Biometrika (2017), **104**, 4, *pp.* 953–969
Printed in Great Britain
doi: 10.1093/biomet/asx051
Advance Access publication 16 October 2017

Blocking strategies and stability of particle Gibbs samplers

By S. S. SINGH

Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.
sss40@cam.ac.uk

F. LINDSTEN

Department of Information Technology, Uppsala University, Lägerhyddsv. 2, Uppsala 751 05, Sweden
fredrik.lindsten@it.uu.se

and E. MOULINES

Centre de Mathématiques Appliquées, École Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France
eric.moulines@polytechnique.edu

Summary

Sampling from the posterior probability distribution of the latent states of a hidden Markov model is nontrivial even in the context of Markov chain Monte Carlo. To address this, Andrieu et al. (2010) proposed a way of using a particle filter to construct a Markov kernel that leaves the posterior distribution invariant. Recent theoretical results have established the uniform ergodicity of this Markov kernel and shown that the mixing rate does not deteriorate provided the number of particles grows at least linearly with the number of latent states. However, this gives rise to a cost per application of the kernel that is quadratic in the number of latent states, which can be prohibitive for long observation sequences. Using blocking strategies, we devise samplers that have a stable mixing rate for a cost per iteration that is linear in the number of latent states and which are easily parallelizable.

Some key words: Hidden Markov model; Markov chain Monte Carlo; Particle filter; Particle Gibbs sampling.

1. Introduction

1·1. *Notation and background*

Let $\{(X_t, Y_t) \in (\mathcal{X}, \mathcal{Y}) : t \in \mathbb{N}_+\}$ be a hidden Markov model in which $\{X_t : t \in \mathbb{N}_+\}$ is the state process, a Markov chain with state space $\mathcal{X}$. The sequence $\{X_t : t \in \mathbb{N}_+\}$ is not observed and its values have to be inferred using the observed sequence $\{Y_t : t \in \mathbb{N}_+\}$. Conditionally on $\{X_t : t \in \mathbb{N}_+\}$, the observations $\{Y_t : t \in \mathbb{N}_+\}$ are independent. We work under the assumption that a fixed sequence of observations $(y_1, \ldots, y_n)$ is available, where n denotes the final time-point. The key object of interest is the joint smoothing distribution $\phi(\mathrm{d}x_1, \ldots, \mathrm{d}x_n)$, which is the probability distribution of $(X_1, \ldots, X_n)$ conditioned on $(Y_1 = y_1, \ldots, Y_n = y_n)$. Markov chain Monte Carlo simulation can be used to sample from the joint smoothing distribution, for

example by a Gibbs scheme that uses Metropolis–Hastings kernels to update the state variables X_t $(t = 1, \ldots, n)$ individually. However, strong dependence between consecutive states of the state process can cause this method to mix very slowly and the solution is often deemed inefficient (Carter & Kohn, 1994; Frühwirth-Schnatter, 1994).

Recent developments in sequential Monte Carlo methods have had a significant impact on the practice of Markov chain Monte Carlo sampling for hidden Markov models. Sequential Monte Carlo methods are well-established importance sampling techniques for approximating the joint smoothing distribution in general hidden Markov models; see, for example, Doucet et al. (2000), Del Moral (2004) and Doucet & Johansen (2011) for applications and supporting theoretical results. In a seminal paper of Andrieu et al. (2010), this key strength of sequential Monte Carlo simulation was exploited to construct new Markov chain Monte Carlo algorithms which are collectively called particle Markov chain Monte Carlo methods. We will consider a specific particle Markov chain Monte Carlo algorithm called particle Gibbs sampling; see also Chopin & Singh (2015). The particle Gibbs algorithm of Andrieu et al. (2010) defines, via a sequential Monte Carlo construction, a Markov kernel which has the joint smoothing distribution as its invariant distribution. It updates samples of all the state variables $X_1, \ldots, X_n$ as one block, aiming to mimic the behaviour and thus also the efficiency of an ideal sampler. Although standard Gibbs steps can be interleaved to jointly infer the unknown model parameters, we shall focus on the central problem of simulating the latent states from the joint smoothing distribution.

Some recent theoretical results support this approach of mimicking the ideal sampler and its good observed performance: Chopin & Singh (2015) showed that the particle Gibbs kernel is uniformly ergodic. The precise rate of convergence of the iterated particle Gibbs kernel to its stationary distribution was established by Lindsten et al. (2015) and Andrieu et al. (2017), who showed that the number of particles N must increase linearly with the number of observations, n, for its convergence rate not to deteriorate. This effect is also clearly visible in practice (Lindsten & Schön, 2013; Chopin & Singh, 2015) and is related to the path-degeneracy of sequential Monte Carlo samplers. As a result, using $N \propto n$, the particle Gibbs kernel has a computational cost of n^2 per iteration, which is impractical when n is large.

1·2. *Summary of main results and related work*

The particle Gibbs algorithm samples all n hidden states jointly in one block. However, just as Gibbs sampling can update the latent variables one state at a time, particle Gibbs sampling can be used as a partially blocked sampler by jointly updating blocks of L consecutive state variables at a time. This possibility was mentioned in Andrieu et al. (2010, p. 294) but without further elaboration. Specifically, there was no mention of the greatly improved stability of the sampler, which is our main interest here. As we will show, using particle Gibbs sampling in a partially blocked manner results in a stable mixing rate with a cost per iteration that is linear in the number of latent states.

We now give a simplified interpretation of our results. The main insight we exploit is that blocking can also be used to control the convergence properties of the exact, or ideal, blocked Gibbs sampler for a hidden Markov model. In Theorems 1 and 2 we show that, under certain forgetting properties of the model, it is possible to select a blocking scheme with overlapping blocks such that after k complete sweeps, the error between the law of the samples and the target distribution ϕ is

$$\left|\phi(f) - \mu(\mathcal{P}^k_{\text{Ideal}} f)\right| \leqslant \text{constant} \times (\lambda_{\text{Ideal}})^k \sum_{i=1}^{n} \text{osc}_i(f),$$

where $\mathcal{P}_{\text{Ideal}}$ is the Markov kernel defined by one complete sweep of the blocked Gibbs sampler, $f : \mathcal{X}^n \to \mathbb{R}$ is some test function and $\text{OSC}_i(f)$ measures how much f varies when only its ith component is perturbed; see (7). We show that the rate $\lambda_{\text{Ideal}} \in (0, 1)$ is independent of n and improves as the overlap between blocks is increased. If we replace exact Gibbs sampling from each block with particle Gibbs sampling, the rate becomes $\lambda_{\text{PG}} = \lambda_{\text{Ideal}} + \text{constant} \times \epsilon(N, L)$ where $0 \leqslant \epsilon(N, L) \leqslant 1$ quantifies the effect of departing from the ideal blocked Gibbs sampler; see Theorem 3. Specifically, ϵ depends only on N and L but crucially not on the number of latent states n; $\epsilon \downarrow 0$ as the particle number N increases, i.e., as the particle Gibbs kernel better approximates the ideal block sampler, but $\epsilon \uparrow 1$ as the blocks are made larger to include more latent variables in them. In Theorem 3 we also analyse different blocking schemes, one of which is straightforwardly parallelizable. In light of these properties, the number of particles N and the block size L can be chosen independently of the number of observations n so that the convergence rate of blocked particle Gibbs sampling should not deteriorate even as n increases. As increasing n requires more blocks, the cost per complete sweep of blocked particle Gibbs sampling will increase only linearly with n as opposed to quadratically for nonblocked particle Gibbs sampling. Thus, blocking is better for long time series, i.e., when X_t is observed over a longer time rather than at a higher frequency. Our analysis is based on Wasserstein estimates; see, e.g., Wang & Wu (2014) and Rebeschini & van Handel (2014). Wang & Wu (2014) study the convergence properties of a Gibbs sampler in high dimensions under a Dobrushin condition. Our work is in the same vein, but we verify a related condition for blocked Gibbs sampling for hidden Markov models. Furthermore, we study the convergence of the nonideal blocked particle Gibbs sampler, to which the results of Wang & Wu (2014) do not apply.

Refined particle Gibbs algorithms that incorporate explicit updates of the particle ancestry, either as part of the forward sequential Monte Carlo recursion (Lindsten et al., 2014) or in a separate backward recursion (Whiteley, 2010), have been developed. These modified particle Gibbs samplers have been shown to work well empirically with few particles and to be largely robust with respect to n. Nevertheless, to date, no theoretical guarantee of the stability of these algorithms has been given; nor are they easily parallelizable.

2. Hidden Markov models and blocked Gibbs samplers

We assume that the Markov chain (X_t, Y_t) on $\mathcal{X} \times \mathcal{Y}$ has the transition probability kernel

$$R\{(x, y), A\} = \int \nu_1(\mathrm{d}x')\nu_2(\mathrm{d}y')m(x, x')g(x', y')\mathbb{I}_{[(x',y')\in A]},$$

where $\mathbb{I}$ is the indicator function and $m(x, x')$ is the transition density of the Markov chain $X_t : t \in \mathbb{N}_+$ with respect to the dominating measure ν_1. We further assume that the conditional density of Y_t given $X_t = x_t$ with respect to the dominating measure ν_2 is $g(x_t, y_t)$ and denote the initial distribution of X_1 by μ. Recall that $y_1, \ldots, y_n$ is a fixed observation sequence and we seek to sample from the joint smoothing distribution, that is, the conditional distribution of $X_1, \ldots, X_n$ given $Y_1 = y_1, \ldots, Y_n = y_n$. The algorithms and results to be presented also apply to inhomogeneous models where the transition density of (X_t, Y_t) is dependent on time t; Assumption 1 below would have to be modified for this case.

Before giving the algorithmic statements for the blocked Gibbs samplers that are analysed in this article, we introduce some notation. Let $I = \{1, \ldots, n\}$ be the index set of the latent variables $X_1, \ldots, X_n$. Let $\mathcal{X}^n$ be the n-fold Cartesian product of the set $\mathcal{X}$. Given $x = (x_i : i \in I) \in \mathcal{X}^n$ and $J \subset I$, let $x_J = (x_i : i \in J) \in \mathcal{X}^{|J|}$, i.e., the restricted vector. We also write x_{-i} as

shorthand for $x_{I\setminus\{i\}}$. The complement of J in I is denoted by $J^{\mathrm{c}} = I \setminus J$. Given a_J and $b_{J^{\mathrm{c}}}$, let $x = \langle a_J, b_{J^{\mathrm{c}}}\rangle = \langle b_{J^{\mathrm{c}}}, a_J\rangle \in \mathcal{X}^n$ be such that $x_J = a_J$ and $x_{J^{\mathrm{c}}} = b_{J^{\mathrm{c}}}$.

We will analyse the stability of our samplers under the following set of strong, but standard, mixing assumptions (Del Moral, 2004; Lindsten et al., 2015; Andrieu et al., 2017).

Assumption 1. There exist positive constants σ_- and σ_+ and an integer $h \geqslant 1$ such that $m(x, x') \leqslant \sigma_+$ for all $x, x' \in \mathcal{X}$ and

$$\int \nu_1(\mathrm{d}x_2) \cdots \nu_1(\mathrm{d}x_h) \prod_{t=1}^{h} m(x_t, x_{t+1}) \geqslant \sigma_-, \quad x_1, x_{h+1} \in \mathcal{X},$$

and there exists a constant $\delta \geqslant 1$ such that $\sup_x g(x, y) \leqslant \delta^{1/h} \inf_x g(x, y)$ for all $y \in \mathcal{Y}$.

We can write the density of the joint smoothing distribution as

$$\phi(x_1, \ldots, x_n) = \frac{1}{p(y_1, \ldots, y_n)} \mu(x_1) g(x_1, y_1) \prod_{t=2}^{n} m(x_{t-1}, x_t) g(x_t, y_t), \tag{1}$$

where $p(y_1, \ldots y_n) = \int \mu(\mathrm{d}x_1) g(x_1, y_1) \prod_{t=2}^{n} m(x_{t-1}, x_t) g(x_t, y_t) \nu_1(\mathrm{d}x_2) \cdots \nu_1(\mathrm{d}x_n)$ and the same symbol ϕ is used for both the joint smoothing distribution and its density. Let ϕ_x^J be a version of the regular conditional distribution of the variables X_J conditionally on $X_{J^{\mathrm{c}}} = x_{J^{\mathrm{c}}}$ under ϕ in (1). For instance, for $J = \{s, \ldots, u\}$ with $s > 1$ and $u < n$,

$$\begin{aligned} \phi_x^J(x_s, \ldots, x_u) &\propto m(x_u, x_{u+1})\, p(x_s, \ldots, x_u \,|\, x_{s-1}, y_s, \ldots, y_u) \\ &= \frac{m(x_u, x_{u+1})}{p(y_s, \ldots, y_u \,|\, x_{s-1})} \prod_{j=s}^{u} m(x_{j-1}, x_j)\, g(x_j, y_j). \end{aligned} \tag{2}$$

In (2) we have used the symbol p as a generic density function which is identified by its arguments, and we will occasionally use this notation for brevity when no confusion is likely. For instance, $\phi_x^J(x_s, \ldots, x_u) = p(x_s, \ldots, x_u \,|\, x_{s-1}, x_{u+1}, y_s, \ldots, y_u)$.

The Markov property of the hidden Markov model implies that ϕ_x^J depends on x only through the boundary points $x_{\partial J}$, where ∂J denotes the set of indices constituting the boundary of the set J,

$$\partial J = \{t \in J^{\mathrm{c}} : t + 1 \in J \text{ or } t - 1 \in J\}.$$

The Gibbs sampler samples from the joint smoothing distribution ϕ by iteratively sampling from its conditional distributions. Let $\mathcal{J} = \{J_1, \ldots, J_m\}$ be a cover of I. A blocked, deterministic-scan Gibbs sampler proceeds by sampling from the conditional densities ϕ_x^J ($J \in \mathcal{J}$) in turn in some prespecified order. For instance, assume that we apply the blocks in the order J_1, J_2 etc. and that the initial configuration of the sampler is $x_0 \in \mathcal{X}^n$. Then the Gibbs sampler produces the Markov chain $\{X(k) : k \in \mathbb{N}\}$ with $X(0) = x_0$, and given

$$X(lm + k - 1) = x \in \mathcal{X}^n \quad (l \in \mathbb{N};\, k \in \{1, \ldots, m\}) \tag{3}$$

we set $X(lm + k) = X'$ where $X'_{J_k^{\mathrm{c}}} = x_{J_k^{\mathrm{c}}}$ and where we simulate $X'_{J_k} \sim \phi_x^{J_k}(\cdot)$. More generally, we can simulate X'_J from some kernel $Q^J(x, \cdot)$ with the conditional distribution ϕ_x^J as its invariant measure; that is, for any $x \in \mathcal{X}^n$,

$$\int \phi_x^J\left(\mathrm{d}x'_J\right) Q^J(\langle x_{J^{\mathrm{c}}}, x'_J\rangle, A) = \phi_x^J(A). \tag{4}$$

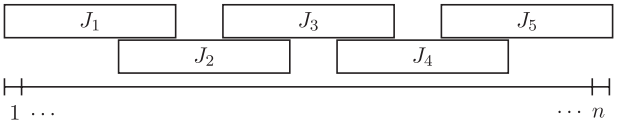

Fig. 1. A blocking scheme satisfying Assumptions 2 and 3.

Since ϕ_x^J depends on x only through the boundary points $x_{\partial J}$, it is natural to assume that $Q^J(x, \cdot)$ depends on x only through $x_{J \cup \partial J}$. For notational convenience, we define $J^+ = J \cup \partial J$ for any subset $J \subset I$. Thus, for any $(x, z) \in \mathcal{X}^n \times \mathcal{X}^n$ we have $Q^J(x, A) = Q^J(\langle x_{J^+}, z_{I \setminus J^+} \rangle, A)$. We write $Q^J(x_{J^+}, \mathrm{d}x'_J)$ in place of $Q^J(x, \mathrm{d}x'_J)$ when we wish to to emphasize the dependence of the kernel on the components in J^+ of the current configuration x. When $Q^J(x, \mathrm{d}x'_J) = \phi_x^J(\mathrm{d}x'_J)$ for every $J \in \mathcal{J}$, we refer to the sampler as an ideal Gibbs sampler. It follows that the Markov kernel that corresponds to updating the block J is

$$P^J(x, \mathrm{d}x') = \begin{cases} \phi_x^J(\mathrm{d}x'_J) \times \delta_{x_{J^c}}(\mathrm{d}x'_{J^c}) & \text{for the ideal Gibbs kernel,} \quad (5) \\ Q^J(x_{J^+}, \mathrm{d}x'_J) \times \delta_{x_{J^c}}(\mathrm{d}x'_{J^c}) & \text{for the nonideal Gibbs kernel.} \quad (6) \end{cases}$$

The subsets in $\mathcal{J}$ can be an arbitrary cover of I. However, certain blocking schemes are likely to be of greater practical interest and these schemes therefore deserve some extra attention. To exemplify this, we consider the following restrictions on the blocks in $\mathcal{J}$.

Assumption 2. Each $J \in \mathcal{J}$ is an interval, i.e., $J = \{s, \ldots, u\}$ for some $1 \leqslant s \leqslant u \leqslant n$. Furthermore, the blocks $J_1, \ldots, J_m$ are ordered in the following way: for any $1 \leqslant j < k \leqslant m$, $\min(J_j) < \min(J_k)$ and $\max(J_j) < \max(J_k)$.

Assumption 3. Consecutive blocks may overlap but nonconsecutive blocks do not overlap and are separated; that is, for $1 \leqslant j < k \leqslant m$ with $k - j \geqslant 2$, $\max(J_j) < \min(J_k) - 1$.

In addition to ordering the blocks according to their minimum element, Assumption 2 avoids the case where one block is a strict subset of some other block. Assumption 3 requires that J_{k-1} and J_{k+1} not cover J_k. Figure 1 illustrates a blocking scheme that satisfies both assumptions.

Definition 1. *When Assumption 2 holds, the left-to-right Gibbs kernel is defined to be* $\mathcal{P} = P^{J_1} \cdots P^{J_m}$. *When Assumptions 2 and 3 hold, the parallel Gibbs kernel is defined to be* $\mathcal{P} = \mathcal{P}_{\mathrm{odd}} \mathcal{P}_{\mathrm{even}}$ *where*

$$\begin{cases} \mathcal{P}_{\mathrm{odd}} = P^{J_1} P^{J_3} \cdots P^{J_m}, \\ \mathcal{P}_{\mathrm{even}} = P^{J_2} P^{J_4} \cdots P^{J_{m-1}}, \end{cases} \quad \textit{or} \quad \begin{cases} \mathcal{P}_{\mathrm{odd}} = P^{J_1} P^{J_3} \cdots P^{J_{m-1}}, \\ \mathcal{P}_{\mathrm{even}} = P^{J_2} P^{J_4} \cdots P^{J_m}, \end{cases}$$

for odd or even m, respectively.

The first Gibbs sampling scheme is a systematic sweep through the blocks from left to right, and $\mathcal{P}$ is the kernel corresponding to one complete sweep. The second blocking scheme updates all the odd-numbered blocks first and then all the even-numbered blocks. It is called parallel Gibbs sampling and is important since it is possible to update all the odd blocks in parallel, followed by a parallel update of all the even blocks, because two consecutive odd or even blocks are separated by at least one element in I. Figure 1 shows this typical scenario for the parallel Gibbs sampler.

In § 4 we use particle Gibbs sampling to define the kernels $Q^J(x_{J^+}, \mathrm{d}x'_J)$ for each J, and the particle Gibbs kernel is known to be reversible (Chopin & Singh, 2015). Hence it is simple to define reversible block samplers.

Lemma 1. *Let $\mathcal{J} = \{J_1, \ldots, J_m\}$ be an arbitrary cover of I, and for each $J \in \mathcal{J}$ let $P^J(x, \mathrm{d}x') = Q^J(x_{J^+}, \mathrm{d}x'_J) \times \delta_{x_{J^c}}(\mathrm{d}x'_{J^c})$ be the Gibbs kernel, possibly nonideal, that updates block J only. Assume that Q^J is reversible with respect to ϕ_x^J for all $J \in \mathcal{J}$. Then*

$$\phi(\mathrm{d}x) \times \left(P^{J_1} \cdots P^{J_m}\right)(x, \mathrm{d}x') = \phi(\mathrm{d}x') \times \left(P^{J_m} \cdots P^{J_1}\right)(x', \mathrm{d}x).$$

For example, for the parallel scheme, the kernel $(\mathcal{P}_{\mathrm{odd}}\mathcal{P}_{\mathrm{even}} + \mathcal{P}_{\mathrm{even}}\mathcal{P}_{\mathrm{odd}})/2$ is reversible. Lemma 1 can be used to define other reversible samplers.

3. Convergence of the ideal block samplers

3·1. *Preliminaries, notation, and definitions*

For a function $f : \mathcal{X}^n \to \mathbb{R}$, the oscillation with respect to the ith coordinate is

$$\mathrm{osc}_i(f) = \sup_{\{x,z \in \mathcal{X}^n : x_{-i} = z_{-i}\}} |f(x) - f(z)|. \tag{7}$$

Let $\mathrm{osc}(f) = \sup_{x,z \in \mathcal{X}^n} |f(x) - f(z)|$ be the oscillation of f. Then

$$|f(x) - f(z)| \leqslant \sum_{i \in I} \mathrm{osc}_i(f) \mathbb{I}_{[x_i \neq z_i]}. \tag{8}$$

For a matrix A with elements $A_{i,j}$, $\|A\|_\infty = \max_i \sum_j |A_{i,j}|$ is a submultiplicative norm, i.e., $\|AB\|_\infty \leqslant \|A\|_\infty \|B\|_\infty$. Let μ and ν be two probability measures on $\mathcal{X}$ and let Ψ be a probability measure on $\mathcal{X} \times \mathcal{X}$; Ψ is a coupling of μ and ν if $\int \Psi(\cdot, \mathrm{d}x) = \mu(\cdot)$ and $\int \Psi(\mathrm{d}x, \cdot) = \nu(\cdot)$.

We review some techniques for the convergence analysis of Markov chains; see, e.g., Follmer (1982). Let P be a Markov kernel on $\mathcal{X}^n$. The nonnegative matrix W is a Wasserstein matrix for P if for any function f of finite oscillation,

$$\mathrm{osc}_j(Pf) \leqslant \sum_{i \in I} \mathrm{osc}_i(f) W_{i,j} \quad (j \in I). \tag{9}$$

If P and Q are Markov kernels with Wasserstein matrices V and W, respectively, then WV is a Wasserstein matrix for the kernel PQ. The convergence rate of a Markov chain Monte Carlo procedure can be characterized in terms of a corresponding Wasserstein matrix for the Markov transition kernel through the following result. For any probability distributions μ and ν and any coupling Ψ of μ and ν, let $\psi_j = \int \Psi(\mathrm{d}x, \mathrm{d}z) \mathbb{I}_{[x_j \neq z_j]}$, which is the probability under the coupling Ψ of elements j not being equal. The function $(x, x') \to \mathbb{I}_{[x \neq x']}$ on $\mathcal{X} \times \mathcal{X}$ is assumed to be measurable; for example, this would be satisfied if the σ-algebra of subsets of $\mathcal{X}$ is countably generated and separable. Then, for any function f of finite oscillation and any $k \geqslant 1$,

$$\left|\mu P^k f - \nu P^k f\right| \leqslant \sum_{i,j \in I} \mathrm{osc}_i(f) (W^k)_{i,j} \psi_j \leqslant \left\{\sum_{i \in I} \mathrm{osc}_i(f)\right\} \|W\|_\infty^k \left(\max_{j \in I} \psi_j\right). \tag{10}$$

This result follows directly from (8), (9) and the remark following it, namely the fact that W^k is a Wasserstein matrix for P^k. It follows from (10) that $\|W\|_\infty < 1$ implies geometric convergence.

3·2. *Wasserstein estimates for the ideal block sampler*

Our convergence results for the ideal and for the particle Gibbs samplers rely on constructions of Wasserstein matrices for the ideal Gibbs kernels. In fact, the particle Gibbs block sampler can be viewed as an ϵ-perturbation of the ideal block sampler and this is exploited in the convergence analysis below. Let P^J denote the ideal Gibbs kernel that updates block J only, as defined in (5), and let W^J be a Wasserstein matrix for P^J. The following lemma, proved in the Appendix, reveals the structure of W^J.

LEMMA 2. *A Wasserstein matrix for the ideal Gibbs kernel updating block J can be chosen to satisfy*

$$W^J_{i,j} = \begin{cases} \mathbb{I}_{[i=j]}, & i \in J^{\mathrm{c}}, j \in I, \\ 0, & i \in J, j \in I \setminus \partial J, \\ \times, & i \in J, j \in \partial J, \end{cases} \tag{11}$$

where $\times$ denotes elements that lie in the interval $[0, 1]$.

For an interval $J = \{s, \ldots, u\}$, it follows that

$$W^J = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \times & 0 \cdots 0 & \times & & \\ & & \vdots & \vdots \ddots \vdots & \vdots & & \\ & & \times & 0 \cdots 0 & \times & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix},$$

with obvious modifications for the boundary blocks, i.e., when $s = 1$ and/or $u = n$. Blank entries of W^J correspond to zeros. The square of zeros corresponds to rows $i \in J$ and columns $j \in J$. The columns of $\times$ entries correspond to $W^J_{i,j}$ with $i \in J$ and $j \in \partial J$. The interpretation of components that are either 0 or 1 is noteworthy, as these cannot be improved further; the smaller the row sum of the Wasserstein matrix the better. For example, Lemma 2 tells us that when the states x and z differ in a single component, say $x_j \neq z_j$, and if $j \in J$, then this error is not propagated when computing the difference $P^J f(x) - P^J f(z)$. In this sense the given Wasserstein matrices are elementwise minimal.

It remains to compute the nonzero, off-diagonal elements of W^J. The strong, but standard, mixing conditions in Assumption 1 allow the remaining undeclared elements of W^J to be expressed in terms of the mixing coefficients of the model.

LEMMA 3. *Suppose that Assumptions 1 and 2 hold, and let $J = \{s, \ldots, u\}$ and*

$$\alpha = 1 - \left\{\delta^{(1-h)/h}\right\} \frac{\sigma_-}{\sigma_+} \in [0, 1),$$

where h, δ, σ_- and σ_+ are defined in Assumption 1. Then the matrix W^J defined in (11) with

$$W^J_{i,j} = \begin{cases} \alpha^{\lfloor h^{-1}(i-j) \rfloor}, & j = s - 1, s > 1, \\ \alpha^{\lfloor h^{-1}(j-i) \rfloor}, & j = u + 1, u < n, \end{cases}$$

for $i \in J$ and $j \in \partial J$ is a Wasserstein matrix for the ideal Gibbs block-transition kernel P^J.

The proof is given in the Supplementary Material.

3·3. *Convergence of the ideal block sampler*

We start with a general geometric convergence result which holds for an arbitrary cover $\mathcal{J}$ of I.

Assumption 4. Let $\partial = \bigcup_{J \in \mathcal{J}} \partial J$ be the set of all boundary points. For all $J \in \mathcal{J}$, $\max_{i \in J \cap \partial} \sum_{j \in \partial J} W^J_{i,j} \leqslant \lambda < 1$.

Theorem 1. *Let $\mathcal{J} = \{J_1, \ldots, J_m\}$ be an arbitrary cover of I and let $\mathcal{P} = P^{J_1} \cdots P^{J_m}$ be the kernel of one complete sweep of the ideal Gibbs sampler. For each $J \in \mathcal{J}$, let W^J be chosen as in Lemma 2 and let $\mathcal{W} = W^{J_m} \cdots W^{J_1}$ be the corresponding Wasserstein matrix for $\mathcal{P}$. If Assumption 4 holds, then for $k \geqslant 1$, $\|\mathcal{W}^k\|_\infty \leqslant \lambda^{k-1} \|\mathcal{W}\|_\infty$.*

See the Supplementary Material for the proof.

The following result now characterizes the convergence of the law of the sampled output of the ideal block sampler.

Corollary 1. *Let μ and ν be two probability distributions on $\mathcal{X}^n$ and let Ψ be an arbitrary coupling of μ and ν. Under the same conditions as in Theorem 1, for any $k \geqslant 1$ and any f of finite oscillation,*

$$\left|\mu \mathcal{P}^k f - \nu \mathcal{P}^k f\right| \leqslant \|\mathcal{W}\|_\infty \lambda^{k-1} \left(\max_{j \in I} \int \Psi(\mathrm{d}x, \mathrm{d}z) \mathbb{I}_{[x_j \neq z_j]} \right) \sum_{i \in I} \mathrm{osc}_i(f).$$

Remark 1. Corollary 1 clarifies two important issues. Firstly, if we are interested only in certain fixed-dimensional marginals of the joint smoothing distribution, then the convergence rate is independent of the dimension n of the joint smoothing distribution. Secondly, for convergence in total variation norm of the law of the sampled process $\{X(k) : k \in \mathbb{N}\}$ in (3) to the full joint smoothing distribution, the bound on the error is $O(n\lambda^k)$, i.e., the error grows only linearly in n. For any other general f, linear-in-n complexity holds only if $\sum_{i \in I} \mathrm{osc}_i(f)$ is bounded in n.

Under the conditions of Lemma 3 we can clearly see the benefit of blocking for verifying Assumption 4. As an illustration, let $h = 1$ in Assumption 1. The condition for contraction in Assumption 4 requires that for any $i \in J \cap \partial$, assuming $J = \{s, \ldots, u\}$ is an internal block,

$$\sum_{j \in \partial J} W^J_{i,j} = \alpha^{i-(s-1)} + \alpha^{(u+1)-i} < 1, \tag{12}$$

where $\alpha \in [0, 1)$ is defined in Lemma 3. First of all, it is possible to ensure $\sum_{j \in \partial J} W^J_{i,j} < 1$ for any $i \in J$ by increasing the block size $L = u - s + 1$. Indeed, the maximum of (12) for $i \in J$ is attained for $i = s$ or $i = u$, for which $\sum_{j \in \partial J} W^J_{i,j} = \alpha + \alpha^L$, which is less than 1 for L large enough. Secondly, Lemma 3 also reveals the benefit of using overlapping blocks. Since we only need to control (12) for $i \in J \cap \partial$, we can select the blocking scheme so that the set of boundary points ∂ excludes indices i close to the boundary of block J. Consequently, by using overlapping blocks we can control both terms in (12) and thus the overall convergence rate of the algorithm by increasing L.

Theorem 1 assumes no specific structure for $\mathcal{J}$ other than it being a cover. As such, it cannot provide a sharper estimate of contraction since it caters for all blocking structures. In order to refine the estimate we impose the blocking structure of Assumptions 2 and 3, as illustrated in Fig. 1, and study the effect of block size and overlap on convergence. Theorem 2 improves the estimate of the decay of errors per complete sweep from λ in Theorem 1 to λ^2 for the blocking structure of Fig. 1.

THEOREM 2. *Suppose that Assumptions 2, 3 and 4 hold. For the ideal parallel Gibbs sampler,*

$$\|\mathcal{W}^k\|_\infty \leqslant \|\mathcal{W}\|_\infty \lambda^{2(k-1)},$$

$\|\mathcal{W}\|_\infty \leqslant 2$ *and* λ *is defined as in Assumption 4. For the ideal left-to-right Gibbs sampler,*

$$\|\mathcal{W}^k\|_\infty \leqslant \|\mathcal{W}\|_\infty \beta^{k-1},$$

$\|\mathcal{W}\|_\infty \leqslant 1+\lambda$, $\beta = \max_{k\in\{2,\ldots,m\}} \lambda a_k + b_k$, $a_k = W^{J_k}_{\partial_+ J_{k-1},\partial_- J_k}$, *and* $b_k = W^{J_k}_{\partial_+ J_{k-1},\partial_+ J_k}$.

The proof can be found in the Supplementary Material.

For example, let each block be the same length L, and let the overlap between all adjacent blocks $J_{k-1}, J_k \in \mathcal{J}$ be fixed, $|J_{k-1} \cap J_k| = p$. When Assumption 1 holds, $b_k = \alpha^{\lfloor h^{-1}(L-p)\rfloor}$, $a_k = \alpha^{\lfloor h^{-1}(p+1)\rfloor}$ and $\lambda = \alpha^{\lfloor h^{-1}(L-p)\rfloor} + \alpha^{\lfloor h^{-1}(p+1)\rfloor}$. There is parity in the two rates of Theorem 2 as L increases while p is fixed, since $\beta/\lambda^2 \to 1$.

4. CONVERGENCE OF THE BLOCKED PARTICLE GIBBS SAMPLER

4·1. *The particle Gibbs construction of Q^J*

The particle Gibbs sampler of Andrieu et al. (2010) is a Markov chain Monte Carlo sampler for simulating from the joint state and parameter posterior distribution of a state space model. It does so by iteratively simulating the model parameter from its conditional distribution, i.e., a standard Gibbs sampling step, and simulating the system states from the particle Gibbs kernel, which is a Markov kernel that preserves the invariance of the full joint smoothing distribution for a fixed value of the model parameter. We omit the routine step that updates the static parameter and in Algorithm 1 describe how the standard particle Gibbs algorithm needs to be modified to target the conditional density ϕ_x^J; Algorithm 1 defines the nonideal block transition kernel (6).

Algorithm 1 is a sequential Monte Carlo-based construction of a Markov kernel on $\mathcal{X}^{|J|}$. The algorithm associates with each x_{J^+} a probability distribution on $\mathcal{X}^{|J|}$, denoted by $Q_N^J(x_{J^+}, \cdot)$, where N is the number of particles used in the underlying sequential Monte Carlo sampler; that is, $Q_N^J(x_{J^+}, A) = \mathrm{pr}(X'_J \in A)$ where X'_J is the output of Algorithm 1. A straightforward extension of Theorem 5 of Andrieu et al. (2010) shows that Q_N^J has ϕ_x^J as its invariant distribution in the sense of (4). Invariance holds for any $N \geqslant 1$, although $N \geqslant 2$ is required for the kernel to be ergodic (Andrieu et al., 2010; Lindsten et al., 2015; Andrieu et al., 2017). We briefly explain Algorithm 1; for more discussion of the particle Gibbs algorithm see Andrieu et al. (2010).

Algorithm 1. Particle Gibbs kernel $Q_N^J(x_{J^+}, \mathrm{d}x'_J)$ with invariant distribution ϕ_x^J for non-boundary block $J = \{s, \ldots, u\}$.

Input: Observations y_J, fixed boundary states $x_{\partial J}$ and input block states x_J.

Draw $X_s^i \sim r_s(x_{s-1}, \cdot)$ for $i = 1, \ldots, N-1$ and set $X_s^N = x_s$.

Set $W_s^i = w_s(x_{s-1}, X_s^i)$ for $i = 1, \ldots, N$; see (13).

For $t = s+1$ to u
  Draw A_t^i with $\mathrm{pr}(A_t^i = j) = W_{t-1}^j / \sum_{\ell=1}^N W_{t-1}^\ell$ for $i = 1, \ldots, N-1$.
  Draw $X_t^i \sim r_t(X_{t-1}^{A_t^i}, \cdot)$ for $i = 1, \ldots, N-1$. Set $X_t^N = x_t$ and $A_t^N = N$.
  Set $W_t^i = w_t(X_{t-1}^{A_t^i}, X_t^i)$ for $i = 1, \ldots, N$.
  Set $X_{s:t}^i = (X_{s:t-1}^{A_t^i}, X_t^i)$ for $i = 1, \ldots, N$.
Set $\tilde{W}_u^i = W_u^i \times m(X_u^i, x_{u+1})$ for $i = 1, \ldots, N$.
Draw K with $\mathrm{pr}(K = j) = \tilde{W}_u^i / \sum_{\ell=1}^N \tilde{W}_u^\ell$ for $j \in \{1, \ldots, N\}$.
Output $X'_J = X_J^K$.

The internal steps of the particle Gibbs kernel, specifically the initialization and for-loop of Algorithm 1, can be interpreted as approximating the sequence of target distributions $p(x_s, \ldots, x_t \mid x_{s-1}, y_s, \ldots, y_t)$ $(t = s, \ldots, u)$ sequentially by constructing the collections of particles $X_{s:t}^i$ and weights W_t^i $(i = 1, \ldots, N;\ t = s, \ldots, u)$. We use $X_{s:t}$ to denote a trajectory in state space from time s to time t, i.e., $X_{s:t} = (X_s, \ldots, X_t)$. The initialization simulates particles $X_s^i \sim r_s(x_{s-1}, \cdot)$ $(i = 1, \ldots, N-1)$ independently from the proposal density $r_s(x_{s-1}, \cdot)$. The proposal density may depend on the fixed observation sequence as well as the fixed endpoint x_{u+1}, but we omit the explicit dependence from the notation for brevity. The Nth particle is set deterministically to the kernel's input value: $X_s^N = x_s$. This N-particle empirical measure is meant to approximate $p(x_s \mid x_{s-1}, y_s)$. To correct for the discrepancy between the proposal density and this target density, importance weights are computed as in standard sequential Monte Carlo sampling: $W_s^i = w_s(x_{s-1}, X_s^i)$ $(i = 1, \ldots, N)$, where the weight function is

$$w_t(x_{t-1}, x_t) = \frac{g(x_t, y_t) m(x_{t-1}, x_t)}{r_t(x_{t-1}, x_t)}. \tag{13}$$

The weighted particles (X_s^i, W_s^i) $(i = 1, \ldots, N)$ now form an approximation of the target density $p(x_s \mid x_{s-1}, y_s)$. The remaining steps of Algorithm 1 can be understood similarly via induction. That is, assume that the weighted samples $(X_{s:t-1}^i, W_{t-1}^i)$ $(i = 1, \ldots, N)$ approximate the target $p(x_s, \ldots, x_{t-1} \mid x_{s-1}, y_s, \ldots, y_{t-1})$ at time $t-1$. This empirical approximation is then sampled and extended in the first two lines of the for-loop and then reweighted, in the third line of the for-loop, to correct for the discrepancy between the importance sampling proposal and the target $p(x_s, \ldots, x_t \mid x_{s-1}, y_s, \ldots, y_t)$. At the final iteration of block J, we need to take into account the fact that the target distribution is $\phi_x^J(x_s, \ldots, x_u) = p(x_s, \ldots, x_u \mid x_{s-1}, x_{u+1}, y_s, \ldots, y_u)$ and not $p(x_s, \ldots, x_u \mid x_{s-1}, y_s, \ldots, y_u)$, the two being related through (2). In other words, we need to take into account the conditioning on the fixed boundary state x_{u+1}, which contributes via the term $m(X_u^i, x_{u+1})$ to the final weight $\tilde{W}_u^i$. After completing the for-loop, components in block J of the particle Gibbs sampler input, namely x_J, are updated with a draw from the particle approximation of ϕ_x^J, which is denoted by X'_J.

Algorithm 1 is a basic particle Gibbs sampler and can be made more efficient (Andrieu et al., 2010; Chopin & Singh, 2015). Specific to the blocked particle Gibbs sampler is that the algorithm may be adapted to the fixed boundary points x_{s-1} and x_{u+1}. Both the proposal distributions and the intermediate target distributions may depend on these boundary states, as long as the final target distribution is ϕ_x^J; we use this type of adaptation in the numerical illustration in § 5. Additionally, the mixing of the particle Gibbs kernel can be improved significantly by updating the ancestor indices A_t^N $(t = s+1, \ldots, u)$ either as part of the for-loop in Algorithm 1 (Lindsten et al., 2014) or in a separate backward recursion (Whiteley, 2010). Although we do not elaborate on the use of

these modified particle Gibbs algorithms in this work, the stability results of the blocked particle Gibbs sampler presented in the subsequent sections also hold when the particle Gibbs kernel is replaced by one of these modified algorithms, which might result in better empirical performance. The reason is that our results follow from the uniform minorization of the particle Gibbs kernel, which, as pointed out by Lindsten et al. (2015, § 3), also holds for the algorithms of Lindsten et al. (2014) and Whiteley (2010).

4·2. *Convergence of the particle Gibbs block sampler*

We now discuss the convergence properties of the particle Gibbs block sampler. In Theorem 3 we state a main result that parallels Theorem 2 for the blocked particle Gibbs sampler. For the sake of interpretability, we specialize the result to the case of a common block size L and a common overlap p between successive blocks. A version of Theorem 3 with neither this assumption nor strong mixing is presented in the Appendix; see Theorems A1 and A2, of which Theorem 3 is a corollary.

Assumption 5. For all $J \in \mathcal{J}$, $|J| = L$; and for all consecutive $J_{k-1}, J_k \in \mathcal{J}$, $|J_{k-1} \cap J_k| = p$. The number of states n of the joint smoothing distribution satisfies $n = (L-p)m + p$.

THEOREM 3. *Suppose that Assumptions 1, 2, 3 and 5 hold. Let $\mathcal{P}$ denote the Markov kernel corresponding to one complete sweep of either the parallel sampler or the left-to-right sampler, and assume that each block is updated by simulating from the particle Gibbs kernel as detailed in Algorithm 1 using the proposal $r(x, x') = m(x, x')$. Let μ and ν be two probability distributions on $\mathcal{X}^n$, and let Ψ be an arbitrary coupling of μ and ν. Then, for any f of finite oscillation and any $k \geqslant 1$,*

$$\left|\mu\mathcal{P}^k f - \nu\mathcal{P}^k f\right| \leqslant \lambda_{\mathrm{PG}}^k \times \left\{\int \Psi(\mathrm{d}x, \mathrm{d}z)\mathbb{I}_{[x \neq z]}\right\} \sum_{i \in I} \mathrm{osc}_i(f), \tag{14}$$

where for the parallel sampler,

$$\lambda_{\mathrm{PG}} \leqslant \lambda(\beta \vee 1) + \epsilon\{2\lambda + 25\epsilon + 8(\beta \vee 1)\}, \tag{15}$$

and for the left-to-right sampler,

$$\lambda_{\mathrm{PG}} \leqslant \lambda + \alpha^{\lfloor h^{-1}(L-p+1)\rfloor} + 2\epsilon\,\frac{3(\beta \vee 1) + 1 + \lambda}{1 - 2\epsilon - \alpha^{\lfloor h^{-1}(p+1)\rfloor}},$$

with

$$\beta = \alpha^{\lfloor h^{-1}\rfloor} + \alpha^{\lfloor h^{-1}(L-p+1)\rfloor} \leqslant 2, \quad \epsilon = 1 - \left\{1 - \frac{1}{c(N-1)+1}\right\}^L,$$

provided that $\lambda = 2\alpha^{\lfloor h^{-1}(p+1)\rfloor} < 1$ and $2\epsilon + \alpha^{\lfloor h^{-1}(p+1)\rfloor} < 1$. The constant c in the definition of ϵ is specified in Proposition A1 and is independent of n, N, L and p.

The proof is given in the Appendix.

Theorem 3 also applies to the ideal sampler upon setting $\epsilon = 0$. In terms of sufficiency for contraction, the requirement that the non-ϵ terms of this theorem be less than 1 is stronger than Assumption 4; this should not be surprising since the analysis is tailored to the nonideal particle

Gibbs kernel and thus is inherently more conservative. For common block length L and overlap p, Assumption 4 requires $\alpha^{\lfloor (p+1)/h \rfloor} + \alpha^{\lfloor (L-p)/h \rfloor} < 1$. Nevertheless, the non-$\epsilon$ terms of Theorem 3 can be controlled by increasing the overlap of blocks and then increasing the block size with the overlap fixed. Alternatively, if p is a constant fraction of L, then λ tends to zero as L increases. Once L and p are fixed, N can be increased to ensure that $\lambda_{\mathrm{PG}} < 1$ independently of n.

5. Numerical illustration

We illustrate the blocked particle Gibbs samplers on a time-varying autoregressive model commonly used for audio processing (Godsill et al., 2004). The model, which violates the strong mixing assumption and has a nontrivial latent state structure, is intentionally chosen to illustrate that the intuition and conclusions from the theory presented here may generalize to more challenging situations. Let the signal $\{Z_t : t \in \mathbb{N}_+\}$ be a Pth order Gaussian autoregressive process

$$Z_t = \sum_{j=1}^{P} a_{t,j} Z_{t-j} + E_t, \quad E_t \sim \mathcal{N}\{0, \exp(2\xi_t)\},$$

with time-varying coefficients $a_t = (a_{t,1}, \ldots, a_{t,P})^{\mathrm{T}}$ and log standard deviation ξ_t where ξ_t follows a first-order Gaussian autoregressive model, $p(\xi_t \mid \xi_{t-1}) = \mathcal{N}(\xi_t \mid \eta \xi_{t-1}, \sigma_\xi^2)$. The coefficients a_t are parameterized using partial correlation coefficients $\rho_t \in \mathbb{R}^P$ (Friedlander, 1982) with truncated Gaussian first-order autoregressive dynamics,

$$p(\rho_t \mid \rho_{t-1}) \propto \mathcal{N}(\rho_t \mid \theta \rho_{t-1}, \sigma_\rho^2 I) \mathbb{I}_{[\max_j \{|\rho_{t,j}|\} < 1]}.$$

Constraining each component to the interval $(-1, 1)$ ensures the stability of the model and there is a one-to-one mapping between a_t and ρ_t. The signal Z_t is observed in noise, $Y_t = Z_t + V_t$, with $V_t \sim \mathcal{N}(0, \sigma_v^2)$. The static parameters of the model are assumed to be known, with values given in the Supplementary Material. The latent process $X_t = \left(Z_t, \rho_t, \xi_t\right)^{\mathrm{T}} \in \mathbb{R}^{P+2}$ is Pth-order Markov. The Supplementary Material details a straightforward generalization of the blocked particle Gibbs samplers to take into account this lag-P dependence at the block boundaries.

We use a simulated dataset with $n = 2000$ and $P = 4$. The methods considered are: (i) standard particle Gibbs by Andrieu et al. (2010); (ii) particle Gibbs with ancestor sampling by Lindsten et al. (2014); and (iii) the proposed parallel block sampler with different block sizes L and overlaps p, where $(L, p) \in \{(10, 0), (50, 0), (50, 10)\}$. The Supplementary Material contains full implementation details, as well as results for the right-to-left block sampler, which performs similarly to the parallel sampler. All methods used $N = 100$ particles in the underlying particle filters and have similar computational costs per complete sweep, except for the block sampler with overlap $p = 10$, which costs approximately 25% more. Sampling efficiency is measured by the mean squared jump distance (Pasarica & Gelman, 2010), which is computed for each component of x_t and at each time-point, based on 10 000 iterations. The results for ξ_t are shown in Fig. 2. The other variables were found to behave similarly; see the Supplementary Material.

The particle Gibbs sampler suffers from path-degeneracy and is stuck for all time-points $t < 1600$; clearly $N = 100$ is insufficient for this problem. However, the other samplers appear to have stable mixing. For the block sampler without overlap, i.e., $p = 0$, the jump distance drops close to the block boundaries, owing to the performance limitation of the ideal block sampler without overlapping blocks. This issue is mitigated by taking $p > 0$: comparing $L = 50$ and $p = 0$

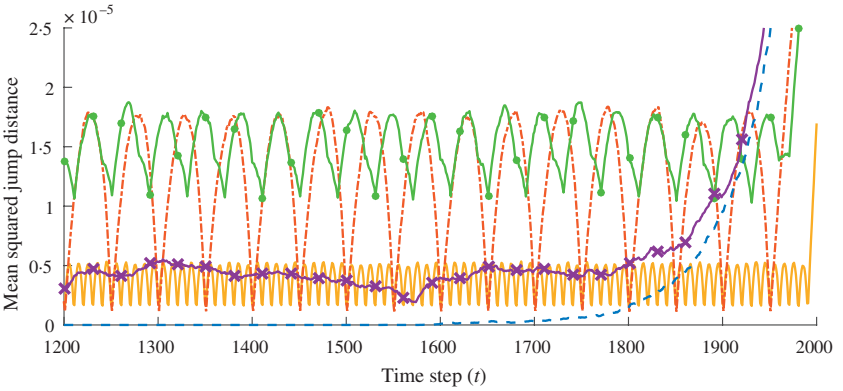

Fig. 2. Mean squared jump distance for ξ_t for particle Gibbs (- - -) with ancestor sampling (—✖—). Parallel block particle Gibbs: $L = 10, p = 0$ (——); $L = 50, p = 0$ (·---·); $L = 50, p = 10$ (—●—). Axis zoomed in to $1200 \leqslant t \leqslant 2000$ for clarity.

with $L = 50$ and $p = 10$ shows, as expected, that increasing the overlap will enhance mixing, owing to the better mixing of the ideal sampler; the error ϵ of (15) is not expected to improve since N is unchanged. The block samplers perform significantly better than particle Gibbs with ancestor sampling, probably due to the use of a more efficient proposal mechanism, not available to the unblocked implementation, as explained in the Supplementary Material. This is in addition to the benefits of a more explicit convergence theory and the possibility for parallelization. Assuming N is given by a fixed computational budget, the parameters L and p can be chosen to maximize the ratio of the effective sample size, which is the total number of samples produced divided by the integrated autocorrelation time, to the total running time. The search could be started by first tuning L with $p = 0$ based on the efficient sample size for the block conditionals, and then increasing p until there is no noticeable improvement. Alternatively, L could be selected a priori and N tuned independently for each block.

Acknowledgement

S. S. Singh and F. Lindsten are joint first authors. The authors thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme Monte Carlo Inference for Complex Statistical Models when work on this paper was undertaken. This work was supported by the U.K. Engineering and Physical Sciences Research Council and the Swedish Research Council.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of Theorems 1, 2, A1 and A2, Lemma 3 and Proposition A1, as well as additional details and results on the numerical illustration in § 5.

Appendix

Proof of Lemma 2

From (5) we have $P^J(x, \mathrm{d}x') = \phi_x^J(\mathrm{d}x'_J)\delta_{x_{J^c}}(\mathrm{d}x'_{J^c})$. A candidate Wasserstein matrix W^J can be found via a coupling argument. For $j \in I$ and any pair (x, z) such that $x_{-j} = z_{-j}$ and $x_j \neq z_j$, let $\Psi^J_{j,x,z}$ be a coupling

of ϕ_x^J and ϕ_z^J. Then

$$\begin{aligned}\left|P^J f(x) - P^J f(z)\right| &\leqslant \int \Psi_{j,x,z}^J(\mathrm{d}x'_J, \mathrm{d}z'_J)\delta_{x_{J^c}}(\mathrm{d}x'_{J^c})\delta_{z_{J^c}}(\mathrm{d}z'_{J^c})|f(x') - f(z')| \\ &\leqslant \sum_{i\in I} \mathrm{OSC}_i(f) \int \Psi_{j,x,z}^J(\mathrm{d}x'_J, \mathrm{d}z'_J)\delta_{x_{J^c}}(\mathrm{d}x'_{J^c})\delta_{z_{J^c}}(\mathrm{d}z'_{J^c})\mathbb{I}_{[x'_i \neq z'_i]} \\ &= \sum_{i\in J} \mathrm{OSC}_i(f) \int \Psi_{j,x,z}^J(\mathrm{d}x'_i, \mathrm{d}z'_i)\mathbb{I}_{[x'_i \neq z'_i]} + \sum_{i\in J^c} \mathrm{OSC}_i(f)\mathbb{I}_{[x_i \neq z_i]},\end{aligned}$$

where the second line follows from (8). Since $x_{-j} = z_{-j}$, at most one term from the sum over J^c will be nonzero and hence $\sum_{i\in J^c} \mathrm{OSC}_i(f)\mathbb{I}_{[x_i\neq z_i]} = \mathbb{I}_{[j\in J^c]}\mathrm{OSC}_j(f)$. Therefore, for $i \in J$ set

$$W_{i,j}^J = \sup_{\{x,z\in\mathcal{X}^n : x_{-j}=z_{-j}\}} \int \Psi_{j,x,z}^J(\mathrm{d}x'_i, \mathrm{d}z'_i)\mathbb{I}_{[x'_i\neq z'_i]}, \tag{A1}$$

and for $i \notin J$ set $W_{i,j}^J = \mathbb{I}_{[i=j\in J^c]}$. Furthermore, since ϕ_x^J depends on x only through the boundary points $x_{\partial J}$, it follows that for $j \notin \partial J$ we have $\phi_x^J = \phi_z^J$. Therefore, for $j \notin \partial J$, the coupling $\Psi_{j,x,z}^J$ can be made perfect: $\Psi_{j,x,z}^J(\mathrm{d}x'_J, \mathrm{d}z'_J) = \phi_x(\mathrm{d}x'_J)\delta_{x'_J}(\mathrm{d}z'_J)$, that is, $\int \Psi_{j,x,z}^J(\mathrm{d}x'_i, \mathrm{d}z'_i)\mathbb{I}_{[x'_i=z'_i]} = 1$ for any $i \in J$ and $j \notin \partial J$. Therefore, it is evident from (A1) that $W_{i,j}^J = 0$ for $i \in J$ and $j \in I \setminus \partial J$.

Proof of the convergence of the particle Gibbs block sampler

This subsection is dedicated to the proof of Theorem 3. It provides a more general version of Theorem 3 without the common block length and overlap structure of Assumption 5. Let $\hat{W}^J$ be a Wasserstein matrix for the nonideal block transition kernel defined in (6). By an analogous argument to that in Lemma 2, it follows that $\hat{W}^J$ has a similar structure to W^J but with possibly nonzero entries for rows $i \in J$ and columns $j \in J$. This motivates the following assumed structure.

Assumption A1. For each $J \in \mathcal{J}$ let W^J be a matrix satisfying (11). For some constant $\epsilon \in [0, 1)$ and for all $J \in \mathcal{J}$, let the matrix $\hat{W}^J$ with elements $\hat{W}_{i,j}^J = W_{i,j}^J + \epsilon\,\mathbb{I}_{[i\in J, j\in J_+]}$ $(i,j \in I)$ be a Wasserstein matrix for the nonideal transition kernel (6) which updates block J.

Proposition A1 below shows that Assumption A1 holds for the particle Gibbs kernel with W^J being a Wasserstein matrix for block J of the ideal sampler as in Lemma 3; ϵ depends on the block size $|J|$ and particle number N but is independent of the data length n and the specific observations pertaining to each block, which is the key to the stability of the particle Gibbs block sampler when $n \to \infty$ but with N fixed.

THEOREM A1. *Suppose that Assumptions 2, 3 and A1 hold and that the number of blocks $m = |\mathcal{J}|$ is odd. Let $J_{-k} = \bigcup_{J\in\mathcal{J}\setminus\{J_k\}} J$ and $L = \max_{J\in\mathcal{J}} |J|$. Let $\hat{\mathcal{W}}$ be the Wasserstein matrix of one complete sweep of the nonideal parallel sampler defined analogously to Theorem 1. Then*

$$(\hat{\mathcal{W}}1)_i \leqslant \begin{cases} \lambda^2 + \epsilon\{\lambda(L+4) + \epsilon(L+2)^2 + L(1\vee\beta)\}, & i \in J_{-k}^c,\ k \text{ even}, \\ \lambda + \epsilon\,(L+2), & i \in J_{-k}^c,\ k \text{ odd}, \\ \lambda\beta + \epsilon\{\beta(L+2) + 2\lambda + \epsilon\,(L+2)^2 + L(1\vee\beta)\}, & i \in J_k \cap J_{-k},\ k \text{ even}, \end{cases}$$

where

$$\lambda = \max_{J_k\in\mathcal{J}} \max_{i\in J_{-k}^c} W_{i,\partial_- J_k}^{J_k} + W_{i,\partial_+ J_k}^{J_k}, \quad \beta = \max_{J\in\mathcal{J}} \max_{i\in J} \sum_{j\in\partial J} W_{i,j}^J$$

and $W_{i,\partial_- J_1}^{J_1} = W_{i,\partial_+ J_m}^{J_m} = 0$ by convention.

For the proof see the Supplementary Material.

THEOREM A2. *Suppose that Assumptions 2, 3 and A1 hold, and let* $J_{-k} = \bigcup_{J \in \mathcal{J} \setminus \{J_k\}} J$, $L = \max_{J \in \mathcal{J}} |J|$ *and* $L_1 = \max_{k \in \{2,\ldots,m\}} |J_{k-1} \cap J_k|$. *Let* $\hat{\mathcal{W}}$ *be the Wasserstein matrix of one complete sweep of the nonideal left-to-right sampler. Then*

$$(\hat{\mathcal{W}}1)_i \leqslant \begin{cases} \lambda + c\epsilon, & i \in J_{-k}^{\mathrm{c}},\ k = 1,\ldots,m, \\ \lambda' + 2c\epsilon, & i \in J_{k-1} \cap J_k,\ k = 2,\ldots,m, \end{cases}$$

where

$$\lambda = \max_{J_k \in \mathcal{J}} \max_{i \in J_{-k}^{\mathrm{c}}} \frac{W_{i,\partial_+ J_k}^{J_k}}{1 - W_{i,\partial_- J_k}^{J_k}}, \quad \lambda' = \max_{k \in \{2,\ldots,m\}} \max_{i \in J_{k-1} \cap J_k} \lambda W_{i,\partial_- J_k}^{J_k} + W_{i,\partial_+ J_k}^{J_k},$$

$$\beta = \max_{J \in \mathcal{J}} \max_{i \in J} \sum_{j \in \partial J} W_{i,j}^{J}, \quad \gamma = \max_{k \in \{2,\ldots,m\}} \max_{i \in J_k \cap J_{k-1}^{\mathrm{c}}} W_{i,\partial_- J_k}^{J_k}, \quad c = \frac{L(\beta(\lambda \vee 1) \vee 1) + 1 + \lambda}{1 - (L_1 + 1)\epsilon - \gamma},$$

provided that $\gamma + (L_1 + 1)\epsilon < 1$. *As before the convention is* $W_{i,\partial_- J_1}^{J_1} = W_{i,\partial_+ J_m}^{J_m} = 0$.

The proof is given in the Supplementary Material.

In order to prove Theorem 3 we now use Lemma 3 to identify the constants in Theorems A1 and A2. However, a technical detail is how to handle the dependence on the maximum block size L and, in Theorem A2, the maximum overlap L_1 as well. Indeed, a direct application of Theorems A1 and A2 would suggest that the norm of $\hat{\mathcal{W}}$ grows, respectively, quadratically and linearly with ϵL. To avoid this issue we will make use of the following trick: when applying Theorems A1 and A2 we do not consider the original hidden Markov model formulation but rather an equivalent model that lumps consecutive states together, thus effectively reducing the size of the blocks. As illustrated in Fig. 1, each block can be split into three distinct sections: the left overlap, the middle of the block, and the right overlap. The exceptions are the end blocks, which are split into two sections. By viewing each section as a single lumped state, we reduce the block size to 3 and the maximum overlap to 1. For this scheme, the lumped states $\Xi_1 = (X_1,\ldots,X_{L-p})$ and $\Xi_2 = (X_{L-p+1},\ldots,X_L)$ are the two sections of block 1, while Ξ_2, $\Xi_3 = (X_{L+1},\ldots,X_{2L-2p})$ and $\Xi_4 = (X_{2L-2p+1},\ldots,X_{2L-p})$ are the three lumped states of block 2, and so on.

DEFINITION A1. *The* Ξ*-system groups the random variables* $X_1,\ldots,X_n$ *of the* X*-system into* $\Xi_1,\ldots,\Xi_{2m-1}$, *where Assumption 5 implies* $n = (L-p)m + p$, *the end blocks are* $\Xi_1 = (X_1,\ldots,X_{L-p})$ *and* $\Xi_{2m-1} = (X_{n-(L-p)+1},\ldots,X_n)$, *and the intermediate blocks are*

$$\Xi_{2i} = X_{(L-p)i+1},\ldots,X_{(L-p)i+p} \quad (1 \leqslant i < m),$$
$$\Xi_{2i-1} = X_{(L-p)(i-1)+p+1},\ldots,X_{(L-p)i} \quad (1 < i < m).$$

The index set for the Ξ*-system is* $I_\Xi = \{1,\ldots,2m-1\}$ *and the cover* $\mathcal{J}_\Xi$ *of* I_Ξ *has* m *sets, with set* k *being* $J_{\Xi,k} = \{2k-2, 2k-1, 2k\} \cap I_\Xi$.

To find a Wasserstein matrix for the Ξ-system we note that any conditional density of the states Ξ_i ($i \in J_{\Xi,k}$), conditionally on the boundaries of the block $J_{\Xi,k}$ and the observation pertaining to that block, is coupled analogously to the X-system; see the proof of Lemma 3 in the Supplementary Material. Analogously to Lemma 3, a Wasserstein matrix for block $J_{\Xi,k}$ of the Ξ-system is the $(2m-1) \times (2m-1)$

matrix

$$W_{\Xi}^{k} = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & \alpha^{\lfloor h^{-1} \rfloor} & 0 & 0 & 0 & \alpha^{\lfloor h^{-1}(L-p+1) \rfloor} & \\ & & \alpha^{\lfloor h^{-1}(p+1) \rfloor} & 0 & 0 & 0 & \alpha^{\lfloor h^{-1}(p+1) \rfloor} & \\ & & \alpha^{\lfloor h^{-1}(L-p+1) \rfloor} & 0 & 0 & 0 & \alpha^{\lfloor h^{-1} \rfloor} & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & 1 \end{bmatrix}. \tag{A2}$$

Here the square of zeros corresponds to rows and columns in the set $\{2k-2, 2k-1, 2k\}$, with obvious modifications for $k = 1$ or $k = 2m-1$. The proof is omitted as it is similar to that of Lemma 3.

The final ingredient for proving Theorem 3 is to verify Assumption A1 for the particle Gibbs kernel.

Proposition A1. *Suppose that Assumption 1 holds and that the bootstrap proposal kernel $r(x, x') = m(x, x')$ is used in Algorithm 1. Then for any $J = \{s, \ldots, u\} \subseteq I$ with $u \geqslant s$ and any $N \geqslant 2$,*

$$Q_N^J(x_{J_+}, \mathrm{d}x'_J) \geqslant \{1 - \epsilon(N, |J|)\}\phi_x^J(\mathrm{d}x'_J),$$

where $1 - \epsilon(N, L) = [1 - \{c(N-1)+1\}^{-1}]^L$, $c = (2\delta\sigma_+/\sigma_- - 1)^{-1} \in (0, 1]$, and δ, σ_- and σ_+ are defined in Assumption 1.

If, moreover, Assumptions 2, 3 and 5 also hold, then W_{Ξ}^k $(k = 1, \ldots, m)$ defined in (A2) is a Wasserstein matrix, based on the Ξ-system, for the ideal Gibbs kernel that updates block J_k. A Wasserstein matrix for the particle Gibbs kernel $Q_N^{J_k}$ is the matrix $\hat{W}_{\Xi}^k$ with elements

$$(\hat{W}_{\Xi}^k)_{i,j} = (W_{\Xi}^k)_{i,j} + \epsilon(N, |J_k|)\mathbb{I}_{[2k-2 \leqslant i \leqslant 2k]}\mathbb{I}_{[2k-3 \leqslant j \leqslant 2k+1]}. \tag{A3}$$

See the Supplementary Material for the proof.

We emphasize that the lumping of state variables is used only for the sake of analysis, to improve the contraction rates in Theorems A1 and A2 by avoiding a dependence on the block size. For all practical purposes the lumping has no effect: when implementing Algorithm 1 we still use the original state variables. Consequently, lumping does not affect the ergodicity or the convergence rate of the particle Gibbs kernel. In particular, ϵ in (A3) depends on the size of block k as expressed in the X-system, which is $|J_k|$ and not $|J_{\Xi,k}|$ as for the lumped Ξ-system. We conclude by putting the pieces together to prove Theorem 3.

Proof of Theorem 3

Theorem 3 is proved by applying Theorems A1 and A2 to the Ξ-system. The bound (14) is established by using (10) for the Ξ-system: for any coupling Ψ of μ and ν with $(\Xi, \check{\Xi}) \sim \Psi$,

$$\begin{aligned} &\left|\mu\mathcal{P}^k f - \nu\mathcal{P}^k f\right| \\ &\quad \leqslant \left\{ \sum_{i=1}^{2m-1} \sup_{\{\Xi, \check{\Xi} \in \mathcal{X}^n : \Xi_{-i} = \check{\Xi}_{-i}\}} |f(\Xi) - f(\check{\Xi})| \right\} \|\mathcal{W}_{\Xi}\|_{\infty}^k \max_{j \in \{1, \ldots, 2m-1\}} \int \Psi(\mathrm{d}\Xi, \mathrm{d}\check{\Xi})\mathbb{I}_{[\Xi_j \neq \check{\Xi}_j]}. \end{aligned}$$

Since $(\Xi_1, \ldots, \Xi_{2m-1}) \in \mathcal{X}^n$, we can crudely bound the sum by $\sum_{i=1}^n \sup_{\{x,y \in \mathcal{X}^n : x_{-i} = y_{-i}\}} |f(x) - f(y)|$. The final term is also crudely bounded by $\int \Psi(\mathrm{d}\Xi, \mathrm{d}\check{\Xi})\mathbb{I}_{[\Xi_j \neq \check{\Xi}_j]} \leqslant \int \Psi(\mathrm{d}\Xi, \mathrm{d}\check{\Xi})\mathbb{I}_{[\Xi \neq \check{\Xi}]}$.

Now it remains to derive the expression for λ_{PG} for both the parallel and left-to-right cases. We detail the parallel case only, which uses Theorem A1 for the Ξ-system, as the left-to-right case follows by analogous

arguments. Theorem A1 is applicable to the Ξ-system since the Ξ-system satisfies Assumptions 2, 3 and A1; the fact that it satisfies Assumption A1 follows from Proposition A1. Each block $J_{\Xi,k}$ of the Ξ-system has three elements, except for the initial and final blocks which have two elements each. The specific values of λ and β in Theorem A1 follow from this simple three-element block structure and the declared Wasserstein matrix in (A2). The coefficient of the ϵ-term in (15) follows from a trivial bound on the three separate ϵ-coefficients given in Theorem A1 using $L = 3$.

References

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

Andrieu, C., Lee, A. & Vihola, M. (2017). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, **81**.

Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–53.

Chopin, N. & Singh, S. S. (2015). On particle Gibbs sampling. *Bernoulli* **21**, 1855–83.

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

Doucet, A., Godsill, S. J. & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.* **10**, 197–208.

Doucet, A. & Johansen, A. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering*, D. Crisan & B. Rozovskii, eds. Oxford: Oxford University Press, pp. 656–704.

Follmer, H. (1982). A covariance estimate for Gibbs measures. *J. Funct. Anal.* **46**, 387–95.

Friedlander, B. (1982). Lattice filters for adaptive processing. *Proc. IEEE* **70**, 829–67.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Ser. Anal.* **15**, 183–202.

Godsill, S. J., Doucet, A. & West, M. (2004). Monte Carlo smoothing for nonlinear time series. *J. Am. Statist. Assoc.* **99**, 156–68.

Lindsten, F., Douc, R. & Moulines, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scand. J. Statist.* **42**, 775–97.

Lindsten, F., Jordan, M. I. & Schön, T. B. (2014). Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.* **15**, 2145–84.

Lindsten, F. & Schön, T. B. (2013). Backward simulation methods for Monte Carlo statistical inference. *Foundat. Trends Mach. Learn.* **6**, 1–143.

Pasarica, C. & Gelman, A. (2010). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statist. Sinica* **20**, 343–64.

Rebeschini, P. & van Handel, R. (2014). Comparison theorems for Gibbs measures. *J. Statist. Phys.* **157**, 234–81.

Wang, N.-Y. & Wu, L. (2014). Convergence rate and concentration inequalities for Gibbs sampling in high dimensions. *Bernoulli* **20**, 1698–716.

Whiteley, N. (2010). Discussion of "Particle Markov chain Monte Carlo methods". *J. R. Statist. Soc.* B **72**, 306–7.

[*Received on* 13 *May* 2016. *Editorial decision on* 14 *July* 2017]