



Andrieu, C. (2016). On random- and systematic-scan samplers. *Biometrika*, 103(3), 719-726. <https://doi.org/10.1093/biomet/asw019>

Peer reviewed version

Link to published version (if available):
[10.1093/biomet/asw019](https://doi.org/10.1093/biomet/asw019)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Oxford University Press at <http://biomet.oxfordjournals.org/content/103/3.toc>.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

On the random- versus systematic-scan sampler choice

BY C. ANDRIEU

School of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.

c.andrieu@bristol.ac.uk

SUMMARY

We introduce a simple time-homogeneous Markov embedding of a class of time-inhomogeneous Markov chains widely used in the context of Monte Carlo sampling algorithms, such as systematic-scan Metropolis-within-Gibbs samplers. This allows us to establish that systematic-scan samplers involving two factors are always better than their random-scan counterparts, when asymptotic variance is the criterion of interest. We also show that this embedding sheds some light on the recent result of Maire et al. (2014) and discuss the scenario involving more than two factors.

Some key words: Deterministic-scan sampler; Markov chain Monte Carlo; Metropolis-within-Gibbs algorithm; Peskun order; Random-scan sampler.

1. INTRODUCTION

It is often said that a novice Markov chain Monte Carlo user can easily embarrass an expert with apparently simple questions. One such question is the following. Let π be a probability distribution defined on some measurable space (X, \mathcal{X}) and for some $k \in \mathbb{N}^*$ let $\mathfrak{P} = \{\Pi_i : X \times \mathcal{X} \rightarrow [0, 1], i = 1, \dots, k\}$ be a family of Markov transition kernels assumed to be reversible with respect to π . Markov chain Monte Carlo methods consist of using these Markov transitions in order to simulate realizations of a Markov process $\{X_i : i \geq 0\}$ which may be used to approximate expectations of functions $f : X \rightarrow \mathbb{R}$ with respect to π , $E_\pi \{f(X)\}$ for $X \sim \pi$, with estimators of the form

$$S_M(f) = \frac{1}{M} \sum_{i=0}^{M-1} f(X_i).$$

A natural question is how to best use \mathfrak{P} in order to minimize the variability of this estimator? There are two standard approaches to constructing such Markov chains from \mathfrak{P} . The first consists of considering the homogeneous Markov chain with transition defined as a mixture of the transitions in \mathfrak{P} ,

$$P^{\text{rand}} = \frac{1}{k} \sum_{j=1}^k \Pi_j, \tag{1}$$

which corresponds to choosing one of the kernels at random at each iteration. The second consists of cycling through \mathfrak{P} in a systematic fashion, thus defining an inhomogeneous Markov chain. More precisely, for $k \in \mathbb{N}_*$ define the forward circular permutation $\sigma : \{1, \dots, k\} \rightarrow \{1, \dots, k\}$ such that $\sigma(j) = j + 1$ for $j \in \{1, \dots, k - 1\}$ and $\sigma(k) = 1$, and its powers, starting with $\sigma^0(j) = j$ for $j \in \{1, \dots, k\}$ and $\sigma^i = \sigma^{i-1} \circ \sigma$ for $i \geq 1$. Then we introduce the sequence of

Markov transition probabilities $P^{\text{syst}} = \{P_i^{\text{syst}} = \Pi_{\sigma^{i-1}(1)} : i \geq 1\}$. For simplicity in the two scenarios outlined above we use the same notation $\text{pr}_\mu(\cdot)$ and $E_\mu(\cdot)$ for the probability distribution and expectation of the two Markov chains such that $X_0 \sim \mu$, where μ is a probability distribution on $(\mathsf{X}, \mathcal{X})$. For example, for $i \geq 1$, with $\mathcal{F}_i = \sigma(X_0, \dots, X_i)$ and $A \in \mathcal{X}$,

$$\text{pr}_\mu(X_i \in A \mid \mathcal{F}_{i-1}) = \begin{cases} P^{\text{rand}}(X_{i-1}, A), \\ P_i^{\text{syst}}(X_{i-1}, A), \end{cases}$$

depending on the scenario considered. There should not be room for confusion in what follows. Similarly, letting $\text{var}_\mu(\cdot)$ be the variance operator corresponding to $E_\mu(\cdot)$, we define for $P^* \in \{P^{\text{rand}}\}$ or $P^* = P^{\text{syst}}$ and $f : \mathsf{X} \rightarrow \mathbb{R}$, the asymptotic variance

$$\text{var}(f, P^*) = \lim_{M \rightarrow \infty} \text{var}_\pi \left\{ M^{1/2} S_M(f) \right\},$$

when the limit exists. The novice's question we are interested in here is which of the two schemes one should use in order to minimize the asymptotic variance of the estimator $S_M(f)$? Despite a long interest in ordering Markov chain Monte Carlo methods in terms of this, and other, performance measures (Hastings, 1970; Peskun, 1973; Caracciolo et al., 1990; Liu et al., 1995; Tierney, 1998; Maire et al., 2014), the novice's question is, to the best of our knowledge, still unanswered. Here we provide a partial answer, showing that for $k = 2$ it is always preferable to use the deterministic update; see Theorem 1 for a precise statement. If the measure of performance is the time for convergence to equilibrium then no general conclusions can be made, as shown in Roberts & Sahu (1997) and Roberts & Rosenthal (2015), who consider the homogeneous Markov chain $\{X_{2i} : i \geq 0\}$, and the references therein. Indeed, intractable scenarios involving the Gibbs sampler, it can be established that neither of the schemes dominates the other uniformly and that the dependence structure of the targeted distribution π determines their ordering.

The main idea of our proof consists of embedding the inhomogeneous Markov chain induced by the sequence P^{syst} into an homogeneous Markov chain, of transition kernel \mathcal{T} defined on an extended space, and rewriting the asymptotic variance of the inhomogeneous chain in terms of the resolvent of an associated operator T ; see (3). This leads to a generalization of a well-known identity for the asymptotic variance of a homogeneous Markov chain. In the case $k = 2$, the homogeneous Markov chain defined by P^{rand} can be seen as being associated with the self-adjoint part of the operator T . Together with another variational representation of the asymptotic variance of Markov chains in terms of T , its self-adjoint and skew symmetric parts, this allows us to answer the novice's question when $k = 2$ in Theorem 1. Using this embedding we also present a short proof of the result of Maire et al. (2014), that allows one to compare performance of systematic-scan sampler chains in the case $k = 2$; see Theorem 2. Our proof sheds some light on the developments in Maire et al. (2014) and illustrates the difficulty encountered when trying to extend the result for $k \geq 3$; see Remark 1.

2. HOMOGENEOUS EMBEDDING

Our proofs are based on classical Hilbert space techniques. For any probability measure μ on $(\mathsf{E}, \mathcal{E})$ a family of Markov kernels $\{\Pi_i : \mathsf{E} \times \mathcal{E} \rightarrow [0, 1], i = 1, \dots, k\}$ and any function $f : \mathsf{E} \rightarrow \mathbb{R}$ let, whenever the integrals are well-defined,

$$\mu(f) = \int f(x) \mu(dx), \quad \Pi_q f(x) = \int f(y) \Pi_q(x, dy)$$

and, by induction,

$$\Pi_{\sigma^{0:k}(q)}f(x) = \int \Pi_{\sigma^{0:k-1}(q)}(x, dy) \Pi_{\sigma^k(q)}f(y), \quad (k \geq 2).$$

The latter facilitates the manipulation of terms of the form $\Pi_q \Pi_{q+1} \cdots \Pi_k \Pi_1 \Pi_2 \cdots$, associated to the systematic-scan version of the Markov chain. Consider next the spaces of square integrable and square integrable and centred functions defined respectively as

$$L^2(\mathbb{E}, \mu) = \{f : \mathbb{E} \rightarrow \mathbb{R} : \mu(f^2) < \infty\}, \quad L_0^2(\mathbb{E}, \mu) = \{f \in L^2(\mathbb{E}, \mu) : \mu(f) = 0\},$$

endowed with the inner product defined for any $f, g \in L^2(\mathbb{E}, \mu)$ as $\langle f, g \rangle_\mu = \int f(x)g(x)\mu(dx)$, and the associated norm $\|f\|_\mu = \langle f, f \rangle_\mu^{1/2}$. For $\lambda \in (0, 1)$ and $f \in L^2(\mathbb{E}, \mu)$ we introduce the quantity

$$\text{var}_\lambda(f, P^{\text{sys}}) = \|f - \mu(f)\|_\mu^2 + \frac{2}{k} \sum_{q=1}^k \sum_{s=1}^{\infty} \lambda^s \langle f - \mu(f), \Pi_{\sigma^{0:s-1}(q)}f - \mu(f) \rangle_\mu, \quad (2)$$

which, with an abuse of language, we may refer to as the asymptotic variance. This quantity is well-defined, as for $f \in L^2(\mathbb{E}, \mu)$, $|\langle f, \Pi_{\sigma^{0:s-1}(p)}f \rangle_\mu| \leq \|f\|_\mu^2 < \infty$ ($s \in \mathbb{N}$), while the limit as $\lambda \uparrow 1$ may or may not exist. We first establish an expression for the asymptotic variance of $M^{1/2}S_M(f)$, under minimal conditions, which can be informally thought of as $\lim_{\lambda \uparrow 1} \text{var}_\lambda(f, P^{\text{sys}})$. A similar expression was obtained in Greenwood et al. (1998) for the Gibbs sampler and for $k = 2$ in Maire et al. (2014) under a slightly stronger assumption.

PROPOSITION 1. *Let $f \in L_0^2(\mathbb{X}, \pi)$ and assume that for $q \in \{1, \dots, k\}$ $\sum_{s=1}^{\infty} \langle f, \Pi_{\sigma^{0:s-1}(q)}f \rangle_\pi$ exists. Then for the inhomogeneous chain defined by P^{sys} ,*

$$\lim_{M \rightarrow \infty} \text{var}_\pi \left\{ M^{1/2} S_M(f) \right\} = \|f\|_\pi^2 + \frac{2}{k} \sum_{q=1}^k \sum_{s=1}^{\infty} \langle f, \Pi_{\sigma^{0:s-1}(q)}f \rangle_\pi.$$

The proof can be found in the Supplementary Material. We now embed the inhomogeneous Markov chain of §1 into an homogeneous Markov chain, which facilitates later analysis. This allows us to find a simple expression for $\text{var}_\lambda(f, P^{\text{sys}})$ in Lemma 1 and Corollary 1 and our result is then a direct consequence of Lemma 2. Let $\mathcal{T} : \mathbb{X}^k \times \mathcal{X}^{\otimes k} \rightarrow [0, 1]$ be the Markov transition probability such that

$$\mathcal{T}\{x^{(1)}, \dots, x^{(k)}; d(y^{(1)}, y^{(2)}, \dots, y^{(k)})\} = \prod_{i=1}^k \Pi_i\{x^{(i)}, dy^{(\sigma(i))}\}$$

and $\pi^{\otimes k}\{d(x^{(1)}, \dots, x^{(k)})\} = \prod_{i=1}^k \pi(dx^{(i)})$. We then define the associated time-homogeneous Markov chain $\{(X_i^{(1)}, \dots, X_i^{(k)}) : i \geq 0\}$ such that $(X_0^{(1)}, \dots, X_0^{(k)}) \sim \pi^{\otimes k}$. The sequence $\{X_i^{(\sigma^i(1))} : i \geq 0\}$ coincides with the non-homogeneous chain defined through P^{sys} in §1, a proof is provided in Lemma 4 in the Supplementary Material, and the other embedded chains correspond to the same cycle, but started at different points of the cycle. Our observation is that the asymptotic variance of P^{sys} in (2) can be expressed in terms of the auto-covariance sequence of \mathcal{T} for a specific class of functions in $L_0^2(\mathbb{X}^k, \pi^{\otimes k})$. Indeed one can

check that for any $f \in L_0^2(\mathsf{X}, \pi)$, defining $\tilde{f}(x_1, \dots, x_k) = \sum_{q=1}^k f(x_q)$, we have

$$\langle \tilde{f}, \mathcal{T}^i \tilde{f} \rangle_{\pi^{\otimes k}} = \sum_{q=1}^k \langle f, \Pi_{\sigma^{0:i-1}(q)} f \rangle_{\pi}, \quad (i = 1, 2, \dots);$$

and the sums in (2) can be interchanged. We have found it more convenient to rewrite this in terms of an associated operator T defined on a different functional space. Letting $L^{2,k}(\mathsf{X}, \pi) = (L^2(\mathsf{X}, \pi))^k$ we define $T : L^{2,k}(\mathsf{X}, \pi) \rightarrow L^{2,k}(\mathsf{X}, \pi)$ such that for any $\varphi = (\varphi_1, \dots, \varphi_k) \in L^{2,k}(\mathsf{X}, \pi)$ with $\varphi_i : \mathsf{X} \rightarrow \mathbb{R}$ ($i = 1, \dots, k$), $T\varphi = (\Pi_1 \varphi_{\sigma(1)}, \dots, \Pi_k \varphi_{\sigma(k)})$. We let T^* denote the adjoint of T and endow the vector space $L^{2,k}(\mathsf{X}, \pi)$ with the inner product defined for any $\varphi, \psi \in L^{2,k}(\mathsf{X}, \pi)$ as $\langle \varphi, \psi \rangle = \sum_{i=1}^k \langle \varphi_i, \psi_i \rangle_{\pi}$. The following result allows us to write (2) in terms of the resolvent of T . We use the conventions that $T^0 = I$, the identity operator, and that $\Pi_{\sigma^{0:-1}(j)} = I$ ($j \in \mathbb{N}$), and define for $\lambda \in (0, 1)$ and $\varphi \in L^{2,k}(\mathsf{X}, \pi)$, $(I - \lambda T)^{-1} \varphi = \sum_{i=0}^{\infty} \lambda^i T^i \varphi$. One can show that $\|T\varphi\| \leq \|\varphi\|$, establishing that the sum is absolutely convergent, and that $I - \lambda T$ is a positive operator. For any $f \in L^2(\mathsf{X}, \pi)$ and $(x_1, \dots, x_k) \in \mathsf{X}^k$ we let $\bar{f}(x_1, \dots, x_k) = \{f(x_1), \dots, f(x_k)\} \in L^{2,k}(\mathsf{X}, \pi)$.

LEMMA 1. *Let $f \in L^2(\mathsf{X}, \pi)$. Then for any $\lambda \in (0, 1)$ we have*

$$\langle \bar{f}, (I - \lambda T)^{-1} \bar{f} \rangle = \sum_{i=0}^{\infty} \lambda^i \sum_{q=1}^k \langle f, \Pi_{\sigma^{0:i-1}(q)} f \rangle_{\pi} = \sum_{q=1}^k \sum_{i=0}^{\infty} \lambda^i \langle f, \Pi_{\sigma^{0:i-1}(q)} f \rangle_{\pi}.$$

Proof. Let $\varphi \in L^{2,k}(\mathsf{X}, \pi)$. We first establish that for any $i \geq 1$ we have for all $j \in \{1, \dots, k\}$, $[T^i \varphi]_j = \Pi_{\sigma^{0:(i-1)}(j)} \varphi_{\sigma^i(j)}$. This is clearly true for $i = 1$. If it is true for $i \geq 1$, then

$$[T^{i+1} \varphi]_j = [T^i \circ T \varphi]_j = \Pi_{\sigma^{0:(i-1)}(j)} [T \varphi]_{\sigma^i(j)} = \Pi_{\sigma^{0:(i-1)}(j)} \Pi_{\sigma^i(j)} \varphi_{\sigma^{i+1}(j)},$$

and we conclude that it is true for $i + 1$. This implies that for $i \geq 1$, $\langle \bar{f}, T^i \bar{f} \rangle = \sum_{q=1}^k \langle f, \Pi_{\sigma^{0:i-1}(q)} f \rangle_{\pi}$. Now, with our conventions, we conclude that

$$\langle \bar{f}, (I - \lambda T)^{-1} \bar{f} \rangle = \sum_{i=0}^{\infty} \lambda^i \sum_{q=1}^k \langle f, \Pi_{\sigma^{0:i-1}(q)} f \rangle_{\pi},$$

where the sum is absolutely convergent, and the claimed result follows. \square

COROLLARY 1. *For any $f \in L_0^2(\mathsf{X}, \pi)$ one can write the asymptotic variance as*

$$\text{var}_{\lambda}(f, P^{\text{synt}}) = \frac{2}{k} \langle \bar{f}, (I - \lambda T)^{-1} \bar{f} \rangle - \|f\|_{\pi}^2, \quad (3)$$

which generalizes the expression for the asymptotic variance in the homogeneous case ($k = 1$) in terms of the resolvent.

We have the following general result, which leads to a powerful variational representation of the asymptotic variance associated to general Markov transition probabilities. We have traced this result back at least to Landim et al. (2004, proof of Lemma 3.1) and provide the reader with a proof in the Supplementary Material.

LEMMA 2. *Let \mathcal{H} be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$. Let $\Xi : \mathcal{H} \rightarrow \mathcal{H}$ be an operator such that for any $h \in \mathcal{H}$, $\|\Xi h\| \leq \|h\|$ and define $S = (\Xi + \Xi^*)/2$*

and $A = (\Xi - \Xi^*)/2$, i.e. the self-adjoint and skew symmetric parts of Ξ . Then for any $\lambda \in (0, 1)$ and $f \in \mathcal{H}$

$$\langle f, (I - \lambda\Xi)^{-1} f \rangle = \sup_{g \in \mathcal{H}} 2\langle f, g \rangle - \langle g, (I - \lambda S)g \rangle - \lambda^2 \langle Ag, (I - \lambda S)^{-1} Ag \rangle,$$

where the supremum is attained for $\hat{g} = (I - \lambda\Xi^*)^{-1}(I - \lambda S)(I - \lambda\Xi)^{-1} f$.

COROLLARY 2. As a consequence of Lemma 2, using first the characterization of the supremum and then remarking that the lemma also implies that $\langle f, (I - \lambda S)^{-1} f \rangle = \sup_{g \in \mathcal{H}} 2\langle f, g \rangle - \langle g, (I - \lambda S)g \rangle$,

$$\langle f, (I - \lambda\Xi)^{-1} f \rangle \leq \langle f, (I - \lambda S)^{-1} f \rangle - \lambda^2 \langle A\hat{g}, (I - \lambda S)^{-1} A\hat{g} \rangle \leq \langle f, (I - \lambda S)^{-1} f \rangle.$$

Now we consider a direct application of this result which leads to Theorem 1.

THEOREM 1. Let $k = 2$. For any $f \in L_0^2(\mathcal{X}, \pi)$ and $\lambda \in [0, 1)$,

$$\text{var}_\lambda(f, P^{\text{rand}}) \geq \text{var}_\lambda(f, P^{\text{syst}}). \quad (4)$$

If in addition $f \in L_0^2(\mathcal{X}, \pi)$ satisfies

$$\sum_{q=1}^k \sum_{s=1}^{\infty} |\langle f, \Pi_{\sigma^{0:s-1}(q)} f \rangle_\pi| < \infty, \quad (5)$$

then $\text{var}(f, P^{\text{rand}}) \geq \text{var}(f, P^{\text{syst}})$.

Proof. Let $S = (T + T^*)/2$ and $A = (T - T^*)/2$ denote the self-adjoint and skew-symmetric parts of T . Notice that $T\varphi = (\Pi_1\varphi_2, \Pi_2\varphi_1)$ and $T^*\varphi = (\Pi_2\varphi_2, \Pi_1\varphi_1)$ where the second statement can be established using Lemma 6 in the Supplementary Material. Therefore

$$\frac{1}{2}(T + T^*)\varphi = \left\{ \frac{1}{2}(\Pi_1 + \Pi_2)\varphi_2, \frac{1}{2}(\Pi_1 + \Pi_2)\varphi_1 \right\},$$

which corresponds to two homogeneous chains run in parallel, each with transition probability P^{rand} . We now apply Corollary 2 to T and obtain

$$\langle \bar{f}, (I - \lambda T)^{-1} \bar{f} \rangle - \|f\|_\pi^2 \leq \langle \bar{f}, (I - \lambda S)^{-1} \bar{f} \rangle - \|f\|_\pi^2$$

and remark that

$$\langle \bar{f}, (I - \lambda S)^{-1} \bar{f} \rangle = \sup_{g \in L^{2,k}(\mathcal{X}, \pi)} 2\langle \bar{f}, g \rangle - \langle g, (I - \lambda S)g \rangle = 2\langle f, (I - \lambda P^{\text{rand}})^{-1} f \rangle_\pi$$

where we have used that $2\langle \bar{f}, g \rangle - \langle g, (I - \lambda S)g \rangle = \sum_{i=1}^2 \langle f, g_i \rangle_\pi - \langle g_i, (I - \lambda P^{\text{rand}})g_i \rangle_\pi$ and Lemma 5. Therefore

$$\langle \bar{f}, (I - \lambda T)^{-1} \bar{f} \rangle - \|f\|_\pi^2 \leq 2\langle f, (I - \lambda P^{\text{rand}})^{-1} f \rangle_\pi - \|f\|_\pi^2$$

and the first statement follows from the expression for the asymptotic variance in Corollary 1. For the second statement, since P^{rand} is self-adjoint, $\lim_{\lambda \uparrow 1} \text{var}_\lambda(f, P^{\text{rand}})$ exists and converges to $\text{var}(f, P^{\text{rand}})$, if finite, and the additional summability condition allows us to likewise conclude that $\lim_{\lambda \uparrow 1} \text{var}_\lambda(f, P^{\text{syst}}) = \text{var}(f, P^{\text{syst}})$. \square

For $k = 2$, cycling deterministically through \mathfrak{B} is therefore always better in terms of asymptotic variance than random scanning. A related result was previously known for the Gibbs

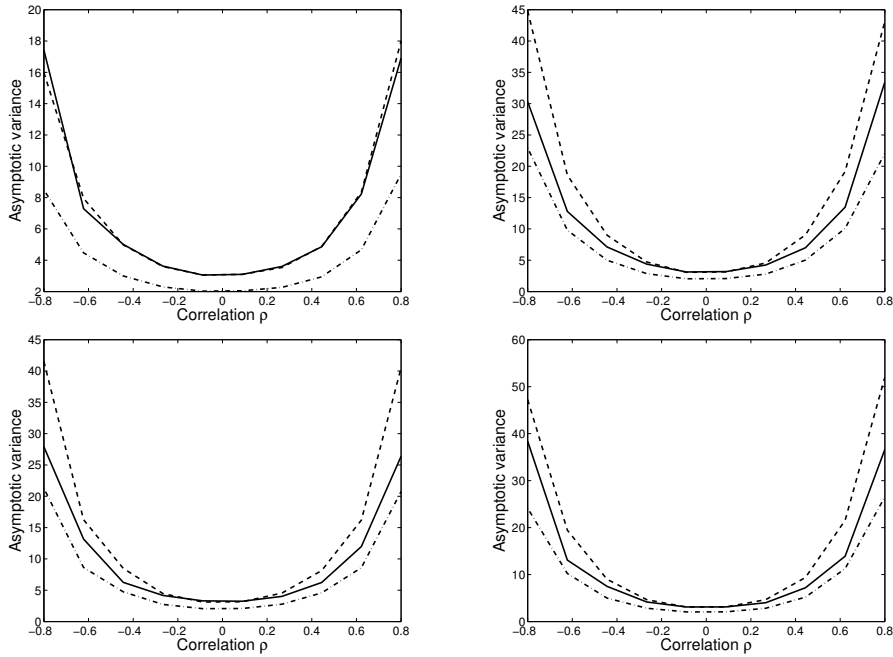


Figure 1. Estimated $\text{var}(f, P^{\text{rand}})$ (solid), $\text{var}(f, P^{\text{syst}})$ (dot-dash) and $2\text{var}(f, P^{\text{syst}}) - \text{var}_{\pi}(f)$ (dashes) for $f(x) = x$ as a function of the target distribution correlation ρ , for the Gibbs sampler (top-left) and our Metropolis-within-Gibbs with parameter $\sigma = 1 \cdot 5$ (top-right), $\sigma = 2 \cdot 5$ (bottom-left) and $\sigma = 4 \cdot 5$ (bottom-right).

150 sampler (Greenwood et al., 1998) when both Π_1 and Π_2 are projections, a property essential in order to establish that $\text{var}(f, P^{\text{rand}}) = 2\text{var}(f, P^{\text{syst}}) - \text{var}_{\pi}(f)$. We remark that in this case the gap $\text{var}(f, P^{\text{rand}}) - \text{var}(f, P^{\text{syst}}) = \text{var}(f, P^{\text{syst}}) - \text{var}_{\pi}(f) \geq 0$ as for $p \in \mathbb{N}_*$ $\langle f, \Pi_{\sigma^{0:2p-1}(q)} f \rangle_{\pi} \geq 0$ from Rosenthal & Rosenthal (2015, Corollary 3.2) and one can check directly that $\langle f, \Pi_{\sigma^{0:2p}(q)} f \rangle_{\pi} \geq 0$, as the two operators involved are positive and self-adjoint.

155 In the more general scenario considered here, Corollary 2 provides us with a theoretical lower bound, $\langle A\hat{g}, (I - \lambda S)^{-1} A\hat{g} \rangle$ with $\hat{g} = (I - \lambda T^*)^{-1} (I - \lambda S) (I - \lambda T)^{-1} \bar{f}$, on the gap in the inequality in (4). We briefly study the impact of the dependence structure of π and the updates $\mathfrak{P} = \{\Pi_1, \Pi_2\}$ used on this gap on a simple example. The target distribution is an exchangeable bivariate normal distribution of marginal variance 1 and correlation ρ . In Figure 1 we present

160 $\text{var}(f, P^{\text{rand}})$, $\text{var}(f, P^{\text{syst}})$ and $2\text{var}(f, P^{\text{syst}}) - \text{var}_{\pi}(f)$, estimated using the remark in the proof of Greenwood et al. (1998, Proposition 2), as a function of ρ for the Gibbs sampler and a Metropolis-within-Gibbs algorithm where one of the Gibbs updates is replaced with a random walk Metropolis with normal increments of variance $(1 - \rho^2)\sigma^2$ for $\sigma \in \{1 \cdot 5, 2 \cdot 5, 4 \cdot 5\}$. While we observe larger asymptotic variances for the hybrid algorithms, the Gibbs sampler is

165 observed to have the worst relative gap.

3. THE ORDERING RESULT OF MAIRE ET AL. (2014)

Maire et al. (2014) established that for $k = 2$, Markov chains of the type P^{syst} satisfy a Peskun type result (Peskun, 1973; Tierney, 1998).

THEOREM 2 (MAIRE, DOUC AND OLSSON, 2014). *Let $k = 2$ and consider two pairs of π -reversible Markov transition probabilities $\mathfrak{P} = \{\Pi_1, \Pi_2\}$ and $\check{\mathfrak{P}} = \{\check{\Pi}_1, \check{\Pi}_2\}$. If for any $g \in L^2(\mathbb{E}, \pi)$ and $i \in \{1, 2\}$, $\langle g, (I - \check{\Pi}_i)g \rangle_\pi \leq \langle g, (I - \Pi_i)g \rangle_\pi$, then for any $f \in L^2(\mathbb{X}, \pi)$, $\text{var}_\lambda(f, P^{\text{sys}}) \leq \text{var}_\lambda(f, \check{P}^{\text{sys}})$. If in addition (5) holds for \mathfrak{P} and $\check{\mathfrak{P}}$ then $\text{var}(f, P^{\text{sys}}) \leq \text{var}(f, \check{P}^{\text{sys}})$.*

170

The proof is a direct consequence of Lemma 3 below. In order to state this key result it is useful to rewrite the operator T as the composition of elementary operators, $\mathfrak{S}^{\pm 1}, \Delta: L^{2,k}(\mathbb{X}, \pi) \rightarrow L^{2,k}(\mathbb{X}, \pi)$ where $\mathfrak{S}^{\pm 1}$ are the forward and backward circular permutation operators and Δ is the diagonal operator such that for any $\varphi \in L^{2,k}(\mathbb{X}, \pi)$

175

$$\mathfrak{S}^{\pm 1}\varphi = (\varphi_{\sigma^{\pm 1}(1)}, \dots, \varphi_{\sigma^{\pm 1}(k)}), \quad \Delta\varphi = (\Pi_1\varphi_1, \dots, \Pi_k\varphi_k).$$

Then $T = \Delta \circ \mathfrak{S}$; see Lemma 6 in the Supplementary Material. Similarly we define the operator $\check{\Delta}$ by replacing Π_1 and Π_2 with $\check{\Pi}_1$ and $\check{\Pi}_2$ respectively, and let $\check{T} = \check{\Delta} \circ \check{\mathfrak{S}}$. We may omit the composition symbol in order to alleviate notation.

180

LEMMA 3. *Let T and \check{T} be the embedding operators associated with \mathfrak{P} and $\check{\mathfrak{P}}$ given in Theorem 2, define $T(\beta) = \beta T + (1 - \beta)\check{T}$ and $\Delta(\beta) = \beta\Delta + (1 - \beta)\check{\Delta}$ for $\beta \in [0, 1]$. For $\lambda \in [0, 1]$ and $f \in L^2(\mathbb{X}, \pi)$ let $\delta_{\lambda,f}(\beta) = \langle \bar{f}, \{I - \lambda T(\beta)\}^{-1} \bar{f} \rangle$. Then*

$$\frac{\partial}{\partial \beta} \delta_{\lambda,f}(\beta) = \lambda \langle \{I - \mathfrak{S}^{-1}\Delta(\beta)\}^{-1} \bar{f}, (\Delta - \check{\Delta}) \{I - \mathfrak{S}\Delta(\beta)\}^{-1} \bar{f} \rangle.$$

Proof. Since $T(\beta) = \Delta(\beta) \circ \mathfrak{S}$,

185

$$\begin{aligned} \frac{\partial}{\partial \beta} \delta_{\lambda,f}(\beta) &= \lambda \langle \bar{f}, \{I - \lambda T(\beta)\}^{-1} (T - \check{T}) \{I - \lambda T(\beta)\}^{-1} \bar{f} \rangle \\ &= \lambda \langle \bar{f}, \{I - \lambda T(\beta)\}^{-1} (\Delta - \check{\Delta}) \mathfrak{S} \{I - \lambda T(\beta)\}^{-1} \mathfrak{S}^{-1} \bar{f} \rangle \\ &= \lambda \langle \{I - \lambda \mathfrak{S}^{-1} T(\beta) \mathfrak{S}^{-1}\}^{-1} \bar{f}, (\Delta - \check{\Delta}) \mathfrak{S} \{I - \lambda T(\beta)\}^{-1} \mathfrak{S}^{-1} \bar{f} \rangle \\ &= \lambda \langle \{I - \lambda \mathfrak{S}^{-1} T(\beta) \mathfrak{S}^{-1}\}^{-1} \bar{f}, (\Delta - \check{\Delta}) \{I - \lambda \mathfrak{S} T(\beta) \mathfrak{S}^{-1}\}^{-1} \bar{f} \rangle. \end{aligned}$$

The first equality follows from standard arguments found in Tierney (1998) and Andrieu & Vi-hola (2015, Lemma 51) where additional details are provided and by noting that reversibility of $T(\beta)$ is not a requirement in those arguments. On the second line we have used the definition of T and \check{T} in terms of $\Delta, \check{\Delta}$ and \mathfrak{S} , and $\mathfrak{S}^{\pm 1} \bar{f} = \bar{f}$. On the third line we have used that for $k \geq 1$ the adjoint of $T(\beta)^k$ is $T^*(\beta)^k = \{\mathfrak{S}^{-1} T(\beta) \mathfrak{S}^{-1}\}^k$ from Lemma 6, from which we deduce that $[\{I - \lambda T(\beta)\}^{-1}]^* = \{I - \lambda \mathfrak{S}^{-1} T(\beta) \mathfrak{S}^{-1}\}^{-1}$. On the fourth line the result follows from $T(\beta) = \Delta(\beta) \mathfrak{S}$ and the identity

190

195

$$\mathfrak{S} \{I - \lambda T(\beta)\} \mathfrak{S}^{-1} = \sum_{k=0}^{\infty} \{\lambda \mathfrak{S} T(\beta) \mathfrak{S}^{-1}\}^k = \{I - \lambda \mathfrak{S} T(\beta) \mathfrak{S}^{-1}\}^{-1}.$$

□

Proof of Theorem 2. The derivative of $\delta_{\lambda,f}(\beta)$ is evidently positive if the resolvent type terms $\{I - \mathfrak{S}\Delta(\beta)\}^{-1} \bar{f}$ and $\{I - \mathfrak{S}^{-1}\Delta(\beta)\}^{-1} \bar{f}$ coincide, since for any $\phi \in L^{2,k}(\mathbb{X}, \pi)$ $\langle \phi, (\Delta - \check{\Delta})\phi \rangle \geq 0$. This is the case for $k = 2$ since then $\mathfrak{S}^{-1} = \mathfrak{S}$, which reduces to a swap operator. □

200

Remark 1. We may wonder whether Theorem 2 applies for $k \geq 3$. To that purpose, assume that the Markov transitions in \mathfrak{P} and $\check{\mathfrak{P}}$ coincide, except for the i th element and notice that with $\Pi_j(\beta) = [\Delta(\beta)]_j$ we have $[\{I - \mathfrak{S}\Delta(\beta)\}^{-1}\bar{f}]_i = \sum_{j=0}^{\infty} \lambda^j \Pi_{\sigma^{1:j}(i)}(\beta) f$ and
 205 $[\{I - \lambda \mathfrak{S}^{-1} \Delta(\beta)\}^{-1}\bar{f}]_i = \sum_{j=0}^{\infty} \lambda^j \Pi_{\sigma^{-j:-1}(i)}(\beta) f$. Then the derivative of $\delta_{\lambda, f}(\beta)$ reduces to

$$\frac{\partial}{\partial \beta} \delta_{\lambda, f}(\beta) = \left\langle \sum_{j=0}^{\infty} \lambda^j \Pi_{\sigma^{-j:-1}(i)}(\beta) f, (\Pi_i - \check{\Pi}_i) \sum_{j=0}^{\infty} \lambda^j \Pi_{\sigma^{1:j}(i)}(\beta) f \right\rangle_{\pi},$$

which is positive if the two sums coincide, and is the case if $k = 2p - 1$ for some $p \in \mathbb{N}^*$, $i = p$ and $(\Pi_1, \dots, \Pi_k) = (Q_p, Q_{p-1}, \dots, Q_2, Q_1, Q_2, \dots, Q_p)$ for a family of transition probabilities $\{Q_i : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1], i = 1, \dots, p\}$. This corresponds to a known strategy to make the Gibbs sampler reversible, albeit for the operator $Q_1 Q_2 \cdots Q_p$.

210

ACKNOWLEDGEMENTS

The author acknowledges the support of the Engineering and Physical Sciences Research Council, U.K. The author is grateful to Balint Tóth and Anthony Lee for useful discussions, as well as the editor, an associate editor and two referees.

SUPPLEMENTARY MATERIAL

215

Supplementary material available at Biometrika online includes proofs of Proposition 1 and Lemma 2.

BIBLIOGRAPHY

- ANDRIEU, C. & VIHOLA, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25**, 1030–1077.
- 220 CARACCILO, S., PELISSETTO, A. & SOKAL, A. D. (1990). Nonlocal Monte Carlo algorithm for self-avoiding walks with fixed endpoints. *J. Stat. Phys.* **60**, 1–53.
- GREENWOOD, P. E., MCKEAGUE, I. W. & WEFELMEYER, W. (1998). Information bounds for Gibbs samplers. *Annals of Statistics*, 2128–2156.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*
 225 **57**, 97–109.
- LANDIM, C., QUASTEL, J., SALMHOFER, M. & YAU, H.-T. (2004). Superdiffusivity of asymmetric exclusion process in dimensions one and two. *Communications in Mathematical Physics* **244**, 455–481.
- LIU, J. S., WONG, W. H. & KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society: Series B (Methodological)*, 157–169.
- 230 MAIRE, F., DOUC, R. & OLSSON, J. (2014). Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Statist.* **42**, 1483–1510.
- PESKUN, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.
- ROBERTS, G. O. & ROSENTHAL, J. S. (2015). Surprising convergence properties of some simple Gibbs samplers under various scans. *International Journal of Statistics and Probability* **5**, 51.
- 235 ROBERTS, G. O. & SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 291–317.
- ROSENTHAL, J. & ROSENTHAL, P. (2015). Spectral bounds for certain two-factor non-reversible MCMC algorithms. *Electron. Commun. Probab.* **20**, no. 91, 1–10.
- TIERNEY, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8**, 1–9.