NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

ANTITHETIC COUPLING OF TWO GIBBS SAMPLER CHAINS

by

Arnoldo Frigessi, Jørund Gåsemyr and Håvard Rue

PREPRINT
STATISTICS NO. 13/1999

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
`http://www.math.ntnu.no/preprint/statistics/1999/S13-1999.ps`
Håvard Rue has homepage: `http://www.math.ntnu.no/~hrue`
E-mail: `hrue@math.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491
Trondheim, Norway.

ANTITHETIC COUPLING OF TWO GIBBS SAMPLER CHAINS

ARNOLDO FRIGESSI
NORWEGIAN COMPUTING CENTER
OSLO, NORWAY

JØRUND GÅSEMYR
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OSLO, NORWAY

HÅVARD RUE
DEPARTMENT OF MATHEMATICAL SCIENCES
NTNU, NORWAY

MARCH 25, 1999

SUMMARY

Two coupled Gibbs sampler chains, both with invariant probability density π, are run in parallel in such a way that the chains are negatively correlated. This allows us to define an asymptotically unbiased estimator of the expectation $\mathrm{E}(f(\boldsymbol{X}))$ with respect to π which achieves significant variance reduction with respect to the usual Gibbs sampler at comparable computational cost. We show that the variance of the estimator based on the new algorithm is always smaller than the variance of a single Gibbs sampler chain, if π is attractive and f is monotone non-decreasing in all components of $\boldsymbol{X}$. For non-attractive targets π, our results are not complete: The new antithetic algorithm outperforms the standard Gibbs sampler by one order of magnitude when π is a multivariate normal density or the Ising model. More generally, non-rigorous arguments and numerical experiments support the usefulness of the antithetically coupled Gibbs samplers also for other non-attractive models. In our experiments the variance is reduced to at most a third and the efficiency also improves significantly.

KEYWORDS: Antithetic Monte Carlo; Associated random variables; Attractive models; Decay of Cross-autocorrelations; Markov Chain Monte Carlo; Variance reduction.

ADDRESSES: A. Frigessi, Norwegian Computing Center, P.O.Box 114 Blindern, N-0314 Oslo. J. Gåsemyr, Department of Mathematics, P.O.Box 1053 Blindern, N-0316 Oslo. H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim.

E-MAIL: `Arnoldo.Frigessi@nr.no`, `gaasemyr@math.uio.no` and `Havard.Rue@math.ntnu.no`

ACKNOWLEDGMENTS: This project was supported by the Universitá di Roma Tre, the EU-TMR project on Spatial Statistics (ERB-FMRX-CT960095) and the ESF program on Highly Structured Stochastic Systems. We thank Dario Gasbarra who has provided us with the proof of Theorem 3.

1 INTRODUCTION

Markov Chain Monte Carlo (MCMC) algorithms allow the approximate calculation of expectations with respect to multivariate probability density functions $\pi(\boldsymbol{x})$ defined up to a normalizing constant. We refer the reader to Gilks, Richardson & Spiegelhalter (1996) as a starting point for a vast literature about MCMC methodology. The underlying idea is to construct an ergodic Markov chain with invariant density function π, whose trajectory is easy to simulate without knowing the normalizing constant of π. In order to approximate the expectation $\mathrm{E}(f(\boldsymbol{X})) < \infty$ of a function $f(\boldsymbol{x})$ with respect to π, one just needs to compute the empirical average of f along the generated trajectory $\boldsymbol{X}^1, \dots, \boldsymbol{X}^T$ of a discrete time Markov chain evolving on Ω and converging to $\pi(\boldsymbol{x}), \boldsymbol{x} \in \Omega$. In practice it is appropriate to drop an initial part of the trajectory in order to avoid strong dependence on the initial conditions. The sample mean with *burn-in* of length T_0

$$\hat{f} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} f(\boldsymbol{X}^t)$$

is used. In this paper we propose a new algorithm for the estimation of $\mathrm{E}(f(\boldsymbol{X}))$. The idea is to simulate two MCMC trajectories in parallel, both invariant with respect to π, which are coupled in such a way that variance reduction can be achieved. We use the Gibbs sampler, which is a particular MCMC scheme where at each transition a sample from an one dimensional conditional density computed from π is generated. The coupling follows the basic idea of antithetic sampling in classical Monte Carlo theory.

After the burn-in we split the simulation into two parallel Gibbs sampler chains, both ergodic with respect to π. Let us denote the two chains $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ for $t = T_0 + 1, T_0 + 2, \dots$. Marginally the two chains are ordinary Gibbs samplers, but their joint probability measure is constructed in such a way that $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{Y}^t)$ have negative covariance. We exploit such antithetic behavior in order to construct another asymptotically unbiased estimator of $\mathrm{E}(f(\boldsymbol{X}))$ with smaller variance than $\mathrm{Var}(\hat{f})$ but with similar computational complexity. The coupling is simple, based on using a common sequence of random numbers. Specifically, if $\boldsymbol{X}^t$ uses an uniform $[0, 1)$ random number U^t to proceed to $\boldsymbol{X}^{t+1}$, then $\boldsymbol{Y}^t$ uses $1 - U^t$ to proceed to $\boldsymbol{Y}^{t+1}$. This coupling is well known to reduce the variance of empirical averages of i.i.d. samples. While the basic idea is simple, a rigorous analysis of the new algorithm needs some care. A pleasant fact of the antithetically coupled Gibbs sampler algorithm is that no significant extra effort is needed to implement the algorithm. Starting from a code of the usual Gibbs sampler, the modifications required in order to implement the new algorithm are simple.

We combine the output of the two coupled chains into the asymptotically unbiased estimator

$$\hat{\hat{f}} = \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \frac{f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t)}{2}. \tag{1}$$

To make a fair comparison between the algorithm based on two coupled Gibbs samplers and the usual, single trajectory Gibbs sampler, we have to take into consideration that each iteration of the new algorithm takes twice the computing time of a single Gibbs sampler iteration. Hence we allow the single

Gibbs sampler to run for twice as many iterations as the new algorithm. This means that $\hat{\hat{f}}$ in (1) has to be compared with

$$\hat{f} = \frac{1}{2T} \sum_{t=T_0+1}^{T_0+2T} f(\boldsymbol{X}^t). \tag{2}$$

We define precisely the new algorithm in Section 2. In Section 3 we assume that $\boldsymbol{X}^{T_0}$ and $\boldsymbol{Y}^{T_0}$ are independent and π-distributed. Hence (1) and (2) are unbiased and we prove that $\mathrm{Var}(\hat{\hat{f}}) \leq \mathrm{Var}(\hat{f})$, for component-wise monotone functions f, attractive π, and for all T. Not surprisingly, the key point is the sign of the cross-autocovariances between the two coupled chains. Under the given conditions, we prove that the cross-autocovariances are all negative. Section 4 is devoted to a study of the multivariate normal density and the Ising model. These distributions are not necessarily attractive but have a certain local symmetry property. If f is linear, then as $T \to \infty$, we have $\mathrm{Var}(\hat{\hat{f}}) = \mathcal{O}(T^{-2})$, while $\mathrm{Var}(\hat{f}) = \mathcal{O}(T^{-1})$ even when π is not attractive.

In Section 5 we discuss the joint asymptotic properties of the coupled chains and the existence of an unique joint stationary measure. In Section 6 we present some heuristic arguments supporting the claim that $\mathrm{Var}(\hat{\hat{f}}) \leq \mathrm{Var}(\hat{f})$ for other non attractive targets π and give some precise results for a non-attractive example that mimics the behavior of the new algorithm. The variance reduction, defined as $\mathrm{Var}(\hat{\hat{f}})/\mathrm{Var}(\hat{f})$, seems to depend only mildly on the mixing property of the single Gibbs sampler chain; if the single Gibbs sampler chain is slowly mixing, then the joint Gibbs sampler will also be slow, however the variance reduction remains roughly the same.

In Section 7 we discuss some practical implementation issues and we test our new algorithm on two data sets: the hierarchical Poisson model (Gelfand & Smith, 1990) and the ordered normal means example (Gelfand, Hills, Racine-Poon & Smith, 1990). The experiments show that the antithetically coupled Gibbs sampler is significantly better than the standard one. The variance reduction is often larger than five and always larger than three. In practice $\boldsymbol{X}^{T_0}$ and $\boldsymbol{Y}^{T_0}$ are not π-distributed. Hence, we include bias in our comparison and find that the ratio

$$\frac{\mathrm{bias}(\hat{f})^2 + \mathrm{Var}(\hat{f})}{\mathrm{bias}(\hat{\hat{f}})^2 + \mathrm{Var}(\hat{\hat{f}})}$$

is larger than ten in our experiments. Looking beyond the Gibbs sampler, we apply the antithetic coupling to more general Metropolis-Hastings updates and show empirically that improvement can be still be achieved, although with a variance reduction of about two. The paper ends with some final comments in Section 8.

2 THE NEW ALGORITHM

Let $\Omega = S \times S \times \cdots \times S = S^n$ be the n-fold product space of a set S, which may be either discrete or continuous. For simplicity we consider two cases: $\Omega = \mathbb{R}^n$ and π is a probability density function that is absolutely continuous with respect to, say, Lebesgue measure; or, S is discrete and π is a discrete probability. Let $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$. The random scan Gibbs sampler for sampling from π is a Markov chain $\boldsymbol{X}^0, \boldsymbol{X}^1, \dots$ constructed as follows. Given $\boldsymbol{X}^t = \boldsymbol{x}^t$, one component in $\{1, 2, \dots, n\}$ is chosen uniformly at random. Denote this component by I^t. Only $X^t_{I^t}$ will be updated by sampling the new value $X^{t+1}_{I^t}$ from the conditional density

$$\pi(x_{I^t} \mid \boldsymbol{X}_{-I^t} = \boldsymbol{x}_{-I^t}), \tag{3}$$

where $\boldsymbol{x}_{-A} = \{x_i : i \notin A\}$, for $A \subset \{1, \dots, n\}$. The remaining components are left unchanged, $\boldsymbol{X}^{t+1}_{-I^t} = \boldsymbol{x}^t_{-I^t}$. We assume (3) to be strictly positive, so that the resulting Markov chain is ergodic and π-invariant. The transition of a random scan Gibbs sampler can be written as

$$\boldsymbol{X}^{t+1} = \boldsymbol{\Phi}(\boldsymbol{X}^t, I^t, U^t), \tag{4}$$

where $U^0, U^1, \dots$ is a sequence of i.i.d. random numbers, uniformly distributed in $[0, 1)$, and $I^0, I^1, \dots$ are i.i.d. random numbers uniform in $\{1, 2, \dots, n\}$, that identify the component to be updated at step $t+1$. The I^t-th component of the vector function $\boldsymbol{\Phi}$ is the inverse distribution function corresponding to the local conditional density (3),

$$\Phi_{I^t}(\boldsymbol{X}^t, I^t, U^t) = \Phi_{I^t}(\boldsymbol{X}^t_{-I^t}, U^t) = \inf\{x \in \mathbb{R} : \pi(X_{I^t} \leq x \mid \boldsymbol{X}^t_{-I^t}) = U^t\}, \tag{5}$$

where the $\inf$ is needed only if S is discrete. The other components of $\boldsymbol{\Phi}$ are identity functions

$$\Phi_j(\boldsymbol{X}^t, I^t, U^t) = X^t_j, \quad \text{for} \quad j \neq I^t. \tag{6}$$

The random number U^t is used to perform the transition from $\boldsymbol{X}^t$ to $\boldsymbol{X}^{t+1}$. We will also give results for another visitation schedule, where each component is updated in a raster scan. Then we shall adopt a similar notation using lowercase letters i^t for the site to be updated at time t, $i^t = (t-1)(\text{mod } n)+1$. Our results are valid for both random and raster scans, but the proofs are sometimes different.

We now define the companion chain. It is marginally a π-stationary Gibbs sampler with the same type of scan and transition rule as (4),

$$\boldsymbol{Y}^{t+1} = \boldsymbol{\Phi}(\boldsymbol{Y}^t, I^t, 1 - U^t), \tag{7}$$

but the common random numbers U^t and I^t couple the two chains and make $\boldsymbol{X}^{t+1}$ and $\boldsymbol{Y}^{t+1}$ dependent. We call the coupling antithetic because we use $1 - U^t$ in (7). The same component I^t is updated in both chains. Looking to the coupled chains jointly, notice that $X^{t+1}_{I^t}$ is conditionally independent of $\boldsymbol{Y}^t_{-I^t}$ given $\boldsymbol{X}^t_{-I^t}$, because of (4) and (7) and since U^t is independent of $\boldsymbol{Y}^t_{-I^t}$. The two coupled Gibbs sampler chains allow us to construct the estimator $\hat{\hat{f}}$ which we shall compare to $\hat{f}$ given in (2) in the rest of this paper.

3 Comparing Variances for Attractive Target Densities

We assume that the two chains are started at time $T_0 = 0$ in the marginal stationary distribution $\boldsymbol{X}^0 \sim \pi, \boldsymbol{Y}^0 \sim \pi$, independently, and then coupled. We shall return to this assumption later in this section and again in Section 7 where we discuss some practical issues. Then both $\hat{\hat{f}}$ and $\hat{f}$ are unbiased. Hence, to evaluate the performance of the antithetically coupled Gibbs sampler, we compare the variance of $\hat{f}$ with the variance of $\hat{\hat{f}}$ (both assumed to be finite). In comparing variances we shall need both autocovariances for the marginal chains, and cross-autocovariances for the two chains jointly. Let

$$\gamma_k = \mathrm{Cov}(f(\boldsymbol{X}^0), f(\boldsymbol{X}^k)), \qquad k = 0, 1, \ldots$$

be the marginal autocovariance at lag k of one of the two components. Because of stationarity, $\gamma_k = \mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{X}^{t+k}))$ for all t. We do not assume stationarity of the joint (bivariate) Markov chain, hence the cross-autocovariances

$$\beta(t, s) = \mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^s))$$

depend on time.

We consider a special class of target distributions π and functions f in order to prove that the variance of $\hat{\hat{f}}$ is smaller than the variance of $\hat{f}$ for all T. The target π is assumed to be attractive. Attractive models are common for instance in spatial statistics, see Møller (1999) for several examples.

DEFINITION 1 *A model π is attractive if*

$$\pi(X_i \leq x_i \mid \boldsymbol{x}_{-i}) \leq \pi(X_i \leq x_i \mid \boldsymbol{x}'_{-i}), \qquad \textit{for} \quad \boldsymbol{x}_{-i} \geq \boldsymbol{x}'_{-i},\ \forall\, \boldsymbol{x}, \boldsymbol{x}' \in \Omega, \tag{8}$$

assuming the partial ordering of Ω given by $\boldsymbol{x}_A \geq \boldsymbol{x}'_A$ if $x_i \geq x'_i$ for all $i \in A$.

We assume from now on and without loss of generality that the expected value of $f(\boldsymbol{X})$ is zero in order to simplify formulae. To be able to study the two estimators $\hat{\hat{f}}$ and $\hat{f}$, we need to restrict the space of functions f, too. Our algorithm induces antithetic dependency between $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$; we want this structure to transfer to $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{Y}^t)$ as well. For this we require $f \in \mathcal{F}$, where:

DEFINITION 2 *Let $\mathcal{F}$ be the class of non-constant functions $f : \Omega \to \mathbb{R}$ which are monotone non-decreasing in all components.*

In practice, often $f(\boldsymbol{x}) = \sum_i g_i(x_i)$ where the $g_i(\cdot)$'s are monotonic increasing functions. If the function of interest is decreasing in, say, component i, we can replace X_i with $-X_i$ to obtain a function in $\mathcal{F}$ and change π accordingly.

THEOREM 1 *Suppose $f \in \mathcal{F}$ and π is attractive. Consider the coupled Gibbs sampler chains given in (4) and (7) using a random scan or a raster scan. If $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ are independent and distributed according to π, then*

$$\mathit{Var}(\hat{f}) - \mathit{Var}(\hat{\hat{f}}) \geq 0 \tag{9}$$

for every $T > 0$.

Proofs are collected in the Appendix. The theorem is based on Lemma 1, which is interesting in itself. It states that under the same assumptions of Theorem 1, $\beta(t,s) \leq 0$, for all t and s. For the raster scan we prove (9) also under the different assumption that all components of $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ are independent, but $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ are not required to be distributed according to π. This condition is more appealing in practice. See the Appendix for details. In the next section we move to non attractive models to see if the variance of $\hat{\hat{f}}$ is still smaller than the variance of $\hat{f}$.

4 Comparing Variances for Some Non-attractive Target Densities

We first consider a multivariate normal target distribution: π is normal with mean vector zero and inverse covariance matrix $\boldsymbol{Q} = (q_{ij})$. We assume without loss of generality that the diagonal of $\boldsymbol{Q}$ consists of ones. When updating component i, the Gibbs Sampler samples from an univariate normal density with mean $-\sum_{j \neq i} q_{ij} x_j^t$ and variance 1. Note that the the off-diagonal terms in $\boldsymbol{Q}$ can be both negative and positive, allowing for non-attractive π.

THEOREM 2 *Let π be the multivariate normal density and let f be a linear function. Assume a deterministic scan for the Gibbs sampler. For T large enough, Var$(\hat{f}) \geq$ Var$(\hat{\hat{f}})$. Moreover, Var$(\hat{\hat{f}}) = \mathcal{O}(T^{-2})$ as $T \to \infty$.*

Theorem 2 is surprising because it shows that coupling two Gibbs sampler chains reduce the variance by a full order of magnitude, since for the single stationary chain $\text{Var}(\hat{f}) = \mathcal{O}(T^{-1})$. The reason is the following. As shown in the proof of Theorem 2, we have that

$$X_{i_t}^{t+1} + Y_{i_t}^{t+1} = -\sum_{j \neq i_t} q_{ij}(X_j^t + Y_j^t). \tag{10}$$

This means that in the limit as $t \to \infty$, the process $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is attracted and trapped in the set $\{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega | \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{0}\}$. Once $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ lies in this set, then $\beta(t, t+k) = -\gamma_k$. We refer to the proof in the Appendix for details.

Theorem 2 holds also for other target densities than the multivariate normal, if they satisfy the following symmetry condition: If $\pi(x_i \mid \boldsymbol{x}_{-i}) = \psi_i(x_i - \widetilde{x}_i)$, $\forall i$, where $\psi_i(\cdot)$ is symmetric around zero, and $\widetilde{x}_i$ is the median in $\pi(x_i \mid \boldsymbol{x}_{-i})$ which can be written as $\widetilde{x}_i = \boldsymbol{a}_i^T \boldsymbol{x}_{-i}$ for some vector $\boldsymbol{a}_i$. Not all models that satisfy this symmetry condition are attractive. The multivariate normal satisfies this condition because the conditional median equals the conditional mean which is linear in $\boldsymbol{x}_{-i}$, and the conditional variance does not depend on $\boldsymbol{x}_{-i}$. Another π, sometimes used for smoothing, that satisfies the symmetry conditions is

$$\pi(\boldsymbol{x}) \propto \exp(-\sum_{i,j} b_{ij}(x_i - x_j)^k),$$

where: $b_{ii} \geq 0$, the off-diagonal coefficients b_{ij} must be chosen appropriately, k is even (say 4), and x_1 is fixed. Although a different proof would be needed, we conjecture that Theorem 2 remains valid for a random scan.

We conclude this section with a second (discrete) example, the two-dimensional Ising model, for which we obtain a similar result as for the multivariate normal case. When π is the Ising model the n variables x_i are positioned on the sites of a finite squared grid, and

$$\pi(\boldsymbol{x}) = \frac{1}{Z} \exp(\delta \sum_{i \sim j} x_i x_j),$$

where $x_i \in \{-1, +1\}$, the sum is taken over the four nearest neighboring pairs and Z is the normalizing constant. The so called inverse temperature δ can be either positive, in which case the model is attractive, or negative which gives a repulsive interaction model. We shall consider our algorithm with a deterministic scan. Define the set $C = \{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega | \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{0}\}$. We observe that C is an absorbing set for the joint chain: if $(\boldsymbol{X}^t, \boldsymbol{Y}^t) \in C$ then also $(\boldsymbol{X}^s, \boldsymbol{Y}^s) \in C$ for $s > t$, because of the antithetic coupling and the form of the conditional distribution $\pi(x_i | \boldsymbol{x}_{-i})$. Furthermore, C is reachable from any initial state within one full sweep with a probability larger than $p = [\exp(-8|\delta|)/(1+\exp(-8|\delta|))]^n > 0$. Hence the random time τ at which C is entered is stochastically dominated by a geometric random variable τ' with mean $1/p$ and finite variance. For any linear f with zero mean, we have, as $T \to \infty$,

$$\text{Var}(\hat{\hat{f}}) = \text{Var}(\frac{1}{2T} \sum_{t=1}^{\min\{T,\tau\}} (f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t))) \leq \frac{1}{T^2} c\text{E}((\tau')^2) = \mathcal{O}(T^{-2}),$$

where c is a finite constant.

5 Joint properties of the coupled Gibbs sampler chains

The coupled Gibbs sampler chains $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ form a Markov chain evolving on $\Omega \times \Omega$ that updates components blockwise, the block being $B_i = (X_i, Y_i)$. Although each marginal component is a Gibbs sampler chain, $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ does not need to be. An algorithm that at each step updates a component B_i using a conditional probability that does not depend of the current value in B_i is not necessarily a Gibbs sampler: It is always possible to produce such an algorithm by adding to the transition matrix of a Gibbs sampler a zero-row-sum matrix (that depend on π).

It is interesting to know if the joint chain $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is ergodic and if so what the properties of the stationary measure are, which of course has π as marginals. The difficulty is well illustrated by the multivariate Gaussian case. As explained in Section 4, if there is a limit distribution $\mu(\boldsymbol{x}, \boldsymbol{y})$ of $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$, as $t \to \infty$, then its support must be

$$\text{supp}(\mu) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega | \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{0}\}. \quad (11)$$

In this case, when $t \to \infty$, the density of $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is attracted towards the subspace $\boldsymbol{x} = -\boldsymbol{y}$. Hence $\Omega \times \Omega$ can be decomposed into a transient class and an ergodic one and μ is singular with respect to

$\pi \times \pi$. For general state space $\Omega \times \Omega$, the picture could be more complicated: it could be that the marginal components converge (to π) while jointly they do not converge, or that there is more than one ergodic class. We are not able to exclude such situations. However the asymptotic behavior of the joint chains does not influence the efficacy of the new algorithm.

The theory in the Appendix of Arjas & Gasbarra (1996) can be used to prove that if the joint chain $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is started in the ergodic class then there exists an unique stationary distribution μ on this class. In the multivariate normal case, this means that if $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ is such that $\boldsymbol{X}^0 = -\boldsymbol{Y}^0$, then there exists an unique stationary distribution μ on the set $\boldsymbol{x} = -\boldsymbol{y}$. We give the precise statement, see Arjas & Gasbarra (1996) for more information on the assumptions.

THEOREM 3 *Let $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ be positive recurrent Markov chains on a complete separable metric space Ω. Let $\boldsymbol{Z}^t = (\boldsymbol{X}^t, \boldsymbol{Y}^t)$ be a φ-irreducible Markovian coupling of $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$. Consider the closure (with respect to the product topology of $\Omega \times \Omega$) $\overline{supp\{\varphi\}}$ with the relative topology inherited from the product topology. Let $(\boldsymbol{X}^0, \boldsymbol{Y}^0) \in \overline{supp\{\varphi\}}$. If, as a Markov chain on $\overline{supp\{\varphi\}}$, $\boldsymbol{Z}^t$ is weakly Feller with respect to the relative topology, and if supp$\{\varphi\}$ contains an open set (with respect to the relative topology), then $\boldsymbol{Z}^t$ is positive recurrent.*

It can be seen that the coupled Gibbs samplers $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ realize a Markovian coupling. In the multivariate normal case φ can be chosen to be the Lebesgue measure. Although we are not able to prove in general that the coupled chains always have a joint unique ergodic class, we have not experienced more than one ergodic class in our numerical experiments. What can we say about the form of the support of μ? In the Gaussian example and in the Ising model it is the symmetry of the conditional density with respect to the median that is linear in the conditioning components that makes the limiting support of the joint chain of the type $\{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} = H(\boldsymbol{x})\}$. It is interesting to note that if there is such a function H and if $\pi(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega$, then this function must act componentwise, i.e. $y_i = h_i(x_i)$ for all i, as happens in the Gaussian case. This is precisely stated in Theorem 5 in the Appendix. Note that if π is a n-fold product measure on $\Omega = S^n$, i.e. $\pi = \pi_1 \times \cdots \times \pi_n$, then $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ has a stationary distribution reached after one single sweep with support $y_i = h_i(x_i), i = 1, \ldots, n$, where the functions h_i are nonlinear. Hence $\boldsymbol{x} + \boldsymbol{y}$ equals some constant is not the only possible form for a degenerate supp(μ).

6 NON RIGOROUS VARIANCE COMPARISON FOR GENERAL NON-ATTRACTIVE TARGET DENSITIES

We would like to extend our theory to more general target densities π, not necessarily attractive, and quantify the gain obtained using the new algorithm. We are however not able to do this rigorously. In this section we present some rough arguments and conjectures. We assume that the coupled chains have an ergodic class in which they have been started and that there exists a joint stationary measure μ on such a class. Denote by $\beta_k = \beta(t, t+k)$ the cross-autocovariances in the stationary regime.

Further let $f \in \mathcal{F}$ and assume a random scan. We first argue that β_k, $k > 0$, all have the same sign as β_0. The heuristic argument is based on the approximations

$$\mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t) \approx \frac{\mathrm{Cov}(f(\boldsymbol{Y}^t), f(\boldsymbol{Y}^{t+k}))}{\mathrm{Var}(f(\boldsymbol{Y}^t))} f(\boldsymbol{Y}^t) = \frac{\gamma_k}{\gamma_0} f(\boldsymbol{Y}^t) \tag{12}$$

and

$$\mathrm{E}(f(\boldsymbol{X}^t) \mid \boldsymbol{Y}^t) \approx \frac{\mathrm{Cov}(f(\boldsymbol{Y}^t), f(\boldsymbol{X}^t))}{\mathrm{Var}(f(\boldsymbol{Y}^t))} f(\boldsymbol{Y}^t) = \frac{\beta_0}{\gamma_0} f(\boldsymbol{Y}^t). \tag{13}$$

Approximation (12) is explained as follows: among all quantities $cf(\boldsymbol{Y}^t)$, linear in $f(\boldsymbol{Y}^t)$, the one given in (12) minimizes the mean squared error,

$$\mathrm{E}_{\boldsymbol{Y}^t}\ \mathrm{E}((cf(\boldsymbol{Y}^t) - f(\boldsymbol{Y}^{t+k}))^2 \mid \boldsymbol{Y}^t).$$

The same argument applies to (13). The $k-$step conditional expectation is the best predictor for $f(\boldsymbol{Y}^{t+k})$ in terms of mean squared error, but is not generally linear in $f(\boldsymbol{Y}^t)$. If f is linear, $f(\boldsymbol{Y}^t) = \boldsymbol{a}^T\boldsymbol{Y}^t$, and if π is multivariate normal, then the k-step conditional expectation is linear in $\boldsymbol{Y}^t$ and approximately linear in $\boldsymbol{a}^T\boldsymbol{Y}^t$ unless the dependency among the Y_i's is very strong.

Using (12), (13) and conditional independence, we obtain the following expression for β_k:

$$\begin{aligned}
\beta_k &= \mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+k})) \\
&= \mathrm{E}_{\boldsymbol{Y}^t}\mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t) \\
&= \mathrm{E}_{\boldsymbol{Y}^t}\left[\mathrm{E}(f(\boldsymbol{X}^t) \mid \boldsymbol{Y}^t)\ \mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t)\right] \\
&= \mathrm{E}_{\boldsymbol{Y}^t}\left[\mathrm{E}(f(\boldsymbol{X}^t) \mid \boldsymbol{Y}^t)\ \mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t)\right] \\
&\approx \mathrm{E}_{\boldsymbol{Y}^t}\left[\frac{\beta_0}{\gamma_0} f(\boldsymbol{Y}^t)\ \frac{\gamma_k}{\gamma_0} f(\boldsymbol{Y}^t)\right] = \beta_0 \frac{\gamma_k}{\gamma_0}.
\end{aligned} \tag{14}$$

The approximations (12) and (13) are only used in the last line of (14). If (12) and (13) are rather precise, so will be (14). For a random scan Gibbs sampler Liu, Wong & Kong (1995) show that $\gamma_k \geq 0$ for all k, regardless of the attractivity. Hence, if (12) and (13) were correct, then β_k would have the same sign as β_0 for all $k > 0$. Figure 3 shows a plot of the estimated values of β_k/γ_0 and the approximation $\beta_0\gamma_k/\gamma_0^2$ for the pump-example described in Section 7.1. The fit is very good.

Using (14), and the expressions for Var($\hat{f}$) and Var($\hat{\hat{f}}$) given in the proof of Theorem 1, we calculate the variance reduction factor of $\hat{\hat{f}}$ with respect to $\hat{f}$ when $T \to \infty$, as

$$\frac{\mathrm{Var}(\hat{f})}{\mathrm{Var}(\hat{\hat{f}})} \sim \frac{1}{1 + \beta_0/\gamma_0}. \tag{15}$$

The antithetic algorithm is always better if the leading factor $\beta_0 \leq 0$ and (12) and (13) (approximately) hold. We conjecture that this is true in many cases. For example, suppose the cross-autocorrelation at lag zero β_0/γ_0 is equal to, say, $-2/3$. Then the variance reduction factor is approximately 3. In the experiments reported in Section 7 we always observe estimated variance reduction factors larger than three. Note further that the ratio (15) does not depend on γ_k, $k > 0$, which may indicate that the efficiency of the new algorithm does not depend on the mixing properties of the marginal chains.

Because of (14), it is natural to try to prove that $\beta_0 \leq 0$ for general non-attractive π and $f \in \mathcal{F}$. We are able to prove $\beta_0 \leq 0$ only for π such that

$$\mathrm{E}_\pi(f(\boldsymbol{X} \mid \boldsymbol{X}_{-i})) \in \mathcal{F} \tag{16}$$

for all i and for all $f \in \mathcal{F}$. Unfortunately, (16) is equivalent to attractivity.

THEOREM 4 *$E_\pi(f(\boldsymbol{X} \mid \boldsymbol{X}_{-i})) \in \mathcal{F}$ for all i and for all $f \in \mathcal{F}$ if and only if π is attractive.*

The if-part is obvious. To prove the only if-part, we construct a counterexample. Suppose π is not attractive. There exists $\boldsymbol{x}_{-j}$, x'_j and $x'_i > x_i$ such that $\pi(X_j \leq x'_j \mid \boldsymbol{x}_{-j}) \leq \pi(X_j \leq x'_j \mid \boldsymbol{x}'_{-j})$ where $\boldsymbol{x}'_{-j}$ denotes $\boldsymbol{x}_{-j}$ with x_i replaced by x'_i. Now put $f(\boldsymbol{x}) = 1_{[x_j > x'_j]}$ to obtain a contradiction.

It remains open to prove that $\beta_0 \leq 0$ (or $\beta_k \leq 0$) for a general non-attractive π.

We conclude this section with a non-attractive example that mimics the behavior of the new algorithm and that allows for rigorous analysis. The two coupled chains are stationary non-Gaussian autoregressive processes. In this case the approximations (12) and (13) are exact; therefore, the sign of β_k follows the sign of β_0 and (15) is valid as a measure of the variance reduction, which is not influenced by the mixing properties of the marginal chains. Furthermore, besides being non-attractive, the process does not satisfy the symmetry condition used in Section 4 to prove variance reduction. Let X^t be the real valued autoregressive process

$$X^t = \phi X^{t-1} + \epsilon_x^t, \qquad t > 0, \tag{17}$$

started in equilibrium at time zero. Here, $\mid \phi \mid < 1$ to ensure stationarity and ϵ_x^t are i.i.d. binary variables with $\mathrm{P}(\epsilon_x^t = 1) = p \geq 1/2$ and $\mathrm{P}(\epsilon_x^t = 0) = 1 - p$. Although this is not a Gibbs sampler, it has the same flavor, see for example (28). We choose $f(x) = x$ with the aim of estimating the mean $\mathrm{E}(X) = p/(1-\phi)$. The variance of $\hat{f}$ is

$$\mathrm{Var}(\hat{f}) = \mathrm{Var}(\frac{1}{2T}\sum_{t=1}^{2T} X^t) \sim \tau_x/(2T), \quad \text{where} \quad \tau_x = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \tag{18}$$

is the integrated auto-covariance time. Also, $\tau_x = \gamma_0(1+2\phi)/(1-\phi)$ where $\gamma_0 = p(1-p)/(1-\phi^2)$. We compare the variance in (18) with that obtained using two realizations of (17), X^t and Y^t, where X^t is sampled (forward in time) using the uniform random variable U^t and Y^t is sampled using $1-U^t$. The estimator $\hat{\hat{f}}$ has variance

$$\mathrm{Var}(\hat{\hat{f}}) = \mathrm{Var}(\frac{1}{T}\sum_{t=1}^{T}\frac{X^t + Y^t}{2}) \equiv \mathrm{Var}(\frac{1}{T}\sum_{t=1}^{T} Z^t) \sim \tau_z/T.$$

Z^t is an autoregressive process of the same form as (17), with ϵ_z^t equal to either one, with probability $2p-1$, or to $1/2$ otherwise. The asymptotic variance of Z^t is $(p-1/2)(1-p)/(1-\phi^2)$. Hence, we obtain the factor of variance reduction of $\hat{\hat{f}}$ w.r.t. $\hat{f}$ as

$$\frac{\text{Var}(\hat{f})}{\text{Var}(\hat{\hat{f}})} \sim \frac{\tau_x}{2\tau_z} = \frac{1}{2-1/p}, \qquad p > 1/2, \tag{19}$$

where we make use of the exponentially decaying auto-covariances of X^t and Z^t. This result shows that the antithetic estimate is always better, and that the variance reduction factor tends to ∞ as the symmetry increases, i.e. $p \to 1/2$. It tends to one as the symmetry decreases, i.e. $p \to 1$. For $p = 1/2$ (perfect symmetry), the variance of $\hat{\hat{f}}$ is again $\mathcal{O}(T^{-2})$. Notice that the joint and marginal chains require a burn-in of similar length, as both are autoregressive processes of the same form (17). For the cross-covariances we get

$$\beta_k = -\frac{(1-p)^2}{1-\phi^2}\phi^{|k|}$$

and we note that (14), and then (15), holds exactly for this example. Furthermore, β_0 is minimal for $p = 1/2$, it takes value $\beta_0 = 0$ if $p = 1$, and the efficiency in (19) increases as β_0 becomes more negative.

7 Practical Implementation and Numerical Experiments

According to Theorem 1, if π is attractive we should start the two marginal chains independently in π. Only if a raster scan is used, can all components of $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ be sampled independently. The latter method is easy, while sampling $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ independently from π requires the independent running to convergence of two ordinary Gibbs Sampler chains. The bias of the two estimators is influenced by the initialization. In general, the asymptotic mean squared error of either estimator is determined by the variance, which is of order T^{-1}, and by the squared bias, which is of order T^{-2}. We discuss more about the bias in our first example. In practice, we run a single Gibbs sampler for T_0 steps. We keep $\boldsymbol{X}^{T_0}$ and discard the rest. We let $\boldsymbol{Y}^{T_0} = \boldsymbol{X}^{T_0}$ and start two (dependent) trajectories, one using (4) and the other (7). We terminate the coupled chains after a further T transitions. In this way we fail to fulfill precisely the requirement of $\boldsymbol{X}^{T_0}$ and $\boldsymbol{Y}^{T_0}$ in Theorem 1 (independence and π-distributed). Nevertheless, we will compare this algorithm, that gives an estimator $\hat{\hat{f}}$ based on a total of $2T$ Gibbs sampler updates, with a single Gibbs sampler chain of length $2T$, started in $\boldsymbol{X}^{T_0}$. As mentioned, the new algorithm requires almost no additional programming compared to the usual simple Gibbs sampler. If the burn-in is long enough, the two estimators will be approximately unbiased.

In the rest of this section we shall apply our new Gibbs sampler algorithm to two well studied data sets, the hierarchical Poisson model (Gelfand & Smith, 1990) and the ordered normal means example (Gelfand et al., 1990). The main purpose is to evaluate the performance of the new algorithm and to quantify its variance reduction and the efficiency w.r.t. the usual Gibbs sampler. We will also introduce antithetically coupled Metropolis-Hastings chains and discuss their performance.

7.1 Hierarchical Poisson Model

Gelfand & Smith (1990) present counts $\boldsymbol{s} = (s_1, \dots, s_n)$ of failures in $n = 10$ pump systems at a nuclear power plant, where the times of operation $\boldsymbol{t} = (t_1, \dots, t_n)$ for each system are known. The hierarchical model assumes $s_k \sim \text{Poisson}(\lambda_k t_k)$, and a common Gamma prior for the failure rate λ_k of each pump, $\lambda_k \sim \Gamma(\alpha, \beta)$. The problem is to infer on α and on the inverse scale β. We take as prior for α the exponential distribution with mean one, and for β a $\Gamma(0.1, 1.0)$ distribution. We shall estimate the posterior means of α and β.

The conjugate priors ensure that λ_1 is Γ-distributed conditional on the remaining variables, as are $\lambda_2 \dots \lambda_n$ and β. It is therefore easy to update each of these variables using a Gibbs sampler. The conditional density for α is however non-standard since

$$\pi(\alpha \mid \lambda_1, \dots, \lambda_{10}, \beta) \propto \exp(\alpha a - n \log \Gamma(\alpha)), \quad \text{where} \quad a = n \log \beta + \sum_{k=1}^{n} \log \lambda_k - 1. \tag{20}$$

In this case it is most natural to perform a Metropolis-Hastings step for the α-parameter update. This means that, using a proposal density, a new value for α is proposed and then accepted or rejected. We suggest three different updating strategies for α.

1. (Gibbs sampler update) To implement the full Gibbs sampler, we compute numerically $F^{-1}(u; a_x)$ and $F^{-1}(1-u; a_y)$, where F is the cumulative conditional distribution function (20) for α.
2. (Hastings update) We approximate the conditional density (20) with a normal ($\tilde{F}$) where the mean and variance match the mode and the curvature in the mode. We update α using a Hastings step, where we propose to move the current values of α to $\tilde{F}_x^{-1}(u)$ and $\tilde{F}_y^{-1}(1-u)$ respectively for the two chains, and accept the proposals using independent uniform variates. We obtain an average acceptance rate for α of 90%.
3. (Metropolis update) We update α using a random walk Metropolis step and propose a new state from an uniform density centered at the old state. The width of the proposal density is determined to obtain an average acceptance rate for α close to 50%. The random variates used in the acceptance-step are again independent.

To verify the robustness with respect to various parameter scanning schedules, we apply each of these three updating rules for α with three different visiting schedules: random scan (RS), where we look to 12 variable updates as one step; random permutation scan (RPS), where at each iteration we update our 12 variables in a random permutation; and deterministic scan (DET), where at each iteration we update $\lambda_1, \dots, \lambda_{10}, \alpha, \beta$ and then $\alpha, \lambda_{10}, \dots, \lambda_1$. All these visitation schedules give a reversible Markov chain.

We run a single Markov chain using $T_0 = 1\,000$ iterations as burn-in, and then we split the chain into two components and run it according to (4) and (7) for $\lambda_1, \dots, \lambda_{10}, \beta$. For α, we run it according to one of the three above methods. The algorithm is set to perform a further $T = 50\,000$ iterations.

Figure 1 shows small parts of the sample paths for the β variables in the two chains, denoted by β_x^t and β_y^t respectively, where we use the Gibbs sampler also for α and RPS. The paths show a clear negative correlation. In Figure 2(a) we plot the sampled points (α_x^t, α_y^t) of the two coupled chains to show the shape of empirical joint density, using $5\,000$ subsequent samples. The second panel of Figure 2 illustrates the empirical joint density of (β_x^t, β_y^t). The negative cross-correlation structure is clearly visible. Figure 3 shows how good the approximation of the cross-autocovariances given in (14) is, for α and β using Gibbs-sampling to update also α.

To give a quantitative measure of the variance reduction using the antithetic chains, we estimate the integrated autocovariance time using all $50\,000$ iterates and the approach of Geyer (1992) for reversible chains. The ratios $\text{Var}(\hat{f})/\text{Var}(\hat{\hat{f}})$, for f projecting on the single components α and β, are listed in Table 1 for the three different updating rules and the three visitation schemes. These ratios do not seem to depend significantly on the visitation schedules. The variance reduction factors for the Gibbs samplers are around 9 and 6 for α and β, respectively. The variance reduction of the two other algorithms (Hastings update and Metropolis update) drops to around $2 - 2.5$ for α. This occurs despite the fact that the acceptance rate was 90% for the Hastings update. A further experiment using a random walk Metropolis update for α with an uniform proposal with larger width and an acceptance rate of 25%, still gave a variance reduction around 2. This reason for this is that the two antithetic chains get out of phase when an antithetic proposal is rejected by one chain but not by the other; the antithetic coupling between the two chains weakens. We do not adjust for this in later iterations, since only shared random numbers are used to introduce antithetic dependency between the two chains and the current states of the two chains are not considered in the proposal.

The antithetic Gibbs sampler is also better than a single hybrid $2T$ long chain, using Gibbs sampling for $\lambda_1, \ldots, \lambda_{10}, \beta$ and a Hastings update for α, as described in item 2 above. The asymptotic variance of such a hybrid sampler is larger than the asymptotic variance of a single pure Gibbs sampler.

We now include the biases in our analysis. Let the efficiency be defined as the squared bias plus the variance, and consider the ratio

$$\frac{\text{bias}(\hat{f})^2 + \text{Var}(\hat{f})}{\text{bias}(\hat{\hat{f}})^2 + \text{Var}(\hat{\hat{f}})}. \tag{21}$$

The estimated ratios (21) regarding the estimation of α and β, for the pure Gibbs sampler and the deterministic visitation scheme are: 30.0 and 13.0 respectively. The true values of α and β, needed in the bias calculation, were estimated with a very long run ($4\,000\,000$ full sweeps). We can conclude that the new antithetic algorithm is still better than a single long chain. This is surprising, but it seems that the bias of the average based on the $\boldsymbol{X}^t$ chain has the opposite sign to the bias of the estimator based on the antithetic $\boldsymbol{Y}^t$ chain, so that these contributions to the bias of $\hat{\hat{f}}$ cancel. This seems to be a further advantage for the new method. In Figure 4 we plot the bias of these two chains. Observe the antithetic sign. In the same figures (one for α and one for β) we have also plotted the total bias of the antithetic Gibbs sampler, which oscillates around zero. To avoid a further figure we have shrunk the time of this total bias, so that it can be compared to the bias of the estimate based on $\boldsymbol{X}^t$ (or on $\boldsymbol{Y}^t$). This now corresponds to the bias of a single, twice as long, run. The bias is significantly smaller.

7.2 The Ordered Normal Mean Problem

Gelfand et al. (1990) use the Gibbs sampler to estimate the mean and precision in normal populations, when the ordering of the means is known in advance. We have repeated their example using the antithetic Gibbs sampler to investigate its variance reduction and efficiency in estimating the posterior mean of the parameters of interest.

Let Y_{ij} be the jth observation ($j = 1, \ldots, n_i$) from the ith group ($i = 1, \ldots, n_g$). Assuming conditional independence throughout, let $Y_{ij} \sim \mathrm{N}(\theta_i, 1/\tau_i)$, $\theta_i \sim \mathrm{N}(\mu, 1/\tau_g)$, $\tau_i \sim \Gamma(a_1, b_1)$, $\tau_g \sim \Gamma(a_2, b_2)$, and $\mu \sim \mathrm{N}(\mu_0, 1/\tau_0)$. Here τ_i, τ_g, τ_0 denote the precision or inverse variance. A priori it is known that the means θ_i satisfy the constraint $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{n_g}$. Gelfand et al. (1990) demonstrate that the Gibbs sampler is easy to implement even in this case. We refer to Gelfand et al. (1990) for details about the Gibbs sampler and for the specific choices of the (flat) priors of the hyperparameters $a_1, a_2, b_1, b_2, \mu_0$ and τ_0.

We simulated data using $n_g = 5$, and sampled from the ith population, $n_i = 2i + 4$ observations from $\mathrm{N}(i, i^2)$. Table 2 lists the empirical mean and variance within each group. Note that the observed ordering of the means is not in agreement with the a priori constraint. We used the deterministic site visitation schedule DET with a burn-in of $1\,000$ cycles. The variance reduction factor $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$ was estimated using the following $50\,000$ iterates of the coupled chains as in Section 7.1. Table 3 displays the estimated ratios for (θ_i, τ_i), $i = 1, \ldots, n_g$. The new antithetic Gibbs sampler again gives a significant speedup with variance reduction between 2.97 and 6.69 with an average of 4.7. Similar results were obtained for the other visiting schedules.

8 Conclusions

We have suggested a simple way to couple two Gibbs sampler chains in order to reduce the variance of the empirical average as an estimator of an expectation. The coupling induces negative cross-autocovariances. The new estimator is also asymptotically unbiased and the reduction of the variance can be remarkable with respect to the simple Gibbs sampler run for the same time. The coding of the proposed algorithm is easy, given a standard Gibbs Sampler implementation.

Other authors have introduced antithetic behaviors into a single MCMC chain. If the density π is symmetric around zero, Geweke (1988) proposes using the estimator $\frac{1}{T}\sum_{t=1}^{T/2}(f(\boldsymbol{X}^t) + f(-\boldsymbol{X}^t))$ and proves that its asymptotic variance is smaller than $\mathrm{Var}(\hat{f})$. Barone & Frigessi (1989) propose a variation of the Gibbs sampler where each step moves antithetically to the current state and show a faster weak convergence rate in some cases. Neal (1998) improves the single updating step further. Green & Han (1992) show that in such a way the asymptotic variance could also be reduced in certain special cases. In the present paper we show that with two chains a more authentic antithetic behavior can be established.

As the example showed, it is not trivial to extend equally successfully the antithetic idea to Metropolis-

Hastings type algorithms. The reason for this is that it is more difficult to induce antithetic correlation when an accept-reject step may well reject a proposed antithetic move. More research is needed in order to understand how to couple such chains properly.

The Gibbs sampler is often not the fastest MCMC algorithm. In fact other Metropolis-Hastings schemes have often a smaller asymptotic variance. However, the new antithetically coupled Gibbs sampler may compete with such algorithms. For example, in the case of the multivariate normal density and the Ising model it should be preferred to other single site updating MCMC for which $\mathrm{Var}(\hat{f}) = \mathcal{O}(T^{-1})$.

Our rigorous results cover the case of attractive target densities π, and we are not able to generalize them to general π's. However, our numerical experiments and our intuitive understanding indicate a broader range of applicability.

REFERENCES

ARJAS, E. & GASBARRA, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints, *Journal of the American Statistical Association* 91(435): 1101–1109.

BARLOW, R. E. & PROSCHAN, F. (1975). *Statistical theory of reliability and life testing: probability models*, New York: Holt, Rinehart and Winston.

BARONE, P. & FRIGESSI, A. (1989). Improving stochastic relaxation for Gaussian random fields, *Probability in the Engineering and Informational Sciences* 3(4): 369–389.

ESARY, J. D., PROSCHAN, J. D. & WALKUP, D. W. (1967). Association of random variables, with applications, *Annals of Mathematical Statistics* 38: 651–655.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* 85: 398–509.

GELFAND, A. E., HILLS, S. E., RACINE-POON, A. & SMITH, A. F. M. (1990). Illustration of Bayesian inference in normal data models using the Gibbs sampler, *Journal of the American Statistical Association* 85(412): 972–985.

GEWEKE, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayseian inference, *Journal of Econometrics* 38: 73–90.

GEYER, C. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* 7: 473–511.

GILKS, W. R., RICHARDSON, S. & SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

GREEN, P. J. & HAN, X. L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables, *in* P. Barone, A. Frigessi & M. Piccioni (eds), *Stochastic Models, Statistical Methods and Algorithms in Image Analysis*, number 74 in *Lecture notes in Statistics*, Springer, Berlin, pp. 142–164.

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* 57(1): 157–169.

MEYN, S. P. & TWEEDIE, R. L. (1993). *Markov chains and stochastic stability*, London: Springer.

MØLLER, J. (1999). Perfect simulation of conditionally specified models, *Journal of the Royal Statistical Society, Series B* 61(1): 251–264.

NEAL, R. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation, *in* M. I. Jordan (ed.), *Learning in Graphical Models*, Kluwer Academic Press.

Estimated $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$	Gibbs sampler			Gibbs/Hastings			Gibbs/Metropolis		
	RS	RPS	DET	RS	RPS	DET	RS	RPS	DET
α	9.10	9.04	9.58	2.26	2.19	2.31	2.14	2.04	2.07
β	5.69	6.25	6.15	2.88	2.86	2.50	2.74	2.50	2.39

TABLE 1: Hierarchical Poisson model: Estimated $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$ for f equal to α or β. Three different ways of updating the parameter α and three different scan strategies are compared using $50\,000$ iterates. The antithetic coupling is very convenient for the pure Gibbs sampler, but the variance reduction decreases using a Hastings-update or a Metropolis-update for α.

Sample values	1	2	3	4	5
n_i	6	8	10	12	14
$\bar{Y}_i$	0.645	2.212	3.576	2.401	4.195
S_i^2	1.473	2.279	3.452	20.186	11.330

TABLE 2: Ordered normal means problem: characteristics of the simulated data. Note the exchange in the empirical ordering of the means.

Estimated $\mathrm{Var}(\hat{f})/\mathrm{Var}(\hat{\hat{f}})$	1	2	3	4	5
θ_i	5.44	4.02	2.97	3.09	4.31
τ_i	4.20	5.08	4.71	6.69	6.53

TABLE 3: The estimated variance reduction of the estimates based on antithetic coupled Gibbs sampler w.r.t. the estimates based on a simple Gibbs sampler in the ordered normal means problem using $50\,000$ iterates.

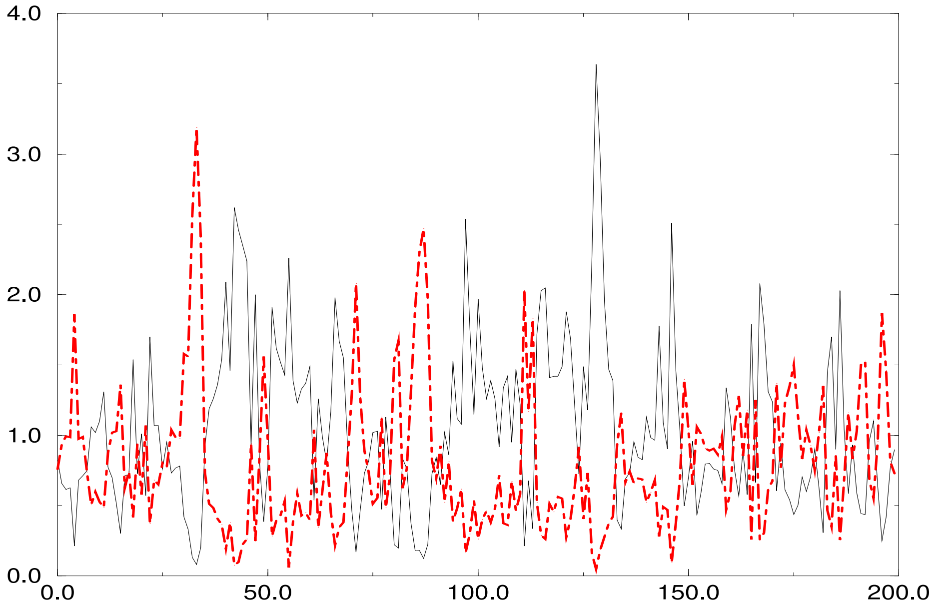


FIGURE 1: The sample-paths of the β component of the two antithetically coupled Gibbs sampler chains show a clear negative correlation. 200 consecutive iterations.

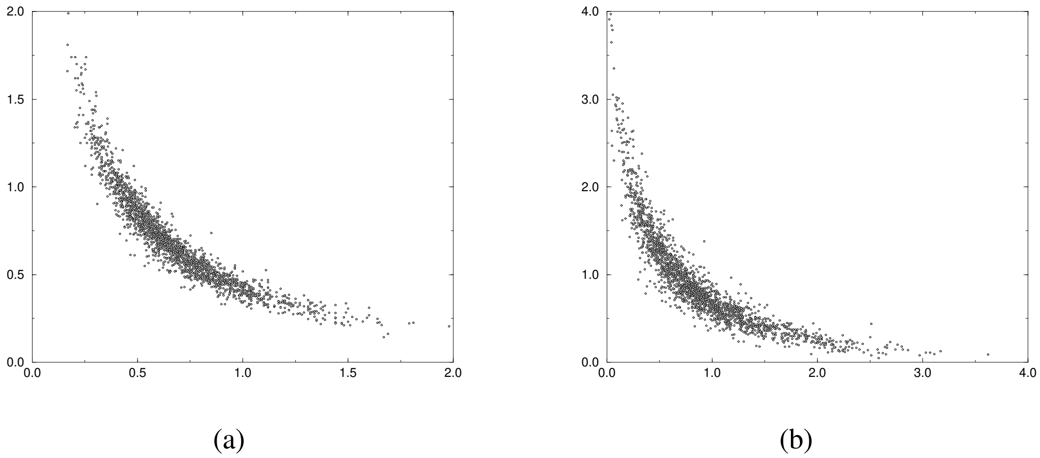
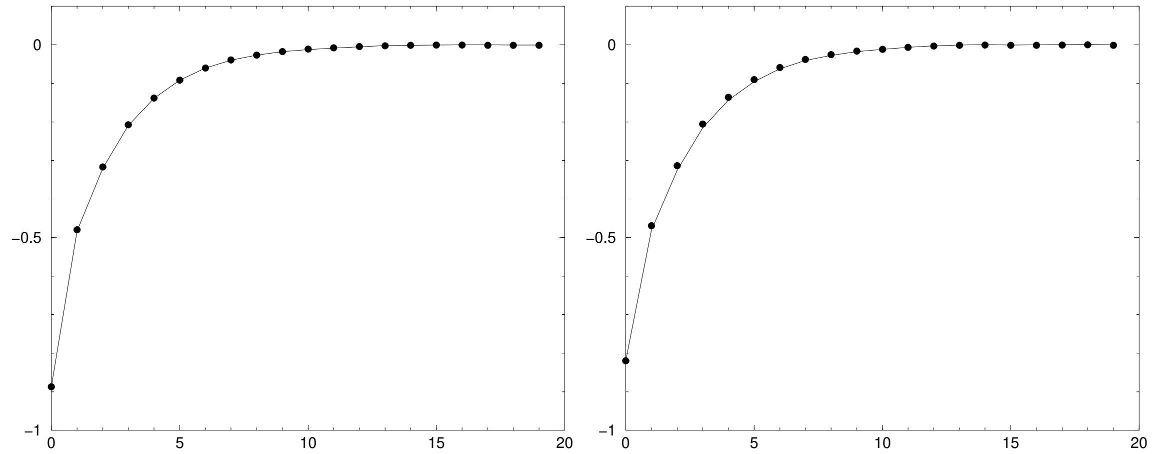
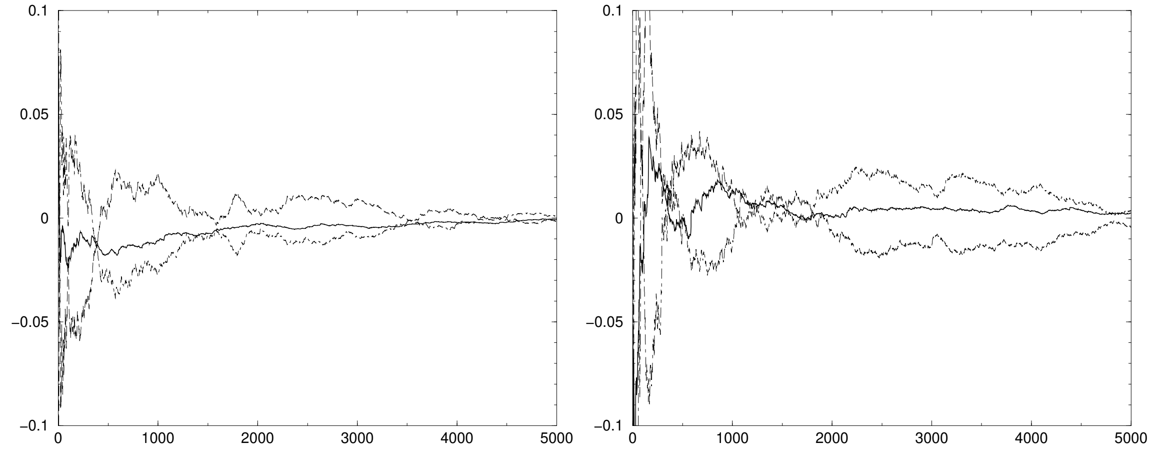


FIGURE 2: Point plots of 5 000 samples from the two antithetic Gibbs sampler chains for (a) α and (b) β. The support of joint density, and a clear negative correlation are illustrated.



(a) (b)

FIGURE 3: The estimated cross-autocorrelation (solid line) for α (in (a)) and β (in (b)) together with the approximated cross-correlation (dots) based on approximation (14), for the pump-example using the Gibbs-sampler and the visitation schedule DET. The approximation is very good.



(a) (b)

FIGURE 4: Bias in estimation of the posterior mean for α (in (a)) and β (in (b)) for each of the two Gibbs-sampler chains $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ (dashed and dot-dashed lines) and the antithetic Gibbs-sampler (solid line) as a function of number of iterations, (counting two for the antithetic Gibbs-sampler). The bias for the antithetic Gibbs-sampler is scaled so the amount of computational work is comparable.

A Proofs

A.1 Proof of Theorem 1

To prove Theorem 1, we need the following lemma.

LEMMA 1 *Suppose $f \in \mathcal{F}$ and let π be attractive. Consider the coupled Gibbs sampler chains given in (4) and (7). If the components $X_1^0, \dots, X_n^0, Y_1^0, \dots, Y_n^0$ are generated independently and if a deterministic raster scan is used, then $\beta(t, t+k) \leq 0$ and $Cov(f(\boldsymbol{X}^t), f(\boldsymbol{X}^{t+k})) \geq 0$ for all $t \geq 0$ and $k \geq 0$. The same assertions hold true if $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ are drawn independently from π and either a deterministic raster or a random scan is used.*

PROOF OF LEMMA 1 The proof relies on the construction of sets of associated random variables. We shall use properties of associated random variables referenced as P1 to P4 in Esary, Proschan & Walkup (1967).

Deterministic scan. First assume that $X_1^0, \dots, X_n^0, Y_1^0, \dots, Y_n^0$ are independent. The component i^t is updated in the transition from $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ to $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1})$, which happens according to $X_{i^t}^{t+1} = \Phi_{i^t}(\boldsymbol{X}_{-i^t}, U^t)$ and $Y_{i^t}^{t+1} = \Phi_{i^t}(\boldsymbol{Y}_{-i^t}^t, 1-U^t)$. $\Phi_{i^t}(\cdot,\cdot)$ is nondecreasing in each variable, by attractivity and because of the monotonicity of inverse conditional distribution functions. Now suppose that

$$\{X_1^t, \dots, X_n^t, -Y_1^t, \dots, -Y_n^t\} \tag{22}$$

is a set of associated random variables. Then also $S^t = \{X_1^t, \dots, X_n^t, -Y_1^t, \dots, -Y_n^t, U^t\}$ are associated, since U^t is independent of the other variables (P2). By the monotonicity of Φ_{i^t} it follows that $X_{i^t}^{t+1}$ and $-Y_{i^t}^{t+1}$ are nondecreasing functions of the variables in S^t. For $j \neq i^t$, X_j^{t+1} and $-Y_j^{t+1}$ are trivially nondecreasing functions of the variables in S^t. Hence $\{X_1^{t+1}, \dots, X_n^{t+1}, -Y_1^{t+1}, \dots, -Y_n^{t+1}\}$ is also a set of associated random variables (P4). Now if $X_1^0, \dots, X_n^0, Y_1^0, \dots, Y_n^0$ are independent, then in particular S^0 is associated (P2), and by induction it follows that (22) is a set of associated random variables for all t. For fixed t, it follows in the same way that $\{X_1^t, \dots, X_n^t, -Y_1^{t+k}, \dots, -Y_n^{t+k}\}$ is associated for each $k \geq 0$. We now use induction on k, changing only $-Y_{i^{t+k-1}}$ in the k-th induction step.

Define two functions, $g(\boldsymbol{x}, -\boldsymbol{y}) = f(\boldsymbol{x})$ and $h(\boldsymbol{x}, -\boldsymbol{y}) = -f(\boldsymbol{y}) = -f(-(-\boldsymbol{y}))$. Because f is nondecreasing, then g and h are nondecreasing functions of $\{x_1, \dots, x_n, -y_1, \dots, -y_n\}$, and it follows by associativity (Esary et al., 1967, Def. 1.1), that

$$\mathrm{Cov}(f(\boldsymbol{X}^t), -f(\boldsymbol{Y}^{t+k})) = \mathrm{Cov}(g(\boldsymbol{X}^t, -\boldsymbol{Y}^{t+k}), h(\boldsymbol{X}^t, -\boldsymbol{Y}^{t+k})) \geq 0.$$

Changing the sign gives the asserted non-positivity of the cross-covariances for each t.

An induction argument similar to the one above shows that the sets $\{X_1^t, \dots, X_n^t, X_1^{t+k}, \dots, X_n^{t+k}\}$ are associated for each $k \geq 0$ and each $t \geq 0$. Replacing the function h in the preceding argument by $h(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y})$, we obtain $\mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{X}^{t+k})) = \mathrm{Cov}(g(\boldsymbol{X}^t, \boldsymbol{X}^{t+k}), h(\boldsymbol{X}^t, \boldsymbol{X}^{t+k})) \geq 0$. Taking the limit as $t \to \infty$ we also have that $\gamma_k \geq 0$.

We move now to the actual assumption of this lemma, that $\{X_1^0, \ldots, X_n^0\}$ and $\{Y_1^0, \ldots, Y_n^0\}$ are π-distributed and independent. Since π is attractive, for $i = 1, \ldots, n$ and arbitrary x_i, we have that $P(X_i \geq x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$ is nondecreasing in $x_1, \ldots, x_{i-1}$ if $\boldsymbol{X}$ is distributed according to π. This means that the variables $X_1, \ldots, X_n$ are conditionally nondecreasing in sequence. By Barlow & Proschan (1975) they are associated. Since association is preserved by multiplying all variables by -1, the same holds for $-X_1, \ldots, -X_n$. Therefore, if the initial state $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ is obtained by drawing $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ independently from π, the set $\{X_1^0, \ldots, X_n^0, -Y_1^0, \ldots, -Y_n^0\}$ is associated. Therefore, the same argument used for the case of independent components in the initial state can be followed.

Random scan. Let the initial states $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$ be π-distributed and independent. Let $\boldsymbol{I}^t = (I^0, \ldots, I^{t-1})$ be the site-updating sequence. Then it holds that

$$\begin{aligned} \mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^{t+k})) = {} & \mathrm{E}\left(\mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{I}^{t+k})\right) \\ & + \mathrm{Cov}\left(\mathrm{E}(f(\boldsymbol{X}^t) \mid \boldsymbol{I}^{t+k}), \mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{I}^{t+k})\right). \end{aligned} \tag{23}$$

If $\boldsymbol{X}^0$ is distributed according to π, then $\boldsymbol{X}^t$ given $\boldsymbol{I}^{t+k}$ is also π-distributed for all t. Hence, $\mathrm{E}(f(\boldsymbol{X}^t) \mid \boldsymbol{I}^{t+k}) = \mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{I}^{t+k}) = 0$ and the second term in (23) is zero. The proof for the deterministic scan shows that the first term in (23) is non-positive. By the same argument, $\mathrm{Cov}(f(\boldsymbol{X}^t), f(\boldsymbol{X}^{t+k})) \geq 0$ for all $k \geq 0, t \geq 0$ if $\boldsymbol{X}^0$ is drawn from π and a random scan is used. Again, letting $t \to \infty$ we get $\gamma_k \geq 0$ for all $k \geq 0$. ■

We are now ready to prove Theorem 1.

PROOF OF THEOREM 1 We compute first $\mathrm{Var}(\hat{f})$ and $\mathrm{Var}(\hat{\hat{f}})$ as functions of $\beta(t,s)$ and γ_k. Then Theorem 1 follows using Lemma 1. The variances are as follows.

$$\mathrm{Var}(\hat{f}) = \mathrm{Var}\left(\frac{1}{2T}\sum_{t=1}^{2T} f(\boldsymbol{X}^t)\right) = \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{2T-1}\gamma_k(1 - \frac{k}{2T}). \tag{24}$$

$$\begin{aligned} \mathrm{Var}(\hat{\hat{f}}) &= \mathrm{Var}\left(\frac{1}{2T}\sum_{t=1}^{T}\left(f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t)\right)\right) \\ &= \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{T-1}\gamma_k(1 - \frac{k}{T}) + \frac{1}{2T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\beta(s,t). \end{aligned} \tag{25}$$

Thus, $T\left(\mathrm{Var}(\hat{f}) - \mathrm{Var}(\hat{\hat{f}})\right) = S - D$, where

$$
\begin{aligned}
S &= \sum_{k=T}^{2T-1} \gamma_k (1 - \frac{k}{2T}) + \frac{1}{2T} \sum_{k=1}^{T-1} k\gamma_k \qquad (26) \\
D &= \frac{1}{2T} \sum_{t=1}^{T} \sum_{s=1}^{T} \beta(s,t). \qquad (27)
\end{aligned}
$$

We can now study the sign of $S - D$ when π is attractive and $f \in \mathcal{F}$, in the case of random or raster scan. Using Lemma 1 we have that $\gamma_k \geq 0$ for all k and $\beta(s,t) \leq 0$ for all s and t, hence $S \geq 0$ and $D \leq 0$ for all T. This concludes the proof. ■

Notice that the proof of Lemma 1 does not require that the components to be updated in the two chains are the same. Hence variance reduction is achieved also if the two chains update different components at each step. However we expect strongest variance reduction when the components are the same.

PROOF OF THEOREM 2

When the i-th component is updated, we can write the usual Gibbs Sampler in matrix notation as

$$\boldsymbol{X}^{t+1} = (\boldsymbol{I} - \boldsymbol{D}_i\boldsymbol{Q})\boldsymbol{X}^t + \boldsymbol{\varepsilon}^t, \qquad (28)$$

where $\boldsymbol{Q}$ is the inverse covariance matrix, $\boldsymbol{D}_i$ is a matrix of zeros, except for a single 1 in i-th position along the diagonal and $\boldsymbol{\varepsilon}^t$ is a vector of zeros except for ε_i^t which is a normal variate with zero mean and variance one. Similarly for the Y chain

$$\boldsymbol{Y}^{t+1} = (\boldsymbol{I} - \boldsymbol{D}_i\boldsymbol{Q})\boldsymbol{y}^t + \boldsymbol{\eta}^t.$$

Due to the antithetic coupling, $\boldsymbol{\eta}^t = -\boldsymbol{\varepsilon}^t$. Hence

$$\boldsymbol{X}^{t+1} + \boldsymbol{Y}^{t+1} = (\boldsymbol{I} - \boldsymbol{D}_i\boldsymbol{Q})(\boldsymbol{X}^t + \boldsymbol{Y}^t). \qquad (29)$$

It is shown in Barone & Frigessi (1989) that the spectral radius of the matrix

$$\boldsymbol{A}_n = (\boldsymbol{I} - \boldsymbol{D}_n\boldsymbol{Q}) \times \cdots \times (\boldsymbol{I} - \boldsymbol{D}_1\boldsymbol{Q}), \qquad (30)$$

that governs a full deterministic raster scan, is strictly smaller than one. The spectral radius of the single components $\boldsymbol{I} - \boldsymbol{D}_i\boldsymbol{Q}$ is smaller or equal to 1. Denote the spectral radius of $\boldsymbol{A}_n$ by q. Consider a linear function f with zero π-mean. We can write (for T multiple of n)

$$\hat{\hat{f}} = \frac{1}{2T} \sum_{k=0}^{T/n-1} \sum_{s=0}^{n-1} \left(f(\boldsymbol{X}^{nk+s}) + f(\boldsymbol{Y}^{nk+s}) \right)$$

so that

$$\mid \hat{\hat{f}} \mid \leq \frac{n}{2T} \sum_{k=0}^{T/n-1} q^k |f(\boldsymbol{X}^0) + f(\boldsymbol{Y}^0)|,$$

and

$$\text{Var}(\hat{\hat{f}}) \leq \frac{cn^2}{4T^2}\left(\frac{1-q^{T/n}}{1-q}\right)^2,$$

for some positive constant c, which gives $\text{Var}(\hat{\hat{f}}) = \mathcal{O}(T^{-2})$. Furthermore, $\text{Var}(\hat{f}) \sim c'/T$ for some constant c'. Hence for T large enough, $\text{Var}(\hat{f}) \geq \text{Var}(\hat{\hat{f}})$. ■

PROOF OF THEOREM 3

We follow Arjas & Gasbarra (1996). A φ-irreducible Markov chain has to be either transient or recurrent. If $\boldsymbol{Z}^t$ is recurrent, it has to be positive recurrent since the marginals are positive recurrent. We show that it cannot be transient.

Fix a starting point $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \overline{\text{supp}\{\varphi\}}$. For each k, the joint k-step transition kernel $K_Z^n(\boldsymbol{z}, \cdot)$ is a coupling of the marginal k-step transition kernels $K_X^k(\boldsymbol{x}, \cdot)$ and ${K_Y}^k(\boldsymbol{y}, \cdot)$. Since the marginal chains are positive recurrent, it follows that the sequences of probability measures $\{K_X^k(\boldsymbol{x}, \cdot), k \in N\}$ and $\{K_Y^k(\boldsymbol{y}, \cdot), k \in N\}$ are tight (see e.g. Meyn & Tweedie (1993)). This implies tightness of the sequence of couplings $\{K_Z^k(\boldsymbol{z}, \cdot), k \in N\}$ in the product space, as well as the tightness in $\overline{\text{supp}\{\varphi\}}$ with the relative topology. This is because if A is a topological space, B is a closed subset, and K is compact in A, then $B \cap K$ is closed w.r.t. the relative topology.

In particular, there is a compact set C (compact w.r.t. the relative topology) such that $K_Z^k(\boldsymbol{z}, C) > 1/2$, $\forall k$. It is not a restriction to take $\boldsymbol{z} \in C$ (add $\boldsymbol{z}$ to C if necessary). Since $\sum_{k=1}^{\infty} K_Z^k(\boldsymbol{z}, C) = \infty$, C is a compact set which is not uniformly transient. By the Lemma in the Appendix of Arjas & Gasbarra (1996), it follows that the Markov chain $\boldsymbol{Z}^t$ is not transient. ■

THEOREM 5

THEOREM 5 *Let $P(\cdot \mid (\boldsymbol{x}, \boldsymbol{y}))$ be a transition kernel on $\Omega \times \Omega$ which updates only one block component (x_i, y_i) at each transition, and which is ergodic with joint stationary measure μ. Suppose that $\pi(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega = S \times \cdots \times S$ and supp$(\mu) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega | \boldsymbol{Y} = H(\boldsymbol{X})\}$. Then this support can be defined componentwise: for each i, y_i is a function of x_i only.*

PROOF We proceed by induction. At step i, consider $(y_1, \ldots, y_i) = h_i(x_1, \ldots, x_i)$ for some function h_i. This is true for $i = n$ with $h_n = H$. Suppose that the chain is stationary at time t, and $(\boldsymbol{X}^t, \boldsymbol{Y}^t) = (\boldsymbol{x}, \boldsymbol{y})$. Assume that $I^t = i$. Then $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1}) = ((x_i^{'}, \boldsymbol{x}_{-i}), (y_i^{'}, \boldsymbol{y}_{-i}))$ belongs to the support of μ, by stationarity. Hence, the induction assumption says that $(y_1, \ldots, y_i^{'}) = h_i(x_1, \ldots, x_i^{'})$. In particular $(y_1, \ldots, y_{i-1}) = h_i(x_1, \ldots, x_i')_{-i}$ is a function of $(x_1, \ldots, x_i')$. But we also have $(y_1, \ldots, y_{i-1}) = h_i(x_1, \ldots, x_i)_{-i}$. By the assumption on π, x_i' could be any point in S, and it follows that $h_i(x_1, \ldots, x_i')_{-i}$ is constant as a function of x_i'. Hence $(y_1, \ldots, y_{i-1})$ is determined by

$(x_1, \dots, x_{i-1})$ only. Proceeding by induction back from $i = n$ to $i = 1$, it follows that y_1 is a function of x_1. The ordering of the components is arbitrary, so that y_k is a function of x_k only, and the theorem follows. ■