plots as in Figures 2 and 3, that is, update rate of X_t versus t, in order to assess the mixing of the algorithm, and in particular to choose a value of N that is a good trade-off between mixing properties and CPU cost. An interesting and important theoretical line of research would be to explain why this update rate seems more or less constant when backward sampling is used, while it deteriorates (while going backward in time) when backward sampling is not implemented. Another line for further research would be to study the effect of replacing the backward sampling step by a *forward-only* ancestor sampling step as recently proposed by [15].

Acknowledgements

We thank the editor and the referees for their insightful comments and helpful suggestions regarding the presentation of the paper. S.S. Singh's research was partly funded by the Engineering and Physical Sciences Research Council (EP/G037590/1) whose support is gratefully acknowledged.

References

[1] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 269–342. MR2758115

[2] Andrieu, C. and Vihola, M. (2012). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. Available at arXiv:1210.1484.

[3] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. New York: Springer. MR2159833

[4] Carpenter, J., Clifford, P. and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar, Sonar Navigation* **146** 2–7.

[5] Chopin, N., Jacob, P.E. and Papaspiliopoulos, O. (2013). SMC^2: An efficient algorithm for sequential analysis of state space models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 397–426. MR3065473

[6] Chopin, N. and Singh, S. (2013). On the particle Gibbs sampler. Available at arXiv:1304.1887.

[7] Del Moral, P. (1996). Nonlinear filtering: Interacting particle solution. *Markov Process. Related Fields* **2** 555–579. MR1431187

[8] Del Moral, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and Its Applications (New York)*. New York: Springer. MR2044973

[9] Doucet, A., de Freitas, N. and Gordon, N., eds. (2001). *Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science*. New York: Springer. MR1847783

[10] Everitt, R.G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Statist.* **21** 940–960. MR3005805

[11] Golightly, A. and Wilkinson, D. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1** 807–820.

[12] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[13] Jacob, P.E., Murray, L. and Rubenthaler, S. (2013). Path storage in the particle filter. Available at arXiv:1307.3180.

[14] Launay, T., Philippe, A. and Lamarche, S. (2013). On particle filters applied to electricity load forecasting. *J. SFdS* **154** 1–36. MR3120434

[15] Lindsten, F., Jordan, M.I. and Schön, T.B. (2012). Ancestor sampling for particle Gibbs. Available at arXiv:1210.6911.

[16] Lindsten, F. and Schön, T.B. (2012). On the use of backward simulation in the particle Gibbs sampler. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 3845–3848. Kyoto: IEEE.

[17] Lindvall, T. (1992). *Lectures on the Coupling Method. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. New York: Wiley. MR1180522

[18] Liu, J.S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. MR1649198

[19] Mira, A. and Geyer, C. (1999). Ordering Monte Carlo Markov chains. Technical report, School of Statistics, Univ. Minnesota.

[20] Peters, G., Hosack, G. and Hayes, K. (2010). Ecological non-linear state space model selection via adaptive particle Markov chain Monte Carlo. Available at arXiv:1005.2238.

[21] Pitt, M.K., Silva, R.d.S., Giordani, P. and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics* **171** 134–151. MR2991856

[22] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[23] Silva, R., Giordani, P., Kohn, R. and Pitt, M. (2009). Particle filtering within adaptive Metropolis–Hastings sampling. Preprint. Available at arXiv:0911.0230.

[24] Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401

[25] Vrugt, J.A., ter Braak, C.J., Diks, C.G. and Schoups, G. (2014). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources*. To appear.

[26] Whiteley, N. (2010). Discussion of "Particle Markov chain Monte Carlo methods" by Andrieu et al. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 306–307.

[27] Whiteley, N., Andrieu, C. and Doucet, A. (2010). Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. Available at arXiv:1011.2437.

[28] Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question – An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Statist.* **20** 531–570. MR2878987

Received January 2014