ANTITHETIC COUPLING OF TWO GIBBS SAMPLER CHAINS

ARNOLDO FRIGESSI
NORWEGIAN COMPUTING CENTER
OSLO, NORWAY

HÅVARD RUE
DEPARTMENT OF MATHEMATICAL SCIENCES
NTNU, NORWAY

MARCH 31, 1998

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 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 costs. We show that the asymptotic variance of the estimator based on the new algorithm is always smaller than the variance of a single Gibbs sampler chain, if π is either attractive or repulsive and f is componentwise monotone. The new antithetic algorithm is shown to outperform the standard Gibbs sampler by one order of magnitude when π is a multivariate normal density or the Ising model. Numerical experiments show that the antithetically coupled Gibbs samplers reduce the finite sample variance in several other models to less than one third, often one fifth, when run for the same time.

KEYWORDS: Antithetic Monte Carlo; Decay of Cross-autocorrelations; Markov Chain Monte Carlo; Parallel Computing; Variance reduction.

ADDRESSES: A. Frigessi, Norwegian Computing Center, P.O.Box 114 Blindern, N-0314 Oslo. H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7034 Trondheim.

E-MAIL: Arnoldo.Frigessi@nr.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-CT96-0096) and the ESF program on Highly Structured Stochastic Systems.

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 normalization constant of π. Then 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 collect a sample average along the generated trajectory. Let $\boldsymbol{X}^1, \dots, \boldsymbol{X}^T$ denote the first T steps of a discrete time Markov chain evolving on Ω and converging in law to $\pi(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$. By ergodicity we can use the empirical mean of $f(\boldsymbol{x})$ to estimate $\mathrm{E}(f(\boldsymbol{X}))$ for large T. In practice it is appropriate to drop an initial part of the trajectory in order to avoid strong dependence from the initial conditions. The sample mean

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

is used and we speak of a *burn-in* of length T_0. The variance var$(\hat{f})$, calculated with respect to the probability measure over the trajectory space and assumed to be finite, measures the quality of a such an approximation.

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 then coupled in such a way that variance reduction can be achieved. Our algorithm is based on the Gibbs sampler, which is a particular MCMC scheme where at each transition one samples from a one dimensional conditional density computed from π. The coupling is inspired by the 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 processes will look like the usual Gibbs samplers, but their joint probability measure is constructed in such a way that $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ have negative covariance. The idea is to exploit such antithetic behaviour in order to construct an estimator of $\mathrm{E}(f(\boldsymbol{X}))$ with smaller variance than var$(\hat{f})$ but with similar computational complexity. The coupling we propose here is very simple, based on using a common sequence of random numbers. Specifically, if $\boldsymbol{X}^t$ uses a 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 $\boldsymbol{Y}^{t+1}$. In classic Monte Carlo theory this type of antithetic coupling is well known to reduce the variance of sample averages of i.i.d. samples. Our contribution is to merge this classical technique into the MCMC methodology. While the basic idea is very simple, the rigorous analysis of the new algorithm needs some care. The new antithetically coupled Gibbs sampler algorithm is precisely defined in Section 2. A pleasant fact is that no significant extra effort is needed to implement the algorithm. In fact, starting from the usual Gibbs sampler code, the modifications required in order to implement the new algorithm are usually very easy.

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 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 the variance of $\hat{\hat{f}}$ in (1) has to be compared with the variance of

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

We are able to prove a strict inequality between the asymptotic variances of $\hat{\hat{f}}$ and $\hat{f}$ for component-wise monotone functions f and π in a broad class of densities which includes those which are attractive or repulsive. In particular if π is either the multivariate normal density or the Ising model, then as $T \to \infty$

$$\text{var}(\hat{\hat{f}}) = \mathcal{O}(T^{-2}), \quad \text{while} \quad \text{var}(\hat{f}) = \mathcal{O}(T^{-1}). \tag{3}$$

Assume that instead of running a simple Gibbs sampler, a Metropolis-Hastings algorithm would be used to estimate $\text{E}(f(\boldsymbol{X}))$ when π is the Ising model. If the temperature parameter of the Ising model is large enough, it is known that the Metropolis-Hastings algorithm is faster than the Gibbs sampler, but still it would give rise to an estimator $\hat{f}$ for which var$(\hat{f}) = \mathcal{O}(T^{-1})$. Hence in this case the new antithetic Gibbs sampler should be preferred even to Metropolis-Hastings!

Numerical experiments seem to indicate that the new algorithm provides a benefit with respect to the Gibbs sampler in many cases also after a finite number of iterations. The efficiency does not seem to depend 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 ratio of the variances will not be influenced.

In Section 2 we define the new algorithm and prove that the coupled chains are jointly ergodic and we identify the joint stationary density. (We collect all proofs in the Appendix.) In Section 3 we compare var$(\hat{\hat{f}})$ with var$(\hat{f})$ as $T \to \infty$. It turns out that only the cross-autocovariances between the two coupled chains matter, and we study their sign in Section 4. For component-wise monotone functions f and π which are either attractive or repulsive, we are able to show that all cross-autocovariances are negative, and hence our new algorithm is strictly better than the single chain Gibbs sampler. We give some arguments (although no rigorous proof) supporting this claim for general targets π. In Section 5 we study the multivariate normal density and the Ising model in more detail; these distributions have a certain local symmetry property which yields (3). This is further discussed in Section 6, where we show with an example that our new algorithm can be better even with highly unsymmetrical π's. Section 7 is devoted to a discussion of the burn-in of the joint chains compared to the single chain. We argue that the lengths of both burn-in's have the same magnitude. In Section 8 we provide some experimental results applying our new algorithm to the hierarchical Poisson model (Gelfand & Smith, 1990) and the ordered normal means example (Gelfand, Hills, Racine-Poon & Smith, 1990). The experiments show that the performance of the antithetically coupled Gibbs sampler is significantly better than the standard one. The efficiency, defined as var$(\hat{\hat{f}})$/var$(\hat{f})$, is often larger than five. Looking beyond the Gibbs sampler, we apply the antithetic coupling also to other MCMC methods and show that still an improvement can be achieved, although the efficiency is often only around two. We explain why this is the case. We end the paper with some final remarks.

2 Antithetic Coupling of two Gibbs Sampler Chains

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 take $\Omega = \mathbb{R}^n$ and let π be a probability density function that is absolutely continuous with respect to Lebesgue measure. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. The random scan Gibbs sampler for sampling from π is a Markov chain $\boldsymbol{X}^0, \boldsymbol{X}^1, \ldots$ constructed as follows. Given $\boldsymbol{X}^{t-1} = \boldsymbol{x}^{t-1}$, one component in $\{1, 2, \ldots, n\}$ is chosen uniformly at random. Denote this component by $i = i_t$. Only X_i^{t-1} will be updated by sampling the new value X_i^t from the conditional density

$$\pi(x_i | X_j^{t-1} = x_j^{t-1},\ j \neq i). \tag{4}$$

The resulting Markov chain is ergodic and π-invariant. Note that while it is possible to update the components in many ways, for example in a periodic order (raster scan), our precise results in Section 4.1 apply only for random scan Gibbs sampling. However, we think that our conclusions are valid also for other updating schemes.

The random scan Gibbs sampler transition can be written as

$$\boldsymbol{X}^t = \boldsymbol{\Phi}(\boldsymbol{X}^{t-1}, V^t, U^t), \tag{5}$$

where $U^1, U^2, \ldots$ is a sequence of i.i.d. random numbers, uniformly distributed in $[0, 1)$, and $V^1, V^2, \ldots$ are i.i.d. random numbers uniform in $\{1, 2, \ldots, n\}$. This specific assumption on the distribution of V^t is not needed in our theory, but is made for simplicity. The random number V^t identifies the component i to be updated. The V^t-th component of the vector function $\boldsymbol{\Phi}$ is the inverse distribution function corresponding to the local conditional density in (4). The other components of $\boldsymbol{\Phi}$ are identity functions. The random number U^t is used to obtain the new value for X_i^t.

Let $\boldsymbol{x}_{-A} = \{x_i : i \notin A\}$, $A \subset \{1, \ldots, n\}$. Denote by $\Upsilon_i(\boldsymbol{x}_{-A}, u)$ the inverse distribution function for X_i conditioned on $\boldsymbol{X}_{-A\cup\{i\}} = \boldsymbol{x}_{-A\cup\{i\}}$ evaluated at u. Note that the V^t-th component of $\boldsymbol{\Phi}$ at time t is equal to $\Upsilon_{V^t}(\boldsymbol{X}^{t-1}, U^t)$.

We now define the coupled companion chain. It is marginally a π-stationary Gibbs sampler with the same type of scan and evaluation rule as (5),

$$\boldsymbol{Y}^t = \boldsymbol{\Phi}(\boldsymbol{Y}^{t-1}, V^t, 1 - U^t). \tag{6}$$

Observe that the common random numbers U^t and V^t make $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$ dependent. We call the coupling antithetic because we use $1 - U^t$ in (6). The same component V^t is updated in both chains. After a burn-in of length T_0, we start two dependent trajectories, one using (5) and the other (6), and we terminate both chains after a further T transitions. This algorithm performs $2T$ Gibbs sampler updates and we will compare it with a single Gibbs sampler chain of length $2T$. Notice that the new algorithm requires almost no additional programming compared to the usual simple Gibbs sampler. The two antithetically coupled Gibbs sampler chains allow us to construct the estimator $\hat{\hat{f}}$ in (1) which we shall compare to $\hat{f}$ in (2) for the rest of this paper.

The coupled Gibbs sampler chains jointly form a Markov chain evolving on $\Omega \times \Omega$.

THEOREM 1 *The coupled random scan Gibbs sampler chains given in (5) and (6) are jointly a blockwise random scan Gibbs sampler on $\Omega \times \Omega$ with stationary density*

$$\begin{aligned} \mu(\boldsymbol{x}, \boldsymbol{y}) &\propto \pi(x_i|\boldsymbol{x}_{-i})\pi(y_i|\boldsymbol{y}_{-i})\delta(\boldsymbol{x}, \boldsymbol{y}, i)\mu(\boldsymbol{x}_{-i}, \boldsymbol{y}_{-i}) & (7) \\ &\propto \pi(\boldsymbol{x})\pi(\boldsymbol{y}) \prod_{i=1}^{n} \delta(\boldsymbol{x}_{-\{j:\, j<i\}}, \boldsymbol{y}_{-\{j:\, j<i\}}, j) & (8) \end{aligned}$$

where $\delta(\boldsymbol{x}_{-A}, \boldsymbol{y}_{-A}, i)$ is equal to one if there is a $u \in [0, 1)$ such that $x_i = \Upsilon_i(\boldsymbol{x}_{-A}, u)$ and $y_i = \Upsilon_i(\boldsymbol{y}_{-A}, 1 - u)$. Otherwise $\delta(\boldsymbol{x}_{-A}, \boldsymbol{y}_{-A}, i)$ is zero. The transition kernel is given by

$$p((\boldsymbol{x}, \boldsymbol{y}) \to (\boldsymbol{x}', \boldsymbol{y}')) = \frac{1}{n} \sum_{i=1}^{n} \pi(x_i'|\boldsymbol{x}_{-i})\pi(y_i'|\boldsymbol{y}_{-i})\ \delta(\boldsymbol{x}, \boldsymbol{y}, i)\ 1_{[\boldsymbol{x}_{-i}=\boldsymbol{x}'_{-i}, \boldsymbol{y}_{-i}=\boldsymbol{y}'_{-i}]}. \tag{9}$$

The δ-function describes the fact that common random numbers are used. It is interesting that the joint chain is also a Gibbs sampler with block updates on a larger space, $\Omega \times \Omega$. Each block consists of two indexes that correspond to the same V^t component in the two chains.

We shall use some results of Liu, Wong & Kong (1995) that need the following regularity condition to hold for π

$$\int \frac{\pi(\boldsymbol{x}^0, \boldsymbol{x}^1)^2}{\pi(\boldsymbol{x}^0)}\ \pi(\boldsymbol{x}^1)\ d\boldsymbol{x}^0\ d\boldsymbol{x}^1 < \infty. \tag{10}$$

Here $\pi(\boldsymbol{x}^0, \boldsymbol{x}^1)$ is the joint stationary density of $\boldsymbol{X}^0$ and $\boldsymbol{X}^1$. Condition (10) is quite common and it guarantees that the marginal Gibbs sampler chain is geometrically ergodic in terms of Pearson's χ^2-distance, and that the covariance between $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{X}^{t+k})$ goes to zero geometrically fast as $k \to \infty$. We shall assume that (10) is satisfied. Refer to Liu, Wong & Kong (1994) and Liu et al. (1995) for further discussion on (10).

3 Comparing Asymptotic Variances

To evaluate the performance of the antithetically coupled Gibbs sampler, we will compare the variance of $\hat{f}$ with the variance of $\hat{\hat{f}}$ (assumed to be finite). Let the coupled chains be started in the stationary distribution μ. We shall return to this assumption in Section 7, where we discuss the burn-in period.

Theorem 2 *Assume $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ is distributed according to μ. Let $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ be the stationary Markov chain defined by (5) and (6). Let*

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

be the marginal autocovariance at lag k of one of the two components, and let

$$\beta_k = cov(f(\boldsymbol{X}^0), f(\boldsymbol{Y}^k)), \qquad k = 0, 1, \ldots$$

be the cross-autocovariance at lag k of the two components. Then

$$\begin{aligned} T(var(\hat{f}) - var(\hat{\hat{f}})) &= \sum_{k=T}^{2T-1} \gamma_k + \frac{1}{2T} \sum_{k=1}^{T-1} k\gamma_k - \frac{1}{2T} \sum_{k=T}^{2T-1} k\gamma_k \\ &\quad -\beta_0/2 - \sum_{k=1}^{T-1} \beta_k + \frac{1}{T} \sum_{k=1}^{T-1} k\beta_k, \end{aligned} \tag{11}$$

where $\hat{f}$ is given in (2) and $\hat{\hat{f}}$ in (1), and the variances and covariances are computed with respect to the stationary measure over the appropriate trajectory space. Furthermore, when $T \to \infty$,

$$T(var(\hat{f}) - var(\hat{\hat{f}})) = o(1) - \beta_0/2 - \sum_{k=1}^{T-1} \beta_k. \tag{12}$$

When passing to the limit as $T \to \infty$ in equality (11), using (10), all terms involving the autocorrelations of the marginal chains disappear. The study of the sign of the right hand-side of (11) cannot be done analytically for a finite T, so we consider the case when $T \to \infty$. We want to show that the right hand side of (12) is asymptotically positive. This would be true if $\beta_k \leq 0$ for all k. For this reason we study the sign of the cross-autocovariances β_k in the following section.

Note that if the coupled chains are not started in equilibrium, then (12) still holds if we interpret β_0 as $\frac{1}{T} \sum_{t=1}^{T-1} \text{cov}(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^t))$ and β_k similarly.

4 The cross-autocovariance of the joint chain

We shall assume from now on and without loss of generality that the expected value of $f(\boldsymbol{X})$ is zero in order to simplify formulas. To be able to study in general the right hand side of (12) we need to restrict the space of functions f. Our algorithm induces an antithetic dependency structure between $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$. We want that structure to transfer to $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{Y}^t)$ as well.

DEFINITION 1 *Let $\mathcal{F}$ be the class of non-constant functions $f : \Omega \to \mathbb{R}$ whose i-th component is monotonically either decreasing or increasing in x_i whatever values $\boldsymbol{x}_{-i}$ takes, for each i.*

Some possible choices for $f \in \mathcal{F}$ are $f(\boldsymbol{x}) = \sum_i g_i(x_i)$, and $f(\boldsymbol{x}) = \prod_i g_i(x_i)$, where the $g_i(\cdot)$ are monotonic functions.

We start with the sign of the cross-autocovariance at lag zero, β_0. In order to show that (12) is positive it is important that β_0 is negative. A consequence of Lemma 2, used in the Appendix to prove Theorem 2, is that $|\beta_k| \leq \mathcal{O}(1/k)$ as $k \to \infty$, so that β_0 is the leading term in (12). The following theorem does not assume that the joint chains are in equilibrium.

THEOREM 3 *Consider the coupled random scan Gibbs sampler chains $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$, and assume $f \in \mathcal{F}$. For every $t \geq 1$ it holds that cov$(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^t)) \leq 0$. If the chain $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is stationary, then $\beta_0 \leq 0$. Furthermore, if var$(f(\boldsymbol{X})) > 0$, then $\beta_0 < 0$.*

Notice that $f \in \mathcal{F}$ is not necessary for $\beta_0 \leq 0$ to hold. The important fact is that f preserves antithetic dependency of $\boldsymbol{X}^t$ and $\boldsymbol{Y}^t$. It is not possible to improve the upper bound on β_0 without imposing further assumptions on π (Joe, 1997, pp. 81).

4.1 ATTRACTIVE OR REPULSIVE TARGET DENSITIES

We prove in the case of attractive or repulsive target distributions that $\beta_k \leq 0$, for all k, and hence the variance of $\hat{\hat{f}}$ is always less than the variance of $\hat{f}$ as $T \to \infty$. Our result are based on the technique of iterated conditional expectations introduced by Liu et al. (1994) and Liu et al. (1995).

For simplicity, let us first consider only the $\boldsymbol{X}^t$-chain. Let i_t be the *random variable* describing which site is updated in moving from $\boldsymbol{X}^t$ to $\boldsymbol{X}^{t+1}$. Notice that $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{X}^{t+1})$ are conditionally independent given $(\boldsymbol{X}^t_{-i_t}, i_t)$ (which we in the following write as $\boldsymbol{X}^t_{-i_t}$ for short), and that $f(\boldsymbol{X}^t_{-i_t})$ and $f(\boldsymbol{X}^{t+1}_{-i_{t+1}})$ are conditional independent given $\boldsymbol{X}^{t+1}$. Furthermore $f(\boldsymbol{X}^t)$ and $f(\boldsymbol{X}^t_{-i_t})$ have the same joint distribution as $f(\boldsymbol{X}^{t+1})$ and $f(\boldsymbol{X}^t_{-i_t})$, and this distribution does not depend on time t as a result of the stationarity. To illustrate the use of these facts, we write

$$\gamma_1 = \mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{X}^{t+1})) = \mathrm{E}(\mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{X}^{t+1})|\boldsymbol{X}^t_{-i_t})) = \mathrm{E}([\mathrm{E}(f(\boldsymbol{X})|\boldsymbol{X}_{-i})]^2) \geq 0 \qquad (13)$$

showing that γ_1 is positive for a random scan Gibbs sampler (Liu et al., 1995). Note that expectation with respect to $\boldsymbol{X}_{-i}$ is a shorthand for $\mathrm{E}_i(\mathrm{E}_{\boldsymbol{X}_{-i}}(\cdot|i))$. The expressions for γ_k get more complicated for higher order lags, see the cited references for details. Using this technique on the joint chain, we obtain the following formulas for the cross-autocovariances.

THEOREM 4 *The cross-autocovariances for the joint $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ chain can be expressed as*

$$\beta_k = \begin{cases} E(g^{(k)}(\boldsymbol{X}_{-i})g^{(k)}(\boldsymbol{Y}_{-i})), & \textit{for } k \textit{ odd} \\ E(g^{(k)}(\boldsymbol{X})g^{(k)}(\boldsymbol{Y})), & \textit{for } k \textit{ even} \end{cases} \qquad (14)$$

where

$$g^{(k)}(\cdot) = E(\ldots E(E(f(\boldsymbol{X})|\boldsymbol{X}_{-i})|\boldsymbol{X})|\boldsymbol{X}_{-i}\ldots)$$

is a sequence of k iterated conditional expectations alternating $\boldsymbol{X}_{-i}$ and $\boldsymbol{X}$.

Expression (14) makes it possible to interpret β_k as the cross-autocovariance of $g^{(k)}(\cdot)$ at lag zero. We can then make use of Theorem 3 to prove that under certain conditions $\beta_k \leq 0$ for all k, and hence that the antithetic estimate (1) is always better asymptotically than (2). The required conditions are that $f \in \mathcal{F}$ and that for each i

$$\mathrm{E}(f(\boldsymbol{X})|\boldsymbol{x}_{-i}) \tag{15}$$

is monotonically either increasing in $x_j, \forall j \neq i$, or decreasing in $x_j, \forall j \neq i$.

THEOREM 5 *Assume $f \in \mathcal{F}$ and $g^{(1)}(\boldsymbol{x}_{-i}) = E(f(\boldsymbol{X})|\boldsymbol{x}_{-i})$ is either monotonic increasing in $x_j, \forall j \neq i$, or monotonic decreasing in $x_j, \forall j \neq i$, then $\beta_k \leq 0, \forall k$.*

We can relate the condition of monotonicity of (15) to attractive and repulsive models. In the literature π is called attractive if

$$\pi(X_i \leq x_i|\boldsymbol{x}_{-i}) \leq \pi(X_i \leq x_i|\boldsymbol{x}'_{-i}), \qquad \text{for} \quad \boldsymbol{x}_{-i} \geq \boldsymbol{x}'_{-i},\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \Omega, \tag{16}$$

assuming the partial ordering of Ω given by $\boldsymbol{x}_A \geq \boldsymbol{x'}_A$ if $x_i \geq x'_i$ for all $i \in A$. π is repulsive if (16) holds with "$\geq$" in the first inequality. See Møller (1997) for many examples of such attractive and repulsive models. Define $\mathcal{F}^+$ and $\mathcal{F}^-$ as the set of functions $f \in \mathcal{F}$ that are monotonic increasing in all x_j or decreasing in all x_j, respectively. The monotonicity condition of (15) is guaranteed when $f \in \mathcal{F}^+ \cup \mathcal{F}^-$ and π is either attractive or repulsive.

Note that β_k will be negative for more general densities π and functions f, but our proof requires these assumptions.

4.2 APPROXIMATING THE EFFICIENCY

Although attractive or repulsive models are often encountered, we would like to extend our theory to general target densities and quantify the gain obtained using the new algorithm. We are however not able to do this rigorously. In this section we give some evidence that possibly in general $\beta_k \leq 0$, $\forall k$, for $f \in \mathcal{F}$. We shall show that approximately

$$\beta_k \approx \beta_0 \gamma_k / \gamma_0 \tag{17}$$

provided that

$$\mathrm{E}(f(\boldsymbol{X}^{t+k}) \mid \boldsymbol{x}^t) \approx \frac{\gamma_k}{\gamma_0} f(\boldsymbol{x}^t). \tag{18}$$

In other words, if the best linear k-step ahead predictor (in terms of $f(\boldsymbol{x}^t)$) is close to the k-step ahead conditional expectation, then $\beta_k \leq 0$, $\forall k$, and $|\beta_k|$ decays geometrically to zero. Note that (18) (and hence (17)) is exact for any linear f if π is multivariate normal. This is a significant special case. Because a multivariate normal is often the large sample limit in inference, we might expect (18) to hold more generally for large sample sizes. We will return to the multivariate normal distribution in Section 5.

Equation (17) follows from

$$\begin{aligned}\beta_k &= \mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+k})) = \mathrm{E}(\mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t)) \\ &= \mathrm{E}(f(\boldsymbol{X}^t)\mathrm{E}(f(\boldsymbol{Y}^{t+k}) \mid \boldsymbol{Y}^t)) \approx \mathrm{E}(f(\boldsymbol{X}^t)\, \frac{\gamma_k}{\gamma_0} f(\boldsymbol{Y}^t)) = \beta_0 \frac{\gamma_k}{\gamma_0},\end{aligned}$$

using (18). Using approximation (18), we can calculate the efficiency of $\hat{\hat{f}}$ compared to $\hat{f}$ as $T \to \infty$,

$$\mathrm{eff}(\hat{\hat{f}}, \hat{f}) \equiv \frac{\mathrm{var}(\hat{f})}{\mathrm{var}(\hat{\hat{f}})} \sim \frac{1}{1 + \beta_0/\gamma_0}.$$

As $\beta_0 < 0$, the new antithetic algorithm is always better if (18) holds. Thus, if the cross-autocorrelation at lag zero, β_0/γ_0, is equal to, say, $-2/3$, then the efficiency is approximately 3. In our experiments reported in Section 8, we always obtain efficiencies larger than three. This means that the computational costs can be reduced to at least one third. Note further that the efficiency under approximation (18) does not depend on γ_k, $k > 0$, which indicates that the efficiency does not depend on the specific mixing properties of the marginal chain, which is assumed to be geometrically ergodic by (10).

5 Target Densities with a Certain Local Symmetry

In this section we will present a striking result for the variance of $\hat{\hat{f}}$ when the target density π satisfies certain symmetry conditions fulfilled by the multivariate normal, the Ising model and a few others. Let $\nu(\boldsymbol{x}, \boldsymbol{y})$ be any density defined on $\Omega \times \Omega$ with the same positive support as $\pi \times \pi$.

Theorem 6 *For any linear f, let π be either the multivariate normal or the Ising model*

$$\pi(\boldsymbol{x}) \propto \exp(\beta \sum_{i \sim j} x_i x_j), \qquad x_i \in \{-1, +1\},$$

where the sum is taken over all nearest neighbours on a regular grid $\Lambda \subset \mathbb{Z}^2$. Then, var$(\hat{\hat{f}}) = 0$ if the joint chains start in μ, while var$(\hat{\hat{f}}) \leq \mathcal{O}_p(T^{-2})$ if the joint chains start in $\nu \neq \mu$, where $\mathcal{O}_p$ is with respect to the site updating distribution as $T \to \infty$.

The theorem is surprising because it shows that coupling two Gibbs sampler chains allows us to reduce variances by a full order of magnitude. Furthermore, because var$(\hat{\hat{f}}) = 0$, once the stationary distribution μ is reached, one joint sample is enough to estimate the expected value of $f(\boldsymbol{X})$. When π is a multivariate normal, the proof of the theorem is an immediate consequence of the following lemma, while some more work is needed for the Ising model.

LEMMA 1 *Assume $\pi(x_i|\boldsymbol{x}_{-i}) = \psi_i(x_i - \widetilde{x}_i), \forall i$, where $\psi_i(\cdot)$ is symmetric around zero, and $\widetilde{x}_i$ is the median in $\pi(x_i|\boldsymbol{x}_{-i})$ which can be written as $\widetilde{x}_i = \boldsymbol{A}_i\boldsymbol{x}_{-i}$ for some matrix $\boldsymbol{A}_i$. Then for any linear f, var$(\hat{\hat{f}}) = 0$ if the joint chain starts in μ, and var$(\hat{\hat{f}}) \leq \mathcal{O}_p(T^{-2})$ if the joint chains start in $\nu \neq \mu$, where $\mathcal{O}_p$ is with respect to the site updating distribution.*

It is this specific symmetry of the conditional density with respect to the median that is linear in the conditioning components that makes in some way the joint chain deterministic, as can be seen from the proof. If π is a multivariate normal, then this lemma holds 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 conditions of Lemma 1 is

$$\pi(\boldsymbol{x}) \propto \exp(-\sum_{i,j} a_{ij}(x_i - x_j)^2),$$

where $a_{ij} \geq 0$ and x_1, say, is fixed.

6 Locally Non-symmetric Target Distributions

Lemma 1 may indicate that an important property for achieving variance reduction with the new method is a certain type of symmetry of the conditional distributions. To gain more insight we will now study analytically the same coupling applied to two stationary autoregressive processes. These mimic the behavior of the two Gibbs sampler chains.

Let X^t be the real valued autoregressive process

$$X^t = \phi X^{t-1} + \epsilon_x^t, \qquad t > 0, \tag{19}$$

started in equilibrium at time zero. Here, $|\phi| < 1$ to ensure stationarity, and ϵ_x^t are i.i.d. binary variables with P$(\epsilon_x^t = 1) = p \geq 1/2$ and P$(\epsilon_x^t = 0) = 1 - p$. Although this is not a Gibbs sampler, it has the same flavor. It is known that the Gibbs sampler is a multivariate autoregressive process of order one if π is Gaussian. We chose $f(x) = x$ so that the goal is to estimate the mean E$(X) = p/(1-\phi)$. The variance of $\hat{f}$ is

$$\text{var}(\hat{f}) = \text{var}(\frac{1}{2T}\sum_{t=1}^{2T} X^t) \sim \tau_x/(2T), \quad \text{where} \quad \tau_x = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \tag{20}$$

is the *integrated autocovariance time*. We want to compare the variance in (20) with that obtained using two realizations of (19), X^t and Y^t, where X^t is sampled (forward in time) using the uniform variable U^t's and Y^t is sampled using $1 - U^t$. We compare (20) with the variance of $\hat{\hat{f}}$, where

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

By noting that Z^t is an autoregressive process of the same form as (19), with ϵ_z^t equal to one with probability $2p-1$ and equals otherwise $1/2$, we obtain the efficiency of $\hat{\hat{f}}$ compared to $\hat{f}$ as

$$\text{eff}(\hat{\hat{f}}, \hat{f}) \sim \frac{\tau_x}{2\tau_z} = \frac{1}{2-1/p}, \qquad p > 1/2, \tag{21}$$

where we make use of the exponentially decaying autocovariances of X^t and Z^t. This result shows that the antithetic estimate is always better, and that the efficiency tends to ∞ as the symmetry increases, i.e. $p \to 1/2$, and 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 of $\mathcal{O}(T^{-2})$. Notice that the length of the burn-in of the joint and marginal chain is similar, as both are autoregressive processes of the same form (19).

We compute the cross-autocovariance β_0 by comparing the asymptotic variance in (23) with the difference (12) in Theorem 2, and we obtain further

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

This results shows that (18) holds exactly also for this model. Furthermore β_k is minimal when $p = 1/2$ and $\beta_k = 0$ if $p = 1$. Also note that β_k is decreasing with increasing ϕ, $\phi > 0$. Because a large ϕ in (19) makes the estimation problem harder, one might therefore believe that there is an increasing benefit from using the antithetic idea for strongly positively correlated Gibbs sampler chains. However, the fact that ϕ cancels in (21) indicates that the efficiency of the new algorithm does not depend on the mixing properties of the marginal chain, see also Section 4.2.

7 The Burn-in

We have seen that for the autoregressive example in Section 6, the burn-in of the joint antithetic chain is of the same order as the burn-in of the marginal chain. In this section we will argue that this is a general picture: the burn-in of the joint chain $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ needed to reach the stationary distribution μ is generally of the same order as the burn-in of the single chain $\boldsymbol{X}^t$ converging to π. Assume for simplicity that Ω is finite.

We use a further coupling argument. Consider a new pair of chains $(\tilde{\boldsymbol{X}}^t, \tilde{\boldsymbol{Y}}^t)$ that evolves with the same transition kernel as $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ but is started in equilibrium μ and shares with $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ all random numbers. This means that $\boldsymbol{X}^t$ and $\tilde{\boldsymbol{X}}^t$ both use the random number U^t to evolve to $\boldsymbol{X}^{t+1}$ and $\tilde{\boldsymbol{X}}^{t+1}$, and $\boldsymbol{Y}^t$ and $\tilde{\boldsymbol{Y}}^t$ both use $1-U^t$ to evolve to $\boldsymbol{Y}^{t+1}$ and $\tilde{\boldsymbol{Y}}^{t+1}$. This is the natural coupling of $(\tilde{\boldsymbol{X}}^t, \tilde{\boldsymbol{Y}}^t)$ with $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$. Once the two pairs of chains meet at time t, they coalesce, and we know that $(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim \mu$. For coalescence of the two pairs of chains to happen, both the components have to coalesce, i.e. there are two random times ζ_x, ζ_y, such that $\tilde{\boldsymbol{X}}^{\zeta_x} = \boldsymbol{X}^{\zeta_x}$, and $\tilde{\boldsymbol{Y}}^{\zeta_y} = \boldsymbol{Y}^{\zeta_y}$. The coupling time for the two pair chains is $\tilde{\zeta} = \max\{\zeta_x, \zeta_y\}$. The marginal coupling times ζ_x and ζ_y are dependent but they do have the same marginal distribution. Moreover, if we repeat the whole argument for a single Gibbs sampler targeting π, the coupling time would have the same distribution as ζ_x

(and ζ_y). Therefore, coupling of the joint chains takes a time $\tilde{\zeta}$ that is similar to ζ_x, i.e. $\mathcal{O}_p(\tilde{\zeta}/\zeta_x) = 1$. Furthermore, it is easy to show that $\mathrm{E}(\tilde{\zeta}) \leq 2\mathrm{E}(\zeta_x)$. The motivation for using the coupling argument is that the total variation norm of the difference between μ and the density at time t of $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is less or equal to Pr$(\tilde{\zeta} > t)$. However, often the coupling inequality gives a reasonably good bound of the burn-in period. A further indication that the burn-in of the joint chain is of the same length as the burn-in of the marginal one comes from the exponential decay of the cross-autocovariances. In practice the length of the burn-in period is always taken to be considerably shorter than the T iterations used for averaging.

8 Numerical Experiments

In this section we 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 purposes are to evaluate the performance of the new algorithm for finite T and to quantify the efficiency w.r.t. the usual Gibbs sampler. We will also present antithetically coupled Metropolis-Hastings chains and discuss their performance.

8.1 Hierarchical Poisson Model

Gelfand & Smith (1990) present counts $\boldsymbol{s} = (s_1, \ldots, s_n)$ of failures in $n = 10$ pump systems at a nuclear power plant, where the times of operation $\boldsymbol{t} = (t_1, \ldots, t_n)$ for each system are known. The hierarchical model assumes $s_k \sim$ 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 $\lambda_2 \ldots \lambda_n$ 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

$$\pi(\alpha|\lambda_1, \ldots, \lambda_{10}, \beta) \propto \exp(\alpha a - n \log \Gamma(\alpha)), \quad \text{where} \quad a = n \log \beta + \sum_{k=1}^{n} \log \lambda_k - 1. \tag{22}$$

In this case it is most natural to perform a Metropolis-Hastings step when the α-parameter is updated. This means that, using a proposal density, a new value for α is proposed and then accepted or rejected. We suggest to couple the proposed values, while keeping the acceptance step independent. Here are three different updating strategies for α.

1. (Gibbs sampler update) To implement the full Gibbs sampler, we computed numerically $F^{-1}(u; a_x)$ and $F^{-1}(1-u; a_y)$, where F is the cumulative conditional distribution function (22) for α.

2. (Hastings update) We approximated the conditional density (22) with a normal ($\tilde{F}$) where the mean and variance match the mode and the curvature in the mode. We updated 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 estimated acceptance rate for α of 90%.
3. (Metropolis update) We updated α using a random walk Metropolis step and proposed a new state from a uniform density centred at the old state. The width of the proposal density was determined to obtain an estimated acceptance rate for α close to 50%.

To verify the robustness of our theoretical results with respect to various site visitation schedules, we applied each of these three updating rules for α to three different visiting schedules: RS, the usual random scan assumed above, where we look to 12 variable updates as one step; RPS, where at each iteration we update our 12 variables in a random permutation; and DET, where at each iteration we update $\lambda_1, \ldots, \lambda_{10}, \alpha, \beta$ and then $\beta, \alpha, \lambda_{10}, \ldots, \lambda_1$. All these visitation schedules give rise to a reversible Markov chain.

We ran a single Markov chain using $1\,000$ iterations as burn-in, and then we split the chain into two components and ran, according to (6) and (7) for $\lambda_1, \ldots, \lambda_{10}, \beta$, according to one of the three above methods for α, for a further $50\,000$ iterations. Figure 1 shows a small part of the sample paths for the β variables in the two chains, denoted by β_1^t and β_2^t respectively, where we used the Gibbs sampler also for α and RPS. The paths show a clear negative correlation. The first panel of Figure 2 shows the empirical joint density of the two coupled chains α_1^t and α_2^t, using $5\,000$ subsequent samples. The second panel of Figure 2 shows the empirical joint density of (β_1^t, β_2^t). Again, the negative cross-correlation structure is clearly visible. To give a quantitative measure of the variance reduction using the antithetic chains, we estimated the integrated autocovariance time using all $50\,000$ iterates and the approach of Geyer (1992) for reversible chains. The estimated efficiencies for α and β for different updating rules and visitation schedules, are listed in table 1.

The efficiencies in table 1 do not seem to depend on the visitation schedules. Our theory in Section 4.1 is valid only for the RS schedule, but it seems to be valid in practice for other visitation schedules too. The efficiencies for these Gibbs samplers are around 9 and 6 for α and β, respectively, which gives a significant reduction of the computational costs. However, the efficiencies drop to around $2-2.5$ for other types of update for α (Hastings update and Metropolis update). This occurs despite the fact that the acceptance rate was 90% for the Hastings-step. A further experiment with a random walk Metropolis update for α with increased width and an acceptance rate of 25%, still gave efficiencies around 2. The explanation for this effect is that the two antithetic chains get out of phase immediately when an antithetic proposal is rejected in one chain but not in the other. The sharing of the random numbers U^t is the only way we introduce antithetic dependency between the two chains. When an antithetic proposal is rejected by one of the two chains while being accepted by the other, the antithetic coupling between the two chains weakens. We do not adjust for this in later iterations, since only the random numbers are shared and no consideration is given to the states of the two chains in the proposal.

This could be generalized.

We also notice that in this case a single $2T$ long chain, using Gibbs sampling for $\lambda_1, \dots, \lambda_{10}, \beta$ and a Hastings update for α, as described in item 2 above, has an asymptotic variance that is larger than that-one of a single $2T$ long Gibbs sampler. Hence antithetic Gibbs sampling is better than a hybrid Gibbs sampler-Hastings algorithm.

8.2 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 our antithetic Gibbs sampler to investigate its efficiency in estimating the posterior mean of the parameters of interest.

Let Y_{ij} be the jth observation ($j = 1, \dots, n_i$) from the ith group $(i = 1, \dots, n_g)$. 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 $\tau.$ denotes the precision or inverse variance. The prior ordering constraint of the means θ_i is that $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{n_g}$. Gelfand et al. (1990) demonstrate that the Gibbs sampler is easy to implement even with the ordering constraint. We refer to Gelfand et al. (1990) for details about the Gibbs sampler and choices of the (flat) priors of the hyperparameters $a_1, a_2, b_1, b_2, \mu_0$ and τ_0.

We simulated our data set 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 empirical ordering of the means is not in agreement with the ordering constraint. We used the deterministic site visitation schedule DET with $1\,000$ burn-in's. The efficiency was estimated using the following $50\,000$ iterates of the coupled chains as in Section 8.1. Table 3 displays our estimates of the efficiencies for (θ_i, τ_i), $i = 1, \dots, n_g$. The new antithetic Gibbs sampler gives again a significant speedup with efficiencies between 2.97 and 6.69 with an average of 4.7. Similar results were obtained for the other visiting schedules.

9 Conclusions

We have suggested a simple way to couple two Gibbs sampler chains in order to reduce the variance of the sample average estimator of an expectation. The coupling induces antithetic cross-autocovariances. The reduction of the variance can be remarkable with respect to a simple Gibbs sampler run using the same computational time.

The coding of the proposed algorithm is an easy operation given a standard Gibbs Sampler implementation.

Other authors have tried to introduce antithetic behaviors into a single MCMC chain. Barone & Frigessi

(1989) propose a variation of the Gibbs sampler that moves antithetically to the current state and is shown to have a reduced burn-in in many cases. Neal (1998) complicates the single update further with the same aim of introducing negative correlations. Green & Han (1992) showed that in such a way also the asymptotic variance could be reduced in certain cases. We show, however, that it is with two chains that a complete antithetic behavior can be established.

As the example showed, it is not trivial to extend equally successfully this 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 antithetically.

Although our asymptotic analysis requires a random updating schedule of the variables, there is no reason to doubt that our conclusions can be extended to other types of scans. This is supported by the example in Section 8.1 and by the fact, that follows from the proof of Theorem 3, that β_0 is negative for general scans. Block updates can also be handled.

The Gibbs sampler is often not the fastest MCMC algorithm. In fact other Metropolis-Hastings schemes have 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 any other single site updating MCMC.

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.

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

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.

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*, 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.

JOE, H. (1997). *Multivariate Models and Dependence Concepts*, Monographs on Statistics and Applied Probability 73, Chapman & Hall.

LIU, J. S., WONG, W. H. & KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes, *Biometrica* 81(1): 27–40.

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.

MØLLER, J. (1997). Perfect simulation of conditionally specified models, *Technical report*, Department of Mathematics, Aalborg University, Denmark.

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.

Efficiency	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: The estimated efficiencies using $50\,000$ iterates and different choices for how to update the parameter α and different types of scan (RS: random scan, RPS: random permutation scan, and DET: deterministic scan). The antithetic coupling is highly efficient for the pure Gibbs sampler, but the efficiency decrease using a Hastings-update (with 90% acceptance) or a Metropolis-update (with 50% acceptance) 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: The sample values in the ordered normal means problem. Note the exchange in the empirical ordering of the means.

Efficiencies	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 efficiencies in the ordered normal means problem using $50\,000$ iterates.

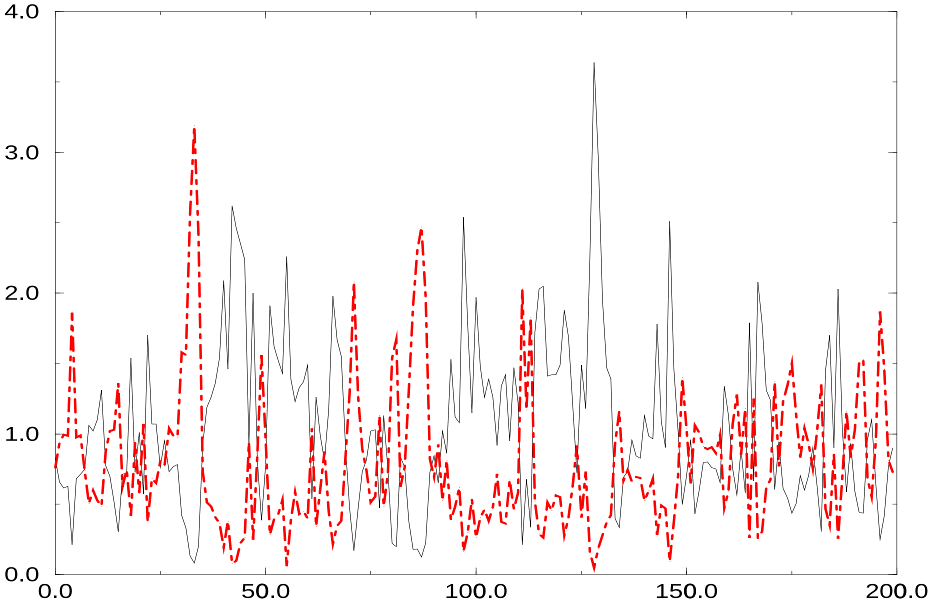
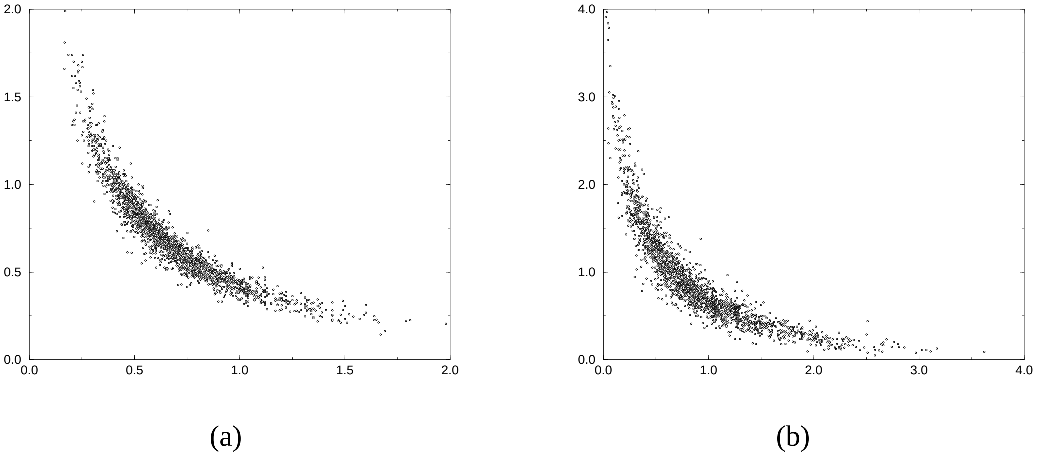


FIGURE 1: The sample-path for the β parameter for 200 iterations, in the two antithetic chains. The sample-paths show a clear negative correlation.



(a)

(b)

FIGURE 2: The empirical joint density of $5\,000$ samples from the two antithetic chains for the α-parameter is shown in (a), and for the β parameter is shown in (b). The negative correlation is clearly visible.

A Proofs

Proof of Theorem 1 The antithetic sampler on $\Omega \times \Omega$ updates components block-wise, where the block corresponds to $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$. Using the Theorem in the Appendix of Arjas & Gasbarra (1996) we conclude that the joint chain $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is ergodic and converges for each starting configuration $(\boldsymbol{X}^0, \boldsymbol{Y}^0) \in \Omega \times \Omega$ as long as $\pi(\boldsymbol{X}^0)\pi(\boldsymbol{Y}^0) > 0$. Note that $\mu(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ can be zero. The form of the transition kernel (9) follows directly using (5) and (6). It is easy to see that (9) is a block-wise Gibbs sampler step with respect to (7), which can be rewritten as (8) by iterating from the last factor. ■

Proof of Theorem 2 By stationarity it holds that

$$\text{var}(\hat{f}) = \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{2T-1}\gamma_k - \frac{1}{2T^2}\sum_{k=1}^{2T-1}k\gamma_k.$$

For the coupled chains we get

$$\text{var}(\hat{\hat{f}}) = \frac{1}{2T}\gamma_0 + \frac{1}{T}\sum_{k=1}^{T-1}\gamma_k - \frac{1}{T^2}\sum_{k=1}^{T-1}k\gamma_k + \frac{1}{2T}\beta_0 + \frac{1}{T}\sum_{k=1}^{T-1}\beta_k - \frac{1}{T^2}\sum_{k=1}^{T-1}k\beta_k, \tag{23}$$

using stationarity of the joint processes, the knowledge of the marginals and that β_k is in our case even in k. Some simplifications give the expression for the difference of variances (11).

To obtain (12) first observe that Corrolary 1 in Liu et al. (1995) and assumption (10) ensures that $\gamma_k \geq 0, \forall k$, and $\gamma_k = o(1/k)$ as $k \to \infty$. Hence all sums in (11) involving γ_k are of order $o(1/T)$ as $T \to \infty$. By applying the following lemma, (12) follows.

Lemma 2 $\frac{1}{T}\sum_{k=1}^{T-1} k\beta_k = o(1)$, *as* $T \to \infty$.

Proof of Lemma 2 Define $\boldsymbol{Z}^t = (f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t))/2$ and write $T\text{var}(\hat{\hat{f}})$ in two ways, in terms of $\boldsymbol{Z}^t$ and as (23). So for $T \to \infty$

$$\tilde{\gamma}_0 + 2\sum_{k=1}^{T-1}\tilde{\gamma}_k = \gamma_0/2 + \sum_{k=1}^{T-1}\gamma_k + \beta_0/2 + \sum_{k=1}^{T-1}\beta_k - \frac{1}{T}\sum_{k=1}^{T-1}k\beta_k,$$

where $\tilde{\gamma}_k = \text{cov}(f(\boldsymbol{Z}^0), f(\boldsymbol{Z}^k))$. Repeat this equality for $T+1$ and subtract term by term. We get as $T \to \infty$,

$$2\tilde{\gamma}_T = \gamma_T - \frac{1}{T(T+1)}\sum_{k=1}^{T-1}k\beta_k + \frac{1}{T+1}\beta_T.$$

From Liu et al. (1995) and Theorem 1, we know that $\tilde{\gamma}_T = o(1/T)$ and $\gamma_T = o(1/T)$, hence

$$2\beta_T - \sum_{k=1}^{T}k\beta_k/T = o(1) \tag{24}$$

Since $|\beta_k| \leq \gamma_0$, we have to rule out that $\sum_{k=1}^{T} k\beta_k/T =$ constant $= c$. If this was the case, write (24) for $T + 1$ to conclude that $c = o(1)$. Hence $\sum_{k=1}^{T} k\beta_k/T = o(1)$. Note that $|\beta_k| \leq \mathcal{O}(1/k)$ as $k \to \infty$. ■

PROOF OF THEOREM 3 We will show that

$$\mathrm{cov}(f(\boldsymbol{X}^t), f(\boldsymbol{Y}^t) \mid i_{t-1}, \boldsymbol{X}^{t-1} = \boldsymbol{x}^{t-1}, \boldsymbol{Y}^{t-1} = \boldsymbol{y}^{t-1}) \leq 0 \tag{25}$$

for every t, where conditioning on i_{t-1} indicate that we update site i_{t-1} in $\boldsymbol{X}^t$. This is then sufficient, integrating (25) with respect to the density of $(\boldsymbol{X}^{t-1}, \boldsymbol{Y}^{t-1})$, which does not need to be in equilibrium, and the uniformly density for i_{t-1} over the n sites. We simplify the notation. Let $A = f(\boldsymbol{X}^t)$, $B = f(\boldsymbol{Y}^t)$ and interpret A and B as functions of $X^t_{i_{t-1}}$ and $Y^t_{i_{t-1}}$ only. Further let all probabilities be conditioned on $(i_{t-1}, \boldsymbol{X}^{t-1} = \boldsymbol{x}^{t-1}, \boldsymbol{Y}^{t-1} = \boldsymbol{y}^{t-1})$. If A does not depend on $X^t_{i_{t-1}}$ (or similar with B) then the conditional covariance is zero. Assume now that A and B depend on $X^t_{i_{t-1}}$ and $Y^t_{i_{t-1}}$, respectively, and let A (B) have cumulative probability distribution $F(a)$ ($G(b)$). Denote by $H(a,b)$ the joint cumulative probability distribution of (A, B). Now we use that $A = F^{-1}(U)$ and $B = G^{-1}(1-U)$, which is valid as $f \in \mathcal{F}$. Then

$$H(a,b) = P(U \leq F(a),\ U \geq 1 - G(b)) = [F(a) + G(b) - 1]^+,$$

where $[\cdot]^+$ indicates the positive part. We insert (A) into the following general result of Hoeffding (see Joe (1997), pp. 55)

$$\mathrm{cov}(A,B) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (H(a,b) - F(a)G(b))\, da\, db,$$

and split the integral into two parts according to the value of $[F(a) + G(b) - 1]^+$. We obtain

$$\begin{aligned} \mathrm{cov}(A,B) \quad &= \quad -\int_{-\infty}^{+\infty}\int_{F^{-1}(1-G(b))}^{+\infty} (F(a)G(b) - F(a) - G(b) + 1)\ da\, db \\ &\quad\quad -\int_{-\infty}^{+\infty}\int_{-\infty}^{F^{-1}(1-G(b))} F(a)G(b)\ da\, db. \end{aligned}$$

The first integral is positive since $F(a)G(b) - F(a) - G(b) + 1 = (F(a) - 1)(G(b) - 1) \geq 0$. The second integral is of course always positive. ■

PROOF OF THEOREM 4 The proof for β_1 is similar to (13). Using that $(\boldsymbol{X}^t, \boldsymbol{Y}^t)$ is conditional independent of $(\boldsymbol{X}^{t+1}, \boldsymbol{Y}^{t+1})$ given $(\boldsymbol{X}^t_{-i_t}, \boldsymbol{Y}^t_{-i_t})$, and that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same marginal density, we obtain

$$\begin{aligned} \beta_1 \quad &= \quad \mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+1})) = \mathrm{E}(\mathrm{E}(f(\boldsymbol{X}^t)f(\boldsymbol{Y}^{t+1})|\boldsymbol{X}^t_{-i^t}, \boldsymbol{Y}^t_{-i^t})) \\ &= \quad \mathrm{E}(\mathrm{E}(f(\boldsymbol{X}^t)|\boldsymbol{X}^t_{-i_t})\mathrm{E}(f(\boldsymbol{Y}^{t+1})|\boldsymbol{Y}^t_{-i_t})) = \mathrm{E}(g^{(1)}(\boldsymbol{X}_{-i})g^{(1)}(\boldsymbol{Y}_{-i})), \end{aligned}$$

where $g^{(1)}(\boldsymbol{X}_{-i}) = \mathrm{E}(f(\boldsymbol{X})|\boldsymbol{X}_{-i})$. The expression for β_k is similar to the one for β_1, and makes repeated use of the iterated structure of $g^{(\cdot)}(\cdot)$. For $k \geq 0$ it holds that

$$g^{(2k+2)}(\boldsymbol{X}) = \mathrm{E}(g^{(2k+1)}(\boldsymbol{X}_{-i})|\boldsymbol{X}), \qquad g^{(2k+1)}(\boldsymbol{X}_{-i}) = \mathrm{E}(g^{(2k)}(\boldsymbol{X})|\boldsymbol{X}_{-i}) \tag{26}$$

where $g^{(0)}(\boldsymbol{X}) = f(\boldsymbol{X})$. Since the marginal distributions for $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the same, we obtain for the $\boldsymbol{Y}$ chain the same g-functions as in (26) using $\boldsymbol{Y}$ as the argument. A formal induction procedure, which we omit, will now give the proof. ■

PROOF OF THEOREM 5 First, $\beta_0 \leq 0$ from Theorem 3. To prove that $\beta_1 \leq 0$ we assume that $g^{(1)}(\boldsymbol{x}_{-i}) = \mathrm{E}(f(\boldsymbol{X})|\boldsymbol{x}_{-i})$ is monotonic increasing in $x_j, \forall j \neq i$, and hence belong of $\mathcal{F}$. If $g^{(1)}(\boldsymbol{x}_{-i})$ is monotonic decreasing then repeat this argument using $-f(\boldsymbol{x})$. We then apply Theorem 3 to show that $\beta_1 \leq 1$. To prove that $\beta_2 \leq 0$ we need to assure that $g^{(2)}(\boldsymbol{x})$ is monotonic increasing in $x_j, \forall j$. Using (26) we obtain

$$g^{(2)}(\boldsymbol{x}) = \mathrm{E}(g^{(1)}(\boldsymbol{X}_{-i})|\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} g^{(1)}(\boldsymbol{x}_{-j}),$$

which is monotonic increasing in all x_j. Hence and then our claim follows again from Theorem 3. In general, the fact that $\beta_k \leq 0$ for all k follows by repeating the above argument: for β_3 use the new function $f'(\boldsymbol{x}) = g^{(2)}(\boldsymbol{x})$ and the iterated structure for $g^{(\cdot)}(\cdot)$ in (26), to show that $\beta_3 \leq 0$ and so on. ■

PROOF OF LEMMA 1 The assumption on $\pi(x_i|\boldsymbol{x}_{-i})$ does also apply to $\pi(y_i|\boldsymbol{y}_{-i})$ since the marginals are the same. Assume $\boldsymbol{x}_{-i} + \boldsymbol{y}_{-i} = \boldsymbol{0}$. Then updating (x_i, y_i) to (x_i', y_i') with the transition kernel in Theorem 1 will ensure that $x_i' + y_i' = 0$ because $\psi_i(\cdot)$ is symmetric around zero and both medians $\widetilde{x}_i$ and $\widetilde{y}_i$ are linear in $\boldsymbol{x}_{-i}$ and $\boldsymbol{y}_{-i}$. The joint chain will therefore be absorbed by the event $\{\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{0}\}$ as soon as $\boldsymbol{x}_{-i}^t + \boldsymbol{y}_{-i}^t = \boldsymbol{0}$. From Theorem 1 we know that the joint chain is ergodic, so it follows that in this case $\mu(\boldsymbol{x}, \boldsymbol{y}) \propto \pi(\boldsymbol{x})\pi(\boldsymbol{y})1_{[\boldsymbol{x}+\boldsymbol{y}=0]}$. If the joint chain starts in μ then for any linear f, $f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t) \equiv 0 = \mathrm{E}f(\boldsymbol{X}), \forall t$, so that the variance is zero. On the other hand, assume that the joint chain starts in $\nu \neq \mu$ and that site i is to be updated. The new values will satisfy $x_i' + y_i' = \boldsymbol{A}_i(\boldsymbol{x}_{-i} + \boldsymbol{y}_{-i})$, due to the assumptions on Ψ_i. This defines a purely deterministic transition rule (conditional on the site updating sequence) for $\boldsymbol{X}^t + \boldsymbol{Y}^t$ and then also for $f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t)$. First notice that the simple Gibbs sampler cannot be ergodic if $\max_i ||\boldsymbol{A}_i|| > 1$. Next assume that $q = \max_i ||\boldsymbol{A}_i|| < 1$. Then $|f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t)|$ tends to zero not slower than q^t. Hence var$(\hat{\hat{f}}) \leq \mathcal{O}(T^{-2})/(1-q)^2$. Assume now that there exists a non empty set $J \subset \{1, 2, \ldots, n\}$ such that $||\boldsymbol{A}_i|| = 1$ for $i \in J$. Then $|f(\boldsymbol{X}^t) + f(\boldsymbol{Y}^t)|$ will stay constant whenever $i \in J$ is updated. Note that $|J| < n$ otherwise no joint equilibrium distribution can exist. Hence the decay to zero is slowed down by waiting times corresponding to updates of sites i in J. However the length of such waiting periods is geometrically distributed with rate $|J|/n < 1$. Hence, our claim var$(\hat{\hat{f}}) \leq \mathcal{O}_p(T^{-2})$ follows. ■

PROOF OF THEOREM 6 FOR THE ISING MODEL Consider the Ising model, whose conditional probabilities are

$$\Pr(x_i = 1|\boldsymbol{x}_{-i}) = 1/(1 + \exp(-2\beta \sum_{i \sim j} x_j)) = p^+_{\boldsymbol{x}_{-i}} = 1 - p^-_{\boldsymbol{x}_{-i}}.$$

Let $p^+_{\boldsymbol{y}_{-i}}$ and $p^-_{\boldsymbol{y}_{-i}}$ denotes the same conditional probabilities for $\boldsymbol{y}$. Assume $\boldsymbol{x}_{-i} + \boldsymbol{y}_{-i} = \boldsymbol{0}$. Then $p^+_{\boldsymbol{x}_{-i}} = p^-_{\boldsymbol{y}_{-i}}$. Hence $x'_i + y'_i = 0$, so that $\mu(\boldsymbol{x}, \boldsymbol{y}) \propto \pi(\boldsymbol{x})\pi(\boldsymbol{y})1_{[\boldsymbol{x}+\boldsymbol{y}=0]}$. The proof now proceeds as in Lemma 1. ■