arXiv:1304.1887v1 [stat.CO] 6 Apr 2013

ON THE PARTICLE GIBBS SAMPLER

N. CHOPIN AND S.S. SINGH

Abstract. The particle Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm which operates on the extended space of the auxiliary variables generated by an interacting particle system. In particular, it samples the discrete variables that determine the particle genealogy. We propose a coupling construction between two particle Gibbs updates from different starting points, which is such that the coupling probability may be made arbitrary large by taking the particle system large enough. A direct consequence of this result is the uniform ergodicity of the Particle Gibbs Markov kernel. We discuss several algorithmic variations of Particle Gibbs, either proposed in the literature or original. For some of these variants we are able to prove that they dominate the original algorithm in asymptotic efficiency as measured by the variance of the central limit theorem's limiting distribution. A detailed numerical study is provided to demonstrate the efficacy of Particle Gibbs and the proposed variants.

1. Introduction

PMCMC (Particle Markov chain Monte Carlo) is a new set of MCMC algorithms discovered by Andrieu et al. (2010) which has attracted considerable attention in Statistics. It has in a relatively short span of time generated a series of diverse research contributions, whether methodological (Silva et al., 2009; Whiteley et al., 2010; Chopin et al., 2012; Lindsten et al., 2012) or applied, the latter in domains as diverse as Ecology (Peters et al., 2010), Electricity Forecasting (Launay et al., 2012), Finance (Pitt et al., 2012), systems Biology (Golightly and Wilkinson, 2011), study of social networks (Everitt, 2012), Hydrology (Vrugt et al., 2012), among other fields. 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 most cases a non demanding requirement, in contrast to previous approaches based on standard MCMC.

The specificity of PMCMC is that each MCMC step includes the generation of a complete interacting particle system, that is, the set of random variables generated when running a particle filtering (also known as Sequential Monte Carlo (SMC)) algorithm; see Doucet et al. (2001) Del Moral (2004) and Cappé et al. (2005) for general references on particle filtering. Several instances of PMCMC may be understood as *exact Monte carlo approximations* of an ideal algorithm, that is to say they may be analysed as noisy versions of an ideal algorithm where some intractable quantity in the ideal algorithm is replaced by an unbiased Monte carlo estimate, which is computed using the interacting particle system. This is the approach followed in Andrieu and Vihola (2012). The term exact highlights the fact that, despite being an approximation of an ideal algorithm, PMCMC samples exactly from the distribution of

Key words and phrases. Feynmann-Kac formulae, Gibbs sampling, Particle filtering, Particle Markov chain Monte Carlo, Sequential Monte Carlo.

interest. However, this interpretation does not appear to be suitable for those variants of PMCMC that involves a Particle Gibbs (also known as conditional Sequential Monte Carlo (CSMC)) step. While the Particle Gibbs step 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.

One of the objectives of this paper is to undertake a theoretical study of Particle Gibbs. We design a coupling construction between two Particle Gibbs updates that start from two different trajectories and using this coupling construction, we are able to establish that the coupling probability may be made arbitrary large if the number N of particles is large enough. A direct consequence is the uniform geometric ergodicity of the Particle Gibbs Markov transition kernel.

We also derive several variants of the Particle Gibbs algorithm by considering alternative resampling schemes, which is an unexplored and important facet of PMCMC. We also discuss backward sampling, an extra step for Particle Gibbs proposed by Whiteley (2010). We show that all these variants of Particle Gibbs (including the original algorithm) define reversible kernels. Regarding backward sampling, we establish that the Particle Gibbs kernel which incorporates backward sampling dominates Particle Gibbs without it, in asymptotic efficiency as measured by the variance of the CLT. Finally, we present a detailed numerical comparison of all these different algorithms.

The plan of the paper is the following. Sections 2 and 3 set up respectively the notations and the considered Feynman-Kac model. Section 4 gives a probabilistic description of the corresponding particle system, and Section 5 derives the conditional distribution associated to the Particle Gibbs step. Section 6 describes the conditional resampling algorithms corresponding to either residual and systematic resampling. Section 7 develops the coupling construction mentioned above. Section 8 introduces a variant of Particle Gibbs which we call "forced move". Section 9 discusses the BS (backward sampling) step, establishes that Particle Gibbs, either with or without an extra BS step, defines a reversible kernel, and uses this result to establish that Particle Gibbs with BS dominates Particle Gibbs without. Section 10 is a numerical study of the several variants of Particle Gibbs algorithm. Section 11 concludes. The Appendix contains the proofs of technical results not included in the main body of the paper.

2. Notations

For $m \leq n$, we denote by $m:n$ the range of integers $\{m, \ldots, n\}$, and we use extensively the semi-colon short-hand for collections of random variables, e.g. $X_{0:T} = (X_0, \ldots X_T)$, $X_t^{1:N} = (X_t^1, \ldots, X_t^N)$, and even in an nested form, $X_{0:T}^{1:N} = (X_0^{1:N}, \ldots, X_T^{1:N})$; 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, e.g. $x_{0:t}$ or $x_t^{1:N}$. The sub-vector containing the t first components of some vector Z_T is denoted $[Z_T]_t$.

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 distribution 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 notations: $\mu(f) = \int_{\mathcal{X}} d\mu\, f$, Kf is the application $x \to \int_{\mathcal{X}} K(x, dx')f(x')$, and μK is the probability measure $(\mu K)(A) = \int_{\mathcal{X}} \mu(dx)K(x, A)$. The atomic measure at $a \in \mathcal{X}$ is denoted $\delta_a(dx)$. We denote by $\mu \otimes K$ the measure $\mu(dx)K(x, dx')$ on the product space $\mathcal{X} \times \mathcal{X}$.

3. Model

Let $(X_t)_{t\geq 0}$ be a discrete-time $\mathcal{X}$-valued Markov chain, with initial law $m_0(dx_0) = m_0(x_0)\, dx_0$, and transition law $m_t(x_{t-1}, dx_t) = m_t(x_{t-1}, x_t)\, dx_t$. Let $(G_t)_{t\geq 0}$ be a sequence of $\mathcal{X} \to \mathbb{R}^+$ potential functions. In the context of hidden Markov models, $G_t(x_t) = g(x_t, y_t)$, the density (with respect to some dominating measure dy) of observation y_t of the $\mathcal{Y}$-valued random variable Y_t, conditional on state $X_t = x_t$.

It is convenient in this work to rewrite this Feynman-Kac model as a Feynman-Kac path model, that is a model operating on trajectories. Thus, we define $z_t = x_{0:t}$, $Z_t = X_{0:t}$, 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 $q_0(dz_0) = m_0(dx_0)$, and transition

$$q_t(z_{t-1}, dz'_t) = \delta_{z_{t-1}}(dx'_{0:t-1})\, m_t(x_{t-1}, x'_t)dx'_t$$

i.e. keep all of z_{t-1} and append new state x_t, from Markov transition $m_t(x_{t-1}, dx_t)$. The associated Feynman-Kac path measures are then traditionally defined as

$$\mathbb{Q}_t(dz_t) = \mathbb{Q}_t(dx_{0:t}) = \frac{1}{\mathcal{Z}_t} m_0(dx_0) \prod_{s=1}^{t} \{G_{s-1}(x_{s-1})m_s(x_{s-1}, dx_s)\} \tag{3.1}$$

where $\mathcal{Z}_t > 0$ is defined as

$$\mathcal{Z}_t = \int_{\mathcal{X}^{t+1}} m_0(dx_0) \prod_{s=1}^{t} \{G_{s-1}(x_{s-1})m_s(x_{s-1}, dx_s)\}\,.$$

4. Probabilistic description of the interacting particle system

The aim of this section is to derive the joint distribution of all the random variables generated in the course of the execution of an interacting particle algorithm that targets the Feynman-Kac path measures given in the previous section.

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

$$\mathbb{Q}_t^N(dz_t) = \frac{1}{N} \sum_{n=1}^{N} \delta_{Z_t^n}(dz_t)$$

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 $m_0(x_0)dx_0$. 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

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

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 $\varrho_t(z_t^{1:N}, dA_t^{1:N})$ are the ancestor variables, i.e. A_t^n is the label of the particle at time t which generated particle Z_{t+1}^n at time $t+1$.

A common assumption on the resampling distribution ϱ_t is the following.

Assumption (MR)**.** *The resampling distribution $\varrho_t(z_t^{1:N}, da_t^{1:N})$, $t \geq 0$, has probability density (abusing notations and using the same symbol ϱ_t for the probability density):*

$$\varrho_t(z_t^{1:N}, a_t^{1:N}) = \prod_{n=1}^{N} W_t^{a_t^n}(z_t^{1:N})$$

where the W_t^n's are the normalised weights

$$W_t^n(z_t^{1:N}) \stackrel{\Delta}{=} \frac{G_t(z_t^n)}{\sum_{m=1}^N G_t(z_t^m)}, \quad n \in 1:N. \tag{4.1}$$

In other words, the A_t^n's are independent draws from $\mathcal{M}\left(W_t^{1:N}(z_t^{1:N})\right)$.

Analysis under Multinomial resampling, which is an important sampling strategy in the literature on particle filtering, is much simpler than under the more general strategies below. That is, the validity of the particle Gibbs sampler may be established for other resampling schemes, provided they follow the more general assumption (R).

An $N-$cycle $c : (1:N) \to (1:N)$ is a permutation such that $c(n) \equiv (n+k) \pmod N$, for all $n \in 1:N$, and some $k \in 1:N$. Note that if c is a cycle then so is c^{-1}.

Assumption (R)**.** *The discrete distribution $\varrho_t(z_t^{1:N}, da_t^{1:N}) = \varrho_t(z_t^{1:N}, a_t^{1:N}) da_t^{1:N}$, which takes values in $(1:N)^{\otimes N}$, is (a) marginally unbiased, that is, the marginal probability that $A_t^n = m$ is $W_t^m(z_t^{1:N})$, for $m, n \in 1:N$; and (b) is cycle-invariant, that for any two $N-$cycles c, c' we have $\varrho_t(z_t^{1:N}, a_t^{c'(1:N)}) = \varrho_t(z_t^{1:N}, a_t^{1:N})$ and $\varrho_t(z_t^{1:N}, a_t^{1:N}) = \varrho_t(z_t^{c(1:N)}, c^{-1}(a_t^{1:N}))$.*

Assumption (Rb) is a technical condition that facilitates the formalisation of Particle Gibbs. The first part of (Rb) states that a sample from $\varrho_t(z_t^{1:N}, \cdot)$ can be arbitarily cycled without changing its law. The second part of (Rb) states that if $A_t^{1:N}$ is a sample from $\varrho_t(z_t^{1:N}, \cdot)$ then $c^{-1}(A_t^1), \ldots, c^{-1}(A_t^N)$ is a sample from $\varrho_t(z_t^{c(1:N)}, \cdot)$. We will discuss both these points in more in detail in Section 6. For now, we note that, under Assumption (R), we can write, for any $n \in 1:N$,

$$\varrho_t(z_t^{1:N}, a_t^{1:N}) = W_t^{a_t^n}(z_t^{1:N}) \varrho_t^c(z_t^{1:N}, a_t^{1:N\setminus n} | A_t^n = a_t^n), \tag{4.2}$$

where $\varrho_t^c(z_t^{1:N}, a_t^{1:N\setminus n} | A_t^n = a_t^n)$ is the probability density of $A_t^{1:N\setminus n}$, the collection of A_t^m variables for $m \neq n$, conditional on $A_t^n = a_t^n$.

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

$$\vartheta_T^N\left(dz_{0:T}^{1:N}, da_{0:T-1}^{1:N}\right) = m_0^{\otimes N}(dz_0^{1:N}) \prod_{t=1}^{T} \left\{ \varrho_{t-1}(z_{t-1}^{1:N}, da_{t-1}^{1:N}) \prod_{n=1}^{N} \left[q_t(z_{t-1}^{a_{t-1}^n}, dz_t^n) \right] \right\}.$$

Proposition 1. *Under Assumption (Ra), one has*

$$\mathbb{E}_{\vartheta_T^N}\left[\prod_{t=0}^{T} \left\{ \frac{1}{N} \sum_{n=1}^{N} G_t(Z_t^n) \right\} \right] = \mathcal{Z}_T. \tag{4.3}$$

This result has been established previously in e.g. Del Moral (2004) under Assumption (MR) but the author only really used the marginal unbiased property as formalised in Assumption (Ra).

5. The Conditional Particle Filter

This section is concerned with the definition of the Markov kernel of the Particle Gibbs algorithm that has $\mathbb{Q}_t(dz_t)$ as its invariant measure; i.e. the Markov kernel operates on the path space in the sense that it maps $\mathcal{X}^{T+1} \to \mathcal{P}(\mathcal{X}^{T+1})$. We call this Markov kernel the CPF (Conditional Particle Filter) kernel for reasons to be made obvious below.

In order to define the CPF kernel and prove it leaves $\mathbb{Q}_t(dz_t)$ invariant, we commence first with the definition of the following extended distribution π_T^N whose sampling space is the sampling space of ϑ_T^N augmented to include a discrete random variable $N^\star \in 1:N$,

$$\begin{aligned}\pi_T^N\left(dz_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star\right) = &\frac{1}{\mathcal{Z}_T} m_0^{\otimes N}(dz_0^{1:N}) \\ &\times \prod_{t=1}^{T}\left\{\left[\frac{1}{N}\sum_{n=1}^{N} G_{t-1}(z_{t-1}^n)\right]\varrho_{t-1}(z_{t-1}^{1:N}, da_{t-1}^{1:N})\prod_{n=1}^{N}\left[q_t(z_{t-1}^{a_{t-1}^n}, dz_t^n)\right]\right\} \\ &\times \frac{1}{N} G_T(z_T^{n^\star}).\end{aligned} \tag{5.1}$$

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 (4.3); note that cycle invariance is not needed to show π_T^N integrates to one.

Proposition 2. *Under Assumption (Ra), 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 has been established by Andrieu et al. (2010) under Assumption (MR), but extension to Assumption (Ra) is trivial. In particular, the expectation of functions of $Z_T^\star$ may be computed by integrating out the variables in the order $n^\star, x_T^{1:N}, a_T^{1:N}, \ldots, x_1^{1:N}, a_0^{1:N}, x_0^{1:N}$. Note that cycle invariance is not needed in this proposition.

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 backwards 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 $Z_T^\star$.

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:

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

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:

$$\pi_T^N(dz_{0:T}^{1:N\setminus\star}, da_{0:T-1}^{1:N\setminus\star}, dz_{0:T}^\star, db_{0:T-1}^\star, dn^\star) = \tag{5.2}$$
$$\frac{1}{N^{T+1}}(db_{0:T-1}^\star dn^\star)\mathbb{Q}_T(dz_T^\star)\prod_{t=0}^{T-1}\delta_{\left([z_T^\star]_{t+1}\right)}(dz_t^\star)$$
$$\times\prod_{n\neq b_0^\star} m_0(dz_0^n)\prod_{t=1}^{T}\left\{\varrho_{t-1}^c(z_{t-1}^{1:N}, da_{t-1}^{1:N\setminus\star}|A_{t-1}^{b_t^\star}=b_{t-1}^\star)\prod_{n\neq b_t^\star}\left[q_t(z_{t-1}^{a_{t-1}^n}, dz_t^n)\right]\right\}.$$

Passage from (5.1) to (5.2) 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 $\pi_T^N(db_{0:T-1}^\star, dn^\star)$; 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

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

Note that Steps 1 and 2 are both Gibbs step with respect to (5.2). Step 3 is to resample the index $N^\star$; hence $N^\star$ is sampled from $\pi_T^N(dn^\star|z_{0:T}^{1:N}, a_{0:T-1}^{1:N}) = \mathcal{M}(W_T^{1:N}(z_T^{1:N}))$, which is also a Gibbs step, but this time with respect to (5.1); 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 $\mathbb{Q}_T$. As we detail below, there is no specific difficulty in performing Step 2, which is pretty much equivalent to the problem of generating a particle filter for T time steps while Step 3 is evidently trivial. We show Step 1 is actually redundant by proving below that Steps 2 and 3 (applied in succession), for a fixed $(B_{0:T-1}^\star, N^\star)$, say $(1,\ldots,1)$, induces a Markov kernel, which we call the CPF kernel, that maps $\mathcal{X}^{T+1} \to \mathcal{P}(\mathcal{X}^{T+1})$ with $\mathbb{Q}_T$ as the invariant measure; Assumption (Rb) is crucial here. We now define the CPF Markov kernel.

Given a realization $z_T^\star$ for $Z_T^\star \sim \mathbb{Q}_T$, we initialize the CPF kernel by arbitrarily choosing a realization $(b_{0:T-1}^\star, n^\star)$ for $(B_{0:T-1}^\star, N^\star)$, which we always set to $(1,\ldots,1)$. Moving $z_T^\star$ through CPF kernel may be decomposed into two steps (which were Steps 2 and 3 above):

CPF-1: The first step where the $N-1$ remaining trajectories are generated, conditional on the trajectory $Z_T^\star$, kept as "frozen", and assigned the ancestry $(B_{0:T-1}^\star, N^\star) = (1,\ldots,1)$. For $t \in 0:T-1$ let $z_t^\star = [z_T^\star]_{t+1}$. By direct inspection of (5.2), one sees that the first step of applying the CPF kernel is equivalent to sampling from the conditional distribution

$$\pi_T^N(dz_{0:T}^{2:N}, da_{0:T-1}^{2:N}|Z_{0:T}^1 = z_{0:T}^\star, A_{0:T-1}^1 = (1,\ldots,1), N^\star = 1) \tag{5.3}$$
$$= m_0^{\otimes(N-1)}(dz_0^{2:N})\prod_{t=1}^{T}\varrho_{t-1}^c(z_{t-1}^{1:N}, da_{t-1}^{2:N}|A_{t-1}^1 = 1)\left[\prod_{n=2}^{N} q_t(z_{t-1}^{a_{t-1}^n}, dz_t^n)\right].$$

CPF-2: A second step where the index $N^\star$ is randomly regenerated to change the selected trajectory, which is the realization $z_T^\star$ we started of with, to $Z^{N^\star}$. The second step is a Gibbs step which updates component $N^\star$ from the distribution π_T^N conditional on all the other variables; hence $N^\star$ is sampled from $\mathcal{M}(W_T^{1:N}(z_T^{1:N}))$.

Sampling from the distribution of step CPF-1 is equivalent to running a particle algorithm with N particles, while maintaining one trajectory (with all labels equal to 1 in the expression above) as "frozen". One difficulty is to sample from the conditional resampling distribution ϱ_{t-1}^c. We return to this point in Section 6.

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 $z_T^\star \in \mathcal{X}^{T+1}$,

$$(P_T^N \varphi)(z_T^\star) = \int P_T^N(z_T^\star, dz_T')\varphi(z_T')$$

$$= \mathbb{E}_{\pi_T^N}\left\{ \frac{G_T(z_T^\star)}{G_T(z_T^\star) + \sum_{m=1}^{N-1} G_T(Z_T^m)}\varphi(z_T^\star) + \sum_{n=2}^{N} \frac{G_T(Z_T^n)}{G_T(z_T^\star) + \sum_{m=2}^{N} G_T(Z_T^m)}\varphi(Z_T^n) \right\}. \tag{5.4}$$

The following result shows that the ancestry $(B_{0:T-1}^\star, N^\star)$ of the starting trajectory in the CPF kernel can indeed be assigned in an arbitrary fashion.

Proposition 3. *Let P_T^N denote the CPF Markov kernel. Under Assumption (R), 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.*

Assumption (Rb), that is cycle invariance, makes it possible to skip the step that would sample $(B_{0:T-1}^\star, N^\star)$. Thus it follows from Proposition 3, and the discussion preceding the definition of steps CPF-1 and CPF-2, that P_T^N has invariant probability measure $\mathbb{Q}_T(dz_T)$, for any $N \geq 2$. To the best of our knowledge, this proposition is original.

Proof. [Proposition 3] The result is proven by showing that the following two choices for $(B_{0:T-1}^\star, N^\star)$ result in the image of $z_T^\star$ under P_T^N being unchanged. Choice one is $(b_{0:T-1}^\star, n^\star) = (1, \ldots, 1)$ while choice two is $(\check{b}_{0:T-1}^\star, \check{n}^\star) = (i_0, \ldots, i_T) \in (1:N)^{T+1}$. For the sake of clarity, all random variables for choice 2 are given the inverted hat superscript. Let $Z_{0:T}^{2:N}, A_{0:T-1}^{2:N}$ denote the random variables sampled in step CPF-1 for choice one. Let $\check{Z}_{0:T}^{1:N\setminus\star}, \check{A}_{0:T-1}^{1:N\setminus\star}$ denote the random variables to be sampled in step CPF-1 for choice two, that is, is a sample from $\pi_T^N(dz_{0:T}^{1:N\setminus\star}, da_{0:T-1}^{1:N\setminus\star} | z_{0:T}^\star, \check{b}_{0:T-1}^\star, d\check{n}^\star)$. We will show a specific deterministic transformation will convert samples from choice one to samples from choice 2 from which we can then straightforwardly obtain the stated result.

Recall, by definition, for $t \in 0:T$ we have $Z_t^1 = [z_T^\star]_{t+1}$. We define $c_0, \ldots, c_T$ be the sequence of cycles satisfying $c_t(i_t) = 1$ for all $t \in 0:T$. It follows that $\check{Z}_0^{1:N} = Z_0^{c_0(1:N)}$ is a sample from $\pi_T^N\left(dz_0^{1:N} | (\check{Z}_0^{i_0}, \ldots, \check{Z}_T^{i_T}) = z_{0:T}^\star, (\check{B}_{0:T-1}^\star, \check{N}^\star) = i_{0:T}\right)$. By (Rb), $\check{A}_0^{1:N} = c_0^{-1}(A_0^{c_1(1)}), \ldots, c_0^{-1}(A_0^{c_1(N)})$ is a sample from $\pi_T^N\left(da_0^{1:N} | \check{Z}_0^{1:N}, (\check{Z}_0^{i_0}, \ldots, \check{Z}_T^{i_T}) = z_{0:T}^\star, (\check{B}_{0:T-1}^\star, \check{N}^\star) = i_{0:T}\right)$. It now follows that $\check{Z}_1^{1:N} = Z_1^{c_1(1:N)}$ is a sample from $\pi_T^N\left(dz_1^{1:N} | \check{Z}_0^{1:N}, \check{A}_0^{1:N}, (\check{Z}_0^{i_0}, \ldots, \check{Z}_T^{i_T}) = z_{0:T}^\star, (\check{B}_{0:T-1}^\star, \check{N}^\star\right.$ Continuing in a similar fashion one shows that $\check{Z}_T^{1:N} = Z_T^{c_T(1:N)}$ is a sample from $\pi_T^N(dz_T^{1:N} | \check{Z}_{0:T-1}^{1:N}, \check{A}_{0:T-1}^{1:N}, (\check{Z}$ $z_{0:T}^\star, (\check{B}_{0:T-1}^\star, \check{N}^\star) = i_{0:T})$. The result now follows since after step CPF-2 the probabilities that either $Z_T^{N^\star}$ or $\check{Z}_T^{\check{N}^\star}$ belong to some measurable set $\mathcal{D}$ is the same by Assumption (Ra). □

6. Conditional resampling algorithms

We have seen in the previous Section that the CPF step involves sampling from the conditional resampling distribution ϱ_t^c. Under Assumption (MR) (multinomial resampling), sampling ϱ_t^c is trivial: draw $N-1$ times from the multinomial distribution that assigns

probability $W_t^n(z_t^{1:N})$ to outcome n, $n \in 1:N$. We describe in this section two alternative conditional resampling algorithms, one corresponding to the residual resampling (Liu and Chen, 1998), and the other corresponding to systematic resampling (Carpenter et al., 1999).

As already explained, it is sufficient to describe how to sample from $\varrho_t^c(z_t^{1:N}, da_t^{2:N}|A_t^1 = 1)$ for some fixed t and we shall use the short-hand W^n for the resampling weights $W_t^n(z_t^{1:N})$, defined in (4.1); note that $\sum_{n=1}^N W^n = 1$.

6.1. **Conditional residual resampling.** The residual resampling algorithm of Liu and Chen (1998) consists in creating N off-springs, based on the weights $W^{1:N}$, as follows. Let the residue $r^n = NW^n - \lfloor NW^n \rfloor$ for $n \in 1:N$, and $R = \sum_{n=1}^N r^n$. First, stack $\lfloor NW^n \rfloor$ "deterministic" copies of particle label n, sequentially in n. This gives a vector of $N - R$ labels. Then, extend that vector by appending R draws from $\mathcal{M}(r^{1:N}/R)$.

To make this algorithm fulfil Assumption (R), it is sufficient to permute randomly the so obtained vector of labels. In this way, the resulting algorithm is both marginally unbiased and cycle-invariant.

We now describe an algorithm to draw from the conditional distribution $\varrho_t^c(z_t^{1:N}, da_t^{2:N}|A_t^1 = 1)$, using the same notations, i.e. $r^n = NW^n - \lfloor NW^n \rfloor$, $R = \sum_{n=1}^N r^n$. We observe first that, under residual resampling, the probability that A_t^1 is set to one of the $\lfloor NW_t^1 \rfloor$ "deterministic" copies of label 1 is $\lfloor NW^1 \rfloor / NW^1$. This observation leads to the following algorithm, which generates a vector $A_t^{1:N}$ of N labels, with label one in slot one, $A_t^1 = 1$.

(1) With probability $\lfloor NW^1 \rfloor / NW^1$, perform standard residual resampling (without random permutation). Note that in that case, the first label is necessarily equal to 1 (since $NW_1 \geq 1$).
(2) Otherwise (with probability r^1/NW^1), perform the following modified residual resampling scheme: put label 1 in slot one, then stack recursively $\lfloor NW_t^n \rfloor$ deterministic copies of particle label n, sequentially in $n \in 1, \ldots, N$. Fill the $(R-1)$ remaining slots by independent draws from $\mathcal{M}(r^{1:N}/R)$.
(3) Randomly permute the $N-1$ labels in positions $2:N$.

Step 1 to Step 3 formally describes how to sample from $\varrho_t^c(z_t^{1:N}, da_t^{2:N}|A_t^1 = 1)$, when $\varrho_t(z_t^{1:N}, da_t^{1:N})$ is the joint distribution of labels obtained by residual resampling, then random permutation. However, Step 3 does not need to be implemented in practice, because (provided residual resampling is used at every iteration) the actual order of particles with labels $2:N$ does not play any role in the operations performed at the following iterations.

6.2. **Conditional systematic resampling.** The systematic resampling algorithm of Carpenter et al. (1999) consists in creating N off-springs, based on the weights $W^{1:N}$, as follows. Let U a uniform variate in $[0, 1]$, and consider a segment of length N, made of N successive segments of length NW^n, $n = 1, \ldots, N$. Take $s = U$, and repeat the following until $s > N$: determine which segment n contain s, add one off-spring to particle n, and increment s, $s \leftarrow s + 1$.

A particular difficulty of this resampling scheme is that it is order-dependent. Even if we would randomly permute the labels at the end of the algorithm (like in residual sampling), we may still have $\varrho_t(z_t^{1:N}, da_t^{1:N}) \neq \varrho_t(z_t^{\sigma(1:N)}, da_t^{1:N})$ for a certain permutation $\sigma : (1:N) \to (1:N)$. Instead, we propose to include a a cycle randomisation step in systematic resampling; that is, we take $A_t^{1:N} = \bar{A}_t^{c(1:N)}$, where c is a $N-$cycle (chosen uniformly among the N possible

cycles) and $\bar{A}_t^{1:N}$ is the vector of labels obtained from systematic resampling. It is trivial to see that that the corresponding algorithm fulfils Assumption (Ra), i.e. it is marginally unbiased, and that (Rb), cycle-invariance, is also satisfied. (Rb) could be established as follows: it is a property of standard systematic sampling without cycling that if $A_t^{1:N}$ is a sample from $\varrho_t(z_t^{1:N}, \cdot)$ then $c^{-1}(A_t^{c(1)}), \ldots, c^{-1}(A_t^{c(N)})$ is a sample from $\varrho_t(z_t^{c(1:N)}, \cdot)$. The result now follows from the cycle randomization step.

We now describe how one may generate $N-1$ labels from the conditional distribution $\varrho_t^c(z_t^{1:N}, da_t^{2:N}|A_t^1 = 1)$. One distinctive property of systematic resampling is that other than cycling the only randomness in the algorithm corresponds to the uniform variate U in $[0,1]$ drawn to commence the sampling. Sampling from $\varrho_t^c(z_t^{1:N}, da_t^{2:N}|A^1 = 1)$ amounts to sampling from the posterior distribution of U conditioned on the fact that after cycling $A_t^1 = 1$ (step 1 below), which is then followed by a step that realizes the Noff-springs from the sampled U as in standard systematic resampling, and finally a step (step 3) that chooses a cycle uniformly from the set of cycles that ensures $A_t^1 = 1$.

The number of off-springs of particle n is either $\lfloor NW^n \rfloor$ or $\lfloor NW^n \rfloor + 1$. Again, we call "deterministic copy" the $\lfloor NW^n \rfloor$ copies obtained by the $\lfloor NW^n \rfloor$ first jumps of cursor s in interval n, and we call "random copy" the extra copy which is generated with probability $r^n = NW^n - \lfloor NW^n \rfloor$. Conditional on $A_t^1 = 1$, the probability that a deterministic copy of 1 was put in slot 1 is $\lfloor NW_t^1 \rfloor / NW_t^1$. This observation leads to the following algorithm for generating from $\varrho_t^c(z_t^{1:N}, dA_t^{2:N}|A_t^1 = 1)$:

(1) If $NW^1 < 1$, sample U from $\mathcal{U}[0, NW^1]$ and go to step 2. Otherwise, with probability $r^1 \left(\lfloor NW^1 \rfloor + 1\right) / NW^1$ sample U from $\mathcal{U}[0, r^1]$, otherwise sample U from $\mathcal{U}[r^1, 1]$.
(2) Perform standard systematic resampling (without cycle randomisation and using the value of U sampled in Step 1). Call $\bar{A}^{1:N}$ the obtained vector of labels.
(3) Choose C uniformly from the set of cycles such that $\bar{A}^{C(1)} = 1$ and set $A^{1:N} = \bar{A}^{C(1:N)}$.

In the second part of step (1), U is being sampled from its law conditional on $A_t^1 = 1$. In step (3) C is sampled from its law conditional on U and $A_t^1 = 1$.

7. A coupling of the Particle Gibbs Markov Kernel

To establish the stability of the Feynman-Kac system, it is common to make the following assumption.

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,*

$$\int m_0(dx_0) G_0(x_0) \geq \frac{1}{g_0}, \qquad \inf_{x_{t-1}\in\mathcal{X}} \int m(x_{t-1}, dx_t) G_t(x_t) \geq \frac{1}{g_t}, \qquad t > 0.$$

The purpose of this section is to establish the following Theorem.

Theorem 4. *Under Assumptions (G) and (MR), for any* $\epsilon \in (0,1)$ *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*

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

This means that, for N large enough, the kernel P_T^N is arbitrary close to the independent kernel that sample from $\mathbb{Q}_T$. A direct corollary is that, again for N large enough, the kernel

P_T^N is uniformly ergodic (see e.g. Roberts and Rosenthal (2004) for a definition), with an arbitrary small ergodicity coefficient.

The proof of Theorem 4 is based on the following coupling construction: let $\bar{\pi}(dz_t^\star, d\check{z}_t^\star)$ be a joint distribution for the couple $(Z_t^\star, \check{Z}_t^\star)$, such that the marginal distribution of $Z_t^\star$, resp. $\check{Z}_t^\star$, is $P_T^N(x_{0:T}, dz_T^\star)$, resp. $P_T^N(\check{x}_{0:T}, d\check{z}_T^\star)$. Then

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

The following section describes the coupling we are using. Section 7.2 then establishes that this particular coupling ensures that

$$\mathbb{P}_{\bar{\pi}}(Z_T^\star \neq \check{Z}_T^\star) \leq \epsilon/2 \tag{7.1}$$

for N large enough, which concludes the proof.

7.1. **Coupling construction.** The coupling operates on the extended space corresponding to the support of the conditional distribution (5.3). The idea is to construct two conditional particle systems generated marginally from (5.3), 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 $\mathcal{C}_t^c = (1:N)\backslash\mathcal{C}_t$; by construction, $\mathcal{C}_t^c$ always contains 1, since the frozen trajectory is relabelled as trajectory 1 in (5.3). 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 $t \geq 0$,

$$\xi_{\mathcal{C}_t} = \sum_{n\in\mathcal{C}_t} \delta_{Z_t^n}(dz_t), \quad \xi_{\mathcal{C}_t^c} = \sum_{n\in\mathcal{C}_t^c} \delta_{Z_t^n}(dz_t), \quad \check{\xi}_{\mathcal{C}_t^c} = \sum_{n\in\mathcal{C}_t^c} \delta_{\check{Z}_t^n}(dz_t),$$

the following probability measures, $\mu_0(dz_0) = m_0(dz_0)$, and for $t \geq 1$,

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

where

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

the measures $\Psi_{G_{t-1}}(m_{C_{t-1}^c})(dz_{t-1})$ and $\Psi_{G_{t-1}}(\check{\xi}_{\mathcal{C}_{t-1}^c})(dz_{t-1})$ being defined similarly, and finally the constants

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

and the measures

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

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 $\left(A_{t-1}^n, \check{A}_{t-1}^n\right)$ 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 $m \in \mathcal{C}_{t-1}$), $Z_{t+1}^n \sim q_{t+1}(Z_t^{A_t^n}, dZ_{t+1})$, and take $\left(\check{A}_{t-1}^n, \check{Z}_t^n\right) = \left(A_{t-1}^n, Z_t^n\right)$. 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, $Z_t^n \sim q_t(Z_{t-1}^{A_{t-1}^n}, dZ_t)$, $\check{Z}_t^n \sim q_t(\check{Z}_{t-1}^{\check{A}_{t-1}^n}, dZ_t)$, 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 $\left(Z_t^n, \check{Z}_t^n\right)$ 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 $P_T^N(x_{0:T}, dz_T^\star)$ and $P_T^N(\check{x}_{0:T}, d\check{z}_T^\star)$ 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* 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.

7.2. **Proof of inequality** (7.1)**.** We now prove that the coupling construction described in the previous section is such that inequality (7.1) 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 $k_1, \ldots, k_T \in 1:N$,

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

Conditional on $Z_{T-1}^{1:N}$, and $n \in \mathcal{C}_T$, Z_T^n is an independent draw from $\mu_T(dz_T)$. Thus, by Hoeffding's inequality (Hoeffding, 1963),

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

where the last line again follows from Assumption (G). Thus

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

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

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

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.

7.3. **Technical lemma.**

Lemma 5. *Under Assumptions (G) and (MR), 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 $x_{0:T}$, $\check{x}_{0:T} \in \mathcal{X}^{T+1}$,*

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

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 $n \in \mathcal{C}_{t+1}$, i.e. that particles Z_{t+1}^n and $\check{Z}_{t+1}^n$ are coupled. Thus, and using the fact that, under assumption (R), the particle system is permutation invariant, one has:

$$\begin{aligned}
&\mathbb{P}\left(|\mathcal{C}_1| = N - k_1, \ldots, |\mathcal{C}_T| = N - k_T\right) \\
&= \left\{\prod_{t=1}^{T} \binom{N-1}{N-k_t}\right\} \mathbb{P}\left(\mathcal{C}_1^c = (1:k_1), \ldots, \mathcal{C}_T^c = (1:k_T)\right) \\
&= \int \prod_{t=0}^{T-1} \binom{N-1}{N-k_{t+1}} \left\{\prod_{n=2}^{k_t} \kappa_t(dz_t^n, d\check{z}_t^n) \prod_{n=k_t+1}^{N} \mu_t(dz_t^n)\right\} (1-\omega_t)^{(k_{t+1}-1)} \omega_t^{(N-k_{t+1})} \\
&\geq \int \prod_{t=0}^{T-1} \left\{\prod_{n=2}^{k_t} \kappa_t(dz_t^n, d\check{z}_t^n) \prod_{n=k_t+1}^{N} \mu_t(dz_t^n)\right\} \qquad (7.2) \\
&\qquad \times \mathbb{I}_{\mathcal{A}_t} \left\{\frac{(N-1)!(1-\omega_t)^{(k_{t+1}-1)}}{(N-k_{t+1})!}\right\} \left\{\frac{\omega_t^{(N-k_{t+1})}}{(k_{t+1}-1)!}\right\} \qquad (7.3)
\end{aligned}$$

with the convention that $k_0 = 1$, that $\mu_0(dz_0) = m_0(dz_0)$, and that empty products equal one, and defining the event $\mathcal{A}_t$ as

$$\mathcal{A}_t = \left\{\frac{\xi_{\mathcal{C}_t}(G_t)}{|\mathcal{C}_t|} \geq \frac{\mu_t(G_t)}{2}\right\}. \qquad (7.4)$$

Note that the integrals above must be understood as conditional probabilities, since μ_t and κ_t are conditional measures. In addition, the conditional distributions (first bracketed factor in the integral above) defining these conditional probabilities 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 general case where $T > 2$.

The key idea is to replace the two factors in the integrand of (7.2) with their large N values which we now define. Let

$$\Lambda_t = |\mathcal{C}_t| \times \frac{\xi_{\mathcal{C}_t^c}(G_t) \vee \check{\xi}_{\mathcal{C}_t^c}(G_t)}{\xi_{\mathcal{C}_t}(G_t)}, \qquad \text{for } |\mathcal{C}_t| > 0 \qquad (7.5)$$

and set $\Lambda_t = 0$ if $|\mathcal{C}_t| = 0$. Using Lemma 6, 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 (7.2) is larger than

$$\begin{aligned}
(1-\delta)^2 \int_{\mathcal{A}_0} \left\{\prod_{n=2}^{N} \mu_0(dz_0^n)\right\} \frac{e^{-\Lambda_0} \Lambda_0^{k_1-1}}{(k_1-1)!} \qquad (7.6) \\
\times \int_{\mathcal{A}_1} \left\{\prod_{n=2}^{k_1} \kappa_1\left(dz_1^n, d\check{z}_1^n\right) \prod_{n=k_1+1}^{N} \mu_1(dz_1^n)\right\} \frac{e^{-\Lambda_1} \Lambda_1^{k_2-1}}{(k_2-1)!}.
\end{aligned}$$

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

$$(7.7)\quad \int_{\mathcal{A}_0}\left\{\prod_{n=2}^{N}\mu_0(dz_0^n)\right\}\sum_{k_1=1}^{L_1}\frac{e^{-\Lambda_0}\Lambda_0^{k_1-1}}{(k_1-1)!}$$
$$\times\int_{\mathcal{A}_1}\left\{\prod_{n=2}^{k_1}\kappa_1(dz_1^n,d\check{z}_1^n)\prod_{n=k_1+1}^{N}\mu_1(dz_1^n)\right\}\sum_{k_2=1}^{L_2}\frac{e^{-\Lambda_1}\Lambda_1^{k_2-1}}{(k_2-1)!}\geq(1-\delta)^4.$$

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 (7.2), 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):

$$(7.8)\qquad 0\leq\Lambda_0\times\mathbb{I}_{\mathcal{A}_0}\leq 2g_0^2\times\mathbb{I}_{\mathcal{A}_0},\qquad 0\leq\Lambda_1\times\mathbb{I}_{\mathcal{A}_1}\leq 2g_1^2(k_1-1)\times\mathbb{I}_{\mathcal{A}_1}$$

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

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

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

$$\int_{\mathcal{A}_1}\left\{\prod_{n=2}^{k_1}\kappa_1(d(z_1^n,\check{z}_1^n))\prod_{n=k_1+1}^{N}\mu_1(dz_1^n)\right\}\sum_{k_2=1}^{L_2}\frac{e^{-\Lambda_1}\Lambda_1^{k_2-1}}{(k_2-1)!}$$
$$\geq(1-\delta)\int_{\mathcal{A}_1}\left\{\prod_{n=2}^{k_1}\kappa_1(d(z_1^n,\check{z}_1^n))\prod_{n=k_1+1}^{N}\mu_1(dz_1^n)\right\}.$$

By Hoeffding's inequality,

$$\int_{\mathcal{A}_1^c}\left\{\prod_{n=k_1+1}^{N}\mu_1(dz_1^n)\right\}\leq\exp\left(-2(N-k_1)\frac{\mu_1(G_1)^2}{4g_1^2}\right)\leq\exp\left(-(N-k_1)\frac{g_1^{-4}}{2}\right)$$

where the last inequality follows from Assumption (G). Using the same calculations for the first integral in (7.7) (that is applying Hoeffding's inequality to $\mathcal{A}_0$ and so on), one obtains eventually:

$$\int_{\mathcal{A}_0}\left\{\prod_{n=2}^{N}\mu_0(dz_0^n)\right\}\sum_{k_1=1}^{L_1}\frac{e^{-\Lambda_0}\Lambda_0^{k_1-1}}{(k_1-1)!}$$
$$\times\int_{\mathcal{A}_1}\left\{\prod_{n=2}^{k_1}\kappa_1(d(z_1^n,\check{z}_1^n))\prod_{n=k_1+1}^{N}\mu_1(dz_1^n)\right\}\sum_{k_2=1}^{L_2}\frac{e^{-\Lambda_1}\Lambda_1^{k_2-1}}{(k_2-1)!}$$
$$\geq(1-\delta)^2\left(1-\exp\left(-(N-L_1)\frac{g_1^{-4}}{2}\right)\right)\left(1-\exp\left(-(N-1)\frac{g_0^{-4}}{2}\right)\right).$$

and therefore, combining this with (7.6), 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 (7.2) by their large-N values.

Lemma 6. *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 arbitrary small, given that $\Lambda_t\mathbb{I}_{\mathcal{A}_t}$ is a bounded quantity; see (7.8). 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 $\mathbb{I}_{\mathcal{A}_t}=1$,

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

which can be clearly made arbitrary 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. □

8. Forced move

Step CPF-2 may be replaced by one that tries to force a selection of a trajectory different from the reference, in the spirit of the metropolized Gibbs sampler of Liu (1995). Specifically, CPF-2 is replaced by a Metropolis-Hastings step on the variable $N^\star$ instead of the Gibbs step. The proposal is concentrated on index $n=2:N$ with probability $W_T^n/(1-W_T^1)$, using the short-hand $W_T^n=W_T^n(z_T^{1:N})$. Thus, set $N^\star=n$ with probability

$$\frac{W_T^n}{1-W_T^1}\min\left(\frac{1-W_T^1}{1-W_T^n},1\right)$$

for $n\neq1$; otherwise set $N^\star=1$. We call this the variation of PG *forced move*. As in Liu (1995), one is able to establish that the forced move strategy always brings an improvement.

Proposition 7. *Assuming $0<G_t(x_t)\leq g_t$ for all $x_t\in\mathcal{X}$ and t, then forced move CPF dominates CPF in Peskun ordering.*

Proof. Recall that for Markov kernels P_1 and P_2, P_1 dominates P_2 in Peskun ordering iff $P_1(x, \mathcal{A}) \geq P_2(x, \mathcal{A})$ for all x and measurable sets $\mathcal{A}$ with $x \notin \mathcal{A}$. The assumption of the potentials are merely to ensure the PG kernel is well defined. Let $Z_{1:T}^{1:N}$ be the sequence of random variables generated in step CPF-1, $N^\star$ the outcome of CPF-2, and let $\check{N}^\star$ be the chosen trajectory of forced move. Then for $z_T \notin \mathcal{A}$

$$\begin{aligned}\mathbb{E}\left(\mathbb{I}_{\mathcal{A}}(Z_T^{\check{N}^\star})\middle| Z_{1:T}^{1:N}\right) &= \sum_{n=2}^{N} \frac{W_T^n}{1-W_T^1} \min\left(\frac{1-W_T^1}{1-W_T^n}, 1\right) \mathbb{I}_{\mathcal{A}}(Z_T^n) \\ &\geq \mathbb{E}\left(\mathbb{I}_{\mathcal{A}}(Z_T^{N^\star})\middle| Z_{1:T}^{1:N}\right).\end{aligned}$$

□

A way to quantify the improvement brought by a forced move is to note that Peskun ordering implies lag-one ordering and, assuming reversibility, efficiency ordering, see the discussion around Theorem 9 in Section 9.4 for a definition of these ordering of Markov kernels.

9. Backward sampling

It is convenient in this section to revert to standard notations based on the initial process X_t, rather than on notations based on trajectories $Z_t = X_{0:t}$. Thus, we now consider the following (extended) invariant distribution for the CPF kernel

$$\begin{aligned}\pi_T^N\left(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star\right) &= \frac{1}{\mathcal{Z}_T} m_0^{\otimes N}(dx_0^{1:N}) \\ &\times \prod_{t=1}^{T}\left\{\left[\frac{1}{N}\sum_{n=1}^{N} G_{t-1}(x_{t-1}^n)\right] \varrho_{t-1}(x_{t-1}^{1:N}, da_{t-1}^{1:N}) \prod_{n=1}^{N}\left[m_t(x_{t-1}^{a_{t-1}^n}, dx_t^n)\right]\right\} \\ &\times \frac{1}{N} G_T(x_T^{n^\star})\end{aligned} \tag{9.1}$$

which, compared to (5.1), represents a simple change of variables. Note that denoting the resampling distribution by $\varrho_t(x_t^{1:N}, da_t^{1:N})$ is an abuse of notations based on the fact that a resampling distribution depends only the weights $G_t(z_t^n) = G_t(x_t^n)$. In this new set of notations, the n-th 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 $B_T^\star = N^\star$, $B_t^\star = A_t^{B_{t+1}^\star}$, i.e. $Z_T^\star$ is a deterministic function of $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N}, N^\star)$.

Whiteley (2010), in his discussion of Andrieu et al. (2010), see also Lindsten and Schön (2012), 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

$$\begin{aligned}&\pi_T^N\left(A_t^b = a_t^b | X_{0:T}^{1:N} = x_{0:T}^{1:N},\, A_t^{-b} = a_t^{-b},\, A_{0:t-1}^{1:N} = a_{0:t-1}^{1:N},\, A_{t+1:T}^{1:N} = a_{t+1:T}^{1:N}, N^\star = n^\star\right) \\ &\qquad \propto \varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right) m_{t+1}(x_t^{a_t^b}, x_{t+1}^b)\end{aligned} \tag{9.2}$$

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 (9.2) , A_t^{-b} is $A_t^{1:N}$ minus A_t^b, a_t^{-b} is defined similarly, $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right)$ stands for the conditional distribution of A_t^b, given that $A_t^{-b} = a_t^{-b}$, relative to the joint distribution $\varrho_t(x_t^{1:N}, da_t^{1:N})$, and finally $m_{t+1}(x_t^{a_t^b}, x_{t+1}^b)$ is the probability density (relative to probability measure dx) of conditional distribution $m_{t+1}(x_t^{a_t^b}, dx_{t+1})$ 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 (9.2) 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 (9.1). As such, this extra step leaves $\mathbb{Q}_T(dx_{0:T})$ 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.

When this density is tractable, and since the support of (4.2) is $1:N$, sampling from this distribution is a $O(N)$ operation, provided one is able to compute simultaneously $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right)$ for all possible values $a_t^b \in 1:N$ in $O(N)$ time. When multinomial resampling is used, this quantity simplifies to $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right) = W_t^b(x_t^{1:N})$, since the A_t^n's are independent. We explain now how to compute this quantity when either residual resampling or systematic resampling is used.

9.1. **Backward sampling for residual resampling.** To describe how to compute the probability $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right)$ when residual resampling is used, we use the same short-hand notations as Section 6: for any $n \in 1:N$, W^n stands for $W_t^n(z_t^{1:N})$, $r^n = NW^n - \lfloor NW^n \rfloor$, $R = \sum_{n=1}^N r^n$, and in addition let $\chi^n(a_t^{1:N}) = \sum_{m=1}^N \mathbb{I}_{\{n\}}(a_t^m)$, $\chi^n(a_t^{-b}) = \sum_{m \neq b} \mathbb{I}_{\{n\}}(a_t^m)$; that is $\chi^n(a_t^{1:N})$, resp. $\chi^n(a_t^{-b})$, counts the number of off-springs of particle x_t^n within $a_t^{1:N}$, resp. a_t^{-b}.

Then, using standard properties of the multinomial distribution, it is easy to see that the algorithm described in Section 6.1 (that is residual resampling, plus random permutation) actually samples from the following joint distribution with density

$$\varrho_t(x_t^{1:N}, a_t^{1:N}) = \frac{R!}{N!} \prod_{n=1}^N \left(\frac{r^n}{R}\right)^{\{\chi^n(a_t^{1:N}) - \lfloor NW^n \rfloor\}} \frac{\{\chi^n(a_t^{1:N})\}!}{(\chi^n(a_t^{1:N}) - \lfloor NW^n \rfloor)!}. \tag{9.3}$$

Thus, provided a_t^{-b} is such that $\chi^n(a_t^{-b}) \geq \lfloor NW^n \rfloor$ for all $n \in 1:N$,

$$\varrho_t(x_t^{1:N}, A_t^b = m | A_t^{-b} = a^{-b}) \propto r^m \times \frac{\chi^m(a^{-b}) + 1}{\chi^m(a^{-b}) - \lfloor NW^m \rfloor + 1}, \tag{9.4}$$

for any $m \in 1:N$. In case there exists $n \in 1:N$ such that $\chi^n(a_t^{-b}) = \lfloor NW^n \rfloor - 1$, then one simply has

$$\varrho_t(x_t^{1:N}, A_t^b = m | A_t^{-b} = a^{-b}) = \mathbb{I}_n(m).$$

This reflects the fact that residual resampling always generates at least $\lfloor NW^n \rfloor$ off-springs of particle n for each $n \in 1:N$.

During the forward pass of the algorithm, one may compute quickly (that is, without changing the $O(N)$ complexity of the algorithm) quantities such as r^n, $\lfloor NW^n \rfloor$ and $\chi^n(a_t^{1:N})$

for all $n \in 1:N$. Then, since $\chi^n(a_t^{-b}) = \chi^n(a_t^{1:N}) - \mathbb{I}_{\{n\}}(a_t^b)$, computing (9.4) for all $m \in 1:N$ is also a $O(N)$ operation.

9.2. **Backward sampling for systematic resampling.** One difficulty in computing the probability $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right)$ that is specific to systematic resampling plus cycle randomisation, as described in Section 6.2, is that one needs to consider several cases, depending whether one or more than one cycles may have generated $A_t^{-b} = a_t^{-b}$. To outline the main ideas, we consider here only the case where the cycle is uniquely determined by inspection of a_t^{-b}, and defer other cases to Appendix A. This amounts to assuming that $a_t^{b-1} < a_t^{b+1}$. To avoid treating separately $b = 1$ and $b = N$, we abuse notations as follows: a_t^{b-1} refers to a_t^N when $b = 1$, and a_t^{b+1} to a_t^1 for $b = N$. This can be seen by recalling that $A_t^{1:N} = \bar{A}_t^{c(1:N)}$ (see Section 6.2) where $\bar{A}^{1:N}$ is the non-decreasing sequence taking values in $1:N$. Thus if $a_t^{b-1} < a_t^{b+1}$, c is determined by $c(m) = 1$, where a_t^m is the unique component of a_t^{-b} such that $a_t^{m-1} > a_t^m$. In that case, $\varrho_t\left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right) = \bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{b'} = a_t^{b'} | \overline{A}_t^{-b'} = \overline{a}_t^{-b'}\right)$, where the latter quantity is the probability that $\bar{A}_t^{b'} = a$, conditional on $\bar{A}_t^{-b'} = \bar{a}_t^{-b'}$, with $b' = c(b)$, which, in turn, may be deduced from the joint distribution of the labels before cycle randomisation $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{1:N} = \overline{a}_t^{1:N}\right)$ as follows:

$$\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{b'} = a_t^{b'} | \overline{A}_t^{-b'} = \overline{a}_t^{-b'}\right) \propto \bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{1:N} = \overline{a}_t^{1:N}\right),$$

i.e. compute the right hand side for all $a_t^{b'} \in 1:N$, and a fixed $\bar{a}_t^{-b'}$, then renormalise it to obtain the left hand side for all $a_t^{b'} \in 1:N$.

Let $S(0) = 0$, $S(n) = N\sum_{m=1}^n W^m$ for $n \in 1:N$. By construction, see Section 6.2, the probability $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{1:N} = \overline{a}_t^{1:N}\right)$ equals the probability that a uniform random variate $u \sim \mathcal{U}[0,1]$ simultaneously fulfil the following N inequalities, for $n \in 1:N$,

$$S(\bar{a}_t^n - 1) - (n-1) \leq u < S(\bar{a}_t^n) - (n-1). \tag{9.5}$$

In other words, $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{1:N} = \overline{a}_t^{1:N}\right)$ is the length of the interval defined as the intersection of $[0,1]$ and the N intervals defined above.

To efficiently compute $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{b'} = a_t^{b'} | \overline{A}_t^{-b'} = \overline{a}_t^{-b'}\right)$ for all $a_t^{b'} \in 1:N$, one may precompute the intersection $J(\bar{a}_t^{-b'})$ of $[0,1]$ and the $N-1$ intervals above for $n \neq b'$. Then computing $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{1:N} = \overline{a}_t^{1:N}\right)$ for all $a_t^b \in 1:N$ reduces to computing the intersection of $J(\bar{a}_t^{-b'})$ and the interval (9.5) for $n = b'$ and the considered value of $\bar{a}_t^b$. Thus, this backward step may be implemented in $O(N)$ time. Note that the $S(n)$'s have already been computed during the forward pass of the algorithm. Also, in practice $\bar{\varrho}_t\left(x_t^{1:N}, \overline{A}_t^{b'} = a_t^{b'} | \overline{A}_t^{-b'} = \overline{a}_t^{-b'}\right)$ needs to be computed only for a_t^b in the range $a_t^{b-1} \leq a_t^b \leq a_t^{b+1}$, as this probability is zero by construction outside this range. This remarks allows for an extra speed-up.

Again, we refer to Appendix A for how to implement backward sampling when $a_t^{b-1} \geq a_t^{b+1}$ (and therefore one need to take into account cycle uncertainty).

9.3. **Peskun ordering and backward sampling.** Peskun ordering was useful in establishing that forced move PG is preferable over PG and one might be tempted to think

that Peskun dominance of PG with backward sampling over PG could also be shown. The following counter-example shows otherwise.

Example 8. Let $\mathcal{X} = \mathbb{R}$, $(X_t)_{t\geq 0}$ be an i.i.d. sequence with marginal law $m(dx)$, and let the potentials be unit valued, i.e. $G_t(x_t) = 1$ 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 $\mathcal{B} = \mathcal{B}_0 \times \ldots \times \mathcal{B}_T \subset \mathcal{X}^{T+1}$, where $m(\mathcal{B}_t) = \epsilon$ 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 $x_{0:T}$, i.e.$P_T^N(x_{0:T}, \mathcal{B})$, 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 $\mathcal{B}$.

9.4. **Reversibility, covariance ordering and asymptotic efficiency.** Backward sampling (BS) does in practice bring improvement to the decay of the autocorrelation function of successive samples of $X_{0:T}$ generated by the Particle Gibbs sampler, i.e. more rapid decay compared to not implementing BS; see our numerical experiments in Section 10. However, how much improvement depends on the transition kernel $m_t(x_{t-1}, dx_t)$ of the hidden state process $(X_t)_{t\geq 0}$. If only $X_0 \sim m_0$ is random while $m_t(x_{t-1}, dx_t)$ 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 Roberts and Rosenthal (2004).

The following result due to Tierney (1998), see also Mira and Geyer (1999), 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 9. *(Tierney, 1998) 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 $f \in L^2(\pi) = \left\{f : \int \pi(dx) f(x)^2 < \infty\right\}$ let*

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

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 Tierney (1998) 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 Mira and Geyer (1999).

To prove efficiency odering, 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 autocorellation compared to the CPF kernel without BS, this property if holds is called *lag-one domination*.

Proposition 10. *Assume (R). The CPF kernel is reversible.*

Proof. This result is based on the equivalent representation of the CPF kernel derived in Proposition 3. Consider a $\mathbb{Q}_T(dz_T^\star) \otimes P_T^N(z_{T,}^\star d\check{z}_T^\star)-$measurable function h and let $(Z_{T,}^\star \check{Z}_T^\star) = (X_{0:T}^\star, \check{X}_{0:T}^\star) \sim \mathbb{Q}_T \otimes P_T^N$ then

$$\begin{aligned}
\mathbb{E}\left\{h(Z_{T,}^\star \check{Z}_T^\star)\right\} &= \int \mathbb{Q}_T(dx_{0:T}^\star) \int \pi_T^N(dx_{0:T}^{1:N\setminus\star}, da_{0:T-1}^{1:N\setminus\star}, db_{0:T-1}^\star, dn^\star | x_{0:T}^\star) \\
&\quad \times \int \pi_T^N\left(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}\right) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, d\check{n}^\star) \int \pi_T^N(dn^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \mathbb{E}\left\{h(\check{Z}_T^\star, Z_T^\star)\right\}
\end{aligned}$$

where the first equality follows from Proposition 3 and the remaining steps are a change of variables; again, $z_T^{n^\star}$, resp. $z_T^{\check{n}^\star}$, must be understood as a certain deterministic function of $(x_{0:T}^{1:N}, a_{0:T-1}^{1:N}, n^\star)$, resp. $(x_{0:T}^{1:N}, a_{0:T-1}^{1:N}, \check{n}^\star)$, in the equations above; see notation at the beginning of this Section. □

We now prove a similar result for CPF-BS, the CPF kernel with backward sampling. To that effect, we must first generalise Proposition 3 to the CPF-BS kernel. (Proof of this corollary uses the same construction in the proof of Proposition 3 and is omitted.)

Corollary 11. *Assume (R). Let $P_T^{N,B}$ denote the CPF-BS 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, i.e. Step CPF-1.*

Proposition 12. *Assume (R). The CPF-BS kernel (CPF with backward sampling) is reversible.*

Proof. A simple proof in the multinomial resampling case, i.e. Assumption (MR), is obtained by remarking that 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, i.e. replace $A_{0:T-1}^{1:N}$ by $\check{A}_{0:T-1}^{1:N}$, a sample from $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$. This is because under multinomial resampling, the A_t^n's are independent conditionally on $(X_{0:T}^{1:N}, N^\star)$. We defer the technical proof under the more general Assumption (R) to Appendix B.

Consider a $\mathbb{Q}_T(dz_T^\star)\otimes P_T^N(z_{T,}^\star d\check{z}_T^\star)$-measurable function h and let $(Z_{T,}^\star \check{Z}_T^\star) = (X_{0:T}^\star, \check{X}_{0:T}^\star) \sim \mathbb{Q}_T \otimes P_T^N$ then

$$\begin{aligned}
\mathbb{E}\left\{h(Z_{T,}^\star \check{Z}_T^\star)\right\} &= \int \mathbb{Q}_T(dx_{0:T}^\star) \int \pi_T^N\left(dx_{0:T}^{1:N\setminus\star}, da_{0:T-1}^{1:N\setminus\star} | (b_{0:T-1}^\star, n^\star) = (1,\ldots,1), x_{0:T}^\star\right) \\
&\quad \times \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) \int \pi_T^N(d\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \\
&\quad \times \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) \int \pi_T^N(d\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, d\overline{a}_{0:T-1}^{1:N}) \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) \pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star) \\
&\quad \times \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) \int \pi_T^N(d\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}, \check{n}^\star) h(z_T^{n^\star}, z_T^{\check{n}^\star}) \\
&= \mathbb{E}\left\{h(\check{Z}_T^\star, Z_T^\star)\right\}
\end{aligned}$$

where the second equality is a consequence of Corollary 11, the third 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 $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$, and the fourth equality is a simple change of variables. The simplification of $\pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N})$ into $\pi_T^N(d\check{n}^\star | x_{0:T}^{1:N})$ (second equality onwards) 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 under Assumption (MR).

Proposition 13. *Assume (MR). The CPF-BS kernel, denoted $P_T^{N,B}$, dominates the CPF kernel in lag one autocorrelation, i.e. let h be square integrable function* then

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

Proof. We use again the facts that $\pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N})$ reduces into $\pi_T^N(d\check{n}^\star | x_{0:T}^{1:N})$, 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 12.

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

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

Combining this result with Propositions 10 and 12 (reversibility of both kernels) and Theorem 9 from Tierney (1998), we conclude that the CPF-BS kernel dominates the CPF kernel in efficiency ordering (smaller asymptotic variances), at least when multinomial resampling is used, i.e. Assumption (MR).

For other resampling schemes, establishing lag-one dominance for all test functions does not seem to be easy, so one might consider an alternative measure of efficiency for MCMC algorithms. The Mean Square Jumping Distance (MSJD) is one such measure (e.g. Andrieu and Thoms, 2008) and in the following result we show that incorporating backward sampling will not decrease the MSJD for a certain class of additive test functions. Note that for Multinomial sampling we were able to prove a stronger version of the stated result below (Proposition 13) where the test function need not be of the additive structure required below but just square integrable with respect to π_T. The main reason why we are limited to the following weaker result when one only assume (R) is that for multinomial sampling $\varrho_t\left(x_t^{1:N},\, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}\right) = W_t^b(x_t^{1:N})$, i.e. A_t^n's are independent even when conditioning on $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N} \backslash A_t^n)$, the same though is not true for the systematic and residual sampling schemes.

Proposition 14. *Assume (R) and consider the following additive square integrable (with respect to π_T) function*

$$h(x_{0:T}, b_{0:T-1}, n) = h_0(x_{0:1}, b_{0:1}) + h_1(x_{1:2}, b_{1:2}) \cdots + h_{T-1}(x_{T-1:T}, b_{T-1}, n).$$

The CPF kernel with the additional step CPF-3 (backward sampling) has MSJD (mean square jumping distance) no less than the same kernel without step CPF-3, i.e.

$$E_{\pi_T^N \otimes P_T^{N,B}}\left\{\left(h(Z_T^\star, B_{0:T-1}^\star, N^\star) - h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star)\right)^2\right\} \geq E_{\pi_T^N \otimes P_T^N}\left\{\left(h(Z_T^\star, B_{0:T-1}^\star, N^\star) - h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star)\right)^2\right\}$$

where $\left((Z_T^\star, B_{0:T-1}^\star, N^\star), (\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star)\right)$ is distributed according to $\pi_T^N \otimes P_T^{N,B}$ in the first expectation and $\pi_T^N \otimes P_T^N$ in the second.

For the sake of generality, this results regards $P_T^{N,B}$ and P_T^N as Markov kernels with invariant measure $\pi_T^N(dz_T^\star, db_{0:T-1}^\star, dn^\star)$. By Corollary 11, omitting the dependency of h on $b_{0:T-1}, n$ in Proposition 14 will take us back to the usual setting where $\mathbb{Q}_T \otimes P_T^{N,B}$ is a measure on $\mathcal{X}^{T+1} \times \mathcal{X}^{T+1}$.

It is noteworthy that there are some models for which the update of the parameter θ depends on $x_{0:T}$ precisely through functions h of the above form; see the update for parameters μ, ρ and $1/\sigma^2$ in Section 10. Therefore maximising the MSJD for h in the stationary regime should benefit the exploration of these parameters. The proof of Proposition 14 may be found in Appendix C.

10. Numerical experiments

The focus of our numerical experiments is on comparing the different resampling schemes described in this paper (multinomial, residual, systematic), with or without backward sampling. The forced move step described in Section 8 is implemented.

We consider the following state-space model

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

for $t \in 0:T$, hence one may take $G_t(x_t) = \exp\{-e^{x_t} + y_t x_t\}$, where y_t is the observed value of Y_t. This model is motivated by Yu and Meng (2011) 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 (c) 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 notations $\tilde{x}_t = x_t - \mu$,

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

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.

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

Fig. 10.1 reports the ACF (Autocorrelation function) of certain components of $(\theta, X_{0:T})$ for the six considered variants of our algorithm, for $N = 200$.

One observes the following in this particular case. First, without backward sampling, there is a clear advantage of using systematic resampling over the two alternative resampling schemes. Second, the improvement brought over by backward sampling depends on the resampling scheme: it is small when systematic resampling is used, and quite important when either multinomial or residual resampling is used. Third, the fastest mixing seems to be obtained by multinomial resampling with backward sampling, and residual resampling with backward sampling.

Note that this ACF comparison does not take into account the relative CPU cost of each variant of Particle Gibbs. As a rule of thumb, one may consider that the different resampling

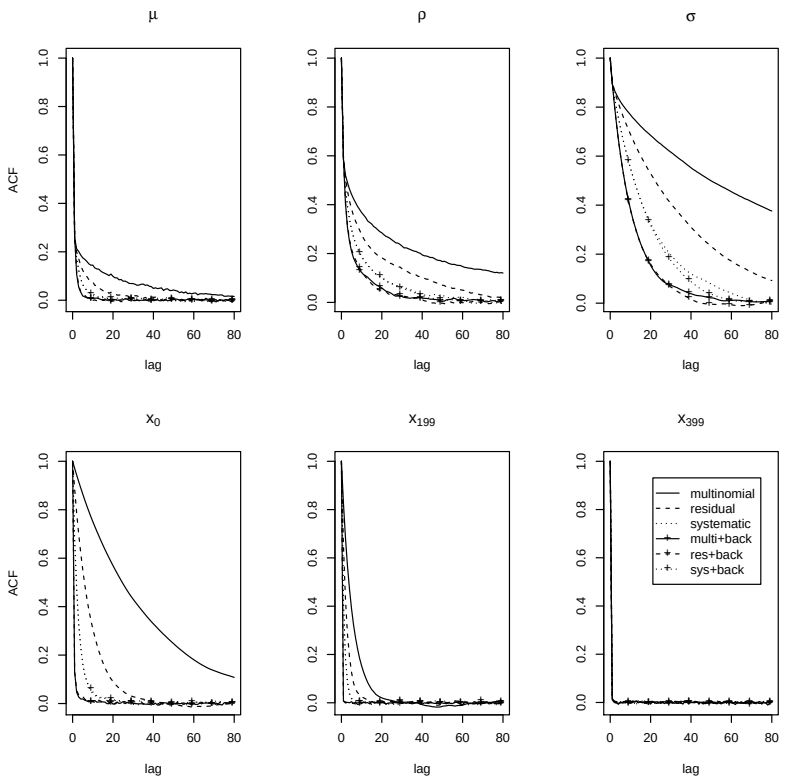


FIGURE 10.1. First dataset: ACF functions for different components of $(\theta, X_{0:T})$ and the six considered variants of Particle Gibbs ($N = 200$.) In all cases performance of multinomial with BS and residual with BS is hardly distinguishable. Also Systematic with and without backwards performs similarly, except maybe for σ.

schemes have essentially the same CPU cost. In fact, it is often the case that the resampling step represents a small part of the total CPU cost of a particle algorithm. On the other hand, backward sampling amounts to a backward pass through the data, which represents a significant extra cost, relative to the forward pass already performed during the generation of the particle system. However, this extra cost is of course highly model and implementation-dependent. In that respect, if CPU cost was taken into account, then systematic resampling without backward sampling would appear be the best choice in this particular exercise.

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 10.2. This figure reveals that backward sampling (when used in conjunction with either multinomial or residual resampling) 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 these two variants might be obtained with a smaller value of N.

To test this idea, we ran the six variants of our Gibbs sampler, but with $N = 20$. The right side of Figure 10.2 reveals the three non-backward sampling algorithms provide useless

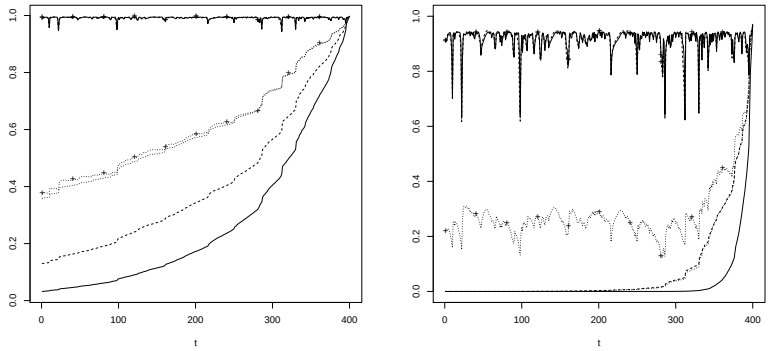


FIGURE 10.2. First dataset and resulting update rates of X_t versus $t \in 0 : 399$. Left plot is for $N = 200$ and right is for $N = 20$, same legend as Fig. 10.1. Note that for $N = 200$ and $N = 20$ it is hard to distinguish between multinomial resampling with backward sampling and residual resampling with backward sampling. For $N = 20$, forward only versions of systematic and residual perform similarly.

results because components of $X_{0:300}$ hardly ever change values. For the same reasons, the ACF's of these variants do not decay at reasonable rate (which is not shown here.)

Figure 10.3 gives the ACF of certain components for the three variants with backward sampling. Although N is quite small, the MCMC chain seems to be mixing well, especially when either multinomial or residual resampling is used.

10.2. **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 Yu and Meng (2011).)

What makes this example interesting relative to the first one is that the 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 10.4. Again, we observe that backward sampling leads to very good performance even when N is small, see also the ACF in Figure 10.3.

11. Discussion and Conclusions

The following guidelines seem to emerge from the numerical results. First, when backward resampling cannot be used (e.g. when the probability density of the Markov transition is not tractable), 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 smaller value while maintaining good mixing; see also Lindsten and

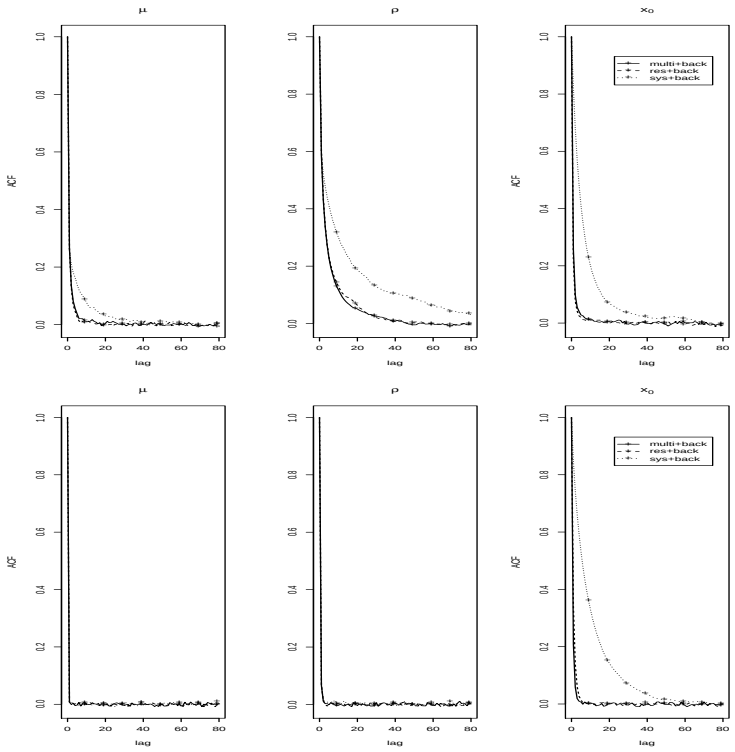


FIGURE 10.3. ACF for certain components of $(\theta, X_{0:T})$ and the three variants of the algorithm that uses backward sampling ($N = 20$); same legend as Fig. 10.1. Top row is for the first dataset, bottom row is for the second.

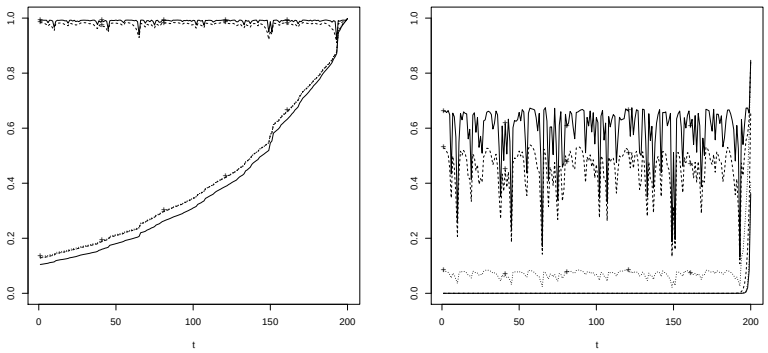


FIGURE 10.4. Second dataset, same plots as Fig. 10.2, with $N = 1000$ (left panel) and $N = 20$ (right panel). Same legend as Fig. 10.1. In the left plot systematic with and without backward and forward only residual largely indistinguishable. In the right plot the three forward only schemes indistinguishable before $t \approx 190$.

Schön (2012); Lindsten et al. (2012) for similar findings. In this case, either multinomial or residual resampling may be used, as they seem to lead to similar performance and they both outperform systematic sampling. It seems more complicated to give a precise justification for this phenomenon. One possible explanation is that the greater variability of e.g. multinomial resampling during the forward pass also results in greater liberty in redefining the ancestry of the chosen trajectory in the backward sampling step; recall that with multinomial resampling one is able to update the index a_t^n of the selected trajectory, without conditioning on $a_t^{n'}$, $n' \neq n$, as opposed to the two other resampling schemes. For example, the update of a_t^n in the backward step of systematic resampling appears to be very constrained.

When backward sampling is used, multinomial and residual resampling seem to lead to very similar ACF. Perhaps an additional advantage of residual resampling is that it may lead to a smaller variance of the estimator of the likelihood of $Y_{0:T}$ given θ, as this quantity depends only on the variables computed in the forward pass. This quantity could be used for instance to compute the evidence of the model (the marginal likelihood of $Y_{0:T}$), using the harmonic mean estimator (Gelfand and Dey, 1994).

In all cases, we recommend inspecting the same type of plots as in Fig. 10.2 and 10.4, 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.

Acknowledgements

S.S. Singh's research was partly funded by the Engineering and Physical Sciences Research Council (EP/G037590/1) whose support is gratefully acknowledged.

Appendix

A. Backward sampling for systematic resampling: remaining cases. We treat in this section the extra cases necessary to complete the description of the backward step under systematic resampling started in Section 9.2, and in particular to compute the probability $\varrho_t(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b})$ for all $a_t^b \in 1:N$ and some fixed set a_t^{-b} of $N-1$ labels. We use exactly the same notations. In particular we avoid treating separately the boundary cases $b = 1$ and $b = N$ by abusing notations: e.g. $a_t^{b-1} > a_t^{b+1}$ should read $a_t^N > a_t^2$ when $b = 1$, and so on. Recall that the case $a_t^{b-1} < a_t^{b+1}$ is already treated in Section 9.2.

Assume first that $a_t^{b-1} > a_t^{b+1}$. Then two cycles are compatible with a_t^{-b}: either $c = c^-$, where c^- is the cycle determined by $c^-(1) = b$, i.e. $A_t^{1:N} = (\bar{A}_t^{b:N}, \bar{A}_t^{1:b-1})$, or $c = c^+$, where c^+ is the cycle such that $c(1) = b+1$, i.e. $A_t^{1:N} = (\bar{A}_t^{b+1:N}, \bar{A}_t^{1:b})$. Since the cycle c is generated uniformly among the N possible cycles, it comes that

$$\varrho_t(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}) \propto \bar{\varrho}_t \left(x_t^{1:N}, \overline{A}_t^{1:N} = a_t^{c^+(1:N)} \right) + \bar{\varrho}_t \left(x_t^{1:N}, \overline{A}_t^{1:N} = a_t^{c^-(1:N)} \right)$$

where $\bar{\varrho}_t \left(z_t^{1:N}, \overline{A}_t^{1:N} = a_t^{c^+(1:N)} \right)$ may be computed as explained in Section 9.2.

Now assume that $a_t^{b-1} = a_t^{b+1} = m$. If at least one component of a_t^{-b} differs from m, then $A_t^b = m$ with probability one.

Now assume $a_t^n = m$ for all $n \neq b$. We consider the three different possible values for $\lfloor NW^m \rfloor$, which must be either $N-2$, $N-1$ or N, separately. If $\lfloor NW^m \rfloor = N$, then $A_t^b = m$ with probability one. If $\lfloor NW^m \rfloor = N-2$, one is in the same situation as the case $a_t^{b-1} > a_t^{b+1}$; that is, only two cycles are compatible with a_t^{-b} which are either c^+ or c^- defined previously, and proceed exactly as was done there. Finally, assume $\lfloor NW^m \rfloor = N-1$. Then all possible cycles are compatible with a_t^{-b}, that is

$$\varrho_t(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b}) \propto \sum_{c \in \mathcal{C}} \bar{\varrho}_t \left(x_t^{1:N}, \overline{A}_t^{1:N} = a_t^{c(1:N)} \right)$$

where $\mathcal{C}$ is the set of the N cycles of $1:N$.

B. Reversibility of the CPF-BS kernel under Assumption (R). Consider the distribution $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$, i.e. the distribution of $A_{0:T-1}^{1:N}$ conditional on $\left(X_{0:T}^{1:N}, N^\star\right) = (x_{0:T}^{1:N}, n^\star)$, relative to the joint distribution defined in (5.1). This conditional distribution factorises as a product of independent conditional distributions for the vectors $A_t^{1:N}$, with probability density:

$$\pi_T^N(a_t^{1:N} | x_{0:T}^{1:N}, n^\star) \propto \varrho_t(x_t^{1:N}, a_t^{1:N}) \prod_{n=1}^N m_{t+1}(x_t^{a_t^n}, x_{t+1}^n)$$

for any $t \in 0 : N-1$, and where $\varrho_t(x_t^{1:N}, a_t^{1:N})$ stands for the probability density of $\varrho_t(x_t^{1:N}, da_t^{1:N})$. The BS step updates, recursively from $t = T-1$ to $t = 0$, a single component A_t^n of the vector $A_t^{1:N}$, by sampling from its distribution conditional on the rest of that vector $A_t^{1:N \setminus n}$; see (9.2). Because of this, the BS steps leaves $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$ invariant. The index of the updated component A_t^n at time t is determined recursively by the value taken by $N^\star$; for $N^\star = n^\star$, one update $A_{T-1}^{n^\star}$, call $b_{T-1}^\star$ the obtained value, then one updates $A_{T-2}^{b_{T-1}^\star}$ and so on.

Thus, one may formalise the BS step as a probability kernel $K_{0:T-1}(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, n^\star, d\overline{a}_{0:T-1}^{1:N})$ that modifies the genealogy $A_{0:T-1}^{1:N}$ of the particle system, while keeping invariant its conditional distribution, $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$. As described below, the resulting genealogy $\bar{A}_{0:T-1}^{1:N}$ differs from the previous genealogy $a_{0:T-1}^{1:N}$ in at most one component for each t.

To establish that the CPF-BS kernel is reversible, we introduce the following generalisation of $K_{0:T-1}(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, d\overline{a}_{0:T-1}^{1:N})$, denoted $K_J(a_{0:T-1}^{1:N} x_{0:T-1}^{1:N}, n^\star, d\overline{a}_{0:T-1}^{1:N})$ for $J = \{t_1, \ldots t_k\} \subseteq 0 : T-1$, with $t_1 > \ldots > t_k$, which corresponds to the following partial BS update.

CPF-3' Let $B_T^\star = N^\star$, then, recursively for $t = T-1$ to 0, do: if $t \notin J$, set $B_t^\star = A_t^{B_{t+1}^\star}$ (deterministically), otherwise (if $t \in J$), sample index $B_t^\star = A_t^{B_{t+1}^\star} \in 1 : N$, conditionally on $B_{t+1}^\star = b$, from the distribution

(11.1)

$$\pi_T^N \left(A_t^b = a_t^b | B_{t+1} = b,\, X_{0:T}^{1:N} = x_{0:T}^{1:N},\, A_t^{-b} = a_t^{-b} \right) \propto \varrho_t \left(x_t^{1:N}, A_t^b = a_t^b | A_t^{-b} = a_t^{-b} \right) m_{t+1}(x_t^{a_t^b}, x_{t+1}^b).$$

Finally, return $Z_T^\star = (X_0^{B_0^\star}, \ldots, X_T^{B_T^\star})$.

Clearly, the kernel $K_J(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, d\overline{a}_{0:T-1}^{1:N})$ is such that the resulting $\overline{A}_{0:T-1}^{1:N}$ differs from $a_{0:T-1}^{1:N}$ by at most one component in those vectors $A_t^{1:N}$ such that $t \in J$.

Recall that the standard CPF kernel (with step CPF-1 and step CPF-2 only) is denoted by P_T^N , and let $P_T^{N,B(J)}$ denote what we call from now on the CPF-BS(J) kernel, which operates Steps CPF-1, CPF-2 and CPF-3', where the latter is implemented for a specific $J \subseteq 0:T-1$ indicated in the superscript.

Proposition 15. *Assume (R). For any $J \subset 0:T-1$, the kernel CPF-BS(J) with the additional step CPF-3' (partial backward sampling) is reversible. In particular, the CPF-BS kernel with Step CPF-3 (complete backward sampling, $J = 0:T-1$) is reversible.*

Proof. The proof will proceed by mathematical induction on the cardinal of J. Let $J = \{t\} \subseteq 0:T-1$, and denote kernel K_J as K_t. One has

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B(J)}} \left\{ h(Z_T^\star, N^\star, B_{0:T-1}^\star, \check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star) \right\}$$
$$= \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$$
$$\times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star)$$

where $\pi_T^N(d\check{n}^\star | x_{0:T}^{1:N})$ corresponds to CPF-2 (and our notations reflect the fact that this update of $N^\star$ does not depend on $a_{0:T-1}^{1:N}$, a property wish we use repeatedly), $K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ corresponds to CPF-3', and $z_T^{n^\star}$ and $b_{0:T-1}^\star$, resp. $\check{z}_T^{\check{n}^\star}$ and $\check{b}_{0:T-1}^\star$, must be understood as deterministic functions of $(x_{0:T}^{1:N}, a_{0:T-1}^{1:N}, n^\star)$, resp. of $(x_{0:T}^{1:N}, \check{a}_{0:T-1}^{1:N}, \check{n}^\star)$, i.e. $z_T^\star = (x_0^{b_0^\star}, \ldots, x_T^{b_T^\star})$, with $b_T^\star = n^\star$, $b_{T-1}^\star = a_{t-1}^{n^\star}$, and so on. The above quantity may be decomposed as

$$\int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$$
$$\times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star = \check{b}_{t+1}^\star\}}$$
$$+ \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$$
$$\times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star \neq \check{b}_{t+1}^\star\}}.$$

The first term of the sum may be rewritten as

$$
\begin{aligned}
&\int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}, a_{0:T-1}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star = \check{b}_{t+1}^\star\}} \\
=&\int \pi_T^N(dx_{0:T}^{1:N}, d\overline{a}_{0:T-1}^{1:N}) \pi_T^N(dn^\star | x_{0:T}^{1:N}) K_t(\overline{a}_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, n^\star, da_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star = \check{b}_{t+1}^\star\}} \\
=&\int \pi_T^N(dx_{0:T}^{1:N}, d\overline{a}_{0:T-1}^{1:N}) \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(\overline{a}_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) K_t(\overline{a}_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, n^\star, da_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star = \check{b}_{t+1}^\star\}} \\
=&\mathbb{E}_{\pi_T^N \otimes P_T^{N, B(J)}} \left\{ h(\check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star, Z_T^\star, N^\star, B_{0:T-1}^\star) \mathbb{I}_{\{B_{t+1}^\star = \check{B}_{t+1}^\star\}} \right\}
\end{aligned}
$$

where, in the first equality, we have used the fact that kernel K_t leaves invariant $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}, n^\star)$, and thus a valid generative mechanism for π_T^N is $\left(X_{0:T}^{1:N}, \overline{A}_{0:T-1}^{1:N}, N^\star\right) \sim \pi_T^N$, then $A_{0:T-1}^{1:N} | \overline{A}_{0:T-1}^{1:N} = \overline{a}_{0:T-1}^{1:N}, X_{0:T}^{1:N} = x_{0:T}^{1:N}, N^\star = n^\star \sim K_t(\overline{a}_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, n^\star, da_{0:T-1}^{1:N})$, and, in the second equality, we have use the following property: since the genealogies $a_{0:T-1}^{1:N}$ and $\overline{a}_{0:T-1}^{1:N}$ differ only with respect to one single component a_t^n, with $n = b_{t+1}^\star$, then for any pair $(n^\star, \check{n}^\star)$ such that $b_{t+1}^\star = \check{b}_{t+1}^\star$ (an event that is entirely determined by $n^\star$, $\check{n}^\star$ and $a_{t+1:T-1}^{1:N} = \overline{a}_{t+1:T-1}^{1:N}$), the distributions $K_t(a_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ and $K_t(\overline{a}_{0:T-1}^{1:N}, x_{0:T-1}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ are identical, as they both reduce to updating that particular component a_t^n with $n = b_{t+1}^\star = \check{b}_{t+1}^\star$ which is the only difference between $a_{0:T-1}^{1:N}$ and $\overline{a}_{0:T-1}^{1:N}$, and therefore sample from a distribution conditional on either $a_t^{1:N \setminus n}$ or $\overline{a}_t^{1:N \setminus n}$, which are equal. The last equality is a simple change of variable based on the symmetry of the previous expression.

Now for the second part of the sum:

$$
\begin{aligned}
&\int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star \neq \check{b}_{t+1}^\star\}} \\
=&\int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}) \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, z_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star \neq \check{b}_{t+1}^\star\}} \\
=&\int \pi_T^N(dx_{0:T}^{1:N}, d\check{a}_{0:T-1}^{1:N}, d\check{n}^\star) \delta_{\check{a}_{0:T-1}^{1:N}}(da_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{t+1}^\star \neq \check{b}_{t+1}^\star\}} \\
=&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ h(\check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star, Z_T^\star, N^\star, B_{0:T-1}^\star) \mathbb{I}_{\{B_{t+1}^\star \neq \check{B}_{t+1}^\star\}} \right\}
\end{aligned}
$$

where in the second inequality we have replaced $\pi_T^N(da_{0:T-1}^{1:N} | x_{0:T}^{1:N}) K_t(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$, the joint distribution of $(A_{0:T-1}^{1:N}, \check{A}_{0:T-1}^{1:N})$ conditional on $X_{0:T}^{1:N}$, by $\pi_T^N(d\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N}) \delta_{\check{a}_{0:T-1}^{1:N}}(da_{0:T-1}^{1:N})$, using the following two properties: (a) the marginal distribution of $\check{a}_{0:T-1}^{1:N}$ relative to that joint distribution is indeed $\pi_T^N(d\check{a}_{0:T-1}^{1:N} | x_{0:T}^{1:N})$, since K_t leaves it invariant; and b) the value taken by the integrand is unchanged is we replace the conditional distribution of $A_{0:T-1}^{1:N}$, given $\check{A}_{0:T-1}^{1:N}$ by $\delta_{\check{a}_{0:T-1}^{1:N}}(da_{0:T-1}^{1:N})$, owning to the fact that, provided $B_{t+1}^\star \neq \check{B}_{t+1}^\star$, the values taken by $(Z_T^\star, B_{0:T-1}^\star)$ depends on components of the genealogy $A_{0:T}^{1:N}$ that are identical in $A_{0:T}^{1:N}$ and $\check{A}_{0:T}^{1:N}$ (or in other words, components that have not been affected by K_t). (Recall that in these calculations, $(z_T^{n^\star}, n^\star, b_{0:T-1}^\star)$, resp. $(\check{z}_T^{n^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star)$, must always be understood as a certain deterministic function of $(x_{0:T}^{1:N}, a_{0:T-1}^{1:N}, n^\star)$,resp. $(x_{0:T}^{1:N}, \check{a}_{0:T-1}^{1:N}, \check{n}^\star)$.)

One may conclude by using the fact the CPF kernel P_T^N itself is reversible, per Proposition 10, and by combining the two terms of the sum above.

Let $k \geq 2$, and assume reversibility for all $J \subseteq 0:T-1$ such that $|J| = k-1$. We show reversibility for $J = \{t_1, \ldots t_k\}$ of cardinality k. Let $I = \{t_{i_1}, \ldots, t_{i_l}\} \subset J$ with $t_{i_1} > \ldots > t_{i_l}$,

$l \leq k$. The idea is to decompose the expectation

$$
\mathbb{E}_{\pi_T^N \otimes P_T^{N,B(J)}} \left\{ h(Z_T^\star, N^\star, B_{0:T-1}^\star \check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star) \right\}
$$

into a sum of 2^k terms, corresponding to the 2^k subsets $I \subset J$:

$$
\begin{aligned}
&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B(J)}} \left\{ h(Z_T^\star, N^\star, B_{0:T-1}^\star, \check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star) \mathbb{I}_{\{B_{r+1}^\star = \check{B}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{B_{s+1}^\star \neq \check{B}_{s+1}^\star : s \in J \setminus I\}} \right\} \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, da_{0:T-1}^{1:N}, dn^\star) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_J(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{r+1}^\star = \check{b}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{b_{s+1}^\star \neq \check{b}_{s+1}^\star : s \in J \setminus I\}} \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, d\overline{a}_{0:T-1}^{1:N}) \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) K_I(\overline{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, da_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_J(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{r+1}^\star = \check{b}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{b_{s+1}^\star \neq \check{b}_{s+1}^\star : s \in J \setminus I\}} \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, d\overline{a}_{0:T-1}^{1:N}) \int \pi_T^N(d\check{n}^\star | x_{0:T}^{1:N}) K_J(\overline{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) \\
&\quad \times \int \pi_T^N(dn^\star | x_{0:T}^{1:N}) K_I(\check{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, da_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{r+1}^\star = \check{b}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{b_{s+1}^\star \neq \check{b}_{s+1}^\star : s \in J \setminus I\}} \\
&= \int \pi_T^N(dx_{0:T}^{1:N}, d\check{a}_{0:T-1}^{1:N}, \check{n}^\star) \pi_T^N(dn^\star | x_{0:T}^{1:N}) K_I(\check{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, da_{0:T-1}^{1:N}) \\
&\quad \times h(z_T^{n^\star}, n^\star, b_{0:T-1}^\star, \check{z}_T^{\check{n}^\star}, \check{n}^\star, \check{b}_{0:T-1}^\star) \mathbb{I}_{\{b_{r+1}^\star = \check{b}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{b_{s+1}^\star \neq \check{b}_{s+1}^\star : s \in J \setminus I\}} \\
&= \mathbb{E}_{\pi_T^N \otimes P_T^{N,B(I)}} \left\{ h(\check{Z}_T^\star, \check{N}^\star, \check{B}_{0:T-1}^\star, Z_T^\star, N^\star, B_{0:T-1}^\star) \mathbb{I}_{\{B_{r+1}^\star = \check{B}_{r+1}^\star : r \in I\}} \mathbb{I}_{\{B_{s+1}^\star \neq \check{B}_{s+1}^\star : s \in J \setminus I\}} \right\}
\end{aligned}
$$

The first equality follows from the definition of $\pi_T^N \otimes P_T^{N,B(J)}$. The second equality uses the fact that K_I leaves invariant π_T^N, and therefore simulating $\left(X_{0:T}^{1:N}, \bar{A}_{0:T}^{1:N}, N^\star\right)$ from π_T^N, then updating $\bar{A}_{0:T}^{1:N}$ into $A_{0:T}^{1:N}$ through K_I ensures that $(X_{0:T}^{1:N}, A_{0:T-1}^{1:N}, N^\star) \sim \pi_T^N$. The third equality is based on the fact that, conditional on $\overline{a}_{0:T-1}^{1:N}$, it is possible to exchange the order in which genealogies $a_{0:T-1}^{1:N}$ and $\check{a}_{0:T-1}^{1:N}$ are generated without changing the value taken by the integrand; or in other words to replace $K_I(\overline{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, da_{0:T-1}^{1:N}) K_J(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ by $K_J(\overline{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N}) K_I(\check{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, n^\star, da_{0:T-1}^{1:N})$. To verify this, given $(\overline{a}_{0:T-1}^{1:N}, n^\star, \check{n}^\star)$, consider applying the kernels in the order K_I and then K_J which gives rise to the sequences $a_{0:T-1}^{1:N}$ and $\check{a}_{0:T-1}^{1:N}$ such that the constraints are satisfied by the genealogies $b_{0:T-1}^\star$ and $\check{b}_{0:T-1}^\star$. Then these same genealogies $b_{0:T-1}^\star$ and $\check{b}_{0:T-1}^\star$ can also be generated by reversing the order of the application of the kernels K_I and K_J while retaining the same probability of realizing them. To see why this is so, firstly, we note that the probability of realizing $\check{b}_{0:T-1}^\star$ is the same under $K_J(a_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ or $K_J(\overline{a}_{0:T-1}^{1:N}, x_{0:T}^{1:N}, \check{n}^\star, d\check{a}_{0:T-1}^{1:N})$ because K_I does not alter the variables $\overline{a}_{0:T-1}^{1:N}$ at the times $t \in J \setminus I$; and at times $t \in I$ when K_I does make changes the constraints $b_{t+1}^\star = \check{b}_{t+1}^\star$ for $t \in I$ enforces $\check{a}_t^{-\check{b}_{t+1}^\star} = \overline{a}_t^{-\check{b}_{t+1}^\star} = a_t^{-\check{b}_{t+1}^\star}$. Finally, whether K_I operates on $\overline{a}_{0:T-1}^{1:N}$ or $\check{a}_{0:T-1}^{1:N}$ does not alter the probability of realizing genealogy $b_{0:T-1}^\star$. This is in part because at the times $t \in J \setminus I$ where K_J alters $\overline{a}_{0:T-1}^{1:N}$ the constraints $b_{t+1}^\star \neq \check{b}_{t+1}^\star$ for $t \in J \setminus I$ are in force, which ensures the same sequence of ancestors $b_{0:T-1}^\star$of $n^\star$ can be generated if K_I acted on $\check{a}_{0:T-1}^{1:N}$ by only resampling $n^\star$'s genealogy at times $t \in I$.

Furthermore, the constraints $b^\star_{t+1} = \check{b}^\star_{t+1}$ for $t \in I$ enforces $a_t^{-b^\star_{t+1}} = \overline{a}_t^{-b^\star_{t+1}} = \check{a}_t^{-b^\star_{t+1}}$ which means the probability of realizing $b^\star_{0:T-1}$ is unchanged whether K_I acts on $\overline{a}^{1:N}_{0:T-1}$ or $\check{a}^{1:N}_{0:T-1}$.

The fourth identity uses that π_T^N is invariant through K_J, and marginalises out $\bar{a}^{1:N}_{0:T-1}$. One concludes by using the fact that $P_T^{N,B(I)}$ is reversible for any $I \subsetneq J$, and combining the 2^k terms corresponding to the above decomposition. □

C. Lag-one dominance under Assumption (R). We establish first the following lemma.

Lemma 16. *Assume (R). Then for $0 \le k \le T-1$ and any function g and h of their respective arguments, one has*

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ h(X^\star_{k:T}, B^\star_{k:T-1}, N^\star)\, g(\check{X}^\star_{0:T}, \check{B}^\star_{0:T-1}, \check{N}^\star)\, \mathbb{I}_{[B^\star_i \neq \check{B}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[N^\star \neq \check{N}^\star]} \right\}$$
$$= \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ g(X^\star_{0:T}, B^\star_{0:T-1}, N^\star) h(\check{X}^\star_{k:T}, \check{B}^\star_{k:T-1}, \check{N}^\star)\, \mathbb{I}_{[B^\star_i \neq \check{B}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[N^\star \neq \check{N}^\star]} \right\}$$

with the convention that $\mathbb{I}_{[B^\star_i \neq \check{B}^\star_i : T-1 \ge i > T-1]} = 1$, where $\big((X^\star_{0:T}, N^\star, B^\star_{0:T-1}), (\check{X}^\star_{0:T}, \check{N}^\star, \check{B}^\star_{0:T-1})\big)$ is distributed according to $\pi_T^N \otimes P_T^{N,B}$ in the first expectation and $\pi_T^N \otimes P_T^N$ in the second. Also,

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ h(X^\star_{0:T}, B^\star_{0:T-1}, N^\star)\, g(\check{X}^\star_{0:T}, \check{B}^\star_{0:T-1}, \check{N}^\star)\, \mathbb{I}_{[B^\star_i \neq \check{B}^\star_i : T-1 \ge i \ge 0]} \mathbb{I}_{[N^\star \neq \check{N}^\star]} \right\}$$
$$= \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ g(X^\star_{0:T}, B^\star_{0:T-1}, N^\star) h(\check{X}^\star_{0:T}, \check{B}^\star_{0:T-1}, \check{N}^\star)\, \mathbb{I}_{[B^\star_i \neq \check{B}^\star_i : T-1 \ge i \ge]0} \mathbb{I}_{[N^\star \neq \check{N}^\star]} \right\}$$

and

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ g(\check{X}^\star_{0:T}, \check{B}^\star_{0:T-1}, \check{N}^\star)\, \mathbb{I}_{[N^\star = \check{N}^\star]} \right\} = E_{\pi_T^N \otimes P_T^N} \left\{ g(X^\star_{0:T}, B^\star_{0:T-1}, N^\star)\, \mathbb{I}_{[N^\star = \check{N}^\star]} \right\}.$$

Proof. Expanding the first integral gives

$$\int \pi_T^N(dx^{1:N}_{0:T}, da^{1:N}_{0:T-1}) \int \pi_T^N(dn^\star | x^{1:N}_{0:T}) \int \pi_T^N(d\check{n}^\star | x^{1:N}_{0:T}) \int K_{0:T-1}(a^{1:N}_{0:T-1}, x^{1:N}_{0:T}, \check{n}^\star, d\check{a}^{1:N}_{0:T-1})$$
$$\times\, h(x^\star_{k:T}, b^\star_{k:T-1}, n^\star)\, g(\check{x}^\star_{0:T}, \check{b}^\star_{0:T-1}, \check{n}^\star)\, \mathbb{I}_{[b^\star_i \neq \check{b}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[n^\star \neq \check{n}^\star]}$$
$$= \int \pi_T^N(dx^{1:N}_{0:T}, d\overline{a}^{1:N}_{0:T-1}) \int \delta_{\overline{a}^{1:N}_{0:T-1}}(da^{1:N}_{0:T-1}) \int \pi_T^N(d\check{n}^\star | x^{1:N}_{0:T}) \int K_{0:T-1}(\overline{a}^{1:N}_{0:T-1}, x^{1:N}_{0:T}, \check{n}^\star, d\check{a}^{1:N}_{0:T-1})$$
$$\times \int \pi_T^N(dn^\star | x^{1:N}_{0:T}) h(x^\star_{k:T}, b^\star_{k:T-1}, n^\star)\, g(\check{x}^\star_{0:T}, \check{b}^\star_{0:T-1}, \check{n}^\star)\, \mathbb{I}_{[b^\star_i \neq \check{b}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[n^\star \neq \check{n}^\star]}$$
$$= \int \pi_T^N(dx^{1:N}_{0:T}, d\overline{a}^{1:N}_{0:T-1}) \int \pi_T^N(d\check{n}^\star | x^{1:N}_{0:T}) \int K_{0:T-1}(\overline{a}^{1:N}_{0:T-1}, x^{1:N}_{0:T-1}, \check{n}^\star, d\check{a}^{1:N}_{0:T-1})$$
$$\times \int \pi_T^N(dn^\star | x^{1:N}_{0:T}) \int \delta_{\check{a}^{1:N}_{0:T-1}}(da^{1:N}_{0:T-1}) h(x^\star_{k:T}, b^\star_{k:T-1}, n^\star)\, g(\check{x}^\star_{0:T}, \check{b}^\star_{0:T-1}, \check{n}^\star)\, \mathbb{I}_{[b^\star_i \neq \check{b}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[n^\star \neq \check{n}^\star]}$$

Recall that in the first line we write $\pi_T^N(dn^\star | x^{1:N}_{0:T})$ instead of $\pi_T^N(dn^\star | x^{1:N}_{0:T}, a^{1:N}_{0:T-1})$ as the update of $N^\star$ does not depend on $a^{1:N}_{0:T-1}$, a property we use repeatedly. The second line introduces the variables $\overline{a}^{1:N}_{0:T-1}$ and changes the order of integration for $N^\star$. The final equality introduces the measure $\delta_{\check{a}^{1:N}_{0:T-1}}(da^{1:N}_{0:T-1})$, which can be done without altering the value of the integrand of the outer most measure $\pi_T^N(dx^{1:N}_{0:T}, d\overline{a}^{1:N}_{0:T-1})$ since $h(x^\star_{k:T}, b^\star_{k:T-1}, n^\star)$ does not depend on $(x^\star_{0:k-1}, b^\star_{0:k-1})$ and the event $\mathbb{I}_{[b^\star_i \neq \check{b}^\star_i : T-1 \ge i > k]} \mathbb{I}_{[n^\star \neq \check{n}^\star]}$ holds. The proof of the second and third assertions follow similarly. □

We know proceed with the proof of Proposition 14.

Proof. It is enough to show that

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ h(Z_T^\star, B_{0:T-1}^\star, N^\star) h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star) \right\} \\ \leq \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ h(Z_T^\star, B_{0:T-1}^\star, N^\star) h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star) \right\}$$

since both $P_T^{N,B}$ and P_T^N have π_T as their invariant measure. To establish the inequality, we will compare both the expectations above on the mutually exclusive events $\mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > k]}$ for $k = 1, \ldots, T$ and $\mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i \geq 0]}$ and show that the stated inequality holds on all these events separately. We adopt the convention that $B_T^\star = N^\star$, $\check{B}_T^\star = \check{N}^\star$ and $\mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > T]} = 1$.

For the event $\mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i \geq 0]}$, by Lemma 16,

$$\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ h(Z_T^\star, B_{0:T-1}^\star, N^\star) h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star) \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i \geq 0]} \right\} \\ = \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ h(Z_T^\star, B_{0:T-1}^\star, N^\star) h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star) \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i \geq 0]} \right\}.$$

Consider now the remaining events. For the sake of terseness, let $h = f_k + g_k$ with

$$f_k(X_{0:k}^\star, B_{0:k}^\star) = \sum_{i=0}^{k-1} h_i(X_{i:i+1}^\star, B_{i:i+1}^\star), \quad g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) = \sum_{i=k}^{T-1} h_i(X_{i:i+1}^\star, B_{i:i+1}^\star)$$

with empty sums regarded as zero, i.e. $f_0 = g_T = 0$. Thus

$$h(Z_T^\star, B_{0:T-1}^\star, N^\star) h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star) \\ = \left(f_k(X_{0:k}^\star, B_{0:k}^\star) + g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) \right) \left(f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) + g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) \right)$$

For the product $f_k(X_{0:k}^\star, B_{0:k}^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star)$, note that

$$f_k(X_{0:k}^\star, B_{0:k}^\star)^2 + f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star)^2 \geq 2 f_k(X_{0:k}^\star, B_{0:k}^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star)$$

From Lemma 16 and the reversibility of $P_T^{N,B}$ and P_T^N (Propositions 10 and 12),

$$\begin{aligned} &\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star)^2 \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > k]} \right\} \\ =&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star)^2 \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > k]} \right\} \\ =&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star)^2 \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > k]} \right\} \\ =&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star)^2 \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : T \geq i > k]} \right\} \end{aligned}$$

where the first equality uses reversibility, the second Lemma 16 (using that $\mathbb{I}_{[B_k^\star = \check{B}_k^\star]} = \sum_n \mathbb{I}_{[B_k^\star = n]} \mathbb{I}_{[\check{B}_k^\star = n]}$, and taking $h(x_{k:T}, b_{k:T}) = \mathbb{I}_{[b_k = n]}$, $g(x_{0:T}, b_{0:T}) = f_k(x_{0:k}, b_{0:k})^2 \mathbb{I}_{[b_k = n]}$) and the final reversibility again. Thus

$$\begin{aligned} &\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i > k]} \right\} \\ \leq&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star)^2 \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i > k]} \right\} \\ =&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i > k]} \right\} \end{aligned}$$

For the 3 remaining terms of the product $h(Z_T^\star, B_{0:T-1}^\star, N^\star)h(\check{Z}_T^\star, \check{B}_{0:T-1}^\star, \check{N}^\star)$, we again invoke the established reversibility of $\pi_T^N \otimes P_T^{N,B}$ and $\pi_T^N \otimes P_T^N$ and Lemma 16, but this time to show that the expectation is the same under both $\pi_T^N \otimes P_T^{N,B}$ and $\pi_T^N \otimes P_T^N$. Details are as follows:

$$\begin{aligned}
&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star) g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(X_{0:k}^\star, B_{0:k}^\star) g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=&\mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\}
\end{aligned}$$

then

$$\begin{aligned}
&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=& \, \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) f_k(X_{0:k}^\star, B_{0:k}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=& \, \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) f_k(\check{X}_{0:k}^\star, \check{B}_{0:k}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\pi_T^N \otimes P_T^{N,B}} \left\{ g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\} \\
=& \, \mathbb{E}_{\pi_T^N \otimes P_T^N} \left\{ g_k(X_{k:T}^\star, B_{k:T-1}^\star, N^\star) g_k(\check{X}_{k:T}^\star, \check{B}_{k:T-1}^\star, \check{N}^\star) \, \mathbb{I}_{[B_k^\star = \check{B}_k^\star]} \mathbb{I}_{[B_i^\star \neq \check{B}_i^\star : i>k]} \right\}
\end{aligned}$$

which concludes the proof. □

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 72(3):269–342.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373.

Andrieu, C. and Vihola, M. (2012). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *ArXiv:1210.1484*.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar, Sonar Navigation*, 146(1):2–7.

Chopin, N., Jacob, P., and Papaspiliopoulos, O. (2012). SMC^2: an efficient algorithm for sequential analysis of state space models. *J. R. Statist. Soc. B*, (in press).

Del Moral, P. (2004). *Feynman-Kac Formulae*. Springer.

Doucet, A., de Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Statist.*, 21(4):940–960.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc. B*, 56(3):501–514.

Golightly, A. and Wilkinson, D. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Am. Statist. Assoc.*, pages 13–30.

Launay, T., Philippe, A., and Lamarche, S. (2012). On particle filters applied to electricity load forecasting. *arXiv:1210.0770*.

Lindsten, F., Jordan, M. I., and Schön, T. B. (2012). Ancestor Sampling for Particle Gibbs. *ArXiv:1210.6911*.

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), Kyoto, Japan*.

Liu, J. (1995). Metropolized Gibbs sampler: An improvement. Technical report, Dept. of Statistics, Stanford U., California.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Am. Statist. Assoc.*, 93(443):1032–1044.

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

Peters, G., Hosack, G., and Hayes, K. (2010). Ecological non-linear state space model selection via adaptive particle Markov chain Monte Carlo. *ArXiv:1005.2238*.

Pitt, M., Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, 171(2):134–151.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Silva, R., Giordani, P., Kohn, R., and Pitt, M. (2009). Particle filtering within adaptive Metropolis Hastings sampling. *Arxiv preprint arXiv:0911.0230*.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Prob.*, 8:1–9.

Vrugt, J. A., ter Braak, C. J., Diks, C. G., and Schoups, G. (2012). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources*, (in press).

Whiteley, N. (2010). Discussion of 'Particle Markov chain Monte Carlo methods' by Andrieu et al. *J. R. Statist. Soc. B*, 72(3):306–307.

Whiteley, N., Andrieu, C., and Doucet, A. (2010). Efficient Bayesian Inference for Switching State-Space Models using Discrete Particle Markov Chain Monte Carlo Methods. *ArXiv:1011.2437*.

Yu, Y. and Meng, X. (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(3):531–570.

CREST-ENSAE
E-mail address: nicolas.chopin@ensae.fr

Cambridge University
E-mail address: sss40@cam.ac.uk