Uniform Ergodicity of the Particle Gibbs Sampler

Fredrik Lindsten, Randal Douc and Eric Moulines

Original Publication:

Fredrik Lindsten, Randal Douc and Eric Moulines, Uniform Ergodicity of the Particle Gibbs Sampler, 2015, Scandinavian Journal of Statistics, (42), 3, 775-797.
http://dx.doi.org/10.1111/sjos.12136

Uniform ergodicity of the Particle Gibbs sampler

FREDRIK LINDSTEN
Department of Engineering, The University of Cambridge, and
Division of Automatic Control, Linköping University

RANDAL DOUC
Department CITI, Telecom Sudparis

ERIC MOULINES
Department LTCI, Telecom Paristech

November 21, 2014

Abstract

The particle Gibbs (PG) sampler is a systematic way of using a particle filter within Markov chain Monte Carlo (MCMC). This results in an off-the-shelf Markov kernel on the space of state trajectories, which can be used to simulate from the full joint smoothing distribution for a state space model in an MCMC scheme. We show that the PG Markov kernel is uniformly ergodic under rather general assumptions, that we will carefully review and discuss. In particular, we provide an explicit rate of convergence which reveals that: *(i)* for fixed number of data points, the convergence rate can be made arbitrarily good by increasing the number of particles, and *(ii)* under general mixing assumptions, the convergence rate can be kept constant by increasing the number of particles superlinearly with the number of observations. We illustrate the applicability of our result by studying in detail a common stochastic volatility model with a non-compact state space.

Keywords: particle Gibbs, particle Markov chain Monte Carlo, conditional sequential Monte Carlo, particle smoothing, state space models

1 Introduction

Statistical inference in general state space hidden Markov models involves computation of the posterior distribution of a set $X_{s:t} := [X_s, \ldots, X_t]$ of hidden state variables conditionally on a record $Y_{0:T}$ of observations, which we denote as $\phi_{s:t}\langle Y_{0:T}\rangle$. Of particular interest is the so called *joint smoothing distribution* (JSD) $\phi_{0:T}\langle Y_{0:T}\rangle$. Any marginal or fixed-interval smoothing distribution can be obtained from the JSD by marginalization. The JSD can be expressed in closed-form only in very specific cases, principally, when the state space model is linear and Gaussian or when the state space of the hidden Markov chain is a finite set. In the vast majority of cases, nonlinearity or non-Gaussianity render analytic solutions intractable.

This limitation has lead to an increase of interest in computational strategies handling more general state and measurement equations. Among these, *sequential Monte Carlo* (SMC) methods play a central role. SMC methods refer to a class of algorithms approximating a sequence of probability distributions, defined on a sequence of probability spaces. This is done by updating recursively a set of random *particles* with associated nonnegative importance weights. The SMC methodology has emerged as a key tool for approximating JSD flows in general state space models; see Doucet et al. (2000); Del Moral (2004); Doucet and Johansen (2011) for general introductions as well as applications and theoretical results for SMC methods.

However, a well known problem with SMC methods is that the particle approximation of any marginal smoothing distribution $\phi_{t:t}\langle Y_{0:T}\rangle$ becomes inaccurate for $t \ll T$. The reason is that the particle trajectories degenerate gradually as the interacting particle system evolves (Godsill et al., 2004; Fearnhead et al., 2010). To address this problem, several methods have been proposed; see Lindsten and Schön (2013) and the references therein. Among these methods, the recently introduced particle Markov chain Monte Carlo (PMCMC) framework, proposed in the seminal paper by Andrieu et al. (2010), plays a prominent role. PMCMC samplers make use of SMC (or variants thereof) to construct efficient, high-dimensional MCMC kernels which are reversible with respect to the JSD. These methods can then be used as components of more general sampling schemes relying on Markov kernels, for instance enabling joint state and parameter inference in general state space models.

Coupling SMC and MCMC is very useful since the distribution of the state sequence given the stream of observations is generally both high-dimensional and strongly dependent, rendering the design of alternative MCMC procedures, such as single-state Gibbs samplers and Metropolis-Hastings samplers, problematic. PMCMC has already found many applications in

areas such as hydrology (Vrugt et al., 2013), finance (Pitt et al., 2012), systems biology (Golightly and Wilkinson, 2011), and epidemiology (Rasmussen et al., 2011), to mention a few. Several methodological developments of the framework have also been made; see e.g. Whiteley et al. (2010); Lindsten et al. (2014); Chopin and Singh (2014); Pitt et al. (2012).

PMCMC algorithms can, broadly speaking, be grouped into two classes of methods: those based on particle independent Metropolis-Hastings (PIMH) kernels and those based on particle Gibbs (PG) kernels. The two classes of kernels are motivated in different ways and they have quite different properties. The former class, PIMH, exploits the fact that the SMC method defines an unbiased estimator of the likelihood, which is used in place of the intractable likelihood in the MH acceptance probability. This method can thus be viewed as a special case of the pseudo-marginal method introduced by Beaumont (2003); Andrieu and Roberts (2009) and later analyzed by Andrieu and Vihola (2012); Lee and Latuszynski (2012). The latter class, PG, on the other hand relies on conditioning the underlying SMC sampler on a reference trajectory to enforce the correct limiting distribution of the kernel; see Section 3. This algorithm can be interpreted as a Gibbs sampler for an extended model where the random variables generated by the SMC sampler are treated as auxiliary variables.

One of the main practical issues with PMCMC algorithms is the choice of the number, N, of particles. Using fewer particles will result in faster computations at each iteration, but can at the same time result in slower mixing of the resulting Markov kernel. For a fixed computational budget, there is a trade-off between taking the number of particles N large to get a faster mixing kernel, and to run many iterations of the MCMC sampler. Andrieu and Roberts (2009); Andrieu and Vihola (2012); Lee and Latuszynski (2012) investigate the rate of convergence of the pseudo-marginal method and characterize the approximation of the marginal algorithm by the pseudo-marginal algorithm in terms of the variability of their respective ergodic averages. Doucet et al. (2012) and Pitt et al. (2012) conclude, using partially heuristic arguments, that it is close to optimal to let N scale at least linearly with T.

The theoretical properties of the PG kernel, however, are not as well understood. Andrieu et al. (2010) establish under weak conditions that the PG kernel is ϕ-irreducible and aperiodic for any $N \geq 2$ (see Meyn and Tweedie (2009) for definitions). However, this does not provide a control for the rate of convergence of the iterates of the PG kernel to stationarity. In this work, we establish that the PG kernel is, under mild assumptions, uniformly ergodic. This interesting property has already been established in an earlier work by Chopin and Singh (2014), but we give here a direct proof under weaker conditions, which in addition provides an explicit lower

bound for the convergence rate.

During the preparation of this manuscript, a preprint was made available by Andrieu et al. (2013), who, independently, have found similar results as presented here. Indeed, they establish basically the same lower bound on the minorizing constant for the PG kernel (which they refer to as the iterated conditional SMC kernel), though using a different proof technique based on a "doubly conditional" SMC algorithm. There are, however, several differences between these two contributions. We focus in particular on analyzing the minorizing constant under mixing conditions for the state space model which hold very generally, even if the state space is not compact (see Section 4.3). We then study how the number of particles N should be increased with the number of observations T. We show that under weak assumptions, it suffices to increase the number of particles N as T^δ where $\delta \geq 1$ can be determined explicitly. This is in contrast with Andrieu et al. (2013) who, effectively, assume a compact state space; see Remark 3 and Section 4.2. On the other hand, Andrieu et al. (2013) study necessary (i.e., not only sufficient) conditions for uniform ergodicity and translate the convergence results for the PG kernel to a composite MCMC scheme for simulating both states and parameters of a state space model. Given these differences, we believe that the two contributions complement each other.

This paper is organized as follows: In Section 2 we introduce our notation, and in Section 3 we review the PG sampler and formally define the PG Markov kernel. In Section 4 we state the main results, starting with a minorization condition for the PG kernel followed by mixing conditions that allow for time uniform control of the convergence rate. In Section 5 we study, in detail, a commonly used stochastic volatility model with a non-compact state space to illustrate how the conditions of our results can be verified in practice. The proofs of the main theorems are postponed to Sections 6 and 7. An additional example as well as some of the proofs are given in the online supplementary material.

2 Notations and problem statement

We write $\mathbb{N}$ and $\mathbb{N}^+$ for the sets of non-negative and positive integers, respectively. Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces and let $\mathcal{P}(\mathsf{X})$ be the set of all probability measures on $(\mathsf{X}, \mathcal{X})$. Let M be a kernel on $(\mathsf{X}, \mathcal{X})$ and G a kernel on $(\mathsf{X}, \mathcal{Y})$. Assume that for all $x \in \mathsf{X}$, $G(x, \cdot)$ is dominated by some common nonnegative measure κ on $(\mathsf{Y}, \mathcal{Y})$ and denote by $g(x, \cdot)$ its Radon-Nikodym derivative, i.e., for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$: $g(x, y) = \frac{\mathrm{d}G(x,\cdot)}{\mathrm{d}\kappa(\cdot)}(y)$. Let $\{(X_t, Y_t), t \in \mathbb{N}\}$ be a hidden Markov chain associated to the pair (M, G). That is, $\{(X_t, Y_t), t \in \mathbb{N}\}$ is a Markov

chain with transition kernel defined by: for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ and all $C \in \mathcal{X} \otimes \mathcal{Y}$,

$$((x, y), C) \mapsto \iint_C M(x, \mathrm{d}x')G(x', \mathrm{d}y') \ .$$

The sequence $\{X_t,\ t \in \mathbb{N}\}$ is usually not observed and inference should be carry out on the basis of the observations $\{Y_t,\ t \in \mathbb{N}\}$ only. With $\mu \in \mathcal{P}(\mathsf{X})$ being the initial distribution of the hidden state process, for all $t \geq 0$, denote by

$$y_{0:t} \mapsto p_\mu(y_{0:t}) := \int \mu(\mathrm{d}x_0)g(x_0, y_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s)g(x_s, y_s) \ ,$$

the density of the observations $Y_{0:t}$ with respect to $\kappa^{\otimes(t+1)}$. In what follows, we set, by abuse of notation, for all $x \in \mathsf{X}$, $p_x(y_{0:t}) = p_{\delta_x}(y_{0:t})$, where δ_x is the Dirac measure at x.

For all $y \in \mathsf{Y}$, define the (unnormalized) kernel $Q\langle y \rangle$ on $(\mathsf{X}, \mathcal{X})$ by

$$Q\langle y \rangle(x, A) = \int M(x, \mathrm{d}x')g(x', y)\mathbb{1}_A(x') \ , \tag{1}$$

and for all $s \leq t$ and all $y_{s:t} \in \mathsf{Y}^{t-s+1}$, define the kernel $Q\langle y_{s:t} \rangle$ on $(\mathsf{X}, \mathcal{X})$ by

$$Q\langle y_{s:t} \rangle(x, A) = Q\langle y_s \rangle Q\langle y_{s+1} \rangle \dots Q\langle y_t \rangle(x, A) \ . \tag{2}$$

With these notations, $p_\mu(y_{0:t}) = \mu(g(\cdot, y_0)Q\langle y_{1:t} \rangle \mathbf{1}(\cdot))$ where $\mathbf{1}$ is the constant function, $\mathbf{1}(x) = 1$ for all $x \in \mathsf{X}$. In what follows, we set by convention $Q\langle y_{s:t} \rangle \mathbf{1}(x) \equiv 1$ for all $s > t$.

For all $\mu \in \mathcal{P}(\mathsf{X})$ and for all $0 \leq s \leq t$, denote $p_\mu(y_{s:t}|y_{0:s-1}) := p_\mu(y_{0:t})/p_\mu(y_{0:s-1})$ if $p_\mu(y_{0:s-1}) \neq 0$ and $p_\mu(y_{s:t}|y_{0:s-1}) := 0$ otherwise, with the convention $p_\mu(y_{0:t}|y_{0:-1}) = p_\mu(y_{0:t})$.

A quantity of central interest is the JSD, given by

$$\phi_{\mu,0:t}\langle y_{0:t} \rangle(D) := \frac{1}{p_\mu(y_{0:t})} \int \mu(\mathrm{d}x_0)g(x_0, y_0) \prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s)g(x_s, y_s)\mathbb{1}_D(x_{0:t}) \ , \tag{3}$$

for all $D \in \mathcal{X}^{\otimes(t+1)}$. With T being some final time point, the PG sampler (reviewed in the subsequent section) defines a Markov kernel which leaves the JSD $\phi_{\mu,0:T}\langle y_{0:T} \rangle$ invariant. Samples drawn from the PG kernel can thus be used to draw inference about the states (and/or parameters) of the state space model.

3 The particle Gibbs sampler

Consider first an SMC sampler targeting the sequence of JSDs defined in (3). Let $y_{0:T}$ be a fixed sequence of observations. The SMC sampler approximates $\phi_{\mu,0:t}\langle y_{0:t} \rangle$ by a collection of

weighted samples $\{(X^i_{0:t}, \omega^i_t)\}_{i=1}^N$, in the sense that

$$\phi^N_{\mu,0:t}\langle y_{0:t}\rangle(h) := \sum_{i=1}^N \frac{\omega^i_t}{\sum_{\ell=1}^N \omega^\ell_t} h(X^i_{0:t})$$

is an estimator of $\phi_{\mu,0:t}\langle y_{0:t}\rangle(h)$ for a measurable function $h : \mathsf{X}^{t+1} \to \mathbb{R}$. These weighted samples can be generated in many different ways, see e.g. Doucet et al. (2000); Del Moral (2004); Doucet and Johansen (2011); Cappé et al. (2005) and the references therein. Here we review a basic method, though it should be noted that the PG sampler can be generalized to more advanced procedures, see Andrieu et al. (2010); Chopin and Singh (2014).

Initially, $\phi_{\mu,0:0}\langle y_0\rangle$ is approximated by importance sampling. That is, we simulate independently $\{X^i_0\}_{i=1}^N$ from a proposal distribution: $X^i_0 \sim r_0\langle y_0\rangle(\cdot)$. The samples, commonly referred to as *particles*, are then assigned importance weights, $\omega^i_0 = w_0\langle y_0\rangle(X^i_0)$, where $w_0\langle y_0\rangle(x) = g(x, y_0)\frac{\mathrm{d}\mu}{\mathrm{d}r_0\langle y_0\rangle}(x)$ (provided that $r_0\langle y_0\rangle$ is such that $\mu \ll r_0\langle y_0\rangle$).

We proceed inductively. Denote by $\mathcal{F}^N_t$ the filtration generated by the particles and weights up to the current time instant t: $\mathcal{F}^N_t := \sigma\left(\{(X^i_{0:s}, \omega^i_s)\}_{i=1}^N, 0 \le s \le t\right)$. Assume that we have at hand a weighted sample $\{(X^i_{0:t-1}, \omega^i_{t-1})\}_{i=1}^N$ approximating the JSD $\phi_{\mu,0:t-1}\langle y_{0:t-1}\rangle$ at time $t-1$. This weighted sample is then propagated sequentially *forward in time*. This is done by sampling, conditionally independently given the particle history $\mathcal{F}^N_{t-1}$, for each particle $i \in \{1, \ldots, N\}$ an *ancestor index* A^i_t with probability

$$\mathbb{P}\left(A^i_t = j \,\middle|\, \mathcal{F}^N_{t-1}\right) = \frac{\omega^j_{t-1}}{\sum_{\ell=1}^N \omega^\ell_{t-1}}\,, \quad j \in \{1, \ldots, N\}\,, \tag{4}$$

and then by sampling a new particle position from the proposal kernel $R\langle y_t\rangle$:

$$X^i_t \sim R\langle y_t\rangle(X^{A^i_t}_{t-1}, \cdot)\,. \tag{5}$$

Finally, the particles are assigned importance weights given by

$$\omega^i_t = w\langle y_t\rangle(X^{A^i_t}_{t-1}, X^i_t) := \frac{\mathrm{d}Q\langle y_t\rangle(X^{A^i_t}_{t-1}, \cdot)}{\mathrm{d}R\langle y_t\rangle(X^{A^i_t}_{t-1}, \cdot)}(X^i_t)\,, \tag{6}$$

where $Q\langle y\rangle$ is defined in (1) and, as before, it is assumed that $Q\langle y\rangle(x, \cdot) \ll R\langle y\rangle(x, \cdot)$. The particle trajectories (i.e., the ancestral paths of the particles X^i_t, $i \in \{1, \ldots, N\}$) are constructed sequentially by associating the current particle X^i_t with the particle trajectory of its ancestor: $X^i_{0:t} := (X^{A^i_t}_{0:t-1}, X^i_t)$. This results in a weighted particle system $\{(X^i_{0:t}, \omega^i_t)\}_{i=1}^N$ targeting $\phi_{\mu,0:t}\langle y_{0:t}\rangle$, completing the induction. Two classical choices for the proposal kernel $R\langle y\rangle$ are:

$$R\langle y\rangle(x, \mathrm{d}x') = \begin{cases} M(x, \mathrm{d}x') & \text{bootstrap filter,} \\ \frac{M(x,\mathrm{d}x')g(x',y)}{\int M(x,\mathrm{d}x')g(x',y)} & \text{fully-adapted filter.} \end{cases} \tag{7}$$

The former is commonly used due to its simplicity, whereas the latter is preferable whenever it can be implemented since it is known to minimize the incremental weight variance; see Doucet and Johansen (2011) for further discussion.

Assume now that T is some final time point and that we are interested in simulating from the JSD $\phi_{\mu,0:T}\langle y_{0:T}\rangle$ using an MCMC procedure. For that purpose, it is required to define a Harris positive recurrent Markov kernel on the path space $(\mathsf{X}^{T+1}, \mathcal{X}^{\otimes(T+1)})$ having the JSD $\phi_{\mu,0:T}\langle y_{0:T}\rangle$ as its unique invariant distribution. The PG sampler accomplishes this by making use of SMC. From an algorithmic point of view, the difference between PG and a standard SMC sampler is that in the PG sampler, one particle trajectory, denoted as $x'_{0:T} = (x'_0, \ldots, x'_T) \in \mathsf{X}^{t+1}$, is specified *a priori*. This trajectory is used as a reference for the sampler, as discussed below.

The reference trajectory is taken into account by simulating only $N-1$ particles in the usual way. The Nth particle is then set deterministically according to the reference. At the initialization, we thus simulate independently $\{X_0^i\}_{i=1}^{N-1}$ with $X_0^i \sim r_0\langle y_0\rangle(\cdot)$ and set $X_0^N = x'_0$. We then compute importance weights for all particles: $\omega_0^i = w_0\langle y_0\rangle(X_0^i)$, for $i = 1, \ldots, N$.

Analogously, at any consecutive time point t, we sample the first $N-1$ particles $\{(A_t^i, X_t^i)\}_{i=1}^{N-1}$ conditionally independently given $\mathcal{F}_{t-1}^N$ according to (4)–(5). Note that these particles will depend on the reference trajectory through the resampling step (4). The Nth particle and its ancestor index are then set deterministically: $X_t^N = x'_t$ and $A_t^N = N$. Finally, importance weights are then computed for all the particles according to (6). Note that, by construction, the Nth particle trajectory will coincide with the reference trajectory for all t, $X_{0:t}^N = x'_{0:t}$.

After a complete pass of the above procedure, a trajectory $X_{0:T}^\star$ is sampled from among the particle trajectories at time T, with probabilities given by their importance weights, i.e.,

$$\mathbb{P}\left(X_{0:T}^\star = X_{0:T}^i \,\middle|\, \mathcal{F}_T^N\right) = \frac{\omega_T^i}{\sum_{\ell=1}^N \omega_T^\ell}\,, \quad i \in \{1, \ldots, N\}\,. \tag{8}$$

The PG sampling procedure (summarized in Algorithm 1 in the online supplementary material) thus associates each trajectory $x'_{0:T} \in \mathsf{X}^{T+1}$ to a probability distribution on $(\mathsf{X}^{T+1}, \mathcal{X}^{\otimes(T+1)})$, defining a Markov kernel on $(\mathsf{X}^{T+1}, \mathcal{X}^{\otimes(T+1)})$. More specifically, this kernel is given by

$$P_{T,N}(x'_{0:T}, D) := \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_T^i \mathbb{1}_D(X_{0:T}^i)}{\sum_{\ell=1}^N \omega_T^\ell}\right], \tag{9}$$

for $(x'_{0:T}, D) \in \mathsf{X}^{T+1} \times \mathcal{X}^{\otimes(T+1)}$, where $\mathbb{E}$ refers to expectation with respect to the random variables generated by the PG algorithm. We refer to $P_{T,N}$ as the PG kernel.

As shown by Andrieu et al. (2010), the conditioning on a reference trajectory implies that

the PG kernel leaves the JSD invariant:

$$\phi_{\mu,0:T}\langle y_{0:T}\rangle(D) = \int P_{T,N}(x'_{0:T}, D)\phi_{\mu,0:T}\langle y_{0:T}\rangle(\mathrm{d}x'_{0:T})\,, \qquad \forall D \in \mathcal{X}^{(T+1)}.$$

Quite remarkably, this invariance property holds for any $N \geq 1$.

Empirically, it has been found that the mixing of the PG kernel can be improved significantly by updating the ancestor indices A_t^N for $t \in \{1, \ldots, T\}$, either as part of the forward recursion (Lindsten et al., 2014) or in a separate backward recursion (Whiteley et al., 2010). We shall not specifically analyze these modified PG algorithms in this work. However, as detailed below, our analysis uses a worst-case bound to completely remove the influence of the reference trajectory (and thus also the corresponding ancestor indices) and therefore our uniform ergodicity result applies straightforwardly to these modified algorithms as well.

4 Main result

In this section we state the main results. First, in Section 4.1, we give a minorization condition for the PG kernel. For a fixed observation sequence $y_{0:T}$, the result of Section 4.1 implies uniform ergodicity of the PG kernel under weak conditions. Thereafter, in Section 4.2 and 4.3, we discuss how to increase the number of particles $N = N_T$ as a function of the number of observations T in order to obtain a non-degenerate lower-bound as $T \to \infty$.

4.1 Minorization condition

Define the sequence of nonnegative variables $\{B_{t,T}\}_{t=0}^T$ by

$$B_{t,T} = \sup_{0 \leq \ell \leq T-t} \frac{|w\langle y_t\rangle|_\infty |Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty}{p_\mu(y_{t:t+\ell}|y_{0:t-1})} \tag{10}$$

where $|w\langle y\rangle|_\infty = \sup_{(x,x')} w\langle y\rangle(x, x')$, $|Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty = \sup_x Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}(x)$ and, by convention, $|Q\langle y_{t+1:t}\rangle \mathbf{1}|_\infty = 1$. We then have the following minorization condition for the PG kernel.

Theorem 1. *For all $x'_{0:T} \in \mathsf{X}^{T+1}$ and $D \in \mathcal{X}^{\otimes(T+1)}$,*

$$P_{T,N}(x'_{0:T}, D) \geq \epsilon_{T,N}\ \phi_{\mu,0:T}\langle y_{0:T}\rangle(D)\,, \tag{11}$$

where

$$\epsilon_{T,N} := \prod_{t=0}^{T} \frac{N-1}{2B_{t,T} + N - 2}\,. \tag{12}$$

Proof. The proof is postponed to Section 6. However, to provide some intuition for the result, the main ideas of the proof are outlined below. From (9) we have

$$P_{T,N}(x'_{0:T}, D) \geq (N-1)\mathbb{E}\left[\frac{\omega_T^1 \mathbb{1}_D(X_{0:T}^1)}{\sum_{i=1}^N \omega_T^i}\right] \geq (N-1)\mathbb{E}\left[\mathbb{E}\left[\frac{\omega_T^1 \mathbb{1}_D(X_{0:T}^1)}{2|w\langle y_t\rangle|_\infty + \sum_{i=2}^{N-1} \omega_T^i} \,\middle|\, \mathcal{F}_{T-1}^N\right]\right] ,$$

where, for the first inequality, we have simply discarded the Nth term (corresponding to the reference particle) and used the fact that the $N-1$ weighted particles $\{(X_{0:T}^i, \omega_T^i)\}_{i=1}^{N-1}$ are equally distributed. For the second inequality, we bound the first and the last term of the sum in the denominator by $|w\langle y_t\rangle|_\infty$. This has the effect that the random variables entering the numerator and the denominator of the expression are conditionally independent given $\mathcal{F}_{T-1}^N$. By convexity of $x \mapsto 1/x$ and using Jensen's inequality we therefore obtain the bound

$$P_{T,N}(x'_{0:T}, D) \geq (N-1)\mathbb{E}\left[\frac{\mathbb{E}\left[\omega_T^1 \mathbb{1}_D(X_{0:T}^1) \,\middle|\, \mathcal{F}_{T-1}^N\right]}{2|w\langle y_t\rangle|_\infty + (N-2)\mathbb{E}\left[\omega_T^2 \,|\, \mathcal{F}_{T-1}^N\right]}\right] .$$

The inner conditional expectations can be computed explicitly. Essentially, the result follows by repeating this procedure for time $T-1$, then for $T-2$, etc. □

Corollary 2. *Assume that $g(x,y) > 0$ for all $(x,y) \in \mathsf{X} \times \mathsf{Y}$ and $|w\langle y\rangle|_\infty < \infty$ for all $y \in \mathsf{Y}$. Then, for fixed a fixed observation sequence $y_{0:T}$,*

$$\epsilon_{T,N} \geq 1 + \frac{1}{N-1}\sum_{t=0}^{T}(1 - 2B_{t,T}) + O(N^{-2}) ,$$

and $\lim_{N\to\infty} \epsilon_{T,N} = 1$.

Proof. From the definition (12) we have

$$\epsilon_{T,N} = \exp\left\{-\sum_{t=0}^{T} \ln\left(1 + \frac{2B_{t,T} - 1}{N-1}\right)\right\} \geq \exp\left\{\frac{1}{N-1}\sum_{t=0}^{T}(1 - 2B_{t,T})\right\} .$$

For a fixed T, we thus obtain the result provided that $B_{t,T} < \infty$ for all $t \in \{0, \ldots, T\}$. However, the positivity of g implies that $p_\mu(y_{t:t+\ell}|y_{0:t-1}) > 0$ for all $\ell \geq 0$, and since $|w\langle y\rangle|_\infty < \infty$ for all $y \in \mathsf{Y}$, it can be easily checked that $|Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty \leq \prod_{s=t+1}^{t+\ell} |w\langle y_s\rangle|_\infty < \infty$, which immediately implies that $B_{t,T} < \infty$ for all $t \in \{0, \ldots, T\}$. □

For a fixed observation sequence $y_{0:T}$ (with finite T), Theorem 1 and Corollary 2 ensure uniform ergodicity of the PG kernel under weak conditions. In the following two sections we analyze the minorizing constant of Theorem 1 when T is taken to infinity. Specifically, we discuss how to increase the number of particles $N = N_T$ as a function of the number of observations T in order to obtain a non-degenerate lower-bound.

Remark 3. *The minorization condition of Theorem 1 is similar to Proposition 6 by Andrieu et al. (2013). However, they express the minorizing constant in terms of the expectation of a likelihood estimator with respect to the law of a "doubly conditional SMC" algorithm. They do not pursue an analysis of the effect on the minorization condition by the forgetting of the initial condition of the state space model. To obtain a non-degenerate rate of convergence when $T \to \infty$ they assume (in our notation) that the triangular array of variables $\{B_{t,T}\}_{0 \leq t \leq T}$ is uniformly bounded for $T \geq 0$. This is the case, basically, only when the model satisfies strong mixing conditions, as we discuss in the subsequent section. Indeed, Andrieu et al. (2013, Proposition 14 and Lemma 17) is the same as our Proposition 5 below.*

4.2 Strong mixing condition

We first assume a strong mixing condition for the kernel M:

(S-1) There exist positive constants (σ_-, σ_+), a nonnegative measure γ and an integer $m \in \mathbb{N}^+$ such that for all $x \in \mathsf{X}$: $\sigma_- \gamma(\mathrm{d}x') \leq M^m(x, \mathrm{d}x') \leq \sigma_+ \gamma(\mathrm{d}x')$.

This condition has been introduced by Del Moral and Guionnet (1999) to establish the uniform-in-time convergence of the particle filter. This condition, which typically requires that the state space is compact, is overly restrictive but it is often used in the analysis of state space models because it implies a form of uniform forgetting of the initial condition of the filter, which is key to obtaining long-term stability of the particle filter.

Proposition 4. *Assume that $g(x, y) > 0$ for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, that (S-1) holds with $m = 1$ and that the proposal kernel is fully-adapted as defined in (7). Then, taking $N_T \sim \lambda T$ for some $\lambda > 0$, we have*

$$\liminf_{T \to \infty} \epsilon_{T, N_T} \geq \exp\left(\frac{1 - 2(\sigma_+/\sigma_-)^2}{\lambda}\right) > 0 \,.$$

Proof. See the online supplementary material. □

Note that Proposition 4 holds whatever the observation sequence $\{y_t,\ t \in \mathbb{N}\}$ is. This is a consequence of the strong mixing condition (S-1) which provides a simple result, but at the expense of an assumption which is rarely met in practice. If instead of the fully adapted case, we consider the bootstrap filter (see (7)), we may also obtain a uniform-in-time bound. However, this requires an even stronger assumption of the existence of a lower and an upper bound for the observation likelihood.

(S-2) There exists a constant $\delta \geq 1$, such that for all $y \in \mathsf{Y}$, $\sup_{x \in \mathsf{X}} g(x, y) \leq \delta \inf_{x \in \mathsf{X}} g(x, y)$.

Proposition 5. *Assume that (S-1)-(S-2) hold and that the bootstrap proposal is used:* $R\langle y\rangle(x,\cdot) = M(x,\cdot)$. *Then, taking* $N_T \sim \lambda T$ *for some* $\lambda > 0$, *we have*

$$\liminf_{T\to\infty} \epsilon_{T,N_T} \geq \exp\left(\frac{1 - 2\delta^m \sigma_+/\sigma_-}{\lambda}\right) > 0 \,.$$

where m *is defined in (S-1).*

Proof. See the online supplementary material. □

4.3 Moment assumption

Under the restrictive assumptions (S-1) and (S-2), we obtained non-degenerate *uniform* convergence bounds when $N_T \propto T$. However, these conditions are hardly ever satisfied when the state space is non-compact. We now turn to the analysis of the minorization condition under a much weaker moment assumption.

When the strong mixing assumption is relaxed, however, we are no longer able to obtain bounds that hold uniformly with respect to the observations. Instead, we take a probabilistic approach and analyze the minorizing constant with respect to the distribution of the observation sequence $\{Y_t,\ t \in \mathbb{N}\}$. For this reason, it is also of interest to carry out the analysis for a parametric family of state space models $\{(M^\theta, G^\theta), \theta \in \Theta\}$, where Θ is a compact subset of a Euclidean space. Informally, this allows us to analyse the ergodicity of the PG kernel, even when the algorithm is executed using a misspecified model. That is, assume that $\{Y_t,\ t \in \mathbb{N}\}$ is generated according to $(M^{\theta_\star}, G^{\theta_\star})$ where $\theta_\star$ is some "true" parameter of the model. Since $\theta_\star$ is typically unknown we assume that the PG algorithm is implemented using some estimate θ instead. The question is then: with respect to the law $\bar{\mathbb{P}}^{\theta_\star}$ (see definition below) of the observation sequence $\{Y_t,\ t \in \mathbb{N}\}$, can we expect that the convergence rate of the PG algorithm (which depends on the realization of the observations) will be non-degenerate as $T \to \infty$?

In Theorem 6 below, we show that this is indeed the case provided that N_T is a power of T. We consider a sequence of parameters $\{\theta_T,\ T \in \mathbb{N}\}$ that become increasingly close to $\theta_\star$ (in a sense that will be made precise in Theorem 6), converging at a rate $1/\sqrt{T}$. The rationale for this assumption is that we are considering the large T regime and we can therefore expect θ_T to be close to $\theta_\star$. We discuss this condition further below.

We emphasize that the result of this section concerns the dependence of the convergence rate of the PG sampler on the observation sequence for large T. For a fixed observation sequence $Y_{0:T} = y_{0:T}$ (with finite T) we can instead appeal to Corollary 2, which ensures uniform ergodicity of the sampler under weak conditions.

(A-1) For all $\theta \in \Theta$, the kernel M^θ has a unique stationary distribution denoted as π^θ.

In what follows, for $\theta \in \Theta$ we let $\mathbb{E}_\mu^\theta$ and $\mathbb{P}_\mu^\theta$ refer to the expectation and probability, respectively, induced on $((\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}})$ by a Markov chain $\{(X_t, Y_t), t \in \mathbb{N}\}$ evolving according to the state space model (M^θ, G^θ) starting with $X_0 \sim \mu$. For simplicity, we write $\bar{\mathbb{E}}^\theta = \mathbb{E}_{\pi^\theta}^\theta$ and $\bar{\mathbb{P}}^\theta = \mathbb{P}_{\pi^\theta}^\theta$. For $1 \leq s \leq t$, we write $p_{\mu,s}^\theta(y_{s:t}) = \int p_\mu^\theta(y_{0:t}) \kappa^{\otimes s}(\mathrm{d}y_{0:s-1})$ for the marginal probability density function of $Y_{s:t}$ with respect to $\kappa^{\otimes(t-s+1)}$.

(A-2) There exists a constant $\sigma_+ \in \mathbb{R}^+$ and a nonnegative measure λ such that for all $\theta \in \Theta$ and $(x, A) \in \mathsf{X} \times \mathcal{X}$: $M^\theta(x, A) \leq \sigma_+ \lambda(A)$.

Denote by $m^\theta(x, \cdot)$ the Radon-Nikodym derivative

$$m^\theta(x, x') = \frac{\mathrm{d}M^\theta(x, \cdot)}{\mathrm{d}\lambda(\cdot)}(x') \, .$$

Under (A-2), the stationary distribution π^θ is absolutely continuous with respect to λ. Furthermore, for notational simplicity it is assumed that the initial distribution μ is absolutely continuous with respect to λ. By abuse of notation, we write π^θ and μ also for the corresponding density functions.

(A-3) For all $\theta \in \Theta$ and $(x, x', y) \in \mathsf{X}^2 \times \mathsf{Y}$, $m^\theta(x, x') > 0$ and $g^\theta(x, y) > 0$.

Define the random variables $\widetilde{B}_t^\theta\langle Y_t \rangle = |w^\theta\langle Y_t \rangle|_\infty / p_{\mu,t}^\theta(Y_t)$ and for all $(t, \ell) \in \mathbb{N} \times \mathbb{N}^+$,

$$\widetilde{B}_t^\theta\langle Y_{t:t+\ell} \rangle := \frac{|w^\theta\langle Y_t \rangle|_\infty |Q^\theta\langle Y_{t+1:t+\ell} \rangle \mathbf{1}|_\infty}{p_{\mu,t}^\theta(Y_{t:t+\ell})} \, , \tag{13}$$

$$\widetilde{C}_t^\theta\langle Y_{t:t+\ell} \rangle := \frac{|w^\theta\langle Y_t \rangle|_\infty \int \lambda(\mathrm{d}x_{t+1}) g^\theta(x_{t+1}, Y_{t+1}) Q^\theta\langle Y_{t+2:t+\ell} \rangle \mathbf{1}(x_{t+1})}{p_{\mu,t}^\theta(Y_{t:t+\ell})} \, , \tag{14}$$

where λ is defined in (A-2).

(A-4) There exist constants $(\ell_\star, \alpha) \in \mathbb{N}^+ \times (0, 1)$ such that,

$$\sup_{t \in \mathbb{N}} \sup_{\theta \in \Theta} \mathbb{E}_\mu^\theta \left[(\widetilde{B}_t^\theta\langle Y_{t:t+\ell} \rangle)^\alpha \right] < \infty \, , \quad \text{for all } \ell \in \{0, \ldots, \ell_\star - 1\} \text{ and } , \tag{15}$$

$$\sup_{t \in \mathbb{N}} \sup_{\theta \in \Theta} \mathbb{E}_\mu^\theta \left[(\widetilde{C}_t^\theta\langle Y_{t:t+\ell_\star} \rangle)^\alpha \right] < \infty \, . \tag{16}$$

Theorem 6. *Assume that (A-1), (A-2), (A-3), and (A-4) hold. Let $\theta_\star \in \Theta$ and let $\{\theta_T, \, T \in \mathbb{N}\}$ be a sequence of parameters such that*

$$\limsup_{T \to \infty} T \bar{\mathbb{E}}^{\theta_\star} \left[\ln \left(\frac{m^{\theta_\star}(X_0, X_1) g^{\theta_\star}(X_1, Y_1)}{m^{\theta_T}(X_0, X_1) g^{\theta_T}(X_1, Y_1)} \right) \right] < \infty \, . \tag{17}$$

Furthermore, assume that

$$\bar{\mathbb{E}}^{\theta_\star}\left[\ln\left(\frac{\pi^{\theta_\star}(X_0)}{\mu(X_0)}\right)\right] < \infty\,, \tag{18}$$

where μ is the initial distribution used in the PG algorithm. Then, for all $0 \leq \gamma < \alpha$ (where α is defined in (A-4)) and for all sequences of positive integers $\{N_T\}_{T\geq 1}$ such that $N_T \sim T^{1/\gamma}$, the sequence $\{\epsilon^{-1}_{T,N_T}(\theta_T)\}_{T\geq 1}$, defined in (12) is $\bar{\mathbb{P}}^{\theta_\star}$-tight (bounded in probability).

Proof. The proof is postponed to Section 7. □

To understand condition (17), note that for any $\theta \in \Theta$,

$$\mathrm{D}(\theta_\star||\theta) := \bar{\mathbb{E}}^{\theta_\star}\left[\ln\left(\frac{m^{\theta_\star}(X_0,X_1)g^{\theta_\star}(X_1,Y_1)}{m^{\theta}(X_0,X_1)g^{\theta}(X_1,Y_1)}\right)\right]\,,$$

is the expectation under the stationary distribution $\pi^{\theta_\star}$ of the Kullback-Leibler divergence between the conditional distribution of $p^{\theta_\star}(X_1,Y_1|X_0)$ and $p^{\theta}(X_1,Y_1|X_0)$. Hence, $\mathrm{D}(\theta_\star||\theta) \geq 0$ for all $\theta \in \Theta$ and $\mathrm{D}(\theta_\star||\theta_\star) = 0$. Assuming that $\theta_\star$ belongs to the interior of Θ and that the function $\theta \mapsto \mathrm{D}(\theta_\star||\theta)$ is twice differentiable at $\theta_\star$, a Taylor expansion at $\theta_\star$ yields

$$\mathrm{D}(\theta_\star||\theta) = \frac{1}{2}(\theta_\star - \theta)^t H^{\theta_\star}(\theta_\star - \theta) + o(\|\theta_\star - \theta\|^2)\,,$$

where H^θ is the Hessian of $\theta \mapsto \mathrm{D}(\theta_\star||\theta)$. Consequently, for regular statistical models, (17) holds provided that θ_T converges to $\theta_\star$ at a rate $1/\sqrt{T}$, i.e., $\theta_T = \theta_\star + \varrho_T/\sqrt{T}$, where the sequence $\{\varrho_T,\ T \in \mathbb{N}\}$ is bounded: $\sup_{T\geq 0}\|\varrho_T\| < \infty$.

Remark 7. *It should be noted that our results do not cover explicitly the case when the sequence $\{\varrho_T,\ T \in \mathbb{N}\}$ is stochastic. Still, we believe that our results hint at the possibility of obtaining a non-degenerate lower bound on the minorizing constant also in the stochastic case, given that $\{\varrho_T,\ T \in \mathbb{N}\}$ is tight, under conditions that are much weaker than the previously considered strong mixing assumption.*

Remark 8. *It is interesting to note that we do not require the initial distribution μ to be equal to $\pi^{\theta_\star}$, but only that the Kullback-Leibler divergence (18) is bounded. Hence, we may use a quite arbitrary initial distribution and still obtain a sequence of inverse minorization constants that is tight with respect to $\bar{\mathbb{P}}^{\theta_\star}$.*

A straightforward generalization of the above result is to let the initial distribution belong to a parametric family of distributions, $\{\mu^\theta : \theta \in \Theta\}$. The condition (18) should then be replaced by $\limsup_{T\to\infty}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\left(\frac{\pi^{\theta_\star}(X_0)}{\mu^{\theta_T}(X_0)}\right)\right] < \infty$. Allowing for the initial distribution to depend on θ can be useful in some cases. For instance, if the stationary distribution π^θ is known it may serve as a natural choice for the initial distribution used in the algorithm.

5 Example – A stochastic volatility model

In this section we consider a concrete example to illustrate how the conditions of Theorem 6 can be verified in practice. An additional example is provided in the online supplementary material. We start by stating a technical lemma which will be very useful for checking the assumptions.

Lemma 9. *Let $(\mathsf{Z}, \mathcal{Z})$ be a measurable set and ξ be a measure on $(\mathsf{Z}, \mathcal{Z})$. Let $\alpha \in (0,1)$ and let φ, ψ and q be nonnegative measurable functions, such that*

$$\int \psi(z)\varphi(z)\xi(\mathrm{d}z) < \infty \,, \tag{19}$$

$$\int \varphi^{-\frac{\alpha}{1-\alpha}}(z)q(z)\xi(\mathrm{d}z) < \infty \,. \tag{20}$$

Then, $\int \psi^{\alpha}(z)q^{1-\alpha}(z)\xi(\mathrm{d}z) < \infty$.

Proof. The result follows from Hölder's inequality; see the online supplementary material. □

The canonical model in stochastic volatility for discrete-time data has been introduced by Taylor (1982) and worked out since then by many authors; see Shephard and Andersen (2009) for an up-to-date survey. In this model, the hidden volatility process, $\{X_t,\ t \in \mathbb{N}\}$, follows a first order autoregression,

$$X_{t+1} = \phi X_t + \sigma W_{t+1} \,, \tag{21}$$

$$Y_t = \beta \exp(X_t/2) U_t \,. \tag{22}$$

where $\{W_t,\ t \in \mathbb{N}\}$ and $\{U_t,\ t \in \mathbb{N}\}$ are mutually independent white Gaussian noises with mean zero and unit variance. We denote by $\theta = (\phi, \sigma, \beta) \in \Theta$, where Θ is a compact subset of $(-1,1) \times (0,\infty)^2$. For $\delta > 0$, denote by $V_\delta(x) = \mathrm{e}^{\delta|x|}$ and let μ an arbitrary distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, for which $\mu(V_\delta) < \infty$.

For this model the transition kernel and the likelihood of the observation are given by

$$m^\theta(x,x') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x' - \phi x)^2\right) \quad \text{and} \quad g^\theta(x,y) = \frac{1}{\sqrt{2\pi\beta^2}} \mathrm{e}^{-(x/2 + (y^2/2\beta^2)\mathrm{e}^{-x})} \,,$$

respectively. For any $\theta \in \Theta$, the autoregressive process $\{X_t,\ t \in \mathbb{N}\}$ has a unique stationary distribution π^θ, which is zero-mean Gaussian with variance $\sigma^2/(1-\phi^2)$. Hence, (A-1) is satisfied.

We consider the bootstrap proposal kernel as defined in (7), in which case $w^\theta\langle y\rangle(x,x') = g^\theta(x',y)$ for all $(x,x') \in \mathbb{R}\times\mathbb{R}$ and $y \in \mathbb{R}$. Note that, $|w^\theta\langle y\rangle|_\infty = |g^\theta(\cdot,y)|_\infty$. Assumptions (A-2)

and (A-3) are readily satisfied. We finally check (A-4). It is easily shown that, for all $\theta \in \Theta$,

$$\int_{-\infty}^{\infty} g^\theta(x,y)\mathrm{d}x = \frac{D_1}{|y|}\,, \qquad D_1 := \frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{\mathrm{e}^{-u/2}}{\sqrt{u}}\mathrm{d}u\,, \tag{23}$$

$$\sup_{x\in\mathbb{R}} g^\theta(x,y) = \frac{D_2}{|y|}\,, \qquad D_2 := \frac{1}{\sqrt{2\pi\mathrm{e}}}\,. \tag{24}$$

We will check (A-4) with $\ell_\star = 1$, i.e., we show that

$$\sup_{t\in\mathbb{N}}\sup_{\theta\in\Theta}\mathbb{E}_\mu^\theta\left[(\widetilde{B}_t^\theta\langle Y_t\rangle)^\alpha\right] < \infty\,, \qquad \text{and} \qquad \sup_{t\in\mathbb{N}}\sup_{\theta\in\Theta}\mathbb{E}_\mu^\theta\left[(\widetilde{C}_t^\theta\langle Y_{t:t+1}\rangle)^\alpha\right] < \infty\,, \tag{25}$$

for any $\alpha \in (0,1)$. Note that we cannot expect (25) to hold with $\alpha = 1$ since,

$$\mathbb{E}_\mu^\theta\left[\widetilde{B}_t^\theta\langle Y_t\rangle\right] = \mathbb{E}_\mu^\theta\left[\frac{|w^\theta\langle Y_t\rangle|_\infty}{p_{\mu,t}^\theta(Y_t)}\right] = \int \sup_{x\in\mathbb{R}} g^\theta(x,y)\mathrm{d}y = D_2\int\frac{1}{|y|}\mathrm{d}y = \infty\,.$$

We now turn to the proof of (25). Note first that $\limsup_{|x|\to\infty}\sup_{\theta\in\Theta}\frac{\mathbb{E}_x^\theta[V_\delta(X_1)]}{V_\delta(x)} = 0$, and, for any $K < \infty$, $\sup_{|x|\le K}\sup_{\theta\in\Theta}\mathbb{E}_x^\theta\left[V_\delta(X_1)\right] < \infty$. Therefore, there exist constants $\lambda_\delta \in (0,1)$ and $b_\delta < \infty$ such that, for all $x \in \mathsf{X}$, $\sup_{\theta\in\Theta}\mathbb{E}_x^\theta\left[V_\delta(X_1)\right] \le \lambda_\delta V_\delta(x) + b_\delta$. This implies that, for all $\delta > 0$, $\sup_{\theta\in\Theta}\mathbb{E}_x^\theta\left[V_\delta(X_t)\right] \le \lambda_\delta^t V_\delta(x) + b_\delta(1+\lambda_\delta+\cdots+\lambda_\delta^{t-1}) \le V_\delta(x) + b_\delta/(1-\lambda_\delta)$, and thus

$$\sup_{t\in\mathbb{N}}\sup_{\theta\in\Theta}\mathbb{E}_\mu^\theta\left[V_\delta(X_t)\right] \le \mu(V_\delta) + b_\delta/(1-\lambda_\delta) < \infty\,. \tag{26}$$

Using the fact that $w^\theta\langle y\rangle(x,x') = g^\theta(x',y)$ and (24), we get

$$\mathbb{E}_\mu^\theta\left[(\widetilde{B}_t^\theta\langle Y_t\rangle)^\alpha\right] = \int |w^\theta\langle y_t\rangle|_\infty^\alpha\{p_{\mu,t}^\theta(y_t)\}^{1-\alpha}\mathrm{d}y_t \le D_2^\alpha\int|y_t|^{-\alpha}\{p_{\mu,t}^\theta(y_t)\}^{1-\alpha}\mathrm{d}y_t\,. \tag{27}$$

We apply Lemma 9 to establish a bound for (27). Consider the functions φ and ψ given by

$$\psi(y) = \frac{1}{|y|}\,, \qquad \text{and} \qquad \varphi(y) = \frac{|y|^\gamma}{|y|^2\vee 1}\,,$$

with $\gamma\alpha/(1-\alpha) < 1$ and $\gamma \in (0,1)$. With these definitions, we get $\int\varphi(y)\psi(y)\mathrm{d}y = \int\frac{1}{|y|}\frac{|y|^\gamma}{|y|^2\vee 1}\mathrm{d}y < \infty$, showing that the first condition, (19), is satisfied. We now check (20):

$$\int\varphi^{-\frac{\alpha}{1-\alpha}}(y_t)p_{\mu,t}^\theta(y_t)\mathrm{d}y_t = \mathbb{E}_\mu^\theta\left[\mathbb{E}_{X_t}^\theta\left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0)\right]\right]\,. \tag{28}$$

Since $Y_0 = \beta\exp(X_0/2)U_0$ it follows that $\mathbb{E}_x^\theta\left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0)\right] = \mathbb{E}\left[\varphi^{-\frac{\alpha}{1-\alpha}}(\beta\mathrm{e}^{x/2}U)\right]$ where U is standard normal. We have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\beta^2\mathrm{e}^x U^2\vee 1}{\beta^\gamma\mathrm{e}^{\gamma x/2}|U|^\gamma}\right)^{\frac{\alpha}{1-\alpha}}\right] &\le \mathbb{E}\left[(\beta\mathrm{e}^{x/2}|U|)^{-\frac{\gamma\alpha}{1-\alpha}}\mathbb{1}_{\{\beta|U|\mathrm{e}^{x/2}\le 1\}}\right] + \mathbb{E}\left[(\beta^2\mathrm{e}^x U^2)^{\frac{\alpha}{1-\alpha}}\mathbb{1}_{\{\beta|U|\mathrm{e}^{x/2}>1\}}\right]\\ &\le (\beta\mathrm{e}^{\frac{x}{2}})^{-\frac{\gamma\alpha}{1-\alpha}}\mathbb{E}\left[|U|^{-\frac{\gamma\alpha}{1-\alpha}}\right] + (\beta^2\mathrm{e}^x)^{\frac{\alpha}{1-\alpha}}\mathbb{E}\left[|U|^{\frac{2\alpha}{1-\alpha}}\right]\,.\end{aligned}$$

Since $\gamma\alpha/(1-\alpha) < 1$ it holds that $\mathbb{E}\left[|U|^{-\frac{\gamma\alpha}{1-\alpha}}\right] < \infty$ and, additionally, $\mathbb{E}\left[|U|^{\frac{2\alpha}{1-\alpha}}\right] < \infty$. Therefore, there exist constants $D_3 < \infty$ and $\delta > 0$ such that, for all $x \in \mathbb{R}$ and $\theta \in \Theta$,

$$\mathbb{E}_x^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0)\right] \leq D_3 \mathrm{e}^{\delta|x|} = D_3 V_\delta(x) \; . \tag{29}$$

Using (28), (29) and (26) verifies the second condition in (20). Lemma 9 can thus be used to conclude that $\mathbb{E}_\mu^\theta \left[(\widetilde{B}_t^\theta \langle Y_t \rangle)^\alpha\right] < \infty$ for all $\alpha \in (0,1)$. Since this holds for any $t \in \mathbb{N}$ and $\theta \in \Theta$, we establish the first part of (25).

Next we check that, for all $\alpha \in (0,1)$, $\mathbb{E}_\mu^\theta \left[(\widetilde{C}_t^\theta \langle Y_{t:t+1} \rangle)^\alpha\right] < \infty$. Using (23) and (24), we get

$$\begin{aligned}
\mathbb{E}_\mu^\theta \left[(\widetilde{C}_t^\theta \langle Y_{t:t+1} \rangle)^\alpha\right] &= \mathbb{E}^\theta \left[\frac{|w^\theta \langle Y_t \rangle|_\infty^\alpha \left(\int g^\theta(x_{t+1}, Y_{t+1}) \mathrm{d}x_{t+1}\right)^\alpha}{(p_{\mu,t}^\theta(Y_{t:t+1}))^\alpha}\right] \\
&= \iint |w^\theta \langle y_t \rangle|_\infty^\alpha \left(\int g^\theta(x_{t+1}, y_{t+1}) \mathrm{d}x_{t+1}\right)^\alpha (p_{\mu,t}^\theta(y_{t:t+1}))^{1-\alpha} \mathrm{d}y_{t:t+1} \; , \\
&\leq D_1^\alpha D_2^\alpha \iint |y_t y_{t+1}|^{-\alpha} (p_{\mu,t}^\theta(y_{t:t+1}))^{1-\alpha} \mathrm{d}y_{t:t+1} \; .
\end{aligned} \tag{30}$$

We use again Lemma 9 with

$$\psi(y_0, y_1) = \frac{1}{|y_0||y_1|} \; , \qquad \text{and} \qquad \varphi(y_0, y_1) = \frac{|y_0|^\gamma |y_1|^\gamma}{(y_0^2 \vee 1)\,(y_1^2 \vee 1)} \; ,$$

with $\gamma\alpha/(1-\alpha) < 1$ and $\gamma \in (0,1)$. Note first that

$$\iint \psi(y_0, y_1)\varphi(y_0, y_1) \mathrm{d}y_{0:1} = \iint \left\{(|y_0||y_1|)^{1-\gamma}(y_0^2 \vee 1)(y_1^2 \vee 1)\right\}^{-1} \mathrm{d}y_{0:1} < \infty \; . \tag{31}$$

Hence, (19) is satisfied. We finally check (20). Using the conditional independence of the observations given the states and (29),

$$\begin{aligned}
&\iint \varphi^{-\frac{\alpha}{1-\alpha}}(y_{t:t+1}) p_{\mu,t}^\theta(y_{t:t+1}) \mathrm{d}y_{t:t+1} = \mathbb{E}_\mu^\theta \left[\mathbb{E}_\mu^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_{t:t+1}) \,\middle|\, X_{t:t+1}\right]\right] \\
&= \mathbb{E}_\mu^\theta \left[\mathbb{E}_{X_t}^\theta \left[\left(\frac{\beta^2 \mathrm{e}^{X_0} U^2 \vee 1}{\beta \mathrm{e}^{\gamma X_0/2} |U|^\gamma}\right)^{\frac{\alpha}{1-\alpha}}\right] \mathbb{E}_{X_{t+1}}^\theta \left[\left(\frac{\beta^2 \mathrm{e}^{X_0} U^2 \vee 1}{\beta \mathrm{e}^{\gamma X_0/2} |U|^\gamma}\right)^{\frac{\alpha}{1-\alpha}}\right]\right] \leq D_3^2 \mathbb{E}_\mu^\theta \left[\mathrm{e}^{\delta|X_t|} \mathrm{e}^{\delta|X_{t+1}|}\right] \; .
\end{aligned}$$

From the Cauchy-Schwarz inequality we get,

$$\mathbb{E}_\mu^\theta \left[\mathrm{e}^{\delta|X_t|} \mathrm{e}^{\delta|X_{t+1}|}\right] \leq \left(\mathbb{E}_\mu^\theta \left[\mathrm{e}^{2\delta|X_t|}\right] \mathbb{E}_\mu^\theta \left[\mathrm{e}^{2\delta|X_{t+1}|}\right]\right)^{1/2} \; ,$$

Applying (26) with δ replaced by 2δ yields (20). Using Lemma 9 thus establishes (25) and thereby, (A-4) holds.

Provided that θ_T converges to $\theta_\star$ at a rate $1/\sqrt{T}$, we may therefore apply Theorem 6 which shows that for any $\gamma \in (0,1)$, $\{\epsilon_{T,N_T}^{-1}(\theta_T)\}_{T \geq 1}$ is tight with $N_T = T^{1/\gamma}$.

6 Proof of Theorem 1

We will now turn to the proof of the minorization condition in Theorem 1. As in the statement of the theorem, we will not explicitly indicate any possible dependence on unknown model parameters in the notation in this section. This is done for notational convenience and is without loss of generality. Throughout this section, $\mathbb{P}$ and $\mathbb{E}$ refer to probability and expectation, respectively, with respect to the random variables generated by the PG algorithm (the observation sequence $y_{0:T}$ is treated as fixed). The proof is inductive and follows from a series of lemmas.

Lemma 10. *Let $X \geq 0$ and $Y > 0$ be independent random variables. Then, $\mathbb{E}\left[\frac{X}{Y}\right] \geq \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}$.*

Proof. Since $f(y) = 1/y$ is convex on $y > 0$ the result follows by Jensen's inequality. □

Lemma 11. *Let f and h be nonnegative measurable functions. For $t \in \{0, \ldots, T-1\}$, we have*

$$\mathbb{E}\left[\left.\frac{\sum_{i=1}^N \omega_{t+1}^i f(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h(X_{t+1}^i)}\right| \mathcal{F}_t^N\right] \geq \frac{\sum_{i=1}^N \omega_t^i \int Q\langle y_{t+1}\rangle(X_t^i, \mathrm{d}x_{t+1}) f(X_{0:t}^i, x_{t+1})}{\sum_{i=1}^N \omega_t^i \left[\frac{N-2}{N-1} Q\langle y_{t+1}\rangle h(X_t^i) + \frac{2}{N-1}\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h(x')\right]}, \quad (32)$$

and

$$\mathbb{E}\left[\frac{\sum_{i=0}^N \omega_0^i f(X_0^i)}{\sum_{i=0}^N \omega_0^i h(X_0^i)}\right] \geq \frac{(N-1)\mu(g(\cdot, y_0) f(\cdot))}{(N-2)\mu(g(\cdot, y_0) h(\cdot)) + 2\sup_x [w\langle y_0\rangle(x) h(x)]}. \quad (33)$$

Proof. Using that $\omega_{t+1}^1 h(X_{t+1}^1) + \omega_{t+1}^N h(X_{t+1}^N) \leq 2\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h(x')$, and that the weighted particles $\{(X_{t+1}^i, \omega_{t+1}^i)\}_{i=1}^{N-1}$ are conditionally i.i.d. with respect to $\mathcal{F}_t^N$, we get

$$\begin{aligned}
\mathbb{E}\left[\left.\frac{\sum_{i=1}^N \omega_{t+1}^i f(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h(X_{t+1}^i)}\right| \mathcal{F}_t^N\right] &\geq \mathbb{E}\left[\left.\frac{\sum_{i=1}^{N-1} \omega_{t+1}^i f(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h(X_{t+1}^i)}\right| \mathcal{F}_t^N\right] \\
&\geq (N-1)\mathbb{E}\left[\left.\frac{\omega_{t+1}^1 f(X_{0:t+1}^1)}{\sum_{i=2}^{N-1} \omega_{t+1}^i h(X_{t+1}^i) + 2\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h(x')}\right| \mathcal{F}_t^N\right] \\
&\geq (N-1)\frac{\mathbb{E}\left[\omega_{t+1}^1 f(X_{0:t+1}^1) \,\middle|\, \mathcal{F}_t^N\right]}{\mathbb{E}\left[\left.\sum_{i=2}^{N-1} \omega_{t+1}^i h(X_{t+1}^i)\right| \mathcal{F}_t^N\right] + 2\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h(x')}, \quad (34)
\end{aligned}$$

where the last inequality follows from Lemma 10. Consider first the numerator in the right-hand side of (34). We have

$$\begin{aligned}
\mathbb{E}\left[\omega_{t+1}^1 f(X_{0:t+1}^1) \,\middle|\, \mathcal{F}_t^N\right] &= \frac{1}{\sum_{l=1}^N \omega_t^l} \sum_{j=1}^N \omega_t^j \int R\langle y_{t+1}\rangle(X_t^j, \mathrm{d}x_{t+1}) w\langle y_{t+1}\rangle(X_t^j, x_{t+1}) f(X_{0:t}^j, x_{t+1}) \\
&= \frac{1}{\sum_{l=1}^N \omega_t^l} \sum_{j=1}^N \omega_t^j \int Q\langle y_{t+1}\rangle(X_t^j, \mathrm{d}x_{t+1}) f(X_{0:t}^j, x_{t+1}). \quad (35)
\end{aligned}$$

We now consider the denominator in the right-hand side of (34):

$$\mathbb{E}\left[\sum_{i=2}^{N-1}\omega_{t+1}^i h(X_{t+1}^i)\,\middle|\,\mathcal{F}_t^N\right] = (N-2)\mathbb{E}\left[\omega_{t+1}^1 h(X_{t+1}^1)\,\middle|\,\mathcal{F}_t^N\right] = \frac{(N-2)}{\sum_{l=1}^N \omega_t^l}\sum_{j=1}^N \omega_t^j Q\langle y_{t+1}\rangle h(X_t^j)\;,$$

where the last identity follows from (35) with $f(x_{0:t+1}) = h(x_{t+1})$. The proof of (32) follows.

Consider now (33). Since the particles $\{X_0^i\}_{i=1}^{N-1}$ are i.i.d., we obtain, using again Lemma 10,

$$\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_0^i f(X_0^i)}{\sum_{i=1}^N \omega_0^i h(X_0^i)}\right] \geq \frac{(N-1)\mathbb{E}\left[\omega_0^1 f(X_0^1)\right]}{\mathbb{E}\left[\sum_{i=2}^{N-1}\omega_0^i h(X_0^i)\right] + 2\sup_x w\langle y_0\rangle(x)h(x)}\;.$$

The numerator is given by $\mathbb{E}\left[\omega_0^1 f(X_0^1)\right] = \int r_0\langle y_0\rangle(\mathrm{d}x_0) w\langle y_0\rangle(x_0) f(x_0) = \mu(g(\cdot, y_0) f(\cdot))$. Similarly, we get $\mathbb{E}\left[\sum_{i=2}^{N-1}\omega_0^i h(X_0^i)\right] = (N-2)\mathbb{E}\left[\omega_0^1 h(X_0^1)\right] = (N-2)\mu(g(\cdot, y_0)h(\cdot))$. □

Define a sequence of nonnegative scalars $\{\beta_t\}_{t=0}^T$ by the backward recursion: $\beta_T = |w\langle y_T\rangle|_\infty$, and for $t = T-1, T-2, \ldots, 0$,

$$\beta_t = |w\langle y_t\rangle|_\infty \left\{ \tfrac{2}{N-1}\sum_{\ell=1}^{T-t}\left(\tfrac{N-2}{N-1}\right)^{\ell-1}\beta_{t+\ell}\,|Q\langle y_{t+1:t+\ell-1}\rangle\mathbf{1}|_\infty + \left(\tfrac{N-2}{N-1}\right)^{T-t}|Q\langle y_{t+1:T}\rangle\mathbf{1}|_\infty \right\}\;. \tag{36}$$

Given $\{\beta_t\}_{t=0}^T$, define the functions $\{h_t\}_{t=0}^T$, $h_t : \mathsf{X} \to \mathbb{R}_+$, by the backward recursion: $h_T = \mathbf{1}$ and, for all $t = T-1, T-2, \ldots, 0$,

$$h_t : x \mapsto h_t(x) = \frac{N-2}{N-1}\, Q\langle y_{t+1}\rangle h_{t+1}(x) + \frac{2}{N-1}\beta_{t+1}\;. \tag{37}$$

By solving the backward recursion, (37) implies

$$h_t(x) = \frac{2}{N-1}\sum_{\ell=1}^{T-t}\left(\frac{N-2}{N-1}\right)^{\ell-1}\beta_{t+\ell} Q\langle y_{t+1:t+\ell-1}\rangle\mathbf{1}(x) + \left(\frac{N-2}{N-1}\right)^{T-t} Q\langle y_{t+1:T}\rangle\mathbf{1}(x)\;. \tag{38}$$

For $D \in \mathcal{X}^{\otimes(T+1)}$, set $f_T(x_{0:T}) = \mathbb{1}_D(x_{0:T})$ and, for $t = T-1, T-2, \ldots, 0$,

$$f_t(x_{0:t}) = \int Q\langle y_{t+1}\rangle(x_t, \mathrm{d}x_{t+1}) f_{t+1}(x_{0:t+1})\;, \tag{39}$$

or equivalently, $f_t(x_{0:t}) = \int \prod_{\ell=1}^{T-t} Q\langle y_{t+\ell}\rangle(x_{t+\ell-1}, \mathrm{d}x_{t+\ell})\mathbb{1}_D(x_{0:T})$.

Lemma 12. *For any $x'_{0:T} \in \mathsf{X}^{T+1}$ and $D \in \mathcal{X}^{\otimes(T+1)}$,*

$$P_{T,N}(x'_{0:T}, D) = \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_T^i \mathbb{1}_D(X_{0:T}^i)}{\sum_{i=1}^N \omega_T^i}\right] \geq \frac{(N-1)p_\mu(y_{0:T})}{(N-2)\mu(g(\cdot, y_0)h_0(\cdot)) + 2\beta_0}\phi_{\mu,0:T}\langle y_{0:T}\rangle(D)\;.$$

Proof. Note first that, by construction,

$$\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_T^i \mathbb{1}_D(X_{0:T}^i)}{\sum_{i=1}^N \omega_T^i}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_T^i f_T(X_{0:T}^i)}{\sum_{i=1}^N \omega_T^i h_T(X_T^i)}\right] . \tag{40}$$

We now show that, by backward induction, for all $t \in \{0, \dots, T-1\}$,

$$\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_{t+1}^i f_{t+1}(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h_{t+1}(X_{t+1}^i)}\right] \geq \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_t^i f_t(X_{0:t}^i)}{\sum_{i=1}^N \omega_t^i h_t(X_t^i)}\right] . \tag{41}$$

To obtain (41), note first that the tower property of the conditional expectation, Lemma 11, and (39) imply

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_{t+1}^i f_{t+1}(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h_{t+1}(X_{t+1}^i)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_{t+1}^i f_{t+1}(X_{0:t+1}^i)}{\sum_{i=1}^N \omega_{t+1}^i h_{t+1}(X_{t+1}^i)} \,\middle|\, \mathcal{F}_t^N\right]\right] \\ &\geq \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_t^i f_t(X_{0:t}^i)}{\sum_{i=1}^N \omega_t^i \left[\frac{N-2}{N-1} Q\langle y_{t+1}\rangle h_{t+1}(X_t^i) + \frac{2}{N-1}\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h_{t+1}(x')\right]}\right] .\end{aligned}$$

It follows directly from (36) and (38) that

$$\sup_x w\langle y_0\rangle(x) h_0(x) \leq \beta_0 , \tag{42}$$

$$\sup_{x,x'} w\langle y_{t+1}\rangle(x,x') h_{t+1}(x') \leq \beta_{t+1} , \quad t \in \{0, \dots, T-1\} . \tag{43}$$

Combining the inequality (43) with the definition of h_t in (37) yields

$$\sum_{i=1}^N \omega_t^i \left[\frac{N-2}{N-1} Q\langle y_{t+1}\rangle h_{t+1}(X_t^i) + \frac{2}{N-1}\sup_{(x,x')} w\langle y_{t+1}\rangle(x,x') h_{t+1}(x')\right] \leq \sum_{i=1}^N \omega_t^i h_t(X_t^i) ,$$

showing (41). Combining (41) with (40) and using Lemma 11-(33) establishes that

$$\mathbb{E}\left[\frac{\sum_{i=1}^N \omega_T^i \mathbb{1}_D(X_{0:T}^i)}{\sum_{i=1}^N \omega_T^i}\right] \geq \mathbb{E}\left[\frac{\sum_{i=1}^N \omega_0^i f_0(X_0^i)}{\sum_{i=1}^N \omega_0^i h_0(X_0^i)}\right] \geq \frac{(N-1)\mu(g(\cdot, y_0) f_0(\cdot))}{(N-2)\mu(g(\cdot, y_0) h_0(\cdot)) + 2\beta_0} ,$$

where the last inequality stems from (42). The proof is completed by noting that

$$\mu(g(\cdot, y_0) f_0(\cdot)) = \phi_{\mu,0:T}\langle y_{0:T}\rangle(D) p_\mu(y_{0:T}) .$$

□

Finally, to prove Theorem 1 it remains to show the following.

Lemma 13. *With $B_{t,T}$ defined as in* (10)*, it holds that*

$$(N-2)\mu(g(\cdot, y_0) h_0(\cdot)) + 2\beta_0 \leq (N-1) p_\mu(y_{0:T}) \left[\prod_{t=0}^T \frac{2B_{t,T} + N - 2}{N-1}\right] . \tag{44}$$

Proof. Define for $t \in \{0, \dots, T\}$,

$$\alpha_t = \frac{\beta_t}{p_\mu(y_{t:T}|y_{0:t-1})} , \tag{45}$$

with the convention $p_\mu(y_{0:T}|y_{0:-1}) = p_\mu(y_{0:T})$. In particular, $\alpha_0 = \beta_0/p_\mu(y_{0:T})$.

Eq. (36) implies

$$\alpha_t = |w\langle y_t\rangle|_\infty \left\{ \frac{2}{N-1} \sum_{\ell=1}^{T-t} \left(\frac{N-2}{N-1}\right)^{\ell-1} \alpha_{t+\ell} \left[\frac{|Q\langle y_{t+1:t+\ell-1}\rangle \mathbf{1}|_\infty p_\mu(y_{t+\ell:T}|y_{0:t+\ell-1})}{p_\mu(y_{t:T}|y_{0:t-1})} \right] \right.$$
$$\left. + \left(\frac{N-2}{N-1}\right)^{T-t} \frac{|Q\langle y_{t+1:T}\rangle \mathbf{1}|_\infty}{p_\mu(y_{t:T}|y_{0:t-1})} \right\} .$$

The identity

$$\frac{p_\mu(y_{t+\ell:T}|y_{0:t+\ell-1})}{p_\mu(y_{t:T}|y_{0:t-1})} = \frac{1}{p_\mu(y_{t:t+\ell-1}|y_{0:t-1})}$$

and the definition in (10) imply that

$$\alpha_t \le B_{t,T} \left\{ \frac{2}{N-1} \sum_{\ell=1}^{T-t} \left(\frac{N-2}{N-1}\right)^{\ell-1} \alpha_{t+\ell} + \left(\frac{N-2}{N-1}\right)^{T-t} \right\} . \tag{46}$$

By a backward induction, define the sequence $\{\tilde{\alpha}_t\}_{t=0}^T$ as follows: set $\tilde{\alpha}_T = B_{T,T}$ and

$$\tilde{\alpha}_t = B_{t,T} \left[\frac{2}{N-1} \sum_{\ell=1}^{T-t} \left(\frac{N-2}{N-1}\right)^{\ell-1} \tilde{\alpha}_{t+\ell} + \left(\frac{N-2}{N-1}\right)^{T-t} \right] . \tag{47}$$

Since by construction, $\alpha_T \le B_{T,T} = \tilde{\alpha}_T$, an elementary backward recursion using (46) shows that $\alpha_t \le \tilde{\alpha}_t$ for all $t \in \{0, \dots, T\}$. However

$$\begin{aligned}
\tilde{\alpha}_{t-1} &= B_{t-1,T} \left[\frac{2}{N-1} \sum_{s=1}^{T-t+1} \left(\frac{N-2}{N-1}\right)^{\ell-1} \tilde{\alpha}_{t+\ell-1} + \left(\frac{N-2}{N-1}\right)^{T-t+1} \right] \\
&= B_{t-1,T} \left[\frac{2}{N-1} \tilde{\alpha}_t + \frac{2}{N-1} \sum_{k=1}^{T-t} \left(\frac{N-2}{N-1}\right)^{k} \tilde{\alpha}_{t+k} + \left(\frac{N-2}{N-1}\right)^{T-t+1} \right] \\
&= \frac{2B_{t-1,T}}{N-1} \tilde{\alpha}_t + B_{t-1,T} \frac{N-2}{N-1} \frac{\tilde{\alpha}_t}{B_{t,T}} = \frac{B_{t-1,T}}{B_{t,T}} \left[\frac{2B_{t,T}}{N-1} + \frac{N-2}{N-1} \right] \tilde{\alpha}_t .
\end{aligned}$$

Therefore

$$\tilde{\alpha}_0 = B_{0,T} \prod_{t=1}^{T} \frac{2B_{t,T} + N - 2}{N-1} . \tag{48}$$

Now, since

$$h_0(x) = \frac{2}{N-1} \sum_{s=1}^{T} \left(\frac{N-2}{N-1}\right)^{s-1} \beta_s Q\langle y_{1:s-1}\rangle \mathbf{1}(x) + \left(\frac{N-2}{N-1}\right)^{T} Q\langle y_{1:T}\rangle \mathbf{1}(x) ,$$

we have

$$\mu(g(\cdot, y_0) h_0(\cdot)) = \frac{2}{N-1} \sum_{s=1}^{T} \left(\frac{N-2}{N-1}\right)^{s-1} \beta_s p_\mu(y_{0:s-1}) + \left(\frac{N-2}{N-1}\right)^{T} p_\mu(y_{0:T}) .$$

Plugging (45) into this equation and using that $p_\mu(y_{0:s-1}) = p_\mu(y_{0:T})/p_\mu(y_{s:T}|y_{0:s-1})$ yields

$$\mu(g(\cdot, y_0)h_0(\cdot)) = \frac{2}{N-1}\sum_{s=1}^{T}\left(\frac{N-2}{N-1}\right)^{s-1}\alpha_s p_\mu(y_{0:T}) + \left(\frac{N-2}{N-1}\right)^{T} p_\mu(y_{0:T}) \,.$$

Finally, using that $\alpha_t \le \tilde{\alpha}_t$ for all $t \in \{0, \dots, T\}$ and then (47),

$$\begin{aligned}
&(N-2)\mu(g(\cdot, y_0)h_0(\cdot)) + 2\beta_0 \\
&\quad \le p_\mu(y_{0:T})\left\{(N-2)\left[\frac{2}{N-1}\sum_{s=1}^{T}\left(\frac{N-2}{N-1}\right)^{s-1}\tilde{\alpha}_s + \left(\frac{N-2}{N-1}\right)^{T}\right] + 2\tilde{\alpha}_0\right\} \\
&\quad \le p_\mu(y_{0:T})\left((N-2)\frac{\tilde{\alpha}_0}{B_{0,T}} + 2\tilde{\alpha}_0\right) = p_\mu(y_{0:T})(N-2+2B_{0,T})\prod_{t=1}^{T}\frac{2B_{t,T}+N-2}{N-1} \,,
\end{aligned}$$

where the last equality follows from (48). The proof follows. □

7 Proof of Theorem 6

Define

$$\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle := \frac{|w^\theta\langle Y_t\rangle|_\infty |Q^\theta\langle Y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty}{p_\mu^\theta(Y_{t:t+\ell}|Y_{0:t-1})} \,, \tag{49}$$

$$\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle := \frac{|w^\theta\langle Y_t\rangle|_\infty \int \lambda(\mathrm{d}x_{t+1}) g^\theta(x_{t+1}, Y_{t+1}) Q^\theta\langle Y_{t+2:t+\ell}\rangle \mathbf{1}(x_{t+1})}{p_\mu^\theta(Y_{t:t+\ell}|Y_{0:t-1})} \,. \tag{50}$$

Note that

$$\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle = \widetilde{B}_t^\theta\langle Y_{t:t+\ell}\rangle \frac{p_{\mu,t}^\theta(Y_{t:t+\ell})}{p_\mu^\theta(Y_{t:t+\ell}|Y_{0:t-1})} \quad \text{and} \quad \widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle = \widetilde{C}_t^\theta\langle Y_{t:t+\ell}\rangle \frac{p_{\mu,t}^\theta(Y_{t:t+\ell})}{p_\mu^\theta(Y_{t:t+\ell}|Y_{0:t-1})} \,, \tag{51}$$

where $\widetilde{B}_t^\theta\langle Y_{t:t+\ell}\rangle$ and $\widetilde{C}_t^\theta\langle Y_{t:t+\ell}\rangle$ are defined in (13) and (14), respectively.

Lemma 14. *For all $\theta \in \Theta$, the sequence $\{\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle\}_{\ell\ge 0}$ defined in (50) is a $(\mathbb{P}_\mu^\theta, \{\mathcal{F}_{t+\ell}\}_{\ell\ge 0})$-martingale, where $\mathcal{F}_t = \sigma(Y_{0:t})$.*

Proof. See the online supplementary material. □

Lemma 15. *For all $0 \le \gamma < 1$ and all $\ell \in \mathbb{N}$,*

$$\mathbb{E}_\mu^\theta\left[(\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right] \le \mathbb{E}_\mu^\theta\left[(\widetilde{B}_t^\theta\langle Y_{t:t+\ell}\rangle)^\gamma\right] \,, \quad \textit{and} \quad \mathbb{E}_\mu^\theta\left[(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right] \le \mathbb{E}_\mu^\theta\left[(\widetilde{C}_t^\theta\langle Y_{t:t+\ell}\rangle)^\gamma\right] \,.$$

Proof. See the online supplementary material. □

Lemma 16. *Assume (A-2) and (A-4). Then,* $\sup_{t\geq 0}\sup_{\theta\in\Theta}\mathbb{E}_\mu^\theta\left[\left(\sup_{\ell\geq 0}\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle\right)^\gamma\right]<\infty$ *for all* $0\leq\gamma<\alpha$*, where* α *is defined in (A-4).*

Proof. Under (A-2), we obtain by definitions of $\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle$ and $\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle$,

$$\sup_{\ell\geq 0}\widehat{B}_t^\theta\langle Y_{0:\ell}\rangle\leq\sum_{\ell=0}^{\ell_\star-1}\widehat{B}_t^\theta\langle Y_{0:\ell}\rangle+\sup_{\ell\geq\ell_\star}\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle\leq\sum_{\ell=0}^{\ell_\star-1}\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle+\sigma_+\sup_{\ell\geq\ell_\star}\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle\;,$$

where σ_+ and $\ell_\star$ are defined in (A-2) and (A-4), respectively. Then, by subadditivity of $u\mapsto u^\gamma$,

$$\mathbb{E}_\mu^\theta\left[\left(\sup_{\ell\geq 0}\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle\right)^\gamma\right]\leq\sum_{\ell=0}^{\ell_\star-1}\mathbb{E}_\mu^\theta\left[(\widehat{B}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right]+(\sigma_+)^\gamma\mathbb{E}_\mu^\theta\left[\sup_{\ell\geq\ell_\star}(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right]\;.$$

Applying Lemma 15 and (15), it is thus sufficient to bound $\mathbb{E}_\mu^\theta\left[\sup_{\ell\geq\ell_\star}(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right]$. By Lemma 14, $\{\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle\}_{k\geq 0}$ is a $\{\mathcal{F}_{t+\ell}\}_{\ell\geq 0}$-martingale and since $\alpha\in(0,1)$, we have that $\{(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\alpha\}_{\ell\geq 0}$ is a nonnegative $\{\mathcal{F}_{t+\ell}\}_{\ell\geq 0}$-supermartingale. The Doob maximal inequality then applies: for all $a>0$,

$$a\mathbb{P}_\mu^\theta\left[\sup_{\ell\geq\ell_\star}(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\alpha\geq a\,\middle|\,\mathcal{F}_{t+\ell_\star-1}\right]\leq\mathbb{E}_\mu^\theta\left[(\widehat{C}_t^\theta\langle Y_{0:t+\ell_\star}\rangle)^\alpha\,\middle|\,\mathcal{F}_{t+\ell_\star-1}\right]\;.$$

Take now the expectation in both sides of the previous inequality and set $\delta=a^{\gamma/\alpha}$. We obtain

$$\mathbb{P}_\mu^\theta\left[\sup_{\ell\geq\ell_\star}(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\geq\delta\right]\leq\delta^{-\alpha/\gamma}\mathbb{E}_\mu^\theta\left[(\widehat{C}_t^\theta\langle Y_{0:t+\ell_\star}\rangle)^\alpha\right]\;.$$

Combining this with the inequality $\mathbb{E}[U]\leq 1+\int_1^\infty\mathbb{P}[U>\delta]\,\mathrm{d}\delta$ which holds for all nonnegative random variable U, we obtain under (A-4)

$$\mathbb{E}_\mu^\theta\left[\sup_{\ell\geq\ell_\star}(\widehat{C}_t^\theta\langle Y_{0:t+\ell}\rangle)^\gamma\right]\leq 1+\left(\int_1^\infty\delta^{-\frac{\alpha}{\gamma}}\mathrm{d}\delta\right)\mathbb{E}_\mu^\theta\left[(\widehat{C}_t^\theta\langle Y_{0:t+\ell_\star}\rangle)^\alpha\right]=1+\frac{\gamma}{\alpha-\gamma}\mathbb{E}_\mu^\theta\left[(\widehat{C}_t^\theta\langle Y_{0:t+\ell_\star}\rangle)^\alpha\right]\;.$$

The proof follows by applying again Lemma 15 under (A-4). □

Proof of Theorem 6. For simplicity we will use in this proof the notations $\bar{p}^{\,\theta}(Y_{0:t}):=p_{\pi^\theta}^\theta(Y_{0:t})$ and $\bar{p}^{\,\theta}(Y_{t:s}|Y_{0:t-1})=p_{\pi^\theta}^\theta(Y_{t:s}|Y_{0:t-1})$. First note that

$$\begin{aligned}\bar{\mathbb{P}}^{\theta_\star}\left\{\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}>\rho\right\}&=\bar{\mathbb{P}}^{\theta_\star}\left\{\ln\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}+\frac{p_\mu^{\theta_T}(Y_{0:T})}{\bar{p}^{\,\theta_\star}(Y_{0:T})}-1>\ln\rho+\frac{p_\mu^{\theta_T}(Y_{0:T})}{\bar{p}^{\,\theta_\star}(Y_{0:T})}-1\right\}\\&\leq\bar{\mathbb{P}}^{\theta_\star}\left\{\ln\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}+\frac{p_\mu^{\theta_T}(Y_{0:T})}{\bar{p}^{\,\theta_\star}(Y_{0:T})}-1>\ln\rho-1\right\}\;.\end{aligned}$$

Now, since for all $u>0$, $\ln(u)+u^{-1}-1\geq 0$, we obtain using Markov's inequality, for all $\rho>\mathrm{e}=\exp(1)$:

$$\bar{\mathbb{P}}^{\theta_\star}\left\{\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}>\rho\right\}\leq\frac{1}{\ln\rho-1}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}+\frac{p_\mu^{\theta_T}(Y_{0:T})}{\bar{p}^{\,\theta_\star}(Y_{0:T})}-1\right]=\frac{1}{\ln\rho-1}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\,\theta_\star}(Y_{0:T})}{p_\mu^{\theta_T}(Y_{0:T})}\right]\;.$$

This implies that for all $K > 0$ and all $\rho > \mathrm{e}$,

$$\begin{aligned}\bar{\mathbb{P}}^{\theta_\star}(\epsilon^{-1}_{T,N_T}(\theta_T) > K) &\leq \mathbb{E}^{\theta_T}_{\mu}\left[\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\mathbb{1}_{\left\{\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\leq\rho\right\}}\mathbb{1}_{\left\{\epsilon^{-1}_{T,N_T}(\theta_T)>K\right\}}\right] + \bar{\mathbb{P}}^{\theta_\star}\left\{\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})} > \rho\right\} \\ &\leq \rho\mathbb{P}^{\theta_T}_{\mu}\left[\epsilon^{-1}_{T,N_T}(\theta_T) > K\right] + \frac{1}{\ln\rho - 1}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\right] \\ &\leq \rho\mathbb{P}^{\theta_T}_{\mu}\left[\epsilon^{-1}_{T,N_T}(\theta_T) > K\right] + \frac{1}{\ln\rho - 1}\left(\sup_{T\geq 0}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\right]\right) . \end{aligned} \tag{52}$$

We consider first the last term of the right-hand side. Note first that, by the tower property,

$$\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(X_{0:T}|Y_{0:T})}{p^{\theta_T}_{\mu}(X_{0:T}|Y_{0:T})}\right] = \bar{\mathbb{E}}^{\theta_\star}\left[\int\cdots\int\left(\ln\frac{\bar{p}^{\theta_\star}(x_{0:T}|Y_{0:T})}{p^{\theta_T}_{\mu}(x_{0:T}|Y_{0:T})}\right)\bar{p}^{\theta_\star}(x_{0:T}|Y_{0:T})\prod_{i=0}^{T}\lambda(\mathrm{d}x_i)\right] \geq 0$$

because this quantity is the expectation under the stationary distribution of a Kullback-Leibler divergence. This implies that

$$\begin{aligned}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\right] &\leq \bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\right] + \bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(X_{0:T}|Y_{0:T})}{p^{\theta_T}_{\mu}(X_{0:T}|Y_{0:T})}\right] = \bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(X_{0:T},Y_{0:T})}{p^{\theta_T}_{\mu}(X_{0:T},Y_{0:T})}\right] \\ &= \bar{\mathbb{E}}^{\theta_\star}\left[\ln\left(\frac{\pi^{\theta_\star}(X_0)g^{\theta_\star}(X_0,Y_0)}{\mu(X_0)g^{\theta_T}(X_0,Y_0)}\right)\right] + T\bar{\mathbb{E}}^{\theta_\star}\left[\ln\left(\frac{m^{\theta_\star}(X_0,X_1)g^{\theta_\star}(X_1,Y_1)}{m^{\theta_T}(X_0,X_1)g^{\theta_T}(X_1,Y_1)}\right)\right] .\end{aligned}$$

Hence, we obtain, under (17) and (18), that

$$\sup_{T\geq 0}\bar{\mathbb{E}}^{\theta_\star}\left[\ln\frac{\bar{p}^{\theta_\star}(Y_{0:T})}{p^{\theta_T}_{\mu}(Y_{0:T})}\right] < \infty . \tag{53}$$

Assume first that

$$\limsup_{K\to\infty}\sup_{T\geq 0}\mathbb{P}^{\theta_T}_{\mu}\left[\epsilon^{-1}_{T,N_T}(\theta_T) > K\right] = 0 . \tag{54}$$

The proof of the tightness of $\{\epsilon^{-1}_{T,N_T}(\theta_T)\}_{T\geq 0}$ then follows by plugging (53) into (52) and by noting that (52) holds for all $\rho > \mathrm{e}$, combined with (54). To complete the proof, it thus remains to show (54). Rewriting the definition (12), we obtain

$$\epsilon^{-1}_{T,N_T}(\theta_T) \leq \prod_{t=0}^{T}\frac{2B^{\theta_T}_t + N_T - 2}{N_T - 1} = \exp\left\{\sum_{t=0}^{T}\ln\left(\frac{2B^{\theta_T}_t + N_T - 2}{N_T - 1}\right)\right\} \leq \exp\left\{\sum_{t=0}^{T}\frac{2B^{\theta_T}_t - 1}{N_T - 1}\right\} .$$

where $B^{\theta}_t := \sup_{\ell\geq 0}\widehat{B}^{\theta}_t\langle Y_{0:t+\ell}\rangle$. Using Markov's inequality, this implies that for $K > 1$,

$$\mathbb{P}^{\theta_T}_{\mu}\left[\epsilon^{-1}_{T,N_T}(\theta_T) > K\right] \leq \mathbb{P}^{\theta_T}_{\mu}\left[\sum_{t=0}^{T}\frac{2B^{\theta_T}_t - 1}{N_T - 1} > \ln K\right] \leq \frac{1}{(\ln K)^{\gamma}}\mathbb{E}^{\theta_T}_{\mu}\left[\left(\sum_{t=0}^{T}\frac{2B^{\theta_T}_t - 1}{N_T - 1}\right)^{\gamma}\right] .$$

The proof of (54) follows by noting that $N_T \sim T^{1/\gamma}$ and by using

$$\mathbb{E}^{\theta_T}_{\mu}\left[\left(\frac{\sum_{t=0}^{T}2B^{\theta_T}_t - 1}{T^{1/\gamma}}\right)^{\gamma}\right] \leq \mathbb{E}^{\theta_T}_{\mu}\left[\frac{\sum_{t=0}^{T}(2B^{\theta_T}_t)^{\gamma}}{T}\right] \leq 2^{\gamma}\sup_{t\geq 0}\sup_{\theta\in\Theta}\mathbb{E}^{\theta}_{\mu}\left[(B^{\theta}_t)^{\gamma}\right] < \infty ,$$

where the last inequality follows from Lemma 16. □

Acknowledgement

This work was supported by the project *Learning of complex dynamical systems* (Contract number: 637-2014-466) funded by the Swedish Research Council.

Supporting Information

Additional information for this article is available online including:

Appendix S1 An algorithmic statement of the particle Gibbs sampler.

Appendix S2 An example of how the conditions of Theorem 6 can be verified.

Appendix S3 Proofs of Proposition 4, Proposition 5, Lemma 9, Lemma 14, and Lemma 15.

References

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

Andrieu, C., A. Lee, and M. Vihola (2013, December). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. Preprint, arXiv:1312.6432.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

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

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics 164*(3), 1139–1160.

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

Chopin, N. and S. S. Singh (2014). On particle Gibbs sampling. *Bernoulli*. Forthcoming.

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

Del Moral, P. and A. Guionnet (1999). Central limit theorem for nonlinear filtering and interacting particle systems. *Annals of Applied Probability 9*(2), 275–297.

Doucet, A., S. J. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*(3), 197–208.

Doucet, A. and A. Johansen (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskii (Eds.), *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press.

Doucet, A., M. K. Pitt, and R. Kohn (2012, October). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Preprint, arXiv:1210.1871.

Fearnhead, P., D. Wyncoll, and J. Tawn (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika 97*(2), 447–464.

Godsill, S. J., A. Doucet, and M. West (2004, March). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association 99*(465), 156–168.

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

Lee, A. and K. Latuszynski (2012, October). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. Preprint, arXiv:1210.6703.

Lindsten, F., M. I. Jordan, and T. B. Schön (2014). Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research 15*, 2145–2184.

Lindsten, F. and T. B. Schön (2013). Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning 6*(1), 1–143.

Meyn, S. and R. L. Tweedie (2009). *Markov Chains and Stochastic Stability* (2nd ed.). Cambridge University Press.

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

Rasmussen, D. A., O. Ratmann, and K. Koelle (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biology 7*(8), e1002136.

Shephard, N. and T. Andersen (2009). Stochastic volatility: Origins and overview. In T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Berlin: Springer.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes – A study of daily sugar prices, 1961-79. In O. D. Anderson (Ed.), *Time Series Analysis: Theory and Practice*, Volume 1, pp. 203–226. Elsevier/North-Holland, Amsterdam.

Vrugt, J. A., J. F. ter Braak, C. G. H. Diks, and G. Schoups (2013). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources 51*, 457–478.

Whiteley, N., C. Andrieu, and A. Doucet (2010). Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. Technical report, Bristol Statistics Research Report 10:04.

Fredrik Lindsten, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK.
E-mail: fredrik.lindsten@eng.cam.ac.uk

Uniform ergodicity of the Particle Gibbs sampler - Supplementary material

Fredrik Lindsten
Department of Engineering, University of Cambridge, Cambridge, UK, and
Division of Automatic Control, Linköping University, Linköping, Sweden

Randal Douc
Department CITI, Institut Mines-Telecom/CNRS UMR 5157
Telecom Sudparis, Evry, France

Eric Moulines
Department LTCI, Institut Mines-Telecom/CNRS UMR 5141
Telecom Paristech, Paris, France

September 9, 2014

Section numbers, equation numbers, etc., in this supplementary material are prefixed with "S". References without this prefix refer to the main document.

S1 Particle Gibbs algorithm

The Particle Gibbs (PG) sampling procedure, described in Section 3 of the main document, is summarized in Algorithm S1 below.

Algorithm S1 PG sampler

Input: Observations $y_{0:T}$ and reference trajectory $x'_{0:T}$.

Output: Draw from the PG Markov kernel $X^\star_{0:T} \sim P_{T,N}(x'_{0:T}, \cdot)$.

1: Draw $X_0^i \sim r_0\langle y_0\rangle(\cdot)$ for $i = 1, \ldots, N-1$ and set $X_0^N = x'_0$.
2: Set $\omega_0^i = w_0\langle y_0\rangle(X_0^i)$ for $i = 1, \ldots, N$.
3: **for** $t = 1$ **to** T **do**
4: Draw A_t^i w.p. $\mathbb{P}\left(A_t^i = j \,\middle|\, \mathcal{F}_{t-1}^N\right) = \omega_{t-1}^j / \sum_{\ell=1}^N \omega_{t-1}^\ell$, $j \in \{1, \ldots, N\}$, for $i = 1, \ldots, N-1$.
5: Draw $X_t^i \sim R\langle y_t\rangle(X_{t-1}^{A_t^i}, \cdot)$ for $i = 1, \ldots, N-1$.
6: Set $X_t^N = x'_t$ and $A_t^N = N$.
7: Set $X_{1:t}^i = (X_{1:t-1}^{A_t^i}, X_t^i)$ for $i = 1, \ldots, N$.
8: Set $\omega_t^i = w\langle y_t\rangle(X_{t-1}^{A_t^i}, X_t^i)$ for $i = 1, \ldots, N$.
9: **end for**
10: Draw J w.p. $\mathbb{P}\left(J = j \,|\, \mathcal{F}_T^N\right) = \omega_T^j / \sum_{\ell=1}^N \omega_T^\ell$, $j \in \{1, \ldots, N\}$.
11: **return** $X^\star_{0:T} := X^J_{0:T}$.

S2 Example – A nonlinear model with additive measurement noise

In this section we study a second example (in addition to the one presented in Section 5 of the main document) and show how the conditions of Theorem 6 can be verified. We consider a class of nonlinear state space models where the latent process is observed in additive noise,

$$X_{t+1} = h^{\xi}(X_t) + \sigma_W W_{t+1} \tag{S1}$$

$$Y_t = \phi X_t + \sigma_U U_t \tag{S2}$$

where $\{W_t,\ t \in \mathbb{N}\}$ and $\{U_t,\ t \in \mathbb{N}\}$ are two independent sequences of i.i.d. standard Gaussian random variables and $\{h^{\xi}, \xi \in \Xi\}$ is a parametric family of measurable real-valued functions, where Ξ is a compact subset of a Euclidean space. We denote by $\theta = (\xi, \phi, \sigma_U, \sigma_W)$ the parameters of the model. It is assumed that $\theta \in \Theta$, where Θ is a compact subset of $\Xi \times (0,\infty)^3$. We assume that for all $\xi \in \Xi$, $x \mapsto h^{\xi}(x)$ is continuous and $\sup_{\xi\in\Xi} \limsup_{x\to\infty} |h^{\xi}(x)|/|x| < 1$. For any $\delta > 0$, we set $V_\delta(x) = \mathrm{e}^{\delta|x|}$. It is easily seen that there exist constants $\lambda_\delta \in (0,1)$ and $b_\delta < \infty$ such that

$$\sup_{\theta\in\Theta} \mathbb{E}^{\theta}_x \left[V_\delta(X_1)\right] \leq \lambda_\delta V_\delta(x) + b_\delta \ . \tag{S3}$$

The Markov chain is strong Feller, Harris recurrent, all the compact sets are small, and the Markov chain admits a single invariant distribution. Therefore, (A-1) and (A-2) are satisfied. Since both the transition density and the observation density are Gaussian, (A-3) is also readily satisfied. We will thus focus on verifying the moment assumption (A-4).

First, note that

$$\sup_{\theta\in\Theta} \mathbb{E}^{\theta}_x \left[V_\delta(X_t)\right] \leq \lambda_\delta^t V_\delta(x) + b_\delta(1 + \lambda_\delta + \cdots + \lambda_\delta^{t-1}) \leq V_\delta(x) + b_\delta/(1-\lambda_\delta) \ . \tag{S4}$$

We assume that the initial distribution μ is such that $\mu(V_\delta) < \infty$. Therefore,

$$\sup_{t\in\mathbb{N}} \sup_{\theta\in\Theta} \mathbb{E}^{\theta}_{\mu} \left[V_\delta(X_t)\right] < \infty \ . \tag{S5}$$

Interestingly for the model (S1)–(S2) it is possible to use the fully adapted proposal kernel (Doucet et al., 2000) as defined in Equation (7) of the main document, for which

$$w^{\theta}\langle y\rangle(x,x') = \int m^{\theta}(x,x'')g^{\theta}(x'',y)\mathrm{d}x'' = \frac{1}{\sqrt{2\pi(\phi^2\sigma_W^2+\sigma_U^2)}} \exp\left(-\frac{\left(y-\phi h^{\xi}(x)\right)^2}{2(\phi^2\sigma_W^2+\sigma_U^2)}\right) \ , \tag{S6}$$

for all $(x,x') \in \mathbb{R}\times\mathbb{R}$, $y \in \mathbb{R}$, and $\theta \in \Theta$. It can be seen that, for any $\theta \in \Theta$ and any $y \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} g^{\theta}(x,y)\mathrm{d}x = \frac{1}{\phi} \ , \quad \text{and} \quad |w^{\theta}\langle y\rangle|_\infty \leq \frac{1}{\sqrt{2\pi(\phi^2\sigma_W^2+\sigma_U^2)}} \ ,$$

which implies the existence of constants D_1 and D_2 such that

$$\sup_{\theta\in\Theta} \int_{-\infty}^{\infty} g^{\theta}(x,y)\mathrm{d}x \leq D_1 \ , \quad \text{and} \quad \sup_{\theta\in\Theta} |w^{\theta}\langle y\rangle|_\infty \leq D_2 \ . \tag{S7}$$

Analogous bounds hold also if we would instead consider the bootstrap proposal.

To verify (A-4) we let $\ell_\star = 1$ and show that,

$$\sup_{t\in\mathbb{N}} \sup_{\theta\in\Theta} \mathbb{E}^{\theta}_{\mu} \left[(\widetilde{B}^{\theta}_t\langle Y_t\rangle)^{\alpha}\right] < \infty \, , \quad \sup_{t\in\mathbb{N}} \sup_{\theta\in\Theta} \mathbb{E}^{\theta}_{\mu} \left[(\widetilde{C}^{\theta}_t\langle Y_{t:t+1}\rangle)^{\alpha}\right] < \infty \, , \tag{S8}$$

for some (and actually any) $\alpha \in [0,1)$. Consider first

$$\mathbb{E}_\mu^\theta \left[(\widetilde{B}_t^\theta \langle Y_t \rangle)^\alpha \right] = \int |w^\theta \langle y_t \rangle|_\infty^\alpha \{p_{\mu,t}^\theta(y_t)\}^{1-\alpha} \mathrm{d}y_t \leq D_2^\alpha \int \{p_{\mu,t}^\theta(y_t)\}^{1-\alpha} \mathrm{d}y_t \ , \tag{S9}$$

where the inequality follows from (S7). We apply Lemma 9 to establish a bound for the right-hand side of (S9). Let $\psi(y) = 1$ and $\varphi(y) = 1/(1 \vee |y|^2)$. With these definitions the first condition in (19) is satisfied. To check (20), note that

$$\varphi^{-\frac{\alpha}{1-\alpha}}(y) = (1 \vee |y|^2)^{\frac{\alpha}{1-\alpha}} \leq 1 + |y|^{2\alpha/(1-\alpha)} \ .$$

The integral in (20) may be expressed as

$$\int \varphi^{-\frac{\alpha}{1-\alpha}}(y_t) p_{\mu,t}^\theta(y_t) \mathrm{d}y_t = \mathbb{E}_\mu^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_t) \right] = \mathbb{E}_\mu^\theta \left[\mathbb{E}_{X_t}^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0) \right] \right] \ . \tag{S10}$$

Since $Y_0 = \phi X_0 + \sigma U_0$, we get that for any $x \in \mathsf{X}$, $\mathbb{E}_x^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0) \right] \leq 1 + \mathbb{E}\left[|\phi x + U|^{2\alpha/(1-\alpha)} \right]$, where U is standard normal. This implies that there exists a constant D_3 such that, for all $x \in \mathsf{X}$ and all $\theta \in \Theta$,

$$\mathbb{E}_x^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_0) \right] \leq D_3 (1 + |x|^{2\alpha/(1-\alpha)}) \ . \tag{S11}$$

Plugging this into (S10) and using (S5), this verifies the second condition in (20). Lemma 9 can thus be used to conclude that $\mathbb{E}_\mu^\theta \left[(\widetilde{B}_t^\theta \langle Y_t \rangle)^\alpha \right] < \infty$ for all $\alpha \in (0,1)$. Since this holds for any $t \in \mathbb{N}$ and $\theta \in \Theta$, we obtain the first part of (S8).

Next, we consider

$$\begin{aligned} \mathbb{E}_\mu^\theta \left[(\widetilde{C}_t^\theta \langle Y_{t:t+1} \rangle)^\alpha \right] &= \mathbb{E}_\mu^\theta \left[\frac{|w^\theta \langle Y_t \rangle|_\infty^\alpha \left(\int g^\theta(x_{t+1}, Y_{t+1}) \mathrm{d}x_{t+1} \right)^\alpha}{\{p_{\mu,t}^\theta(Y_{t:t+1})\}^\alpha} \right] \\ &\leq D_1^\alpha D_2^\alpha \iint \{p_{\mu,t}^\theta(y_{t:t+1})\}^{1-\alpha} \mathrm{d}y_{t:t+1} \ . \end{aligned}$$

We will again make use of Lemma 9 to bound this quantity. Proceeding analogously to above, we let $\psi(y_0, y_1) = 1$ and

$$\varphi(y_0, y_1) = \frac{1}{(y_0^2 \vee 1)(y_1^2 \vee 1)} \ , \tag{S12}$$

for which (19) is satisfied. To check (20), we use the conditional independence of the observations given the states and (S10) to get, for any $\theta \in \Theta$,

$$\begin{aligned} &\iint \varphi^{-\frac{\alpha}{1-\alpha}}(y_{t:t+1}) p_{\mu,t}^\theta(y_{t:t+1}) \mathrm{d}y_{t:t+1} = \mathbb{E}_\mu^\theta \left[\mathbb{E}_\mu^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_{t:t+1}) \,\middle|\, X_{t:t+1} \right] \right] \\ &\leq \mathbb{E}_\mu^\theta \left[\mathbb{E}_{X_t}^\theta \left[1 + |Y_0|^{\frac{2\alpha}{1-\alpha}} \right] \mathbb{E}_{X_{t+1}}^\theta \left[1 + |Y_0|^{\frac{2\alpha}{1-\alpha}} \right] \right] \leq D_3^2 \mathbb{E}_\mu^\theta \left[(1 + |X_t|^{\frac{2\alpha}{1-\alpha}})(1 + |X_{t+1}|^{\frac{2\alpha}{1-\alpha}}) \right] \ . \end{aligned}$$

From the Cauchy-Schwarz inequality we get, by using (S5),

$$\sup_{t \in \mathbb{N}} \sup_{\theta \in \Theta} \mathbb{E}_\mu^\theta \left[\varphi^{-\frac{\alpha}{1-\alpha}}(Y_{t:t+1}) \right] \leq D_3^2 \sup_{t \in \mathbb{N}} \sup_{\theta \in \Theta} \mathbb{E}_\mu^\theta \left[(1 + |X_t|^{\frac{2\alpha}{1-\alpha}})^2 \right] < \infty \ .$$

This shows that (20) is satisfied for any $\theta \in \Theta$ and any $t \in \mathbb{N}$ which, by Lemma 9 implies $\sup_{t \in \mathbb{N}} \sup_{\theta \in \Theta} \mathbb{E}^\theta \left[(\widetilde{C}_t^\theta \langle Y_{t:t+1} \rangle)^\alpha \right] < \infty$ for all $\alpha \in [0,1)$, verifying (A-4).

Provided that θ_T converges to $\theta_\star$ at a rate $1/\sqrt{T}$, we may therefore apply Theorem 6 which shows that for any $\gamma \in (0,1)$, $\{\epsilon_{T,N_T}^{-1}(\theta_T)\}_{T \geq 1}$ is tight with $N_T \sim T^{1/\gamma}$.

S3 Proofs

Proof of Proposition 4 First, note that for all $\ell \geq 1$,

$$|Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty \leq \sigma_+ \int \gamma(\mathrm{d}x_{t+1}) g(x_{t+1}, y_{t+1}) Q\langle y_{t+2:t+\ell}\rangle \mathbf{1}(x_{t+1}) \tag{S13}$$

and

$$p_\mu(y_{t:t+\ell}|y_{0:t-1}) \geq \sigma_-^2 \int \gamma(\mathrm{d}x_t) g(x_t, y_t) \int \gamma(\mathrm{d}x_{t+1}) g(x_{t+1}, y_{t+1}) Q\langle y_{t+2:t+\ell}\rangle \mathbf{1}(x_{t+1}) \,. \tag{S14}$$

Now, in the fully-adapted case, we have:

$$R\langle y\rangle(x, \mathrm{d}x') = \frac{M(x, \mathrm{d}x') g(x', y)}{\int M(x, \mathrm{d}u) g(u, y)} \,,$$

so that by the definition of $w\langle y\rangle$,

$$|w\langle y_t\rangle|_\infty = \sup_{x \in \mathsf{X}} \left| \int M(x, \mathrm{d}x_t) g(x_t, y_t) \right| \leq \sigma_+ \int \gamma(\mathrm{d}x_t) g(x_t, y_t) \,.$$

Combining this equality with (S13) and (S14) yields:

$$B_{t,T} = \sup_{0 \leq \ell \leq T-t} \frac{|w\langle y_t\rangle|_\infty |Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty}{p_\mu(y_{t:t+\ell}|y_{0:t-1})} \leq \left(\frac{\sigma_+}{\sigma_-}\right)^2 \,.$$

By the definition of $\epsilon_{T,N}$ (see Equation (12) of the main document) we then obtain:

$$\epsilon_{T,N} = \prod_{t=0}^{T} \frac{N-1}{2B_{t,T} + N - 2} \geq \left(\frac{N-1}{N - 2 + 2(\sigma_+/\sigma_-)^2}\right)^{T+1} \,.$$

Finally, letting $N_T \sim \lambda T$, we obtain the desired result. □

Proof of Proposition 5 For the bootstrap filter, $w\langle y\rangle(x, x') = g(x', y)$. Therefore, $|w\langle y\rangle|_\infty = \sup_{x \in \mathsf{X}} g(x, y)$. On the other hand, for $\ell \geq m$,

$$|Q\langle y_{t+1:t+\ell}\rangle \mathbf{1}|_\infty \leq \sigma_+ \left(\prod_{s=t+1}^{t+m-1} \sup_{x \in \mathsf{X}} g(x, y_s) \right) \int \gamma(\mathrm{d}x_{t+m}) g(x_{t+m}, y_{t+m}) Q\langle y_{t+m+1:l}\rangle \mathbf{1}(x_{t+m}) \,, \tag{S15}$$

and

$$p_\mu(y_{t:t+\ell}|y_{0:t-1}) \geq \sigma_- \left(\prod_{s=t}^{t+m-1} \inf_{x \in \mathsf{X}} g(x, y_s) \right) \int \gamma(\mathrm{d}x_{t+m}) g(x_{t+m}, y_{t+m}) Q\langle y_{t+m+1:t+\ell}\rangle \mathbf{1}(x_{t+m}) \,. \tag{S16}$$

Combining (S15) and (S16) yields

$$B_{t,T} \leq \delta^m \frac{\sigma_+}{\sigma_-} \,. \tag{S17}$$

The result follows as in the proof of Proposition 4. □

Proof of Lemma 9 The result follows from Hölder's inequality:

$$\int \psi^\alpha(z) q^{1-\alpha}(z) \xi(\mathrm{d}z) = \int [\psi(z)\varphi(z)]^\alpha \left[\varphi^{-\frac{\alpha}{1-\alpha}}(z) q(z)\right]^{1-\alpha} \xi(\mathrm{d}z)$$
$$\leq \left(\int \psi(z)\varphi(z)\xi(\mathrm{d}z)\right)^\alpha \left(\int \varphi^{-\frac{\alpha}{1-\alpha}}(z) q(z) \xi(\mathrm{d}z)\right)^{1-\alpha} .$$

□

Proof of Lemma 14 For all $\ell \geq 0$,

$$\mathbb{E}_\mu^\theta \left[\widehat{C}_t^\theta \langle Y_{0:t+\ell+1} \rangle \,\middle|\, \mathcal{F}_{t+\ell} \right] = |w^\theta \langle Y_t \rangle|_\infty \int \left\{ p_\mu^\theta(y_{t+\ell+1} | Y_{0:t+\ell}) \right.$$
$$\left. \times \frac{\int \lambda(\mathrm{d}x_{t+1}) g^\theta(x_{t+1}, Y_{t+1}) Q^\theta \langle Y_{t+2:t+\ell} \rangle (x_{t+1}, \mathrm{d}x_{t+\ell}) Q^\theta \langle y_{t+\ell+1} \rangle \mathbf{1}(x_{t+\ell})}{p_\mu^\theta(Y_{t:t+\ell}, y_{t+\ell+1} | Y_{0:t-1})} \kappa(\mathrm{d}y_{t+\ell+1}) \right\} .$$

Combining this identity with

$$p_\mu^\theta(Y_{t:t+\ell}, y_{t+\ell+1} | Y_{0:t-1}) = p_\mu^\theta(y_{t+\ell+1} | Y_{0:t+\ell}) p_\mu^\theta(Y_{t:t+\ell} | Y_{0:t-1}) ,$$

and $\int Q^\theta \langle y_{t+\ell+1} \rangle \mathbf{1}(x_{t+\ell}) \kappa(\mathrm{d}y_{t+\ell+1}) = M^\theta(x_{t+\ell}, \mathsf{X}) = 1$, we obtain

$$\mathbb{E}_\mu^\theta \left[\widehat{C}_t^\theta \langle Y_{0:t+\ell+1} \rangle \,\middle|\, \mathcal{F}_{t+\ell} \right]$$
$$= \frac{|w^\theta \langle Y_t \rangle|_\infty \int \lambda(\mathrm{d}x_{t+1}) g^\theta(x_{t+1}, Y_{t+1}) Q^\theta \langle Y_{t+2:t+\ell} \rangle \mathbf{1}(x_{t+1})}{p_\mu^\theta(Y_{t:t+\ell} | Y_{0:t-1})} = \widehat{C}_t^\theta \langle Y_{0:t+\ell} \rangle ,$$

which completes the proof. □

Proof of Lemma 15 Using the expressions in Equation (51) of the main document, the proof of Lemma 15 follows from the inequality:

$$\mathbb{E}_\mu^\theta \left[\frac{\psi(Y_{t:t+\ell})}{\left\{ p_\mu^\theta(Y_{t:t+\ell} | Y_{0:t-1}) \right\}^\gamma} \right] \leq \mathbb{E}_\mu^\theta \left[\frac{\psi(Y_{t:t+\ell})}{\left\{ p_{\mu,t}^\theta(Y_{t:t+\ell}) \right\}^\gamma} \right] , \tag{S18}$$

which holds for any nonnegative measurable function $\psi : \mathsf{Y}^{\ell+1} \to \mathbb{R}^+$. We now show (S18). Note first that, by applying the tower property of the conditional expectation and then the Tonelli-Fubini theorem, we get

$$\mathbb{E}_\mu^\theta \left[\frac{\psi(Y_{t:t+\ell})}{\left\{ p_\mu^\theta(Y_{t:t+\ell} | Y_{0:t-1}) \right\}^\gamma} \right] = \mathbb{E}_\mu^\theta \left[\mathbb{E}_\mu^\theta \left[\frac{\psi(Y_{t:t+\ell})}{\left\{ p_\mu^\theta(Y_{t:t+\ell} | Y_{0:t-1}) \right\}^\gamma} \,\middle|\, Y_{0:t-1} \right] \right]$$
$$= \mathbb{E}_\mu^\theta \left[\int \psi(y_{t:t+\ell}) \left\{ p_\mu^\theta(y_{t:t+\ell} | Y_{0:t-1}) \right\}^{1-\gamma} \kappa^{\otimes(\ell+1)}(\mathrm{d}y_{t:t+\ell}) \right]$$
$$= \int \psi(y_{t:t+\ell}) \mathbb{E}_\mu^\theta \left[\left\{ p_\mu^\theta(y_{t:t+\ell} | Y_{0:t-1}) \right\}^{1-\gamma} \right] \kappa^{\otimes(\ell+1)}(\mathrm{d}y_{t:t+\ell}) . \tag{S19}$$

By the Jensen identity, $\mathbb{E}_\mu^\theta \left[\left\{ p_\mu^\theta(y_{t:t+\ell} | Y_{0:t-1}) \right\}^{1-\gamma} \right] \leq \left\{ \mathbb{E}_\mu^\theta \left[p_\mu^\theta(y_{t:t+\ell} | Y_{0:t-1}) \right] \right\}^{1-\gamma}$. On the other hand,

$$\mathbb{E}_\mu^\theta \left[p_\mu^\theta(y_{t:t+\ell} | Y_{0:t-1}) \right] = \int p_\mu^\theta(y_{t:t+\ell} | y_{0:t-1}) p_\mu^\theta(y_{0:t-1}) \kappa^{\otimes t}(\mathrm{d}y_{0:t-1}) = p_{\mu,t}^\theta(y_{t:t+\ell}) . \tag{S20}$$

The proof of (S18) follows by combining the above relations. □

Author address

Fredrik Lindsten
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ, UK
E-mail: fredrik.lindsten@eng.cam.ac.uk

References

Doucet, A., S. J. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*(3), 197–208.