Supplementary material for “Blocking strategies and stability of particle Gibbs samplers”

BY S. S. SINGH

Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.

sss40@cam.ac.uk

F. LINDSTEN

Department of Information Technology, Uppsala University, Uppsala 751 05, Sweden

fredrik.lindsten@it.uu.se

AND E. MOULINES

Institut Télécom/ Télécom ParisTech (ENST), Département TSI / CNRS UMR 5141, 46 Rue Barrault, 75634 Paris, Cedex 13, France

moulines@telecom-paristech.fr

SUMMARY

This supplementary material for the paper Blocking strategies and stability of particle Gibbs samplers contains proofs of Theorems 1, 2, A1 and A2, Lemma 3 and Proposition A1. It also contains additional details and results for the numerical illustration presented in Section 5 of the main manuscript. Labels in this supplementary material have prefix ‘S’ and references (e.g. to equations) without this prefix refer to the main manuscript.

S1. PROOFS OF THE MAIN THEOREMS

For each \(j \in I\) let \(e_j \in \mathbb{R}^n\) be the unit vector \((e_j)_i = I\{i=j\}\). Let \(1\) be the vector of ones. We will use the notation \(x_{s:u} = (x_s, \ldots, x_u)\) to denote sequences of variables. The symbol \(p\) will be used as generic density function which is identified by its arguments and we will occasionally use this notation for brevity when no confusion is possible. For instance, the joint smoothing distribution for a block \(J = \{s, \ldots, u\}\) is \(\phi_J(x_s, \ldots, x_u) = p(x_{s:u} | x_{s-1}, x_{u+1}, y_{s:u})\).

**Lemma S1.** Let \(J = \{J_1, \ldots, J_m\}\) be an arbitrary cover of \(I\). For each \(J \in J\), let \(W^J\) be a matrix with the structure in (11) and let \(W = W^{J_m} \cdots W^{J_1}\). For each \(j \in I\) let \(a_j = \min\{k : j \in J_k\}\) and \(b_j = \min\{k : j \in \partial J_k\}\), with the convention that \(\min\emptyset = \infty\). If \(a_j < b_j\) then \(W e_j = 0\).

**Proof.** Due to the structure of \(W^J\) in (11) it follows that, for \(j \notin \partial J\), \(W^J e_j = \mathbb{I}_{[j \in J]} e_j\). If \(a_j < b_j\) then

\[
W e_j = W^{J_m} \cdots W^{J_{b_j}} \cdots W^{J_{a_j}} \cdots W^{J_1} e_j
= W^{J_m} \cdots W^{J_{b_j}} \cdots W^{J_{a_j}} e_j = 0,
\]

where the first equality follows from the fact that for any \(k < a_j < b_j, j \notin J_k \cup \partial J_k\) by the definition of \(a_j\) and \(b_j\).

© 2016 Biometrika Trust
S1-1. Proof of Theorem 1

From Lemma S1 it is clear that \( \mathcal{W}_{e_j} = 0 \) for any \( j \notin \partial \) (for which \( b_j = \infty \)). Hence, with \( M \) being the binary mask matrix with elements \( M_{i,j} = \mathbb{I}_{[i=j \in \partial]} \), it holds that \( \mathcal{W}_r = \mathcal{W}(Mr) \) for any vector \( r \). Thus \( \mathcal{W}^k \mathcal{W}^{-1} = \mathcal{W}(M)^k \mathcal{W}^{-1} \) and hence \( \|\mathcal{W}^k\| \leq\|\mathcal{W}\|_\infty \|\mathcal{W}^{-1}\|_\infty \|\mathcal{W}\|_{\infty}^{-1} \).

Next, consider \( \| MW \|_\infty = \max_{i,j} e_i^T \mathcal{W} \). Let \( L_0 = \emptyset \) and \( r_0 = 1 \) and define recursively

\[
L_k = L_{k-1} \cup J_k, \quad r_k = W^{J_k}r_{k-1} \quad (k = 1, \ldots, m).
\]

(S1)

We thus have \( r_m = \mathcal{W}1 \). Hence, the result follows if, for \( k = 0, \ldots, m, \)

\[
(r_k)_i = \begin{cases} 1, & i \in I \setminus L_k, \\ \lambda, & i \in L_k \cap \partial. \end{cases}
\]

(S2)

Nothing is said about \( (r_k)_i \) for \( i \in L_k \setminus \partial \). Indeed, the fact that \( J \) is a cover of \( I \) implies that \( L_m = I \). Hence \( \mathcal{W}1 = (r_m)_i \leq \lambda \) for all \( i \in \partial \).

It remains to prove (S2). For \( k = 0 \) the hypothesis is true by construction. We proceed inductively. Hence, assume that the hypothesis is true for \( k - 1 \). Consider

\[
(r_k)_i = (W^{J_k}r_{k-1})_i = \begin{cases} (r_{k-1})_i, & i \in J'_k, \\ \sum_{j \in \partial J_k} W_{i,j}^{J_k} (r_{k-1})_j, & i \in J_k, \end{cases}
\]

(S3)

where we have made use of the structure of the Wasserstein matrix from Lemma 2. We need to consider three different cases: firstly if \( i \in I \setminus L_k \) then \( (r_k)_i = (r_{k-1})_i \leq 1 \), where we first use (S3) and then the induction hypothesis, and the fact that \( L_k = L_{k-1} \cup J_k \). Secondly, if \( i \in (L_{k-1} \setminus J_k) \cap \partial \) then \( (r_k)_i = (r_{k-1})_i \leq \lambda \) where again we use (S3) and the induction hypothesis. Finally, if \( i \in J_k \cap \partial \) then \( (r_k)_i = \sum_{j \in \partial J_k} W_{i,j}^{J_k} (r_{k-1})_j \leq \sum_{j \in \partial J_k} W_{i,j}^{J_k} \leq \lambda \), where we use the fact that for \( j \in \partial \), \( (r_{k-1})_j \leq 1 \) and Assumption 4 for the final inequality. This completes the proof.

S1-2. Proof of Theorem 2

Comming with the left-to-right sampler, the proof is similar to that of Theorem 1, but to exploit the structure of the left-to-right sampler we define the mask matrix \( M \) as \( M_{i,j} = \mathbb{I}_{[i-j \in \partial_+]} \) where \( \partial_+ = \partial J \cup J \partial \) is the set of right boundary points only, of all blocks. If \( j = \partial_- J_k \) is the left boundary of some block \( k \), say, then \( j \in J_{k-1} \) by definition of the left-to-right sampler. Hence, from Lemma S1 it follows that \( \mathcal{W}_r = \mathcal{W}(Mr) \) for any vector \( r \). Thus, \( \|\mathcal{W}^k\| \leq\|\mathcal{W}\|_\infty \|\mathcal{W}^{-1}\|_\infty \|\mathcal{W}\|_{\infty}^{-1} \).

Define \( L_k \) and \( r_k \) as in (S1). Note that (S2) and (S3) hold for any cover \( J \), and in particular for the left-to-right sampler. Thus, for \( i \in J_k, \)

\[
(r_k)_i = W_{i,\partial_- J_k}^{J_k} (r_{k-1})_{\partial_- J_k} + W_{i,\partial_+ J_k}^{J_k} (r_{k-1})_{\partial_+ J_k} \leq W_{i,\partial_- J_k}^{J_k} \lambda + W_{i,\partial_+ J_k}^{J_k},
\]

where the inequality follows from the fact that, for the left-to-right sampler, \( \partial_- J_k \in L_{k-1} \) and \( \partial_+ J_k \in I \setminus L_k \) and by using (S2). Hence, \( (r_k)_{\partial_+ J_{k-1}} \leq \lambda \). For \( k = 1, \ldots, m \) since \( \partial_+ J_{k-1} \) lies in at most one block, namely \( J_k \), we conclude that \( (r_m)_{\partial_+} \leq \beta \) for any \( i \in \partial_+ \). Thus \( \| MW \|_\infty \leq \beta \). The bound \( \| W \|_\infty \leq 1 + \lambda \) follows similarly from (S4).

For the parallel sampler we redefine the mask \( M \) to be \( M_{i,j} = \mathbb{I}_{[i-j \in \partial_\text{odd}]} \) where \( \partial_\text{odd} = \partial_{\text{odd}} J_k \) is the set of all boundary points of odd blocks. Since any boundary point of an even block is the interior of some odd block, it follows from Lemma S1 that \( \mathcal{W}_r = \mathcal{W}(Mr) \) for any vector \( r \): this is most easily seen by noting that the parallel sampler is equivalent to a two-block sampler, comprising the composite blocks \( J_{\text{odd}} = \cup_{\text{odd} J_{k}} J_k \) and \( J_{\text{even}} = \cup_{\text{even} J_{k}} J_k \), the statement
Blocking Strategies and Stability of Particle Gibbs Samplers

then following by applying Lemma S1 to these two blocks. To complete the proof we thus need to bound \( \|MWM\|_\infty \leq \lambda^2 \). Let \( r \in \partial \) and let \( J_k \) be the block such that \( i \in J_k \). Note that \( k \) is odd, separately. We have for \( k \)

Then \( \|W\|_\infty \leq \lambda \) for all \( i \) which are boundaries of even blocks, a fact we will prove next, then

\[
(W_{\text{odd}}^{J_k})_i = W_{i,\partial - J_k}^{J_k} (W_{\text{odd}})_{\partial - J_k} + W_{i,\partial + J_k}^{J_k} (W_{\text{odd}}^{J_k})_{\partial + J_k} \leq \lambda^2.
\]

To conclude: let \( i \) be a boundary of an even block and let \( i \in J_k \), the only odd block containing \( i \). Then \( (W_{\text{odd}}^{J_k})_i = (W_{J_k}^{J_k})_i = W_{i,\partial - J_k}^{J_k} + W_{i,\partial + J_k}^{J_k} \leq \lambda \). \( \square \)

S1.3. Proof of Theorem A1

We prove the following more general version whereas Theorem A1 was stated for \( r = 1 \).

**Lemma S2.** Let the vector \( r^T = (r_1, \ldots, r_n) \in \mathbb{R}^n \) satisfy

\[
r_i = \begin{cases} a, & i \in J_{-k}^e, k \text{ even}, \\ a', & i \in J_{-k}^o, k \text{ odd}, \\ b, & i \in J_k \cap J_{-k}, k \text{ even}, \end{cases}
\]

for some positive constants \( a, a', b \). Then

\[
(\lambda W)_i \leq \begin{cases} \lambda^2 a + \epsilon \{ \lambda c + 2\lambda a + 2\epsilon c + La(1 \lor \beta) + \epsilon Lc \}, & i \in J_{-k}^e, k \text{ even}, \\ \lambda a + \epsilon \{ 2a + L(a' \lor b) \}, & i \in J_{-k}^o, k \text{ odd}, \\ \lambda \beta a + \epsilon \{ \beta c + 2\lambda a + 2\epsilon c + La(1 \lor \beta) + \epsilon Lc \}, & i \in J_k \cap J_{-k}, k \text{ even}, \end{cases}
\]

where \( c = 2a + L(a' \lor b) \) and the remaining constants were defined in Theorem A1.

**Proof.** Note that \( \lambda W = \lambda W_{\text{even}} \lambda W_{\text{odd}} \) where

\[
\lambda W_{\text{odd}} = \lambda W_{J_m} \cdots \lambda W_{J_k} \lambda W_{J_1} \quad \text{and} \quad \lambda W_{\text{even}} = \lambda W_{J_{m-1}} \cdots \lambda W_{J_k} \lambda W_{J_2}.
\]

Let \( r' = \lambda W_{\text{odd}} r \) and \( r'' = \lambda W_{\text{even}} r' \). Recall that for any vector \( s \), \( \lambda W_{\text{odd}} s \) differs from \( s \) only in components indexed by \( J_k \). Thus, by 2, 3, \( \lambda W_{\text{odd}} r \) can be studied by considering each term \( \lambda W_{J_k} r \), \( k \) odd, separately. We have for \( k \) odd and \( i \in J_k \)

\[
(\lambda W_{J_k} r)_i = (W_{i,\partial - J_k}^{J_k} + \epsilon)r_{\partial - J_k} + (W_{i,\partial + J_k}^{J_k} + \epsilon)r_{\partial + J_k} + \epsilon \sum_{j \in J_k} r_j
\]

\[
= (W_{i,\partial - J_k}^{J_k} + \epsilon)r_{\partial - J_k} + 2\epsilon a + \epsilon \sum_{j \in J_k} r_j
\]

\[
\leq (W_{i,\partial - J_k}^{J_k} + \epsilon)r_{\partial + J_k} + 2\epsilon a + \epsilon |J_k|(a' \lor b).
\]

The second line follows by the assumption on \( r \) and the fact that boundaries of an odd numbered block lie in the adjacent even blocks. For \( i \in J_{-k}^e, W_{i,\partial - J_k}^{J_k} + W_{i,\partial + J_k}^{J_k} \leq \lambda \) and since \( |J_k| \leq L \), then

\[
(\lambda W_{J_k} r)_i \leq \lambda a + \epsilon \{ 2a + L(a' \lor b) \}.
\]

For \( i \in J_k \cap J_{-k}, W_{i,\partial - J_k}^{J_k} + W_{i,\partial + J_k}^{J_k} \leq \beta \) and thus

\[
(\lambda W_{J_k} r)_i \leq \beta a + \epsilon \{ 2a + L(a' \lor b) \}.
\]
We separately bound the terms for $J_i$ indexed by $r$. In summary, with $r' = \hat{W}_{\text{odd}} r'$,

$$(r')_i \leq \begin{cases} a, & i \in J_{-k}^c, \text{ } k \text{ even}, \\ \lambda a + \epsilon \{2a + L(a' \lor b)\}, & i \in J_{-k}^c, \text{ } k \text{ odd}, \\ \beta a + \epsilon \{2a + L(a' \lor b)\}, & i \in J_k \cap J_{-k}, \text{ } k \text{ even}, \end{cases}$$

Similarly, $r'' = \hat{W}_{\text{even}} r'$ can be studied by considering each term $\hat{W}_{\text{odd}} r'$, $k$ even, separately. Let $c = 2a + L(a' \lor b)$. For $k$ even and $i \in J_k$,

$$(\hat{W}_{J_k} r')_i = (W_{i,\partial_+ J_k} + \epsilon) r'_{\partial_+ J_k} + (W_{i,\partial_+ J_k} + \epsilon) r'_{\partial_+ J_k} + \epsilon \sum_{j \in J_k} r'_{j}$$

$$\leq (W_{i,\partial_+ J_k} + W_{i,\partial_+ J_k} + 2\epsilon)(\lambda a + \epsilon c) + \epsilon \sum_{j \in J_k} r'_{j}$$

$$\leq (W_{J_k} + W_{i,\partial_+ J_k} + 2\epsilon)(\lambda a + \epsilon c) + \epsilon |J_k|\{a(1 \lor \beta) + \epsilon c\}.$$

For $i \in J_{-k}^c$, $W_{i,\partial_- J_k} + W_{i,\partial_+ J_k} \leq \lambda$ and thus,

$$(\hat{W}_{J_k} r')_i \leq \lambda^2 a + \epsilon \{\lambda c + 2\lambda a + 2\epsilon c + L(1 \lor \beta) + \epsilon L c\}.$$  

For $i \in J_k \cap J_{-k}$, $W_{i,\partial_- J_k} + W_{i,\partial_+ J_k} \leq \beta$ and thus,

$$(\hat{W}_{J_k} r')_i \leq \lambda \beta a + \epsilon \{\beta c + 2\lambda a + 2\epsilon c + L(1 \lor \beta) + \epsilon L c\}.$$  

In summary, with $r'' = \hat{W}_{\text{even}} r'$,

$$(r'')_i \leq \begin{cases} \lambda^2 a + \epsilon \{\lambda c + 2\lambda a + 2\epsilon c + L(1 \lor \beta) + \epsilon L c\}, & i \in J_{-k}^c, \text{ } k \text{ even}, \\ \lambda a + \epsilon \{2a + L(a' \lor b)\}, & i \in J_{-k}^c, \text{ } k \text{ odd}, \\ \lambda \beta a + \epsilon \{\beta c + 2\lambda a + 2\epsilon c + L(1 \lor \beta) + \epsilon L c\}, & i \in J_k \cap J_{-k}, \text{ } k \text{ even}. \end{cases}$$

S1.4. Proof of Theorem A2

The proof is inductive. For any vector $r$, observe that $\hat{W}_{J_k} r$ differs from $r$ only in components indexed by $J$. In particular, for any $i \in J_1$, $(\hat{W}^{J_1} r)_i = W_{i,\partial_+ J_1} + (|J_1| + 1) \epsilon$. Thus,

$$(\hat{W}^{J_1} r)_i \leq \begin{cases} \lambda + \epsilon \lambda, & i \in J_{-1}^c, \\ \beta + \epsilon \lambda, & i \in J_1 \cap J_2. \end{cases}$$

We separately bound the terms for $i \in J_{-1}^c$ and $i \in J_1 \cap J_2$ since $W_{i,\partial_+ J_1}$ can approach 1 as $i$ approaches the boundary $\partial_+ J_1$; recall that $\hat{W}_{i,\partial_+ J_1} = W_{i,\partial_+ J_1} + \epsilon$ for $i \in J_1$. Now assume

$$(\hat{W}^{J_{k-1}, \ldots, J_1} r)_i \leq \begin{cases} \lambda + \epsilon \lambda, & i \in J_{- (k-1)}, \\ \lambda' + 2\epsilon \lambda, & i \in J_{k-2} \cap J_{k-1}, \\ \beta(\lambda \lor 1) + \epsilon \lambda, & i \in J_{k-2} \cap J_{k-1}; \end{cases}$$
For $i$ in the first case indexes the middle of block $J_{k-1}$, its left overlap in the second case and its right overlap in the final one. We have

$$
(W^{J_k} \tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_i = \sum_{j \in J_k^c} W^{J_k}_{i,j} (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j
$$

$$
= \sum_{j \in \partial J_k} (W^{J_k}_{i,j} + \epsilon) (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j + \sum_{j \in J_k} \epsilon (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j.
$$

The first term of the sum can be simplified to

$$
\sum_{j \in \partial J_k} (W^{J_k}_{i,j} + \epsilon) (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j
$$

$$
= (W^{J_k}_{i,\partial J_k} + \epsilon) (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_{\partial J_k} + (W^{J_k}_{i,\partial J_k} + \epsilon) (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_{\partial J_k}
$$

$$
= (W^{J_k}_{i,\partial J_k} + \epsilon) (\lambda + \epsilon c) + (W^{J_k}_{i,\partial J_k} + \epsilon)
$$

$$
= W^{J_k}_{i,\partial J_k} \lambda + W^{J_k}_{i,\partial J_k} + \epsilon (W^{J_k}_{i,\partial J_k} c + \lambda + \epsilon c + 1).
$$

The second term of (S5) is

$$
\sum_{j \in J_k} \epsilon (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j
$$

$$
= \sum_{j \in J_{k-1} \cap J_k} \epsilon (\tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_j + \sum_{j \in J_{k-1} \cap J_k} \epsilon
$$

$$
\leq |J_{k-1} \cap J_k| \epsilon \{\beta(\lambda \lor 1) + c\} + \epsilon |J_{k-1} \cap J_k|
$$

$$
\leq \epsilon \{\beta(\lambda \lor 1) \lor 1\} |J_k| + c \epsilon |J_{k-1} \cap J_k|.
$$

Thus

$$
(\tilde{W}^{J_k} \tilde{W}^{J_{k-1}} \ldots \tilde{W}^{J_1})_i \leq W^{J_k}_{i,\partial J_k} \lambda + W^{J_k}_{i,\partial J_k}
$$

$$
+ \epsilon [W^{J_k}_{i,\partial J_k} c + \lambda + \epsilon c + 1 + \{\beta(\lambda \lor 1) \lor 1\} |J_k| + c \epsilon |J_{k-1} \cap J_k|]
$$

For $i \in J_k \cap J_{k+1}$, $W^{J_k}_{i,\partial J_k} \lambda + W^{J_k}_{i,\partial J_k} \leq \beta(\lambda \lor 1)$. For $i \in J_{k-1} \cap J_k$, $W^{J_k}_{i,\partial J_k} \lambda + W^{J_k}_{i,\partial J_k} \leq \lambda$ by the definition of $\lambda$. For $i \in J_{k-1} \cap J_k$, $W^{J_k}_{i,\partial J_k} \lambda + W^{J_k}_{i,\partial J_k} \leq \lambda'$ by the definition of $\lambda'$.

For $i \in J_k \cap J_{k-1}$, the coefficient of $\epsilon$ is by itself bounded by $c$ since

$$
\frac{\lambda + 1 + \{\beta(\lambda \lor 1) \lor 1\} |J_k|}{1 - W^{J_k}_{i,\partial J_k} - \epsilon - \epsilon |J_{k-1} \cap J_k|} \leq c.
$$

For $i \in J_{k-1} \cap J_k$, the coefficient of $\epsilon$ is

$$
W^{J_k}_{i,\partial J_k} c + \lambda + \epsilon c + 1 + \{\beta(\lambda \lor 1) \lor 1\} |J_k| + c \epsilon |J_{k-1} \cap J_k|
$$

$$
\leq c + \lambda + \epsilon c + 1 + \{\beta(\lambda \lor 1) \lor 1\} |J_k| + c \epsilon |J_{k-1} \cap J_k| \leq 2c.
$$

**S2. Additional Proofs**

**S2.1. Proof of Lemma 3**

**Lemma S3.** For each $i \in 1, \ldots, t$, let $P_i(x_{i-1}, \dd x_i)$ be a Markov transition kernel on $X$. For some integer $h > 0$, assume that the composite transition kernels $Q_i(x_{(i-1)h}, \dd x_{ih}) = \int x_{ih} P_i(x_{i-1}, \dd x_i) Q(x_{(i-1)h}, x_{ih}) \dd x_i$ put enough mass on $x_{ih}$.
\( (P_{(i-1)h+1} \cdots P_{ih})(x_{(i-1)h}, dx_{ih}) \) satisfy the following minorization condition: there exists probability measures \( \nu_i(dx) \) on \( X \) and a common constant \( \alpha \in [0, 1] \) such that \( Q_i \geq \alpha \nu_i \).

Consider the probability measures \( \prod_{i=1}^{t} P(x_{i-1}, dx_i) \) and \( \prod_{i=1}^{t} P(y_{i-1}, dy_i) \) on the product space \( X^t \). For any \( x_0 \) and \( y_0 \), there exists a coupling \( \Psi \) such that if \( (X_{1:t}, Y_{1:t}) \sim \Psi \) then \( \text{pr}(X_t \neq Y_t) \leq (1 - \alpha)^{\lfloor t/h \rfloor} \).

**Proof.** When \( h = 1 \) the results follows standard arguments (7) and is repeated here for the sake of completeness. The coupling \( \Psi \) attempts to couple \( (X_1, Y_1) \), followed by \( (X_2, Y_2) \) etc. Specifically, if \( X_{i-1} = Y_{i-1} \), then draw \( X_i \) from \( P_1(X_{i-1}, dx_i) \) and set \( Y_i = X_i \). If \( X_{i-1} \neq Y_{i-1} \), draw \( (X_i, Y_i) \) from the distribution

\[
\alpha \nu(dx) \delta_x(dx_y) + (1 - \alpha)^{-1} (P_1(X_{i-1}, dx_i) - \alpha \nu(dx_i)) (P_1(Y_{i-1}, dy_i) - \alpha \nu(dy_i))
\]

where \( \delta_x \) is the atom measure. It now follows that \( \text{pr}(X_t \neq Y_t) \leq (1 - \alpha) \text{pr}(X_{i-1} \neq Y_{i-1}) \leq (1 - \alpha)^{\lfloor t/h \rfloor} \).

For \( h > 1 \) we couple the skeleton process \( \{(X_h, Y_h), (X_{2h}, Y_{2h}), \ldots \} \) using a similar scheme as for \( h = 1 \). The \( h \)-skeleton \( X \)-system has transition kernels \( Q_i \) that satisfy a minorization condition which is to be used in the same manner as in the proof for \( h = 1 \). Let \( i \) be the first instance that \( X_{ih} = Y_{ih} \), i.e. \( X_{ih} \neq Y_{ih} \) for \( j < i \). Then, simulate the future \( X \)-process, \( X_k \) \((k = ih + 1, ih + 2, \ldots)\) from \( P_k \) and set \( Y_k = X_k \). The non-skeleton terms \( X_k \) and \( Y_k \) for \( k < ih \) are simulated independently from their respective conditional laws. Thus for \( i \geq h \), \( \text{pr}(X_t \neq Y_t) \leq \text{pr}(X_{ih} \neq Y_{ih}) \) which is bounded above by \( (1 - \alpha)^{\lfloor t/h \rfloor} \). \( \square \)

**Proof.** (Lemma 3) For notational brevity we write \( dx' \) instead of \( \nu_1(dx) \), where \( \nu_1(dx') \) is the dominating finite measure of the transition probability density \( m(x, x') \). For integers \( s < u \), integration with respect to the product measure \( \nu_1(dx_s) \cdots \nu_1(dx_u) \) is similarly abbreviated to \( dx_s \cdots dx_u \), or even more tersely as \( dx_{su} \). Recall that Assumption 2 stipulates all blocks \( J \) are intervals, let \( J = \{s, \ldots, u\} \). To compute \( W_{i,j} \) for \( i \in J \) we use a coupling as in (A1):

\[
W_{i,j} = \sup_{\{x,z \in X^u: x = z\} \cup \{y \neq z\}} \int_{x=x_0}^{x=x_n} \Psi_{j,i,x,z}(dx') \|_{x' \neq z'}
\]

where for any \( j \in J \) and \( x, z \in X^n \) such that \( x = z \), \( \Psi_{j,i,x,z} \) is a coupling of \( \phi_{x}^j \) and \( \phi_{z}^j \). We know from Lemma 2 that we only need to consider the cases \( j = s - 1 \) and \( j = u + 1 \). Consider first \( j = s - 1 \) assuming \( s > 1 \).

Write the density of \( \phi_{x}^j \) as \( p(x_s, \ldots, x_u | x_{s-1}, x_{u+1}, y_{s:u}) = \prod_{i=s}^{u} p(x_i | x_{i-1}, x_{u+1}, y_{s:i}) \), which is a product of inhomogeneous Markov transition kernels. A similar expression follows if written backwards, which is \( \prod_{i=1}^{u-1} p(x_i | x_{i+1}, y_{s:i}) \).

Using integer \( h \) defined in Assumption 1, we show that the composite kernel formed by any \( h \) successive kernels of the given inhomogeneous product, \( \int_{x_t+1}^{x_t+h} p(x_t+1, y_{t+h+1} | x_{t+1}, y_{t+h+1}) dx_{t+1} \cdots h^{-1} \), satisfies a minorization condition with respect to some probability measure \( \delta^{(1-h)/h} \sigma / \sigma_z \). Thus the coupling \( \Psi_{j,i,x,z} \) can be defined as in Lemma S3 to complete the proof.

Let \( t \leq u - h \) be some time index and consider

\[
p(x_t+h | x_t, x_{u+1}, y_{t+1} : u) = \frac{p(x_{u+1} | x_t+h, y_{t+h+1}) p(x_t+h | x_t, y_{t+h+1})}{\int p(x_{u+1} | x_t+h, y_{t+h+1}) p(x_t+h | x_t, y_{t+h+1}) dx_{t+h}}.
\]
We have
\[
p(x_{t+h} \mid x_t, y_{t+1:t+h-1}) = \frac{\int \prod_{j=t+1}^{t+h-1} \{g(x_j, y_j)m(x_{j-1}, x_j)\} m(x_{t+h-1}, x_{t+h}) \, dx_{t+1:t+h-1}}{\int \prod_{j=t+1}^{t+h-1} \{g(x_j, y_j)m(x_{j-1}, x_j)\} m(x_{t+h-1}, x_{t+h}) \, dx_{t+1:t+h}}
\]
which implies, using Assumption 1,
\[
\delta^{1+h}_- \sigma_- \leq p(x_{t+h} \mid x_t, y_{t+1:t+h-1}) \leq \sigma_+.
\]
Plugging these two bounds into the numerator and denominator, respectively, of (S7) gives
\[
p(x_{t+h} \mid x_t, y_{t+1:t+h-1}) \geq \delta^{1+h}_- \frac{\sigma_-}{\sigma_+} \frac{p(x_{u+1}, y_{t+h:u} \mid x_{t+h})}{\int p(x_{u+1}, y_{t+h:u} \mid x_{t+h}) \, dx_{t+h}}.
\]
The proof for \( j = u + 1 \) follows analogously by considering a backward decomposition of \( p(x_t \mid x_{s-1}, x_{u+1}, y_{s:t-1}) \) and minorizing \( p(x_{t-h} \mid x_t, x_{s-1}, y_{s:t-1}) \).

S2.2. Proof of Proposition A1

The first part of the proof concerns the stated uniform minorization of the blocked particle Gibbs kernel. We consider the blocked particle Gibbs kernel \( Q^J_N \) for block \( J = \{s, \ldots, u\} \). The following proof is a slight modification of the proof in Lindsten et al. (2015, Proposition 5), to take into account conditioning on the end-point \( x_{u+1} \) in the definition of \( Q^J_N \). It is known from Lindsten et al. (2015, Theorem 1) that, for any \( x_{J+} \in X^{J+1} \) and measurable \( A \subset X^J \),
\[
Q^J_N(x_{J+}, A) \geq (1 - \epsilon(N, |J|)) \phi^J_N(A),
\]
where
\[
\epsilon(N, |J|) = 1 - \prod_{t=s}^{u} \frac{N - 1}{2B^J_t + N - 2};
\]
\[
B^J_t = \left( \sup_x \frac{p(y_{t:u}, x_{u+1} \mid x_t)}{p(y_{t:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1})} \right) \vee \max_{0 \leq \ell < u-t} \left( \sup_x \frac{p(y_{t+\ell} \mid x_t)}{p(y_{t+\ell} \mid x_{s-1}, y_{s:t-1})} \right).
\]
Using Assumption 1 and from Lindsten et al. (2015, Proposition 5), it follows that the second term in the definition of \( B^J_t \) is bounded by \( \delta \sigma_+ / \sigma_- \). We now bound the first term which is dependent on the boundary point \( x_{u+1} \) unlike the result in Lindsten et al. (2015, Proposition 5).

Consider first \( t \leq u - h + 1 \). We have for the numerator
\[
p(y_{t:u}, x_{u+1} \mid x_t) = \int \prod_{j=t}^{t+h-1} \{g(x_j, y_j)m(x_{j-1}, x_j)\} \prod_{j=t+h}^{u} \{g(x_j, y_j)m(x_{j-1}, x_j)\} \, dx_{t+1:u}
\]
\[
\leq \prod_{j=t}^{t+h-1} \sup_x g(x, y_j) \int \prod_{j=t}^{t+h-1} m(x_{j-1}, x_j) \prod_{j=t+h}^{u} \{g(x_j, y_j)m(x_{j-1}, x_j)\} \, dx_{t+1:u}
\]
\[
\leq \sigma_+ \prod_{j=t}^{t+h-1} \sup_x g(x, y_j) \int \prod_{j=t+h}^{u} \{g(x_j, y_j)m(x_{j-1}, x_j)\} \, dx_{t+1:u}.
\]
Analogously we get for the denominator
\[
p(y_{t:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1}) = \int p(y_{t:u}, x_{u+1} \mid x_t)p(x_t \mid x_{s-1}, y_{s:t-1}) \, dx_t
\]
\[
\geq \sigma_- \prod_{j=t}^{t+h-1} \inf_x g(x, y_j) \int \prod_{j=t+h}^{u} \{g(x_j, y_j)m(x_{j-1}, x_j)\} \, dx_{t+1:u}.
\]
It follows that for $t \leq u-h+1$,
\[
\sup_{x_t} \frac{p(y_{t:u}, x_{u+1} \mid x_t)}{p(y_{t:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1})} \leq \frac{\delta \sigma_+}{\sigma_-}
\]

Next, consider the case $t > u-h+1$. We have for the numerator
\[
p(y_{t:u}, x_{u+1} \mid x_t) = \int \prod_{j=t}^{u} \{g(x_j, y_j)m(x_j, x_{j+1})\} \, dx_{t+1:u}
\]
\[
\leq \int \prod_{j=t}^{u} \left\{ \sup_x g(x, y_j)m(x_j, x_{j+1}) \right\} \, dx_{t+1:u} \leq \sigma_+ \prod_{j=t}^{u} \left\{ \sup_x g(x, y_j) \right\}.
\]

For the denominator we write
\[
p(y_{t:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1}) = \frac{p(y_{u-h+1:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1})}{p(y_{u-h+1:t-1} \mid x_{s-1}, y_{s:t-1})}.
\]

We have
\[
p(y_{u-h+1:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1})
\]
\[
= \int \prod_{j=u-h+1}^{u} \{g(x_j, y_j)m(x_j, x_{j+1})\} \, p(x_{u-h+1} \mid x_{s-1}, y_{s:t-1}) \, dx_{u-h+1:u}
\]
\[
\geq \int \prod_{j=u-h+1}^{u} \left\{ \inf_x g(x, y_j)m(x_j, x_{j+1}) \right\} \, p(x_{u-h+1} \mid x_{s-1}, y_{s:t-1}) \, dx_{u-h+1:u}
\]
\[
\geq \sigma_- \prod_{j=u-h+1}^{u} \left\{ \inf_x g(x, y_j) \right\},
\]

and
\[
p(y_{u-h+1:t-1} \mid x_{s-1}, y_{s:t-1})
\]
\[
= \int \prod_{j=u-h+1}^{t-1} \{g(x_j, y_j)m(x_j, x_{j+1})\} \, p(x_{u-h+1} \mid x_{s-1}, y_{s:t-1}) \, dx_{u-h+1:t}
\]
\[
\leq \sigma_+ \prod_{j=u-h+1}^{t-1} \left\{ \sup_x g(x, y_j) \right\}.
\]

In summary we get
\[
\sup_{x_t} \frac{p(y_{t:u}, x_{u+1} \mid x_t)}{p(y_{t:u}, x_{u+1} \mid x_{s-1}, y_{s:t-1})} \leq \frac{\delta \sigma_+}{\sigma_-}
\]

also for $t > u-h+1$. Thus $B^J_t \leq \delta \sigma_+ / \sigma_-$ for all $t \in J$ and the result follows.

The second part of the proof concerns the given Wasserstein estimate for the blocked particle Gibbs kernel. It is a corollary of the following lemma.

**Lemma S4.** Let $P$ and $Q$ be two Markov kernels on $X^n$ and assume that there exists a constant $\epsilon \in [0, 1]$ such that $Q(x, dy) \geq (1-\epsilon)P(x, dy)$ for all $x \in X^n$. For $x, y \in X^n$ such that $x_{-j} = y_{-j}$, let $\Psi_{j,x,y}$ be a coupling of $P(x, \cdot)$ and $P(y, \cdot)$ and let $W$, a Wasserstein matrix for $P$, 

Thus, we can write

\[ W_{i,j} = \sup_{x,y \in X^n, x_j = y_j} \int \Psi_{j,x,y}(dx', dy') \| x_i' \neq y_i' \].

Then, the matrix \( \hat{W} \) with elements \( \hat{W}_{i,j} = W_{i,j} + \epsilon \) (\( i, j = 1, \ldots, n \)) is a Wasserstein matrix for \( Q \).

**Proof.** For \( \epsilon = 0 \) the result is immediate. Hence, consider \( \epsilon > 0 \). For any \( x, y \in X^n \), define \( r_x = \epsilon^{-1} \{ Q(x, \cdot) - (1 - \epsilon)P(x, \cdot) \} \) and let \( R_{x,y} \) be a coupling of \( r_x \) and \( r_y \). It follows that, for any \( x, y \in X^n \) with \( x_j = y_j \), \( (1 - \epsilon)\Psi_{j,x,y} + \epsilon R_{x,y} \) is a coupling of \( Q(x, \cdot) \) and \( Q(y, \cdot) \). We have,

\[ (1 - \epsilon) \int \Psi_{j,x,y}(dx', dy') \| x_i' \neq y_i' \] + \( \epsilon \int R_{x,y}(dx', dy') \| x_i' \neq y_i' \) \leq (1 - \epsilon)W_{i,j} + \epsilon \leq \hat{W}_{i,j}.

**S3. Details on Implementation and Additional Numerical Results**

**S3.1. Model Setup**

This section contains further details on the model studied in Section 5 of the main manuscript. Matlab code for the numerical implementation is available at https://github.com/freli005/block-pgibbs. We use the same model setup as Godsill et al. (2004). The signal \( (Z_t : t \in \mathbb{N}_+) \) is modelled as a \( P \)th 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 \( a_t = (a_{t,1}, \ldots, a_{t,P})^T \) and log-standard deviation \( \xi_t \). The model is parameterized using partial correlation coefficients \( \rho_t \in \mathbb{R}^P \) (Friedlander, 1982). Transformation between \( a_t \) and \( \rho_t \) is done via a standard Levinson recursion. We put a truncated Gaussian first-order autoregressive prior on \( \rho_t \), i.e.,

\[ p(\rho_t | \rho_{t-1}) \propto \begin{cases} \mathcal{N}(\rho_t | \theta \rho_{t-1}, \sigma^2 I), & \max_j |\rho_{t,j}| < 1, \\ 0, & \text{otherwise}. \end{cases} \]

For later reference we introduce a more explicit notation for this truncated Gaussian prior by defining

\[ f(\rho_t | \rho_{t-1}) = \mathcal{N}(\rho_t | \theta \rho_{t-1}, \sigma^2 I) = \prod_{j=1}^{P} \mathcal{N}(\rho_{t,j} | \theta \rho_{t-1,j}, \sigma^2), \]

\[ I(\rho_t) = \mathbb{I}_{\max_j (|\rho_{t,j}| < 1)} = \prod_{j=1}^{P} \mathbb{I}_{(|\rho_{t,j}| < 1)}, \]

\[ F(\rho_{t-1}) = \int I(\rho_t)f(\rho_t | \rho_{t-1})d\rho_t = \prod_{j=1}^{P} \int_{-1}^{1} \mathcal{N}(\rho_{t,j} | \theta \rho_{t-1,j}, \sigma^2) d\rho_{t,j}. \]

Thus, we can write

\[ p(\rho_t | \rho_{t-1}) = \frac{I(\rho_t)f(\rho_t | \rho_{t-1})}{F(\rho_{t-1})}. \]

The log-standard deviation follows a Gaussian first-order autoregressive process,

\[ p(\xi_t | \xi_{t-1}) = \mathcal{N}(\xi_t | \eta \xi_{t-1}, \sigma^2). \]
The signal $Z_t$ is observed in white Gaussian noise, $Y_t = Z_t + V_t$, where $V_t \sim \mathcal{N}(0, \sigma_v^2)$.

The latent process is $X_t = (Z_t, \rho_t, \xi_t)^T \in \mathbb{R}^{P+2}$, which is $P$th order Markov. The initial state is modelled as,

$$p(z_{-P+1:0}) = \mathcal{N}(z_{-P+1:0} \mid 0, 0I),$$

$$p(\rho_0) \propto \mathcal{N}(\rho_0 \mid 0, 0I), \max_j \{||\rho_{0,j}||\} < 1,$$

$$p(\xi_0) = \mathcal{N} \left(0, \frac{\sigma_2^2}{1 - \eta^2} \right),$$

and for brevity we define $X_0 = (Z_{-P+1:0}, \rho_0, \xi_0)^T \in \mathbb{R}^{2P+1}$.

We use fixed parameter values in the simulation, which are assumed to be known: $P = 4$, $\eta = 0.99$, $\theta = 1$, $\sigma_\rho = 0.01$, $\sigma_\xi = 0.001$, $\sigma_v = 0.02$. Based on a sequence of observations $Y_{1:n} = y_{1:n}$ the task is to simulate from the joint smoothing distribution $p(x_{0:n} \mid y_{1:n})$.

S3.2. Particle filter implementation for non-block samplers

The conditional particle filters used in the particle Gibbs samplers, with and without ancestor sampling, use the same proposal as suggested by Godsill et al. (2004); the variables $\rho_t$ and $\xi_t$ are simulated from their respective priors, whereas the variable $Z_t$ is simulated from its full conditional distribution. Specifically, the particle filter proposal is

$$r_t(x_t \mid x_{t-1:t-P}) = p(z_t \mid z_{t-1:t-P}, \rho_t, \xi_t, y_t)p(\rho_t \mid \rho_{t-1})p(\xi_t \mid \xi_{t-1}),$$

where $p(z_t \mid z_{t-1:t-P}, \rho_t, \xi_t, y_t) = \mathcal{N}(z_t \mid \hat{z}_t, 1/\lambda_t)$ and

$$\hat{z}_t = \frac{1}{\lambda_t} \left( \frac{y_t}{\sigma_v^2} + a_t^T \frac{z_{t-P:t-1}}{\exp(2\xi_t)} \right), \quad \lambda_t = \frac{1}{\sigma_v^2} + \frac{1}{\exp(2\xi_t)}.$$

It follows that the unnormalized importance weights of the particle filter are given by

$$w_t = \mathcal{N} \left(y_t \mid a_t^T \frac{z_{t-P:t-1}}{\exp(2\xi_t)}, \sigma_v^2 + \exp(2\xi_t) \right).$$

For particle Gibbs with ancestor sampling we furthermore need to derive the expression for the so called ancestor sampling probabilities. At time $t + 1$, these depend on the ratio of unnormalized target densities,

$$\frac{p(x_{0:n}, y_{1:n})}{p(x_{0:t}, y_{1:t})} = p(x_{t+1:n}, y_{t+1:n} \mid x_{0:t});$$

see Lindsten et al. (2014, Section 3.2). For the model under study we have

$$p(x_{t+1:n}, y_{t+1:n} \mid x_{0:t}) \propto \left\{ \frac{\prod_{s=t+1}^{t+P} p(z_s \mid z_{s-P:s-1}, \rho_s, \xi_s)}{p(\rho_{t+1} \mid \rho_t) p(\xi_{t+1} \mid \xi_t)} \right\},$$

where the proportionality is with respect to $x_{0:t}$. In the expression above all the factors are Gaussian, except $p(\rho_{t+1} \mid \rho_t)$ which is a truncated Gaussian, see (S11). Evaluating the expression above point-wise is sufficient to compute the ancestor sampling probabilities as described by Lindsten et al. (2014, Section 3.2).

S3.3. Particle filter implementation for block samplers

For the proposed block samplers, the particle Gibbs kernel is targeting a conditional distribution on the form $p(x_{s:u} \mid x_{s-1}, x_{u+1}, y_{s:u})$ for some block $J = \{s, \ldots, u\}$, with obvious modifi-
cations for non-internal blocks. This means that we can allow the underlying conditional particle filter to depend on the fixed boundary states $X_{s-1}$ and $X_{u+1}$.

The left boundary state, $X_{s-1}$, is naturally used as a fixed initial state of the conditional particle filter. The right boundary state, $X_{u+1}$, can however also be used to adapt the proposal distributions of the conditional particle filter. In the numerical implementation we have exploited this fact by constructing a type of bridge proposal for the variables $\xi_t$ and $\rho_t$, by conditioning on the fixed right boundary point.

**Remark 1.** Since the latent process $(X_t : t \in \mathbb{N})$ for the model studied here is $P$th order Markov, we actually need to take the lag-$P$ dependencies into account and target the conditional distribution $p(x_{s:t} \mid x_{s-P:s-1}, x_{u+1:t+P}, y_{s:t})$. This does not, however, affect the suggested proposals for $\xi_t$ and $\rho_t$ as these marginal processes are a priori Markov of order one. The autoregressive nature of $(X_t : t \in \mathbb{N})$ does however affect the weight expression at time $t = u$, as presented below.

For $\xi_t$, which has a Gaussian autoregressive prior, we can exactly compute $p(\xi_t \mid \xi_{t-1}, \xi_{u+1})$ and use this as a proposal at time $t$, instead of simply using the prior $p(\xi_t \mid \xi_{t-1})$. For $\rho_t$ it is not possible to exactly condition on the right boundary point due to the truncation to $(-1, 1)$. However, by simply neglecting this truncation it is still possible to construct an efficient bridge proposal. Recall from (S9) that $f(\rho_t \mid \rho_{t-1}) = \mathcal{N}(\rho_t \mid \theta \rho_{t-1}, \sigma_\rho^2 I)$ denotes a non-truncated autoregressive prior for $\rho_t$. Similarly, let $f(\rho_t \mid \rho_{t-1}, \rho_{u+1})$ denote the Gaussian distribution corresponding to conditioning this process on the right boundary point $\rho_{u+1}$. Furthermore, let $\tilde{F}(\rho_{t-1}, \rho_{u+1}) = \int I(\rho_t) f(\rho_t \mid \rho_{t-1}, \rho_{u+1}) d\rho_t$ denote the probability of hitting the set $(-1, 1)^P$ under this conditional Gaussian distribution, analogously to (S10). We then use

$$r_t(\rho_t \mid \rho_{t-1}, \rho_{u+1}) = \frac{I(\rho_t) f(\rho_t \mid \rho_{t-1}, \rho_{u+1})}{\tilde{F}(\rho_{t-1}, \rho_{u+1})}$$

as particle filter proposal for the variable $\rho_t$ at time $t$. In summary, the full proposal is given by

$$r_t(x_t \mid x_{t-1:t-P}, x_{u+1}) = p(z_t \mid z_{t-1:t-P}, \rho_t, \xi_t, y_t) \frac{I(\rho_t) f(\rho_t \mid \rho_{t-1}, \rho_{u+1})}{\tilde{F}(\rho_{t-1}, \rho_{u+1})} p(\xi_t \mid \xi_{t-1}, \xi_{u+1}).$$

Furthermore, by using the importance weights

$$w_t = \frac{\tilde{F}(\rho_{t-1}, \rho_{u+1})}{\tilde{F}(\rho_{t-1})} \mathcal{N}(y_t \mid a_T z_{t-P:t-1}, \sigma^2_v + \exp(2\xi_t)), \quad (t = s, \ldots, u - 1),$$

$$w_u = \frac{\tilde{F}(\rho_{u+1})}{\tilde{F}(\rho_u)} \mathcal{N}(y_u \mid a_T z_{u-P:u-1}, \sigma^2_v + \exp(2\xi_u)) \prod_{j=1}^P \mathcal{N}(z_{u+j} \mid a_{u+j} z_{u+j-P:u+j-1}, \exp(2\xi_{u+j})),$$

it can be shown that the resulting particle filter targets a sequence of distributions, for $t = s, \ldots, u$, adapted to the fixed boundary points, which coincides with $p(x_{s:t} \mid x_{s-P:s-1}, x_{u+1:t+P}, y_{s:t})$ at time $t = u$. Thus, the particle filter proposals and weight expressions presented above can be used to construct a valid particle Gibbs kernel for the block of states $X_{s:t}$.

### S3.4. Numerical results

We consider two simulated data sets, one with $n = 1000$ and one with $n = 2000$. Various methods are applied to sample from $p(x_{0:n} \mid y_{1:n})$; (i) the standard particle Gibbs sampler by
Andrieu et al. (2010), (ii) particle Gibbs with ancestor sampling by Lindsten et al. (2014), (iii) the proposed parallel block sampler, and (iv) the proposed right-to-left block sampler. Both block samplers use different block sizes $L$ and overlaps $p$: $(L, p) \in \{(10, 0), (50, 0), (50, 10)\}$. Note that the sequential block sampler considered here operates right-to-left, but this is analogous to the left-to-right block sampler considered in the theoretical analysis. We conjecture that it is better to operate right-to-left than left-to-right when using overlapping blocks and particle Gibbs kernels for updating these blocks. The reason is that the update frequency when applying the particle Gibbs kernel to a specific block will be smallest for the variables furthest to the left in the block due to path degeneracy of the particle Gibbs kernel. Thus, it may be beneficial to overwrite these variables when sampling from the next block in the sequence, suggesting a right-to-left implementation.

All methods used $N = 100$ particles in the underlying particle filters. The block samplers with overlap $p = 10$ have around 25% higher computational costs than the other methods, but apart from that all methods have similar running costs. We compute the mean squared jump distance for each state coordinate and for each time point, based on $R = 10000$ iterations of each sampler, i.e.,

$$
\text{ESJD}(x_{j,t}) = \frac{1}{R} \sum_{k=1}^{R} (x_{j,t}[k] - x_{j,t}[k-1])^2.
$$

Here, $x_{j,t}[k]$ denotes the state of variable $X_{j,t}$ of the Markov chain Monte Carlo sampler at iteration $k$, where each iteration of the sampler corresponds to one complete sweep.

In Tables 1 and 2 we report the median and lower 5% quantile, respectively, of the mean squared jump distance for each state component, computed over the $n = 1001$ time points. This is a condensed measure of performance, but we believe that it nevertheless provides a good indication of the overall mixing speed of each sampler. From these results we see that the particle Gibbs sampler mixes very poorly. In fact, most of the state variables are not updated at all during the 10000 iterations for which the sampler is run, leading to a median of zero of the mean squared jump distance. This is an effect of path degeneracy, corroborating previous results on the inefficiency of particle Gibbs sampling when $N$ is selected too small with respect to $n$ (Lindsten et al., 2015; Andrieu et al., 2017).

The remaining samplers all appear to quite well. The right-to-left block sampler and the parallel block sampler have more or less identical performance, suggesting that the parallel sampler is preferable in practice since it opens up for computational savings via a parallel implementation. We also note that the performance of the block samplers improve with block size $L$ and overlap $p$. The effect of using overlapping blocks is most clearly seen in Table 2, reporting the lower 5% quantile of the mean jump distance. Indeed, using overlapping blocks has the effect of improving the mixing of variables close to block boundaries, which are otherwise a bottle-neck for the overall mixing speed of the block sampler.

It should be noted that the performance of the block sampler will deteriorate if $L$ is taken too large with respect to $N$. Indeed, if $L = n$ we recover the standard particle Gibbs sampler which performs very poorly with $N = 100$, as discussed above. One way of tuning $L$ in practice is to monitor the path degeneracy of the particle filter and setting $L$ in order to attain a non-zero update frequency for the entire block. In fact, this type of tuning of the block size could be done online for a left-to-right scheme, resulting in an adaptive block sampler. Investigating the performance of such a method is left for future work.

Finally, we note that, in this example, the block samplers with $L = 50$ and $p = 10$ perform better than the particle Gibbs with ancestor sampling method by Lindsten et al. (2014). This
Table 1. Median, over time points $t$, of the mean squared jump distance in each coordinate for $T = 1000$

| Method | $z_1(\times 10^{-4})$ | $\rho_{1,1}(\times 10^{-3})$ | $\rho_{1,2}(\times 10^{-3})$ | $\rho_{1,3}(\times 10^{-3})$ | $\rho_{1,4}(\times 10^{-3})$ | $\xi_l(\times 10^{-3})$ |
|--------|----------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| PG     | 0                    | 0                           | 0                           | 0                           | 0                           | 0                           |
| PGAS   | 3.50                 | 0.35                        | 0.41                        | 0.42                        | 0.41                        | 0.41                        |
| RL (10,0) | 7.55       | 0.40                        | 0.41                        | 0.41                        | 0.41                        | 0.41                        |
| RL (50,0) | 6.07       | 1.08                        | 1.39                        | 1.38                        | 1.37                        | 1.39                        |
| RL (50,10) | 6.43      | 1.27                        | 1.50                        | 1.51                        | 1.51                        | 1.53                        |
| PAR (10,0) | 7.56       | 0.40                        | 0.41                        | 0.41                        | 0.41                        | 0.41                        |
| PAR (50,0) | 6.07       | 1.09                        | 1.38                        | 1.39                        | 1.39                        | 1.39                        |
| PAR (50,10) | 6.46      | 1.25                        | 1.51                        | 1.53                        | 1.52                        | 1.54                        |

Method abbreviations used in the table are PG for particle Gibbs, PGAS for particle Gibbs with ancestor sampling, RL for right-to-left block sampler, and PAR for parallel block sampler. For RL and PAR, the values in parentheses are block size and overlap, i.e., $(L, p)$.

Table 2. Lower 5% quantile, over time points $t$, of the mean squared jump distance in each coordinate for $T = 1000$

| Method | $z_1(\times 10^{-4})$ | $\rho_{1,1}(\times 10^{-3})$ | $\rho_{1,2}(\times 10^{-3})$ | $\rho_{1,3}(\times 10^{-3})$ | $\rho_{1,4}(\times 10^{-3})$ | $\xi_l(\times 10^{-3})$ |
|--------|----------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| PG     | 0                    | 0                           | 0                           | 0                           | 0                           | 0                           |
| PGAS   | 2.98                 | 0.21                        | 0.25                        | 0.26                        | 0.26                        | 0.26                        |
| RL (10,0) | 7.16       | 0.16                        | 0.16                        | 0.17                        | 0.16                        | 0.17                        |
| RL (50,0) | 4.68       | 0.22                        | 0.23                        | 0.23                        | 0.23                        | 0.23                        |
| RL (50,10) | 5.20      | 0.74                        | 1.11                        | 1.12                        | 1.12                        | 1.13                        |
| PAR (10,0) | 7.14       | 0.16                        | 0.16                        | 0.16                        | 0.16                        | 0.17                        |
| PAR (50,0) | 4.66       | 0.22                        | 0.23                        | 0.23                        | 0.23                        | 0.23                        |
| PAR (50,10) | 5.19      | 0.72                        | 1.11                        | 1.12                        | 1.11                        | 1.13                        |

Method abbreviations used in the table are PG for particle Gibbs, PGAS for particle Gibbs with ancestor sampling, RL for right-to-left block sampler, and PAR for parallel block sampler. For RL and PAR, the values in parentheses are block size and overlap, i.e., $(L, p)$.

is likely due to the fact that the block samplers use a better proposal mechanism, described in Section S3.3 above. However, it is not clear how to use this proposal mechanism for the standard particle Gibbs sampling or particle Gibbs with ancestor sampling.

To test the stability of the algorithms, the same experiment was repeated with a longer data set with $n = 2000$. The median and lower 5% quantile of the mean squared jump distance are given in Tables 3 and 4, respectively. The results for all samplers are very similar to the case $n = 1000$ indicating that all algorithms, except the standard particle Gibbs sampler, are indeed stable as $n$ becomes large. This suggests that the stability results for the block samplers given in Theorem 3 of the main manuscript may carry over to challenging models where the conditions of the theorem do not hold.

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. & VIRIOLA, M. (2017). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. Bernoulli (forthcoming).

FRIEDLANDER, B. (1982). Lattice filters for adaptive processing. *Proc. IEEE* **70**, 829–867.

GODSILL, S. J., DOUCET, A. & WEST, M. (2004). Monte Carlo smoothing for nonlinear time series. *J. Am. Statist. Assoc.* **99**, 156–168.
Table 3. Median, over time points $t$, of the mean squared jump distance in each coordinate for $T = 2000$

| Method | $z_t (\times 10^{-4})$ | $\rho_{t,1} (\times 10^{-3})$ | $\rho_{t,2} (\times 10^{-3})$ | $\rho_{t,3} (\times 10^{-3})$ | $\rho_{t,4} (\times 10^{-3})$ | $\xi_t (\times 10^{-5})$ |
|--------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|--------------------------|
| PG     | 0                      | 0                             | 0                             | 0                             | 0                             | 0                        |
| PGAS   | 3.47                   | 0.35                          | 0.43                          | 0.42                          | 0.42                          | 0.42                     |
| RL (10, 0) | 7.56             | 0.40                          | 0.41                          | 0.41                          | 0.41                          | 0.41                     |
| RL (50, 0) | 6.12             | 1.11                          | 1.37                          | 1.36                          | 1.36                          | 1.38                     |
| RL (50, 10) | 6.45            | 1.29                          | 1.51                          | 1.50                          | 1.51                          | 1.51                     |
| PAR (10, 0) | 7.56            | 0.39                          | 0.41                          | 0.41                          | 0.41                          | 0.41                     |
| PAR (50, 0) | 6.12            | 1.11                          | 1.38                          | 1.37                          | 1.37                          | 1.39                     |
| PAR (50, 10) | 6.46           | 1.30                          | 1.51                          | 1.50                          | 1.50                          | 1.53                     |

Method abbreviations used in the table are PG for particle Gibbs, PGAS for particle Gibbs with ancestor sampling, RL for right-to-left block sampler, and PAR for parallel block sampler. For RL and PAR, the values in parentheses are block size and overlap, i.e., $(L, p)$.

Table 4. Lower 5% quantile, over time points $t$, of the mean squared jump distance in each coordinate for $T = 2000$

| Method | $z_t (\times 10^{-4})$ | $\rho_{t,1} (\times 10^{-3})$ | $\rho_{t,2} (\times 10^{-3})$ | $\rho_{t,3} (\times 10^{-3})$ | $\rho_{t,4} (\times 10^{-3})$ | $\xi_t (\times 10^{-5})$ |
|--------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|--------------------------|
| PG     | 0                      | 0                             | 0                             | 0                             | 0                             | 0                        |
| PGAS   | 2.99                   | 0.22                          | 0.28                          | 0.28                          | 0.28                          | 0.27                     |
| RL (10, 0) | 7.14             | 0.16                          | 0.16                          | 0.16                          | 0.16                          | 0.17                     |
| RL (50, 0) | 4.74             | 0.22                          | 0.23                          | 0.23                          | 0.23                          | 0.23                     |
| RL (50, 10) | 5.24             | 0.75                          | 1.13                          | 1.12                          | 1.12                          | 1.13                     |
| PAR (10, 0) | 7.15            | 0.16                          | 0.16                          | 0.16                          | 0.16                          | 0.17                     |
| PAR (50, 0) | 4.76            | 0.22                          | 0.23                          | 0.23                          | 0.23                          | 0.23                     |
| PAR (50, 10) | 5.25            | 0.72                          | 1.13                          | 1.12                          | 1.12                          | 1.13                     |

Method abbreviations used in the table are PG for particle Gibbs, PGAS for particle Gibbs with ancestor sampling, RL for right-to-left block sampler, and PAR for parallel block sampler. For RL and PAR, the values in parentheses are block size and overlap, i.e., $(L, p)$.

LINDSTEN, F., DOUC, R. & MOULINES, E. (2015). Uniform ergodicity of the particle Gibbs sampler. Scand. J. Statist. 42, 775–797.

LINDSTEN, F., JORDAN, M. I. & SCHÖN, T. B. (2014). Particle Gibbs with ancestor sampling. J. Machine Learning Research 15, 2145–2184.

[Received April 2016]