MULTIPROCESS PARALLEL ANTITHETIC COUPLING FOR BACKWARD AND FORWARD MARKOV CHAIN MONTE CARLO[1]

BY RADU V. CRAIU AND XIAO-LI MENG

University of Toronto and Harvard University

Antithetic coupling is a general stratification strategy for reducing Monte Carlo variance without increasing the simulation size. The use of the antithetic principle in the Monte Carlo literature typically employs two strata via antithetic quantile coupling. We demonstrate here that further stratification, obtained by using $k > 2$ (e.g., $k = 3$–10) antithetically coupled variates, can offer substantial additional gain in Monte Carlo efficiency, in terms of both variance and bias. The reason for reduced bias is that antithetically coupled chains can provide a more dispersed search of the state space than multiple independent chains. The emerging area of perfect simulation provides a perfect setting for implementing the k-process parallel antithetic coupling for MCMC because, without antithetic coupling, this class of methods delivers genuine independent draws. Furthermore, antithetic backward coupling provides a very convenient theoretical tool for investigating antithetic forward coupling. However, the generation of $k > 2$ antithetic variates that are *negatively associated*, that is, they preserve negative correlation under monotone transformations, and *extremely antithetic*, that is, they are as negatively correlated as possible, is more complicated compared to the case with $k = 2$. In this paper, we establish a theoretical framework for investigating such issues. Among the generating methods that we compare, Latin hypercube sampling and its iterative extension appear to be general-purpose choices, making another direct link between Monte Carlo and quasi Monte Carlo.

1. Paired antithetic coupling. Monte Carlo estimation of the expectation of an estimand function f with respect to a probability measure π, $\mu = \int f(x)\pi(dx)$, can be done in many ways. The simplest method, known as "crude Monte Carlo," proceeds by simulating a (not necessarily independent) sample $X_1, X_2, \ldots, X_n$ from π and estimating μ by the sample average $\hat{\mu}_n = \frac{1}{n}\sum_i f(X_i)$. More refined methods, collectively known as *swindles*, take advantage of well-known statistical principles to construct more efficient designs and/or more efficient estimators. In this paper we focus on antithetic coupling, which can be viewed as a way to induce

stratification. For other type of swindles see Hammersley and Handscomb (1965), Simon (1976), Kennedy and Gentle (1980) and Gentle (1998), among others.

In the context of classic independent sampling, antithetic coupling [Hammersley and Morton (1956)] is commonly described as a method of producing two negatively correlated copies of an unbiased estimator. The average of the two, each based on m draws, will then have smaller variance compared to the estimator based on the $n = 2m$ independent draws. Substantial reductions have been documented throughout the literature from early works, such as Page (1965) and Fishman (1972), to most recent ones such as Frigessi, Gåsemyr and Rue (2000), who demonstrate the power of the antithetic principle in the more complicated (forward) MCMC setting.

The negative correlation between the two copies is typically induced via antithetic quantile coupling by using a pair $\{U, 1 - U\}$, where $U \sim \text{Uniform}(0, 1)$, in the sample generating process. The amount of variance reduction is governed by the degree of symmetry in the distribution of our estimator. In the extreme case when this distribution is symmetric, the use of paired antithetic variates can entirely eliminate the Monte Carlo variance when the underlying draws are independent or reduce the variance from the usual n^{-1} rate to n^{-2} rate when the draws are realizations of a genuine MCMC algorithm, as observed by Frigessi, Gåsemyr and Rue (2000). This emphasizes the stratification aspect of the antithetic principle as a way to divide the sample space into a "negative" stratum and a "positive" stratum. The equal amount of draws from each stratum ensured through *pairing* brings in further variance reduction compared to using simple random sampling within each stratum. Generalizing the antithetic coupling from $k = 2$ to $k > 2$ processes is quite natural from this stratification point of view, as often more than two strata can produce substantial additional gain. A main goal of this paper is to demonstrate such gains in the context of MCMC, as well as the additional benefit of improving the mixing of the original Markov chains.

The paired quantile coupling via $\{U, 1 - U\}$ has the following *extreme antithesis* (EA) property that is usually not emphasized in the literature. Specifically, if F is an *arbitrary* univariate cumulative distribution function (CDF), and $X_1 = F^{-1}(U)$, $X_2 = F^{-1}(1 - U)$, where $U \sim \text{Uniform}(0, 1)$, then $\text{Corr}(X_1, X_2)$ achieves *the minimal possible value* subject to the constraint that $X_1, X_2 \sim F$. The proof [Moran (1967)] relies on the elegant Hoeffding identity

$$(1.1) \qquad \text{Cov}(X_1, X_2) = \int \int [F(x_1, x_2) - F(x_1)F(x_2)]\mu(dx_1 \, dx_2),$$

and the equally elegant Fréchet (1951) inequality

$$(1.2) \qquad \max\{F(x_1) + F(x_2) - 1, 0\} \le F(x_1, x_2) \le \min\{F(x_1), F(x_2)\},$$

where $F(x_1, x_2)$ is the joint CDF. The fact that the single strategy of using $\{U, 1 - U\}$ achieves EA *simultaneously for all F* also implies that it achieves EA for $\text{Corr}(g(X_1), g(X_2))$, where g is any monotone function such that

$\int g^2(x) F(dx) < \infty$. It is known that negative correlation is not even preserved by monotone transformations, so the above discussion implies that $\{U, 1 - U\}$ must satisfy a stronger condition, *negative association* (NA) [Joag-Dev and Proschan (1983)], which requires exactly that the negative correlation be preserved by monotone functions of the variates (see Section 3.1). The fact that $\{U, 1 - U\}$ is automatically NA appears to be responsible for the general silence of the NA requirement in the literature of antithetic coupling, but once we move beyond $k = 2$, NA (as well as EA) becomes a key notion in our general theoretical foundation.

The rest of the paper is divided into three sections. Section 2 presents the general k-process parallel antithetic coupling technique, illustrated with a backward MCMC mixture sampling and two forward MCMC sampling schemes. However, for $k > 2$, as we demonstrate in Section 3, there is no general strategy that achieves EA simultaneously even just for uniform and normal distributions. As a result, it is harder to ensure NA and EA for a given problem, especially in theory. Section 3 is thus devoted to a general theory for ensuring NA and EA for arbitrary k, which is our main theoretical contribution. Section 4 discusses several common methods for generating k antithetic variates and their general properties with respect to achieving EA and NA.

2. k-process parallel antithetic coupling and illustration.

2.1. *Perfect simulation and time-backward dual sequence.* Since the seminal work by Propp and Wilson (1996) on *coupling from the past* (CFTP), there has been an array of methodological and theoretical papers on how to use backward coupling methods for exact MCMC sampling. David Wilson's web site (http://dimacs.rutgers.edu/~dbwilson/exact) is the most comprehensive single source for learning about the fast-growing field of *exact sampling* or *perfect simulation*, so named because, for this class of sampling methods, the thorny issue of deciding the running time for an acceptable error in approximating the distribution of interest, π, completely vanishes. A CFTP algorithm, or Fill's algorithm (1998), or many of their variations and generalizations [e.g., Fill, Machida, Murdoch and Rosenthal (2000), Meng (2000) and Wilson (2000b)] will terminate itself with probability 1 in a finite amount of time and yet delivers authentic (and hence *exact/perfect*) independent draws from the limiting distribution.

This can be achieved by constructing, for a Markov chain $\{X_t\}_{t \geq 0}$ with stationary distribution π, the "dual" sequence $\tilde{X}_t$ such that it has the same distribution as X_t but it purposely violates the Markovian property. This is perhaps most easy to illustrate by first considering *a time-inhomogeneous* Markov chain defined by $X_t = \psi_t(X_{t-1}, U_t)$, where $\{U_t, t \in \mathcal{N}\}$ are i.i.d. random variables. For any $t > 0$ and any starting value X^*, we can compute a *time-forward* sequence

$$(2.1) \qquad X_t = \psi_t(\psi_{t-1}(\ldots \psi_1(X^*, U_1) \ldots, U_{t-1}), U_t),$$

as well as a *time-backward* sequence

$$\tilde{X}_t = \psi_1(\psi_2(\ldots \psi_t(X^*, U_t)\ldots, U_2), U_1). \tag{2.2}$$

Clearly, X_t is Markovian because $X_t = \psi_t(X_{t-1}, U_t)$, but $\tilde{X}_t$ is not.

Evidently, when ψ_t is time-homogeneous, X_t and $\tilde{X}_t$ have identical distributions for any t. What is gained by giving up the Markovian property is that the backward sequence $\tilde{X}_t$ can "hit and stay at" the limit [Thorisson (2000), Chapter 1] in a finite amount of time when its forward counterpart X_t cannot. This is easiest to see in an extreme case where $\psi(X, U) = U$, which, say, is uniform on $(0, 1)$. Then it is obvious that $X_t = U_t$, but $\tilde{X}_t = U_1$, for all $t \geq 1$. In other words, while both X_t and $\tilde{X}_t$ converge, trivially, to Uniform$(0, 1)$ in distribution, X_t wanders off from U_t to U_{t+1}, even if it hits the right limiting distribution already at $t = 1$. In contrast, $\tilde{X}_t$ stays at the exact same value U_1 for all $t \geq 1$. In general, assuming X_t is uniformly ergodic, one can conclude that while X_t converges *in distribution*, there exist an updating ψ and a finite stopping time T such that $\tilde{X}_T$ hits the same limit *with probability* 1 [Foss and Tweedie (1998)]. By mapping t to $-t$ in (2.2), and thereby creating a convenient forward execution of the time-backward sequence given in (2.2), Propp and Wilson (1996) devise the CFTP method for finding such a T.

At the crux of CFTP implementation is the ability to follow a large, possibly infinite, number of paths in time and to assess whether these paths have all merged after a certain time T. For a good introduction to CFTP we refer to Casella, Lavine and Robert (2001). In real applications, the most challenging problem is that for many routine statistical problems, such as posterior sampling, the state space is both uncountable and unbounded. Although intense work has been done in this area, in terms of both general strategies and specific implementation [e.g., Kendall (1998), Murdoch and Green (1998), Møller and Nicholls (1999), Møller (1999), Murdoch (2000), Hobert, Robert and Titterington (1999), Casella, Mengersen, Robert and Titterington (2002) and Philippe and Robert (2003)], the difficulties exhibited in Murdoch and Meng (2001) in the context of posterior sampling with a t likelihood and normal mixture priors indicate that much more research is needed before CFTP and its variants and extensions can become a method of choice in common statistical applications.

2.2. *Multiprocess antithetically coupled CFTP.* Since CFTP delivers i.i.d. draws from the target density, when viewed as a "black box," it is no different from many classical Monte Carlo sampling methods, such as inverse CDF transform or rejection sampling. It is then natural to consider antithetic coupling for CFTP. However, unlike classical methods, the CFTP black box is a mapping from an infinite product space for $\{U_t, t \leq 0\}$ to the state space S, which makes the underlying theory for guaranteeing NA (and EA) of the samples more complicated than those available in the literature. A theoretical foundation is therefore needed, and this is established in Section 3. Here we focus on the description of the method

itself.

At the core of our method is the generation of k negatively associated $\{U^{(1)}, \ldots, U^{(k)}\}$, each having the same marginal distribution as the U required by $\psi(\cdot, U)$. There are many ways for doing so for a given distribution of U; see Section 4. Given such a method, one can run k CFTP processes in parallel, the jth one using $\{U_t^{(j)}, t \leq 0\}$, $j = 1, \ldots, k$, where $\{U_t^{(1)}, \ldots, U_t^{(k)}\}$, $t \leq 0$, are i.i.d. copies of $\{U^{(1)}, \ldots, U^{(k)}\}$, as sketched in Figure 1. Within the jth process of CFTP all paths are positively coupled because they use the same $\{U_t^{(j)}, t \leq 0\}$. At each update, $\{U_t^{(1)}, \ldots, U_t^{(k)}\}$ are NA, a property that clearly does not alter the validity of each individual CFTP process.

To obtain $n = km$ draws, we repeat the above procedure *independently m* times and collect $\{X_i^{(j)}, 1 \leq i \leq m; 1 \leq j \leq k\}$, where i indexes the replication, as our sample $\{X_1, X_2, \ldots, X_n\}$. Let $\sigma_f^2 = \mathrm{Var}_\pi[f(X)]$ and $\rho_k^{(f)} = \mathrm{Corr}_\pi(f(X_1^{(1)}), f(X_1^{(2)}))$, which is intended to be negative. Then

$$(2.3) \qquad \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \frac{\sigma_f^2}{n}[1 + (k-1)\rho_k^{(f)}].$$

Consequently, the *variance reduction factor* (VRF) relative to the independent sampling *with the same simulation size*, is

$$(2.4) \qquad S_k^{(f)} = 1 + (k-1)\rho_k^{(f)}.$$

We emphasize here the dependence of $S_k^{(f)}$ on k and more importantly on f, and
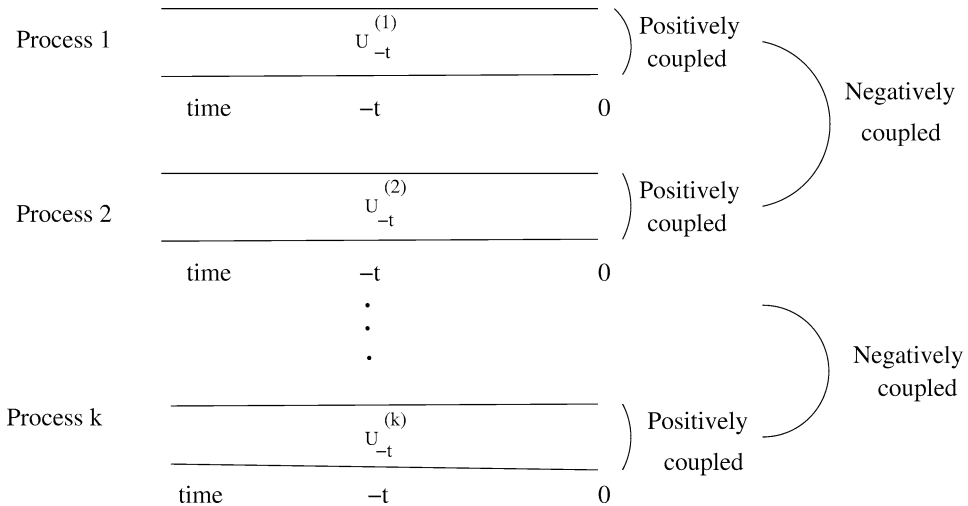

FIG. 1. *Parallel antithetic backward CFTP processes.*

thus the actual gain in reduction can be of practical importance for some f but not for others.

The size-fixed VRF $S_k^{(f)}$ ignores the possible different computational requirements between generating k independent draws and k antithetically coupled draws. A completely fair comparison is typically impossible with any simulation study because the overall computational time is seldom exactly linear in size and it can depend critically on software, hardware, programming skill and the actual implementation. Nevertheless, a useful approximation can be derived by assuming linearity and by ignoring any "overhead." Specifically, let τ_k be the average CPU time needed to make a *joint* draw from a particular antithetic k-process. Then, under our assumptions, for a given total CPU time T, the average number of dependent draws we can make is $n_{\text{dep}} = Tk/\tau_k$, and the number of independent draws is $n_{\text{indep}} = T/\tau_1$. Consequently, the *time-fixed* VRF is given by

$$(2.5) \qquad T_k^{(f)} = C_k S_k^{(f)} \qquad \text{where } C_k = \frac{\tau_k/k}{\tau_1}.$$

Note C_k is free of f, but a large C_k may offset the gain in $S_k^{(f)}$ (i.e., by making $T_k^{(f)} > 1$), as seen in the next section. Also note that, although we use the τ_1 notation for consistency, there can be a substantial difference between setting $k = 1$ in a general k-process subroutine and using a specifically designed subroutine for making independent draws; the latter was used in all our examples.

2.3. *An illustration with a mixture CFTP.* Hobert, Robert and Titterington [HRT (1999)] considered the mixture model $pf_0(x) + (1 - p)f_1(x)$ where $0 < p < 1$ is the estimand. Given i.i.d. observations $\{x_1, x_2, \ldots, x_n\}$ from this model, and a flat prior on p, the posterior for p is proportional to $\prod_{i=1}^n [pf_0(x_i) + (1 - p)f_1(x_i)]$. To deal with the continuous distribution of p, HRT use the natural discrete data augmentation that comes with the latent mixture indicator $\{z_1, z_2, \ldots, z_n\}$, where $z_i = 0$ if x_i is from f_0 and $z_i = 1$ if x_i is from f_1. Starting from $T = 1$, the following steps define the HRT algorithm:

1. Start a "bottom chain" at $p_{-T}^{(1)} = 0$ and a "top chain" at $p_{-T}^{(2)} = 1$.
2. At each $-T \le t \le -1$ and for $j = 1, 2$: generate n i.i.d. uniform r.v.'s $\{u_{t1}, u_{t2}, \ldots, u_{tn}\} \equiv u_t$, and then set $z_{ti}^{(j)} = 0$ if $u_{ti} \le \dfrac{p_t^{(j)} f_0(x_i)}{p_t^{(j)} f_0(x_i) + (1-p_t^{(j)}) f_1(x_i)}$ and $z_{ti}^{(j)} = 1$ otherwise.

   Let m be the number of $z_{ti}^{(j)}$'s equal to 0 and let $w_t = \{w_{tr}, r = 1, \ldots, n+2\}$, where the w_{tr}'s are i.i.d. samples from an exponential distribution with mean 1. Compute $p_{t+1}^{(j)} = \psi(p_t^{(j)}, u_t, w_t) = \sum_{r=1}^{m+1} w_{tr} / \sum_{r=1}^{n+2} w_{tr}$.

3. If $p_0^{(1)} = p_0^{(2)} = p_0$, then p_0 is our sample. If not, set $T_{\text{old}} = T$ and go to step 2 with $T = 2T_{\text{old}}$ while keeping u_{ti}, w_{tr} for all $-T_{\text{old}} \leq t \leq 0$, $1 \leq i \leq n$, $1 \leq r \leq n + 2$ unchanged.

Although $\psi(p, u, w)$ of step 2 is not monotone (in the same direction) in all its arguments, the variance reductions obtained with antithetic variates are still significant in all the simulation examples we have looked at. Figure 2 gives an illustration based on $0.33 \cdot \mathcal{N}(3.2, 3.2) + (1 - 0.33) \cdot \mathcal{N}(1.4, 1.4)$. The first two columns in Figure 2 plot the size-fixed VRF $S_k^{(f)}$ against k. The plots on each row correspond to a different generating method: the permuted displacement (PD) method, Latin hypercube sampling (LHS) and iterative Latin hypercube sampling (ILHS); see Section 4. The simulation size here is 7500 for each k, and the Monte Carlo variance for estimating $S_k^{(f)}$ is on the order of 10^{-6}. We see that $S_k^{(f)}$ decreases from 0.5–0.6 with $k = 2$ to 0.2–0.3 when $k \geq 6$ (all numbers are from the ILHS method, which performs best). It appears that $S_k^{(f)}$ stabilizes after about $k = 10$. The second column shows the results for three nonmonotone
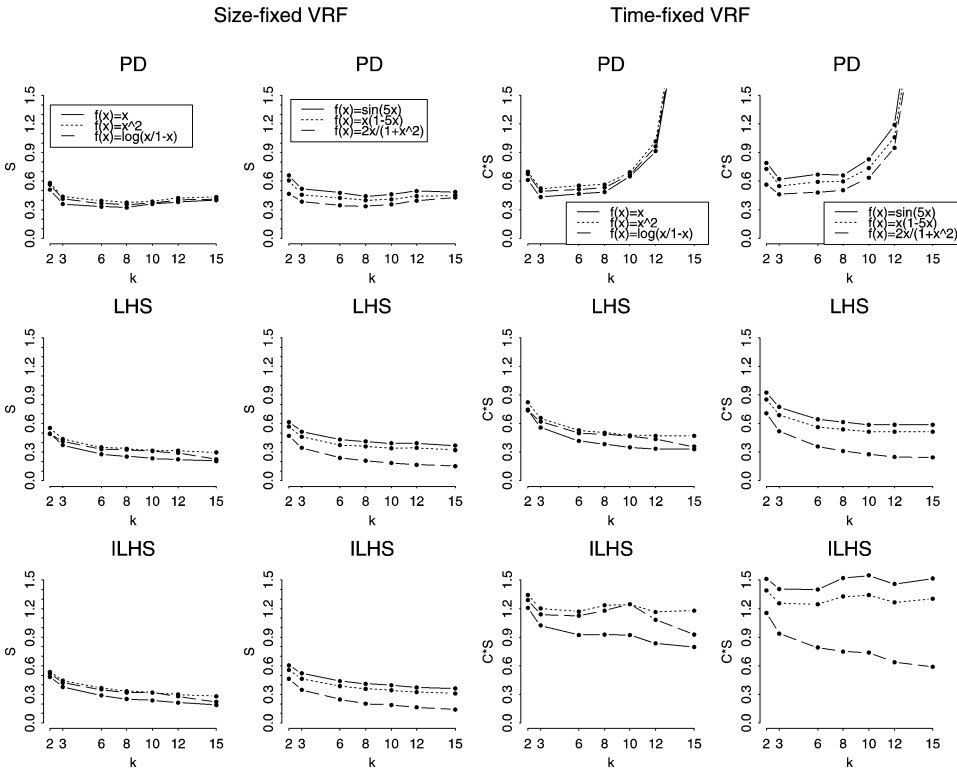

FIG. 2. *Normal mixture example. Size-fixed variance reduction factor* (left) *and time-fixed variance reduction factor* (right) *plotted against the number of parallel chains for different functions. Note the values of the time-fixed VRF for PD at $k = 15$ are too large for the plotting range.*

estimand functions. The theoretical guarantee given in Section 3 does not apply to such functions, but nevertheless all $S_k^{(f)} < 1$. Even for $f(x) = \sin(5x)$, $S_k^{(f)}$ is around 0.4, implying a 60% reduction in variance when $k \geq 6$.

The last two columns in Figure 2 plot the time-fixed VRF, $T_k^{(f)}$ of (2.5). From the first row, it is seen that using $k \geq 10$ becomes too costly for PD, at least in our implementation. In contrast, the monotone patterns for the second and third rows resemble that for $S_k^{(f)}$, albeit the reduction is less because the C_k factor tends to be more than 1. Nevertheless, we still see that $T_k^{(f)}$ reduces from 0.8 with $k = 2$ to about 0.5 when $k \geq 6$, and the empirical finding that using $k \geq 10$ is not practically beneficial remains true. These conclusions are based on the best performing method, LHS (ILHS is obviously more costly because of the iteration). However, we emphasize that in our implementation we have made no attempt to optimize our code. It is thus important to separate the gain in statistical efficiency, as measured by $S_k^{(f)}$, which does not depend on the particular implementation, from the possible offset, as measured by C_k, which depends critically on the particular implementation and thus could be further reduced with a more refined code/implementation.

2.4. *Forward antithetic coupling and slice sampling.* The next application used for illustration is a forward slice sampler. A graphical scheme of the antithetic principle applied to forward chains is shown in Figure 3. In this situation the parallel chains are started from different points in the sample space. After the so-called *burn-in period*, the realizations of each path are used, typically as positively correlated samples from the stationary distribution of the chain. Like CFTP, we
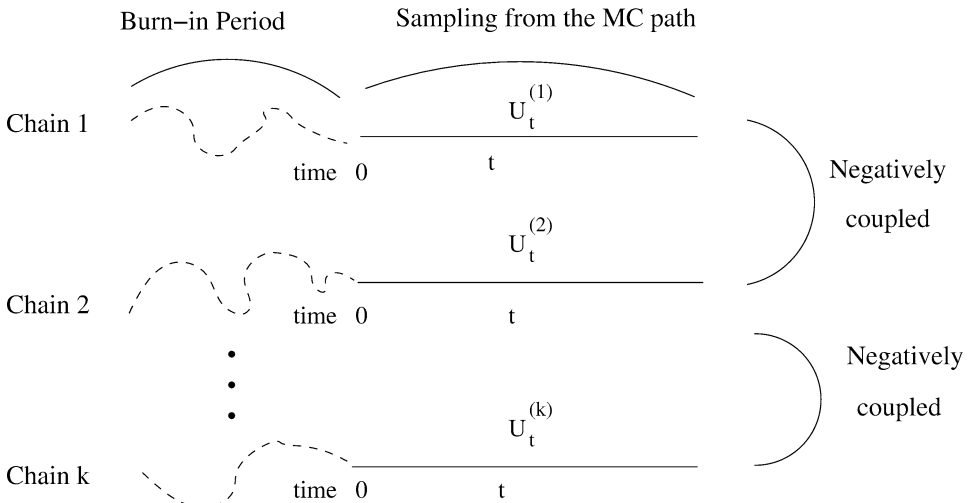

FIG. 3.   *Parallel antithetic forward MCMC chains.*

can update k chains using k antithetically coupled variates. However, unlike CFTP, the within-chain autocorrelation for the forward implementation makes the determination of the VRF a bit more complicated. Specifically, we need to generalize (2.4) to

$$S_k^{(f)} = 1 + \frac{\sum_{j_1 \neq j_2}^{k} \text{Cov}(\sum_{t=1}^{m} f(X_t^{(j_1)}), \sum_{t=1}^{m} f(X_t^{(j_2)}))}{\sum_{j=1}^{k} \text{Var}(\sum_{t=1}^{m} f(X_t^{(j)}))}$$

(2.6)

$$= 1 + \frac{\sum_{j_1 \neq j_2}^{k} \sum_{t_1,t_2}^{m} \beta_{t_1,t_2;j_1,j_2}^{(f)}}{\sum_{j=1}^{k} \sum_{t_1,t_2}^{m} \gamma_{t_1,t_2;j}^{(f)}},$$

where $\gamma_{t_1,t_2;j}^{(f)}$ and $\beta_{t_1,t_2;j_1,j_2}^{(f)}$ are respectively, the *within-chain* and *between-chain* autocovariance, that is,

(2.7)
$$\gamma_{t_1,t_2;j}^{(f)} = \text{Cov}(f(X_{t_1}^{(j)}), f(X_{t_2}^{(j)})),$$
$$\beta_{t_1,t_2;j_1,j_2}^{(f)} = \text{Cov}(f(X_{t_1}^{(j_1)}), f(X_{t_2}^{(j_2)})).$$

In many applications expression (2.6) can be greatly simplified because of the ergodicity of the antithetically coupled joint chain (which is *not* an automatic consequence of the ergodicity of the marginal chains; see Section 3.2). Here we retain the full generality of (2.6) to show that as long as all the between-chain covariances are nonpositive, the use of antithetic coupling will always guarantee $S_k^{(f)} \leq 1$ regardless of the mixing properties of the individual chains (e.g., still in burning periods) or its within-covariance structure—that is, using antithetic coupling cannot hurt as long as $\beta_{t_1,t_2;j_1,j_2}^{(f)} \leq 0$ for all (t_1, t_2, j_1, j_2); see Section 3 for conditions to guarantee this. The definition of time-fixed VRF remains the same as in (2.5) because C_k is unaffected by the autocorrelation.

To illustrate possible gains, we adopt an example of Damien, Wakefield and Walker (1999), who use slice sampling [Neal (2003)] to draw from $\pi(x) \propto x^2 e^{-e^x} I_{\{x \geq 0\}}$. The resulting Gibbs sampler has the updating function,

(2.8) $$X_{t+1} = \psi(X_t, \xi_1, \xi_2) = \xi_1^{1/3} \log(e^{X_t} - \log(1 - \xi_2)),$$

where ξ_1 and ξ_2 are i.i.d. Uniform(0, 1). Since $\psi(x, \xi_1, \xi_2)$ is nondecreasing, Theorem 1 of Section 3 guarantees that $S_k^{(f)} \leq 1$ for any monotone function f. Note (2.8) is an example where ψ is *made* to be nondecreasing by using $1 - \xi_2$ instead of ξ_2 inside the second logarithm.

Figure 4 is the counterpart of Figure 2 for the forward case, with similar general patterns (the simulation size here is 5000). In particular, for monotone functions $S_k^{(f)}$ decreases from between 0.35–0.45 with $k = 2$ to 0.1–0.15 when $k \geq 6$. For highly nonmonotone $f(x) = \sin(5x)$, in contrast to $k = 2$, $S_k^{(f)} \leq 1$ when $k \geq 3$. For less variable nonmonotone functions $f(x) = 2x(1 + x^2)^{-1}$ and $f(x) = x(1 - 5x)$, $S_k^{(f)}$ goes below 0.5. Partial theoretical support for this phenomenon can be
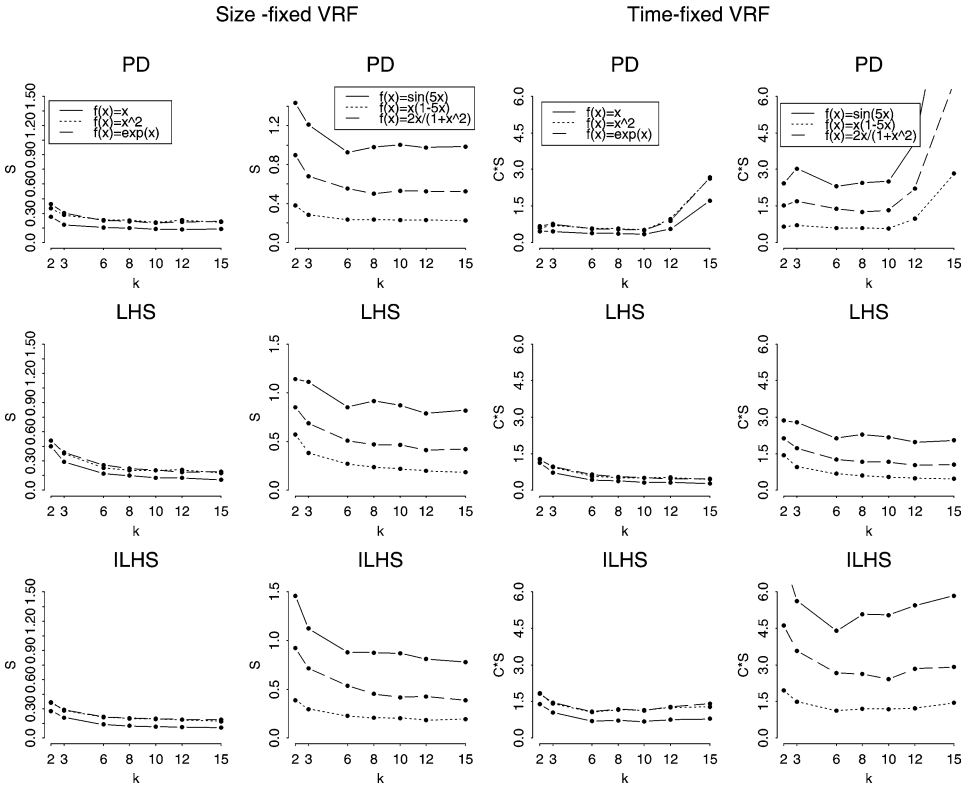
FIG. 4.   *Slice sampling example. Size-fixed variance reduction factor* (left) *and time-fixed variance reduction factor* (right) *plotted against the number of parallel chains for different functions. Note the change of scale in the vertical axis from the sized-fixed VRF to time-fixed VRF.*

found in Owen (1997), whose result implies that, under LHS, $S_k^{(f)} \leq k/(k-1)$ for any square-integrable f. This suggests that there is more room for $S_2^{(f)}$ to exceed 1 than for $S_k^{(f)}$ with $k > 2$. However, in terms of the time-fixed VRF, the clear winner is the LHS method with monotone functions. For highly nonmonotone functions, our implementation becomes too costly, suggesting that if such estimand functions are of main interest, then it would be generally safer to just use the independent implementation of multiple chains, as recommended in Gelman and Rubin (1992).

2.5. *A real-data application with Bayesian probit regression.*   We conclude our empirical investigation by presenting a multidimensional real-data application. The data are taken from van Dyk and Meng (2001) and consist of measurements on 55 patients, of whom 18 have been diagnosed with latent membranous lupus. Table 1 shows the data with two clinical covariates, IgA and IgG, that measure the levels of immunoglobulin of type A and of type G, respectively. Of interest

TABLE 1
*The number of latent membranous lupus nephritis cases, the numerator,
and the total number of cases, the denominator, for each combination
of the values of the two covariates*

IgG3 − IgG4	IgA				
	0	**0.5**	**1**	**1.5**	**2**
−3.0	0/1	—	—	—	—
−2.5	0/3	—	—	—	—
−2.0	0/7	—	—	—	0/1
−1.5	0/6	0/1	—	—	—
−1.0	0/6	0/1	0/1	—	0/1
−0.5	0/4	—	—	1/1	—
0	0/3	—	0/1	1/1	—
0.5	3/4	—	1/1	1/1	1/1
1.0	1/1	—	1/1	1/1	4/4
1.5	1/1	—	—	2/2	—

is the prediction of disease occurrence using the two covariates IgG3 − IgG4 and IgA. As in van Dyk and Meng (2001), we consider a probit regression model. For each patient i, we model the disease indicator variables as independent $Y_i \sim \text{Bernoulli}(\Phi(x_i^T \beta))$, where $\Phi(\cdot)$ is the CDF of $N(0, 1)$, x_i is the vector of the covariates and β is a 3×1 vector of parameters. We want to sample from the posterior distribution corresponding to the flat prior for β.

To illustrate the impact of antithetic coupling on mixing, we adopt the standard Gibbs sampler with the latent variables $\psi_i \sim N(x_i^T \beta, 1)$, of which only the sign Y_i is observed, as the augmented data [e.g., Albert and Chib (1993)]. Let X be the $n \times p$ matrix whose ith row is x_i and $\psi = (\psi_1, \ldots, \psi_n)$; for our example $n = 55$ and $p = 3$. The resulting Gibbs sampler alternates between sampling from (i) $\beta|\psi \sim N(\tilde{\beta}, (X^T X)^{-1})$ with $\tilde{\beta} = (X^T X)^{-1} X^T \psi$ and from (ii) $\psi_i|\beta, Y_i \sim TN(x_i^T \beta, 1, Y_i)$, where $TN(\mu, \sigma^2, Y)$ is $N(\mu, \sigma^2)$ truncated to be positive if $Y > 0$ and negative if $Y < 0$. Our parallel chains are then coupled antithetically at each of these two updating steps.

In the first step we want to generate exchangeable $\{\beta^{(1)}, \ldots, \beta^{(k)}\}$ with $\beta^{(j)} \in \mathbf{R}^p$ for all $1 \leq j \leq k$, such that marginally $\beta^{(j)} \sim N(\tilde{\beta}, \Sigma = (X^T X)^{-1})$ and $\text{Corr}(\beta_i^{(j)}, \beta_i^{(j')})$ is minimum possible, for all $1 \leq i \leq p$ and all $1 \leq j \neq j' \leq k$. This can be realized by generating p i.i.d. multivariate normal $(Z_i^{(1)}, \ldots, Z_i^{(k)})^\top$ such that $Z_i^{(j)} \sim N(0, 1)$ and $\text{Corr}(Z_i^{(j)}, Z_i^{(j')}) = -1/(k - 1)$. We then let $\beta^{(j)} = \Sigma^{1/2} Z^{(j)} + \tilde{\beta}$, where $Z^{(j)} = (Z_1^{(j)}, \ldots, Z_p^{(j)})^\top$ and $\Sigma^{1/2}$ is the Choleski decomposition of Σ.

In the second step, we use the inverse CDF method suggested by Gelfand, Smith and Lee (1992) to sample from $TN(\mu, \sigma^2, 1)$. Namely, we simulate

$U \sim \text{Uniform}(0, 1)$ and then take

$$(2.9) \qquad Z = \mu + \sigma \Phi^{-1}\big[\Phi(-\mu/\sigma) + U\big(1 - \Phi(-\mu/\sigma)\big)\big].$$

The methods of Lew (1981) and Bailey (1981) were used to approximate $\Phi(x)$ and $\Phi^{-1}(x)$ needed for (2.9). The antithetic coupling is then realized via n independent vectors of NA uniform random variables $\{U_i^{(1)}, \ldots, U_i^{(k)}; 1 \le i \le n\}$ so that for the jth chain $\psi_i^{(j)}$ is generated by using $U = U_i^{(j)}$ in (2.9). We used only two iterations of ILHS, to increase the speed of the antithetic sampler. The computational overhead is small even in the first step as all the matrix inversions required there were performed once outside the inner loop of the stochastic algorithm.

The improvement brought about by the antithetic coupling is twofold because it can reduce both the variance and the bias of the original sampler. Van Dyk and Meng (2001) demonstrate the slow mixing of the standard algorithm and propose a much more reliable marginal data augmentation algorithm. Our antithetic implementation does not, and cannot, remedy all the problems with the original sampler, but it can provide noticeable improvement without requiring a new algorithm or a substantial increase in the computational load. This is seen in Figure 5, which contains scatter plots of (β_0, β_1) using draws from k independent chains of the original sampler (left column) and from k antithetic chains (right column). The true contour plots, obtained via numerical methods, are from van Dyk and Meng (2001). Each scatter plot is based on 9000 sample points, divided equally among the k chains used. All the starting points were set to the same MLE; we did this on purpose to see to what extent the antithetic coupling can "spread out" despite the fact that all chains are started at the same point. Although the antithetic chains still have the serious problem of missing a good part of the "right tail," a problem that was avoided by van Dyk and Meng's (2001) algorithm, in all rows the scatter plot in the right column extends to the "right tail" no less, and sometime substantially more, than the scatter plot in the left column.

To represent quantitatively the benefit of antithetic coupling, Figure 6 plots, against k, the relative bias, relative standard error (SD) and relative root mean square error (RMSE) from antithetic chains, in estimating six posterior quantities, with the results from independent chains as the baseline (e.g., a relative bias 0.4 means the antithetic implementation reduces the bias, in magnitude, by $1 - 0.4 = 60\%$). The six quantities we choose are, in the order of the rows in Figure 6, $E[\beta_0|\mathcal{D}]$, $E[\beta_1|\mathcal{D}]$, $\text{Var}[\beta_0|\mathcal{D}]$, $\text{Var}[\beta_1|\mathcal{D}]$, $E[-\beta_0/\beta_1|\mathcal{D}]$ and $E[Q|\mathcal{D}]$, where $Q = [\Phi(\beta_0 - 0.5\beta_2 + 1.5\beta_2)]/[1 - \Phi(\beta_0 - 0.5\beta_2 + 1.5\beta_2)]$ and $\mathcal{D}$ denotes the observed data (X, Y). Whereas the first four are usually required in a Bayesian analysis, the last two are specific to the example at hand: $-\beta_0/\beta_1$ is the so-called LD50 level (i.e., with 50% chance the corresponding dosage becomes lethal) for covariate x_1 when $x_2 = 0$, and Q represents the odds of having the disease when $x_1 = -0.5$ and $x_2 = 1.5$. In particular, four out of these six estimand functions are nonmonotone.
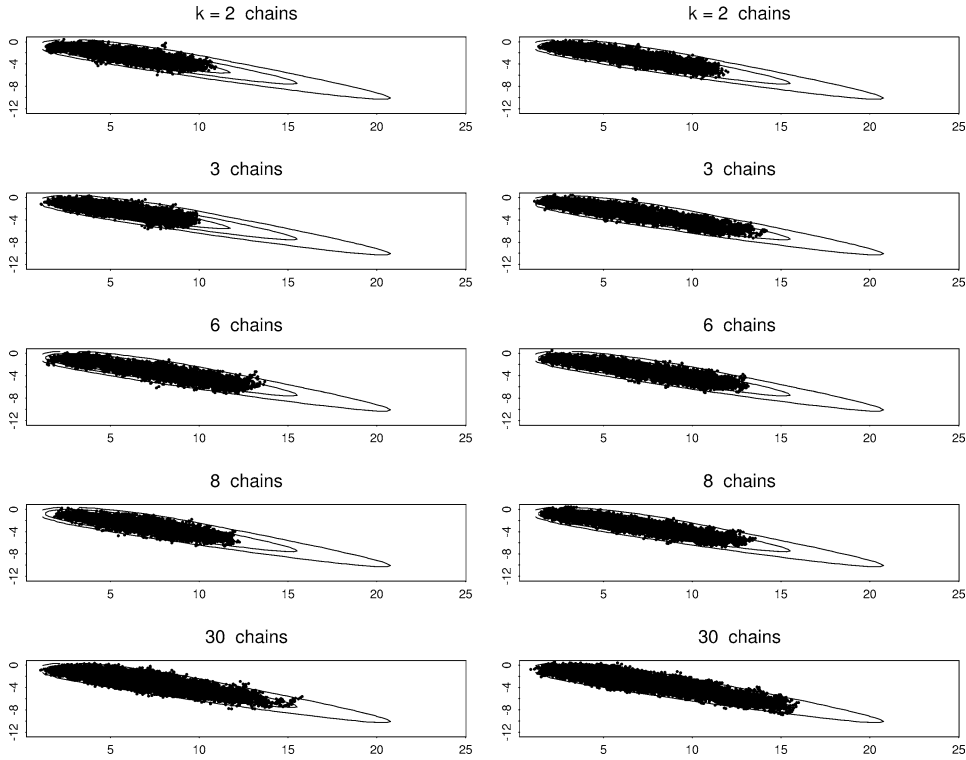
FIG. 5. *Bayesian probit regression example. Scatter plots of* (β_0, β_1) *generated by using k independent chains* (left) *and k antithetic chains* (right) *for different values of k; the contour lines are from the targeted bivariate posterior distribution. Each plot contains* 9000 *draws.*

The simulation sizes in Figure 6 are the same as in Figure 5, and the reduction factors are then computed using 1000 replicates of the simulation process. No burn-in period was discarded because we are interested in comparing the mixing properties of the two implementations. Because the total sample size is fixed at 9000 for each plot, using a larger k means a smaller within-chain sample size m, and hence possibly more strong influence of the starting point. To investigate the effect of the starting point, we use three starting strategies: 1. MLE (columns 1–3 in Figure 6). 2. At two standard deviations (SD) from the MLEs, in this case, $\beta_0 + 2SD$ and $\beta_i - 2SD$, $i = 1, 2$ (columns 4–6). 3. From extreme points (more than three SD away from the MLE) in the support of the distribution (columns 7–9). In general it is seen that the RMSE is reduced due to a decrease in variance, bias or both. Best results correspond to $3 \le k \le 6$, reflecting the aforementioned trade-off between k and m. As expected, the reduction factors vary with the starting points, as well as the estimand of interest. Nevertheless, with $k \ge 3$ the antithetic coupling in general does not inflate the RMSE and may result in reduction factors as low as 0.1, that is, up to 90% of saving.
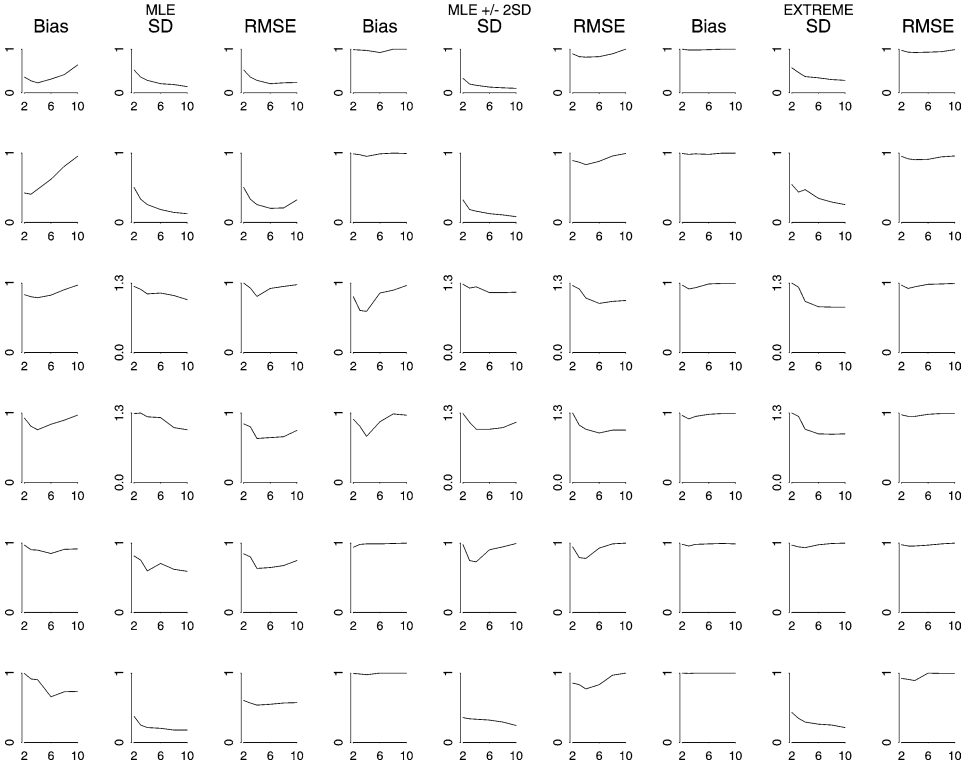
FIG. 6. *Relative Monte Carlo bias, standard error (SD) and root mean squared error (RMSE), from the antithetic chains, relative to that of independent chains, as functions of k, the number of parallel chains. Same simulation configurations as in Figure 5, and with 1000 replication for each k. Each row represents a different estimand function; from top to bottom*: $E[\beta_0|\mathcal{D}]$; $E[\beta_1|\mathcal{D}]$; $Var(\beta_0|\mathcal{D})$; $Var(\beta_1|\mathcal{D})$; $E[-\beta_0/\beta_1|\mathcal{D}]$; $E[\mathcal{Q}|\mathcal{D}]$. *Columns 1–3 use MLE as the starting points, 4–6 use MLE ± 2 SD, and 7–9 use some extreme points situated more than three SD away from the MLE.*

3. A theoretical foundation.

3.1. *Negative association and dependence.* In general, a qualitative measure of negative dependence should adequately reflect the following intuitive behavior among a set of variables: if one subset of the variables is "high," then a disjoint subset of the variables is "low." Different ways to define such negative dependence have received a great deal of attention in the last twenty years or so. Due to the success of the *positive association* concept of Esary, Proschan and Walkup (1967), the main challenge has been to build the negative association concepts as "duals" of the positive ones, but so far there has not been a universally acceptable construction [e.g., Pemantle (2000)]. Specifically, the set of random variables $\{X_1, \ldots, X_n\}$ is said to be *positively associated* (PA) if for any non-decreasing (or nonincreasing—we will not state both hereafter) functions f_1, f_2,

$\operatorname{Cov}(f_1(X_1, X_2, \ldots, X_n), f_2(X_1, X_2, \ldots, X_n)) \geq 0$. Here a function $f : \mathbf{R}^n \to \mathbf{R}$ is called nondecreasing if it is nondecreasing in all of its arguments.

The closest negative counterpart of positive association we can find is the *negative association* concept introduced by Joag-Dev and Proschan (1983).

DEFINITION 1. The random variables $\{X_i\}_{1 \leq i \leq n}$, where each X_i can be of arbitrary dimension, are said to be negatively associated (NA) if for every pair of disjoint finite subsets A_1, A_2 of $\{1, 2, \ldots, n\}$ and for any nondecreasing functions f_1, f_2

(3.1) $$\operatorname{Cov}\big(f_1(X_i, i \in A_1), f_2(X_j, j \in A_2)\big) \leq 0$$

whenever the above covariance function is well defined.

The following equivalence result is important for some of our subsequent theoretical investigation.

PROPOSITION 1. *The random variables $\{X_i\}_{1 \leq i \leq n}$ are NA if and only if* (3.1) *holds for every pair of nonnegative, bounded and nondecreasing functions f_1 and f_2.*

PROOF. The "only if" part is obvious. To prove the "if" part, let $f_m(X_i, i \in A_m)$, $m = 1, 2$, be the two functions in (3.1). For any positive integer l, let $f_m^{(l)}$ be the truncation of f_m to $[-l, l]$, that is, $f_m^{(l)} = f_m$ when $|f_m| \leq l$, and $f_m^{(l)} = \pm l$ depending on $f_m > l$ or $f_m < -l$. Clearly, $|f_m^{(l)}| \leq |f_m|$, and thus by the dominated convergence theorem, $\lim_{l \to \infty} \operatorname{Cov}(f_1^{(l)}(X_i, i \in A_1), f_2^{(l)}(X_j, j \in A_2)) = \operatorname{Cov}(f_1(X_i, i \in A_1), f_2(X_j, j \in A_2))$. This allows us to conclude (3.1) because

$$\operatorname{Cov}\big(f_1^{(l)}(X_i, i \in A_1), f_2^{(l)}(X_j, j \in A_2)\big)$$
$$= \operatorname{Cov}\big(f_1^{(l)}(X_i, i \in A_1) + l, f_2^{(l)}(X_j, j \in A_2) + l\big) \leq 0,$$

where the last inequality holds because $f_m^{(l)} + l$ is a nonnegative, bounded and nondecreasing function for $m = 1, 2$. $\square$

The notion of NA is most useful for our purposes primarily because, like PA, it is closed under the independent union operation as well as monotone transformations, as proved in Joag-Dev and Proschan (1983). Specifically, we have:

PROPOSITION 2. *If $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ are two independent sets of NA (PA) random variables, then their union, $\{X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}\}$, is a set of NA (PA) random variables.*

PROPOSITION 3. *If $\{X_i\}_{1 \leq i \leq n}$ is a sequence of NA (PA) random variables and $(\psi_i)_{1 \leq i \leq n}$ are all nondecreasing functions, then $(\psi_i(X_i))_{1 \leq i \leq n}$ is a sequence of NA (PA) random variables.*

These two results enable us to prove the following fundamental result on antithetic coupling of homogeneous and nonhomogeneous Markov chains.

THEOREM 1. *Suppose we run a k-process antithetically coupled chain, $\mathcal{X}_t^k = \{X_t^{(1)}, \ldots, X_t^{(k)}\}$, forward for T iterations, where $X_t^{(j)} = \psi_t(X_{t-1}^{(j)}, U_t^{(j)})$, $j = 1, \ldots, k$, and $\mathcal{U}_t^k = \{U_t^{(1)}, \ldots, U_t^{(k)}\}$, $t = 1, \ldots, T$, are T independent sets of NA variables. Assuming $\psi_t(X, U)$ is nondecreasing for all $t \leq T$, we have the following results.*

(i) *The k-tuple $\{X_{t_1}^{(1)}, \ldots, X_{t_k}^{(k)}\}$ is a collection of k NA variables for any $\{t_1, \ldots, t_k\} \in \{0, 1, \ldots, T\}^k$ if and only if it is so for $t_1 = \cdots = t_k = 0$.*

(ii) *Assuming $\mathcal{X}_0^k$ is a set of NA variables, then the variance reduction factor, $S_k^{(f)}$, defined by (2.6), is at most 1 for any monotone function f, and it is strictly less than 1 if and only if at least one of the between-chain covariances, $\beta_{t_1,t_2;j_1,j_2}^{(f)}$ of (2.7), is strictly negative.*

PROOF. For (i) the necessity holds by definition. For the sufficiency, because $\{X_0^{(1)}, \ldots, X_0^{(k)}\}$ and $\{U_t^{(1)}, \ldots, U_t^{(k)}\}$, $t = 1, \ldots, T$, are $T + 1$ sets of independent NA variables, by Proposition 2 their union is also NA. Consequently, because $X_{t_j}^{(j)}$ is a nondecreasing function of $Z^{(j)} \equiv \{X_0^{(j)}, U_t^{(j)}, t = 1, \ldots, T\}$ only, and $Z^{(i)}$ and $Z^{(j)}$ do not share any common variable for $i \neq j$, by Proposition 3 $\{X_{t_1}^{(1)}, \ldots, X_{t_k}^{(k)}\}$ are NA for any $t_j \leq T$, $j = 1, \ldots, k$.

For (ii), by (i) all the between-chain autocovariances $\beta_{t_1,t_2;j_1,j_2}^{(f)} \leq 0$, which implies $S_k^{(f)} \leq 1$ because the denominator in (2.6), being a variance, is always positive. (In fact, by using the PA part of Propositions 2 and 3, we can conclude all $\gamma_{t_1,t_2;j}^{(f)} \geq 0$.) Furthermore, because all $\beta_{t_1,t_2;j_1,j_2}^{(f)} \leq 0$, by (2.6), $S_k^{(f)} = 1$ if and only if all $\beta_{t_1,t_2;j_1,j_2}^{(f)}$'s are zero. $\square$

Since for a Gibbs sampler with attractive stationary density the updating function can be expressed as a monotone function [e.g., Propp and Wilson (1996) and Häggström and Nelander (1998)], Theorem 1 covers Frigessi, Gåsemyr and Rue's (2000) Theorem 1 with $k = 2$. For practical purposes, the requirement that $\mathcal{X}_0^k = \{X_0^{(1)}, \ldots, X_0^{(k)}\}$ are NA is immaterial, because being fixed (even with the choice that $X_0^{(1)} = \cdots = X_0^{(k)}$) or more generally being independent is a trivial case of being NA. It is also evident that Theorem 1 does not require $X_0^{(j)}$ to be from the stationary distribution.

Theorem 1, however, does require that the random variables in the k-tuple $\mathcal{U}_t^k = \{U_t^{(1)}, \ldots, U_t^{(k)}\}$ are NA. Typically, for a particular distribution there are many ways of achieving this, and some general recipes are described in Section 4. A useful construction in practice is to use the fact that monotone functions of distinctive subsets of NA variables are NA, a fact that allows us to build upon known NA variables, such as permutation distributions, multinomial, multivariate hypergeometric, Dirichlet and multivariate normals with nonpositive correlations [see Joag-Dev and Proschan (1983)].

We remark that part (i) of Theorem 1 is actually stronger than we need, because from (ii), in order to assure $S_k^{(f)} \leq 1$, we only need to ensure pairwise negative covariance: $\mathrm{Cov}(f(X_{t_1}^{(j_1)}), f(X_{t_2}^{(j_2)})) \leq 0$. This weaker version is sometimes easier to establish, and thus we define the following notion.

DEFINITION 2. The random variables $X_1, X_2, \ldots, X_n$ are said to be pairwise negatively associated (PNA) if $\{X_i, X_j\}$ are NA for any $i \neq j \in \{1, \ldots, n\}$.

Clearly, Propositions 2 and 3 still hold for PNA. In the next section, we will see that only requiring our outcome to be PNA can avoid technical complications that are not critical to MCMC applications in practice. In particular, the PNA property is easier to verify because it is equivalent to the *negative quadrant dependency* (NQD) notion for *a pair* of random variables, as defined by Lehmann (1966). For more than two variables, NQD is a consequence of NA, but not vice versa, as formalized by the following result, also due to Joag-Dev and Proschan (1983).

PROPOSITION 4. *Let* $X_1, X_2, \ldots, X_n$ *be a set of NA random variables. Then they are also negatively lower orthant dependent* (NLOD), *that is,*

$$(3.2) \quad P(X_1 \leq x_1, \ldots, X_n \leq x_n) \leq \prod_{i=1}^{n} P(X_i \leq x_i) \qquad \text{for all } x_1, \ldots, x_n,$$

as well as negatively upper orthant dependent (NUOD), *namely,*

$$(3.3) \quad P(X_1 > x_1, \ldots, X_n > x_n) \leq \prod_{i=1}^{n} P(X_i > x_i) \qquad \text{for all } x_1, \ldots, x_n.$$

3.2. *Negative association in limit and joint convergence.* Although Theorem 1 provides a theoretical foundation for antithetic coupling with forward MCMC, it does not cover backward MCMC such as CFTP. This is because the T in Theorem 1 is a nonrandom constant, whereas for CFTP it is a random variable and, more importantly, it is not independent of the corresponding draw from the CFTP. We thus need to extend Theorem 1 to the case, with $T = +\infty$, namely, we need to prove that the limiting k-tuple $\{X_\infty^{(1)}, \ldots, X_\infty^{(k)}\}$ is still NA, or at least PNA. This would entail that the joint draw from a CFTP, $\{X_0^{(1)}, \ldots, X_0^{(k)}\}$, is NA/PNA

because we can identify its probability structure with that of $\{X_\infty^{(1)}, \ldots, X_\infty^{(k)}\}$ from a forward process.

The extension to $T = \infty$ is not straightforward because the ergodicity of the marginal chain does not guarantee that of the joint chain, so that a limiting argument such as $\mathrm{Cov}(f(X_\infty^{(i)}), f(X_\infty^{(j)})) = \lim_{t \to \infty} \mathrm{Cov}(f(X_t^{(i)}), f(X_t^{(j)})) \leq 0$ needs qualification. The problem has been discussed also by Arjas and Gasbarra (1996) and Frigessi, Gåsemyr and Rue (2000) who have shown that, if the joint chain is ϕ-irreducible and the closure of the support of ϕ contains an open set, then the joint chain is positive recurrent on the closure of the support of ϕ. This can be used to prove that the above convergence of covariances holds. However, the assumption that the support of the irreducibility measure ϕ has nonempty interior is violated even by simple examples such as the following Markov chain: take $X_0 \in S_1$, the unit circle; for $t \geq 1$, draw θ_t uniform on $[0, \pi)$ and construct the line that goes through X_{t-1} and has slope $\tan(\theta_t)$. The intersection of this line with the unit circle is X_t. The left panel of Figure 7 illustrates this construction with $t = 1$. The right panel illustrates a pair of antithetically coupled chains (X_t, Y_t) (for $t \leq 2$) where the Y_t is constructed the same way as X_t except each time X_t is updated using θ_t, Y_t is updated using $\pi - \theta_t$. Algebraically, we have

$$
\begin{aligned}
\theta_t^X &= (-\theta_{t-1}^X + \pi + 2\theta_t) \bmod (2\pi), \\
\theta_t^Y &= (-\theta_{t-1}^Y + \pi - 2\theta_t) \bmod (2\pi),
\end{aligned}
$$
(3.4)

where θ_t^X and θ_t^Y are the polar-coordinate representations of X_t and Y_t, respectively. Marginally, θ_t^X (and hence θ_t^Y) are i.i.d. uniform variables on $[0, 2\pi)$, and hence $\{X_t\}_t$ is trivially uniformly ergodic. However, the joint state space
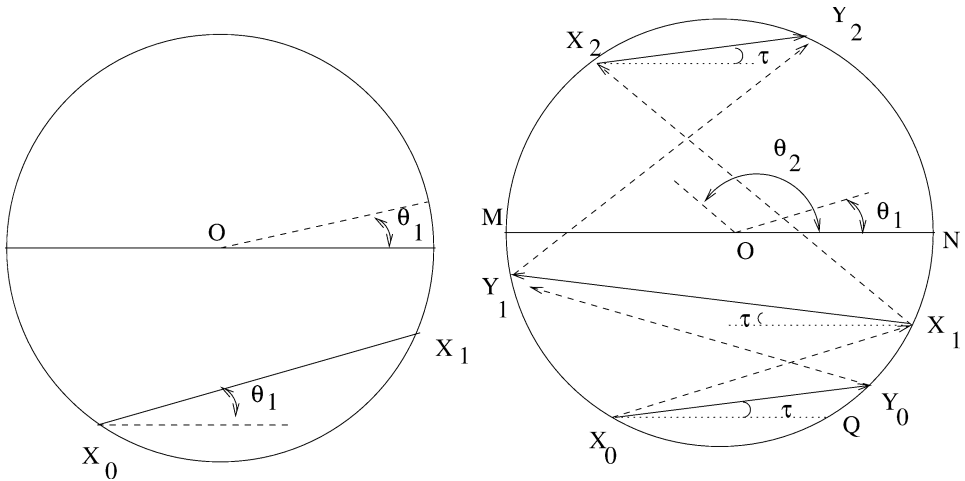

FIG. 7. *The random walk on the circle example. Illustration of construction of X_t (left) and of the antithetic coupling (X_t, Y_t) (right).*

of (θ_t^X, θ_t^Y), $S_2 = [0, 2\pi) \times [0, 2\pi)$, consists of *uncountably* many absorbing subclasses, defined by

$$(3.5) \quad S(\tau) = \{(\theta^X, \theta^Y) \in S_2 : \theta^X + \theta^Y = (2k+1)\pi \pm 2\tau, \text{ for } k = 0, 1\},$$

where $\tau \in [0, \pi/2]$ is the acute angle between the line $\overline{X_0 Y_0}$ and the horizontal diameter (see Figure 7). Here $S(\tau)$ is an absorbing set because the joint updating rule leaves the acute angle between the line $\overline{X_t Y_t}$ and the horizontal diameter, denoted by τ_t, unchanged. Let $\rho_t(\tau) = \mathrm{Corr}(X_t^{(1)}, Y_t^{(1)})$, where $Z^{(1)}$ denotes the X-axis coordinate of Z [i.e., $Z^{(1)} = \cos(\theta^Z)$]. Then, as proved in Appendix A.1, $\rho_t(\tau) = -\cos(2\tau)$, increasing from -1 to 1 as τ increases from 0 to $\pi/2$. In addition, for each of the ergodic classes, the joint chain has an irreducibility measure whose support has an empty interior.

This example shows that the uniform ergodicity of the marginal chain is not sufficient to guarantee the ergodicity of the antithetically coupled chain or to directly justify the extension of the NA properties from $\{X_t^{(1)}, \ldots, X_t^{(k)}\}$ to its "limit" $\{X_\infty^{(1)}, \ldots, X_\infty^{(k)}\}$. Here the "limit" is in quotes because while the marginal distribution of $X_\infty^{(j)}$ does not depend on the starting point, the joint distribution of $\{X_\infty^{(1)}, \ldots, X_\infty^{(k)}\}$ may depend on it, as is the case in the previous example. Fortunately, the time-backward dual sequence approach discussed in Section 2.1 provides a theoretical tool to bypass the issue of the joint chain's ergodicity, and thereby to establish the following counterpart of Theorem 1 for justifying backward antithetic coupling for all monotone CFTP algorithms.

THEOREM 2. *Let $X_t = \psi(X_{t-1}, U_t)$ be a uniformly ergodic Markov chain on state space Γ with π being its invariant distribution and with ψ nondecreasing. Let $\mathcal{X}_t^k = \{X_t^{(1)}, \ldots, X_t^{(k)}\}$ be the corresponding antithetically coupled joint chain as in Theorem 1. Then*:

(i) *For any given starting point $\mathcal{X}_0^k$, $\mathcal{X}_t^k = \{X_t^{(1)}, \ldots, X_t^{(k)}\}$ converges in distribution to some $\mathcal{X}_\infty^k = \{X_\infty^{(1)}, \ldots, X_\infty^{(k)}\}$, whose joint distribution may depend on $\mathcal{X}_0^k$.*

(ii) *For any nondecreasing real functions f_m, $m = 1, 2$, on Γ such that $\mathrm{Var}_\pi(f_m(X)) < \infty$, $m = 1, 2$, and such that their discontinuity points are contained in a π-null set, we have*

$$(3.6) \qquad \mathrm{Cov}\big(f_1(X_\infty^{(j_1)}), f_2(X_\infty^{(j_2)}) | \mathcal{X}_0^k\big) \le 0 \qquad \text{for any } j_1 \ne j_2.$$

PROOF. For each marginal chain $X_t^{(j)}$, let $\tilde{X}_t^{(j)}$ be its dual sequence as in (2.2). Then, by the equivalence of the implementability of CFTP and uniform ergodicity [Foss and Tweedie (1998)], we know that there exists an almost sure finite stopping time S on the infinite-product space $\Gamma_U^\infty = \prod_{t \ge 1} \mathcal{U}_t^k$ such that $\tilde{X}_S^{(j)} \sim \pi$ and $\tilde{X}_t^{(j)} = \tilde{X}_S^{(j)}$ for all $t \ge S$ and all j. Therefore, by re-expressing all relevant variables on their common sampling space Γ_U^∞, we can

define $\mathcal{X}_\infty^k(\omega) = \{\tilde{X}_{S(\omega)}^{(1)}(\omega), \ldots, \tilde{X}_{S(\omega)}^{(k)}(\omega)\}$, which is a well-defined k-component joint random variable on Γ_U^∞ because $P(S(\omega) < \infty) = 1$. Furthermore, because $\tilde{\mathcal{X}}_t^k = \{\tilde{X}_t^{(1)}, \ldots, \tilde{X}_t^{(k)}\}$ converges to $\mathcal{X}_\infty^k$ with probability 1 by the construction, and because $\mathcal{X}_t^k$ and $\tilde{\mathcal{X}}_t^k$ have the same distribution for any t, $\mathcal{X}_t^k$ must converge *in distribution* to $\mathcal{X}_\infty^k$, and hence (i).

For (ii), we only need to prove (3.6) when both f's are bounded, following the same argument as in the proof of Proposition 1. Since $f_m, m = 1, 2$, are almost surely continuous with respect to the distribution of $\{X_\infty^{(i)}, X_\infty^{(j)}\}$ because its margins are π, we can conclude by part (i) that $\{f_1(X_t^{(i)}), f_2(X_t^{(j)})\}$ converges in distribution to $\{f_1(X_\infty^{(i)}), f_2(X_\infty^{(j)})\}$. This implies

$$(3.7) \quad \mathrm{Cov}\big(f_1(X_\infty^{(i)}), f_2(X_\infty^{(j)})|\mathcal{X}_0^k\big) = \lim_{t \to \infty} \mathrm{Cov}\big(f_1(X_t^{(i)}), f_2(X_t^{(j)})|\mathcal{X}_0^k\big) \leq 0,$$

where the limiting argument holds because both f's are bounded. $\quad\square$

Part (ii) of Theorem 2 shows that $\{\tilde{X}_\infty^{(1)}, \ldots, \tilde{X}_\infty^{(k)}\}$ are PNA when the monotone functions are almost surely continuous with respect to the underlying dominating measure, which is the case if the latter is the Lebesgue measure, as is common in Bayesian computation. We emphasize that the condition that ψ is monotone is not needed for part (i), but it is important for part (ii). This is seen in the unit circle example, where (i) holds with uncountably many limiting distributions for the joint chain, depending on the initial τ. However, (ii) does not hold because the ψ function there is obviously not monotone. The fact that $\rho(\tau) < 0$ when $\tau < \pi/4$ in that example also illustrates that the monotonicity is a sufficient but not necessary condition for preserving negative correlation.

3.3. *Extreme antithesis.* The concept of NA provides a qualitative description of negative dependence. Quantitatively, it is desirable to generate $\{X_1, \ldots, X_k\}$ (corresponding to $\{X^{(1)}, \ldots, X^{(k)}\}$ in previous sections) such that $\mathrm{Corr}(f(X_i), f(X_j))$ is as negative as possible. Formally, we define the following notion.

DEFINITION 3. A set of variables $\{X_1, X_2, \ldots, X_k\}$ is said to achieve extreme antithesis (EA) with respect to a (marginal) distribution F if they are exchangeable and

$$\mathrm{Corr}(X_i, X_j) = \min\{\mathrm{Corr}(Y_i, Y_j) : Y_1, \ldots, Y_k \text{ exchangeable}, Y_i \sim F \ \forall i\}.$$

For $k = 2$, the single strategy of using quantile coupling via $X_1 = F^{-1}(U)$ and $X_2 = F^{-1}(1 - U)$ achieves EA for any F, as discussed in Section 1. For $k > 2$, the matter is much more complicated, even just for $F := \mathrm{Uniform}(0, 1)$ and $F := N(0, 1)$. Indeed, it is not hard to establish the following negative result [see Craiu and Meng (2002) for a proof], where Φ is the CDF of $N(0, 1)$.

PROPOSITION 5. *It is impossible to find a joint distribution $F^{(3)}(U_1, U_2, U_3)$ on $(0, 1)^3$ such that all its univariate margins are* Uniform$(0, 1)$ *and the following holds almost surely* (*with respect to joint Lebesgue measure*):

$$(3.8) \quad U_1 + U_2 + U_3 = 3/2 \quad and \quad \Phi^{-1}(U_1) + \Phi^{-1}(U_2) + \Phi^{-1}(U_3) = 0.$$

For k even, say, $k = 4$, one may be tempted to use two pairs of quantile coupled variates, namely, $U_2 = 1 - U_1$, $U_4 = 1 - U_3$, where U_1 and U_3 are i.i.d., to make the $k = 4$ version of (3.8) hold. Consequently, for any f, $\rho_4^{(f)} = \rho_2^{(f)}/3$, where $\rho_2^{(f)} = \text{Corr}(f(U_1), f(1 - U_1))$ and $\rho_4^{(f)}$ is the correlation between any pair of $\{f(U_i), i = 1, \ldots, 4\}$. This implies $S_4^{(f)} = 1 + 3\rho_4^{(f)} = 1 + \rho_2^{(f)} = S_2^{(f)}$, and thus it is just a disguised version of quantile coupling with $k = 2$, not a real generalization to $k = 4$.

For lack of a universal strategy, we seek methods that are effective in common practice. In the exchangeable setting, the quest for EA is the same as that for minimizing the variance of the mean (and sum) $\bar{X}^{(k)} = (X_1 + \cdots + X_k)/k$ because [see (2.3)]

$$(3.9) \qquad \rho_k \equiv \text{Corr}(X_i, X_j) = \frac{1}{k-1}\left[\frac{\text{Var}_k(\bar{X}^{(k)})}{\text{Var}(X_1)} - 1\right] \geq -\frac{1}{k-1},$$

where the subscript k in both ρ_k and $\text{Var}_k(\bar{X}^{(k)})$ emphasizes the dependence on the joint distribution $F^{(k)}(x_1, \ldots, x_k)$. In particular, if an $F^{(k)}$ is constructed such that $\bar{X}^{(k)} = constant$ (almost surely), then EA is achieved [hence (3.8)]. However, although this approach works for some common distributions such as uniform and normal, it is not always possible because the minimal value of $\text{Var}_k(\bar{X}^{(k)})$ may not achieve zero even when $k = 2$. For example, the minimal ρ_2 is $1 - \pi^2/6 = -0.645$ when both X_1 and X_2 are exponentially distributed with mean 1, as reported in Moran (1967).

For specific families of distributions, several methods may be available to achieve EA. Snijders (1984) explores various approaches for binary random variables. For a unimodal symmetric and differentiable density p on $\mathbf{R}$, Rüschendorf and Uckelmann (2000) propose the following. Suppose the center of symmetry is zero and that $p_Q(x) = -xp'(x)$ is also a Lebesgue density on $\mathbf{R}$; let $Q \sim p_Q$. Then $X = QU \sim p$ for any $U \sim$ Uniform$(-1, 1)$ that is independent of Q. Consequently, for any set of $\{U_1, \ldots, U_k\}$, independent of Q, such that $U_i \sim$ Uniform$(-1, 1)$ and $\sum_{i=1}^k U_i = 0$, $\{X_1 = QU_1, \ldots, X_k = QU_k\}$ achieves EA with respect to p because $\sum_{i=1}^k X_i = 0$. In fact, $\text{Corr}(X_i, X_j) = \text{Corr}(U_i, U_j) = -(k-1)^{-1}$ for any $i \neq j$. This construction, however, does not guarantee NA in general. As an alternative, we can draw i.i.d. $\{Q_1, \ldots, Q_k\}$ from p_Q, and then use $X_i = Q_i U_i$, $\forall i \in \{1, \ldots, k\}$. This will guarantee the NA property, but it sacrifices the EA property, because now we have $\text{Corr}(X_i, X_j) = -\frac{1}{k-1}[1 + CV^2(Q)]^{-1}$, where $CV(Q)$ is the coefficient of variation of Q. Nevertheless, when $CV(Q)$ is

small, the loss of EA may be negligible for practical purposes. In general, if one has to make a choice between NA and EA, we recommend choosing NA, for it is preserved by all monotone transformations. With NA in place, the variance reduction factor is guaranteed to be at most 1 when the monotonicity assumption holds.

4. Generating antithetic uniform variates. Since generating uniform variates is, explicitly or implicitly, the most basic component of almost any simulation method, in this section we compare several methods for generating k-antithetically coupled uniform variates. For each method, we investigate whether it leads to NA and/or EA variables, and propose remedies whenever possible if it does not. We emphasize that although there are many different ways of achieving NA and/or EA [e.g., Bondesson (1983) and Gerow and Holbrook (1996)], no method dominates when $k \geq 3$, as demonstrated by Proposition 5. Any of these methods can be more suitable than others in a particular application. Nevertheless, the three methods described below are more or less representative of what has been proposed in the literature.

4.1. *The permuted displacement method.* This is a modified version of the one documented in Arvidsen and Johnsson (1982), which first generates an $r_1 \sim$ Uniform$(0, 1)$, and then constructs

(4.1) $r_i = \{2^{i-2}r_1 + \frac{1}{2}\}$, $i = 2, \ldots, k - 1$,   and $r_k = 1 - \{2^{k-2}r_1\}$,

where $\{x\}$ is the fractional part of x. We find that the binary representation of this method makes it a bit easier to show that $\sum_{i=1}^{k} r_i = k/2$. Specifically, let $r_1 = (a_1, a_2, \ldots, a_m, \ldots)$ denote the (nonterminating) dyadic expansion of r_1, where $a_i = 0$ or 1, that is, $r_1 = \sum_{i=1}^{\infty} a_i/2^i$. Then

$$r_2 = (1 - a_1, a_2, a_3, \ldots, a_m, \ldots),$$
$$r_3 = (1 - a_2, a_3, \ldots, a_{m+1}, \ldots),$$
$$\vdots$$
$$r_{k-1} = (1 - a_{k-2}, a_{k-1}, \ldots, a_{m+k-3}, \ldots),$$
$$r_k = (1 - a_{k-1}, 1 - a_k, \ldots, 1 - a_{m+k-2}, \ldots).$$

Therefore, the method creates negative correlation by displacing the binary digits of r_1. That all r_i's are uniform is a direct consequence of a well-known result of Borel (1924), as discussed in detail by Billingsley (1986). The dyadic expansion representation shows clearly that the r's are not exchangeable. A consequence of that is that they are *not* NA when $k \geq 3$. To prove this, we only need to show that when $k \geq 3$, $\{r_1, r_2\}$ are not NQD, and thus by Proposition 4 they cannot be NA.

To see this, we note that for $k \geq 3$, $r_2 = r_1 + 0.5$ when $r_1 < 0.5$ and $r_2 = r_1 - 0.5$ when $r_1 \geq 0.5$. Consequently, for $0 \leq s < 0.5$ and $s + 0.5 < t \leq 1$, we have

$$P(r_1 \leq t, r_2 \leq s) = P(0.5 \leq r_1 \leq t, r_1 - 0.5 \leq s) = \min\{t, s + 0.5\} - 0.5 = s,$$

which is larger than $ts = P(r_1 \leq t) P(r_2 \leq s)$, and therefore (3.2) is violated.

However, it is easy to fix the nonexchangeability by using the simple permutation method. That is, let S_k be the set of all possible permutations of k objects. Pick a random $\sigma \in S_k$ and define $U_i = r_{\sigma(i)}$. Clearly $\sum_i U_i = \sum_i r_i = k/2$ and thus $\text{Corr}(U_i, U_j) = -(k-1)^{-1}$ for any $i \neq j$. Furthermore, it can be shown that:

THEOREM 3. *For $k = 3$, $\{U_1, U_2, U_3\}$ constructed by the permuted displacement method are PNA.*

The proof is given in Craiu and Meng (2002) and is omitted both because of space limitation and because the approach used was rather brute force. Indeed, we are unable to prove or disprove Theorem 3 for $k \geq 4$. Nevertheless, the result indicates that the exchangeability can play an important role in achieving NA/PNA. An astute reader might wonder how permuting indexes can be helpful since in MCMC our estimates typically are sample averages, which are invariant to the independent permutations of the sample indexes. However, one must keep in mind that in our use of the antithetic variates, the U's are not used just once in the end, but throughout the whole generated sequence and the final estimates are not invariant to the permutations of all the indexes that occur along the sequence. This is perhaps easiest to see for CFTP, as each draw can depend, in principle, on an arbitrarily large number of independent copies of the k-tuple of U's.

4.2. *Multidimensional normal method.* A common way to manipulate correlations, especially in the engineering literature, is through the multivariate normal distribution. For our purposes, we can first generate $(Z_1, \ldots, Z_{k-1})^\top \sim N(\mathbf{0}, \Sigma)$ where $\Sigma_{ij} = -(k-1)^{-1}$ if $i \neq j$ and $\Sigma_{ii} = 1$, and then let $Z_k = -(Z_1 + Z_2 + \cdots + Z_{k-1})$. Finally, if we are interested in uniform deviates, we can use $U_i = \Phi(Z_i)$ for all $i \in \{1, \ldots, k\}$, where Φ is the CDF of $N(0, 1)$. The random variables $\{U_1, \ldots, U_k\}$ are NA following Joag-Dev and Proschan (1983), who proved that multivariate normals with nonpositive pairwise correlations are NA. The result then follows immediately from Proposition 3 because U_i is a monotone function of Z_i.

This method, however, does not achieve EA. Although $\text{Corr}(Z_i, Z_j) = -(k-1)^{-1}$, the nonlinear transformation $U_i = \Phi(Z_i)$ causes an increase in correlation. Specifically, for any $\{Z_1, Z_2\}$ bivariate normal with correlation ρ, the corresponding correlation between $\Phi(Z_1)$ and $\Phi(Z_2)$ is given by

$$(4.2) \qquad \frac{\int \Phi(\rho t) \Phi(t) \phi(t) \, dt - 1/4}{1/12} = \frac{3}{\pi} \rho + \frac{2\sqrt{\pi} - 3}{16\pi} \rho^3 + O(\rho^5)$$

$$= 0.955\rho + 0.01\rho^3 + O(\rho^5),$$

where ϕ is the standard normal density. Consequently, $\mathrm{Corr}(U_1, U_2)$ is larger than the minimum possible value $-(k-1)^{-1}$, though the loss of EA may not be important for many practical applications in view of (4.2). Equation (4.2) is a quantitative version of a result of Lancaster (1957), who proved that for any functions g_1 and g_2, $|\mathrm{Corr}(g_1(Z_1), g_2(Z_2))| \leq |\rho|$ as long as the left-hand side is finite, and the equality holds if and only if $g_i(z) = z$, $i = 1, 2$ (almost surely). Lancaster's result also implies Proposition 5 if we require $\{Z_1, Z_2, Z_3\}$ to be jointly normal. Nevertheless, Proposition 5 is a stronger result, as it shows that it is impossible to find any such trivariate distribution, not just trivariate normal, to simultaneously preserve EA as in (3.8).

An issue with real impact in computation is the requirement for a highly reliable and efficient subroutine to evaluate the function Φ. Otherwise, the use of a not-highly accurate approximation becomes problematic in large replications with arbitrarily many arguments because once in a while $|Z|$ can be too large for $\Phi(Z)$ to be evaluated appropriately. For that reason, we did not use this method to simulate uniform deviates in the simulated examples in Section 2. However, we used it to generate antithetic truncated normal deviates in the probit example presented there.

The normal method is also of interest because it is one that many will likely attempt, especially for generating antithetic normal deviates in high-dimensional settings. An example where such implementation results in acceleration of a classical MCMC method is provided by Craiu (2004) in the context of Multiple-Try Metropolis with antithetic proposals.

4.3. *Iterative Latin hypercube sampling.* The method described in Section 4.1 achieves EA but whether it achieves NA is an open question, and the method given in Section 4.2 achieves NA but not EA (when used for generating uniform variates). To achieve both, we propose to use *iterative Latin hypercube sampling* (ILHS), which is an iterative version of the Latin hypercube sampling (LHS), a well-known scheme in quasi Monte Carlo; see McKay, Beckman and Conover (1979), Stein (1987), Owen (1992), Loh (1996), Iman (1999) and Helton and Davis (2003), among others.

For any given $k \geq 2$, our iterative procedure can be described by the following steps:

1. Set $\mathcal{U}_0^k = (U_0^{(1)}, \ldots, U_0^{(k)})^\top$, where $\{U_0^{(j)}\}_{1 \leq j \leq k}$ are i.i.d. Uniform$(0, 1)$.
2. For $t = 0, 1, 2, \ldots$, let $\mathcal{K}_t = (\sigma_t(0), \ldots, \sigma_t(k-1))^\top$ be a permutation of $\{0, 1, \ldots, k-1\}$, independent of all previous draws, and let

$$(4.3) \qquad \mathcal{U}_{t+1}^k = \frac{1}{k}(\mathcal{K}_t + \mathcal{U}_t^k),$$

where $\mathcal{U}_t^k = (U_t^{(1)}, \ldots, U_t^{(k)})^\top$ for $t \geq 0$.

The case of $t = 1$ corresponds to the original LHS. For general t we have the following result.

THEOREM 4. *For any $t \geq 0$, $i, j \in \{1, \ldots, k\}$, $i \neq j$, we have*:

(i) $U_t^{(i)} \sim \text{Uniform}(0, 1)$.
(ii) $\text{Corr}(U_t^{(i)}, U_t^{(j)}) = -\frac{1}{k-1}(1 - \frac{1}{k^{2t}})$.
(iii) $\{U_{t_1}^{(1)}, \ldots, U_{t_k}^{(k)}\}$ *are NA for any finite* $\{t_1, \ldots, t_k\}$.

PROOF. (i) For $t = 0$, the result obviously holds. Suppose $U_t^{(1)} \sim$ Uniform(0, 1); then

$$P(U_{t+1}^{(1)} \leq s) = P(U_t^{(1)} < ks - \sigma_t(1)) = \frac{1}{k} \sum_{j=0}^{[ks]} P(U < ks - j) = s$$

for any $s \in (0, 1)$. The result thus holds by induction.

(ii) Let $S_t = \sum_j U_t^{(j)} = \mathbf{1}^\top \mathcal{U}_t^k$. Then from the recursion formula (4.3) we have $\text{Var}(S_t) = \text{Var}(S_{t-i})/k^{i+1}$ for all $1 \leq i \leq t$, which implies, by (i) and the exchangeability of $\{U_t^{(1)}, \ldots, U_t^{(k)}\}$,

$$k + k(k-1) \text{Corr}(U_t^{(i)}, U_t^{(j)}) = \frac{k}{k^{2t}},$$

which is just (ii).

(iii) Because any permutation distribution is NA, (4.3) defines an antithetically coupled joint Markov chain with k marginal Markov chains. Furthermore, the marginal updating function from $U_{t-1}^{(j)}$ to $U_t^{(j)}$ is monotone. The result then follows directly from part (ii) of Theorem 1 because $\{U_0^{(1)}, \ldots, U_0^{(k)}\}$ are i.i.d. and thus NA. □

For most practical purposes, Theorem 4 is all we need as it proves that for any T we choose, $\{U_T^{(1)}, \ldots, U_T^{(k)}\}$ are NA. Furthermore, as long as T is not too small (e.g., $T \geq 5$), they practically achieve EA because the relative loss of EA is k^{-2T}, which approaches zero very fast (it is less than 0.02% even for $k = 3$ and $T = 5$). However, in theory $\{U_T^{(1)}, \ldots, U_T^{(k)}\}$ achieves EA only for $T = \infty$. This requires showing first that $\{U_\infty^{(1)}, \ldots, U_\infty^{(k)}\}$ are well defined and second that they are NA. The first question is easy to answer but the second is not, mainly because the support of $\{U_\infty^{(1)}, \ldots, U_\infty^{(k)}\}$ is a "Cantor dust" type of fractal, as investigated in Craiu and Meng (2002).

In practice, one must choose T and k. We have already seen that large T results in more extreme antithesis, while the simulations performed in Section 2 show that as k increases, the efficiency increases too. The following result, proved in Appendix A.2, further shows that, *without* taking into account the computational cost, the larger T the better.

THEOREM 5. *For any monotone* $h \in L^2[0, 1]$, *the correlation* $\mathrm{Corr}(h(U_{T+1}^{(1)}),$ $h(U_{T+1}^{(2)}))$ *is decreasing as a function of* T. *That is, for any* $T \geq 0$,

$$(4.4) \qquad \mathrm{Corr}(h(U_{T+1}^{(1)}), h(U_{T+1}^{(2)})) \leq \mathrm{Corr}(h(U_{T}^{(1)}), h(U_{T}^{(2)})).$$

However, T and k cannot be increased indefinitely in practice. Intuitively, when k becomes large, the choice of T should become less important because a large k means we have a deep stratification even with $T = 1$. Consequently, further stratification within each stratum, which is essentially what each new iteration does, becomes less important. The following result, proved in Appendix A.3, provides a theoretical support of this intuition by showing that for large k, the impact of T is negligible.

THEOREM 6. *For any* $h \in L^2[0, 1]$ *and a fixed* T, $\{h(U_T^{(1)}), \ldots, h(U_T^{(k)})\}$ *achieves EA asymptotically as* $k \to +\infty$, *that is,*

$$(4.5) \qquad \mathrm{Corr}(h(U_T^{(1)}), h(U_T^{(2)})) = -\frac{1}{k-1} + o(k^{-1}).$$

This result implies that asymptotically, as $k \to \infty$, any ILHS iteration achieves EA for any square-integrable function, not just monotone functions. In practice, however, k must be finite, and often quite small for the sake of computational cost. For fixed k, the following result, proved in Appendix A.4, shows that even with k as small as 3, T does not need to be large in order to achieve practically the same result as $T = \infty$, at least for all monotone estimand functions.

THEOREM 7. *Let* $D_{h_1, h_2}(t, t + m)$ *be the Kolmogorov–Smirnov distance between the two-way* (*marginal*) *joint CDF of* $\{h_1(U_{t+m}^{(1)}), h_2(U_{t+m}^{(2)})\}$ *and of* $\{h_1(U_t^{(1)}), h_2(U_t^{(2)})\}$, *where* $h_l, l = 1, 2$, *are nondecreasing functions. Then*

$$(4.6) \qquad D_{h_1, h_2}(t, t + m) \leq k^{-(t-1)}(k-1)^{-(t+2)}.$$

This implies that if we take

$$T \geq \frac{d - 2\log_{10}(k-1) + \log_{10} k}{\log_{10}(k(k-1))},$$

then $D_{h_1, h_2}(T, \infty) \leq 10^{-d}$. In particular, as long as $T \geq 5$, $D_{h_1, h_2}(T, \infty) < 0.0001$ for any $k \geq 3$. Thus, taking T between 5 and 10 is enough for almost any practical purpose. The generality of Theorem 7 should be emphasized since the same bound in (4.6) holds for any monotone h_1 and h_2.

4.4. *Variance reduction factors for indicator functions.* As an attempt to find some theoretical support of the empirical findings reported in Section 2 that the cost-effective choice of k appears to be somewhere between 3 and 10, we report here our findings of the theoretical value of $S_k^{(f)}$ as a function of k when $f = \mathbb{1}_{\{x \leq c\}}$. We choose the indicator function both because of its analytical tractability and its practical relevance (e.g., for estimating CDF), and because it serves as a building block for general functions. We start with the LHS method, for which

$$(4.7) \qquad S_k^{(f)}(c) = 1 + (k-1)\frac{F(c,c) - c^2}{c(1-c)} = \frac{(1 - \{kc\})\{kc\}}{kc(1-c)},$$

where $\{a\}$ denotes the fractional part of a. This result follows from the expression of the joint CDF, $F(c,c) = P(U_1 \leq c, U_2 \leq c)$, given by (A.6) from Appendix A.2; note (A.5) implies that the same $S_k^{(f)}(c)$ holds for any ILHS.

The left panel of Figure 8, realized using the freely available software RGL developed by Duncan Murdoch, plots $S_k^{(f)}(c)$ as a function of both k (up to 30) and $c \in (0, 1)$. Its fascinating shape reveals that as long as c is not too close to 0 or 1, $S_k^{(f)}(c)$ will be rather small. This is more clearly seen in the first two rows of Figure 9, where $S_k^{(f)}(c)$ is plotted against c for given k. It is intuitive that when c approaches 0 or 1, the effect of antithetic coupling fades because the indicator function approaches the constant function. This can be formalized by considering that for $1/k \leq c \leq (k-1)/k$ the maximum of $S_k^{(f)}(c)$ over this range, as shown in
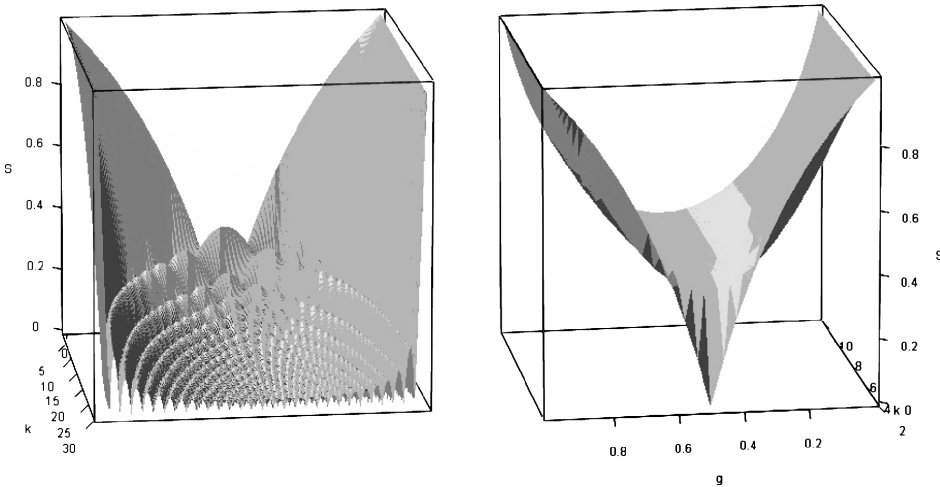

FIG. 8. *Variance reduction for indicator functions for* Uniform(0, 1) *random variables generated using LHS* (left panel) *and for normal variates* (right panel). *Note for better visualization the left panel is seen from back, with the c axis hidden.*

Appendix A.5, is given by

$$(4.8) \qquad S_k^* = \frac{k}{3k - 4 + 2\sqrt{2(k-1)(k-2)}},$$

corresponding to the dashed lines in the first two rows in Figure 9. The use of a large k is seen to be twice advantageous: first, it decreases the S_k^* and second, it shrinks the area (of c) where the antithetic coupling is not effective. The plots also show clearly that $k = 2$ is least effective.

The S_k^* of (4.8) decrease from $S_3^* = 1/3$ to $S_\infty^* = (3 + 2\sqrt{2})^{-1} \approx 0.172$. However, the use of large k also increases the computational cost. Striving for a balance, we consider $R_k = (S_3^* - S_k^*)/(S_3^* - S_\infty^*)$, a measure of the relative gain in efficiency obtained when we increase from $k = 3$ to a finite $k > 3$ instead of $k = \infty$. Table 2 gives, for a few $\alpha \in [0.5, 1]$, $k_\alpha = \min\{k : R_k \geq \alpha\}$. It is seen that with $k = 9$ we already have achieved 80% of the additional gain. The rapid decrease of $S_k^{(f)}(c)$ as a function of k, for fixed c, can also be seen in the last two rows of Figure 9.
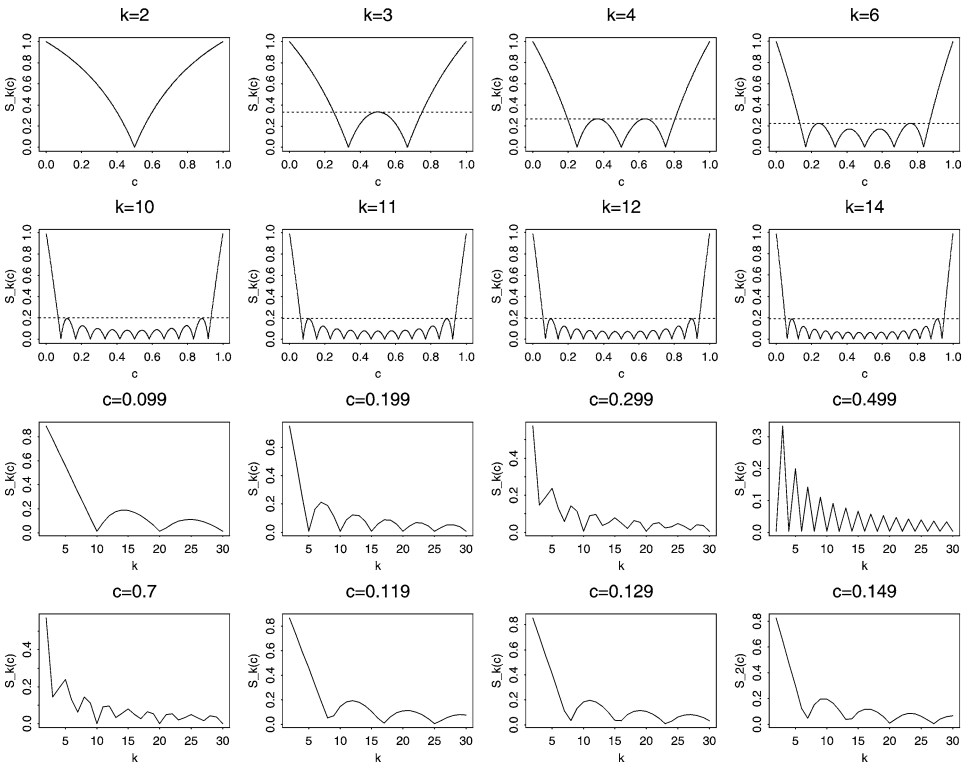

FIG. 9.   *Variance reduction factor for indicator function under LHS method*, *as a function of c for different values of k* (first two rows), *and as a function of k for different values of c* (last two rows).

TABLE 2
*The minimum k_α for achieving α% of maximal possible gain
in efficiency over using $k = 3$*

α	50%	60%	70%	80%	90%	95%	99%
k_α	5	6	7	9	17	32	152

The above exercise can be repeated for the multivariate normal method, that is, when $(X_1, X_2, \ldots, X_k)$ is a multivariate normal vector with $\text{Corr}(X_i, X_j) = -(k-1)^{-1}$ and $N(0, 1)$ margins. The calculation of $S_k^{(f)}$ is a bit more involved than in the LHS case. Specifically, with the orthogonal transformation $Z = X_1 + X_2$, $W = X_1 - X_2$ (i.e., Z and W are independent), we obtain

$$\Phi_k(c, c) \equiv P(X_1 \leq c, X_2 \leq c) = P(Z - 2c \leq W \leq -Z + 2c, Z \leq 2c)$$

(4.9)
$$= 2 \int_{-\infty}^{2c/\sqrt{2(1+\rho_k)}} \Phi\left(\frac{2c - z\sqrt{2(1 + \rho_k)}}{\sqrt{2(1 - \rho_k)}}\right) \phi(z) \, dz$$

$$- \Phi\left(\frac{2c}{\sqrt{2(1 + \rho_k)}}\right),$$

where $\rho_k = -(k - 1)^{-1}$. Consequently, we can compute $S_k^{(f)}(c)$ via

(4.10)
$$S_k^{(f)}(c) = 1 + (k - 1) \frac{\Phi_k(c, c) - \Phi^2(c)}{\Phi(c)(1 - \Phi(c))}.$$

In the normal case, we expect a behavior of $S_k^{(f)}(c)$ similar to that in the uniform case when c is close to the limits of the range, $(-\infty, \infty)$. However, for plotting purposes we use a one-to-one transformation of c, $g = \Phi(c)$, and we plot $S_k^{(f)}(c)$ against g; this is the same as the VRF for the uniform variates via the normal approach, namely, if $U_i = \Phi(X_i)$. One can notice in the right panel of Figure 8 that everywhere but in a region of c around the distribution's center, $S_k^{(f)}(c)$ is decreasing in k. However, the decrease is much less abrupt than in the uniform case, which indicates that it will take a larger k to reach the same relative efficiency as given in Table 2. This suggests that generalizing the findings in Table 2 to other situations is by no means automatic.

This example also shows that it is not always true that the larger k, the better. For example, when $c = 0$, using $k = 2$ completely eliminates the variance because of the symmetry in $f = \mathbb{1}_{\{x \leq 0\}}$; see the right panel of Figure 8. This can also be verified directly from (4.10), which becomes

$$S_k^{(f)}(0) = 1 + (k - 1)\left[8 \int_0^\infty \Phi\left(z\sqrt{1 - \frac{2}{k}}\right) \phi(z) \, dz - 3\right],$$

which is zero when $k = 2$. However, as $k \to \infty$ we can show that $S_k^{(f)}(0) \to$ $1 - 2/\pi = 0.3634$. This reminds us that although there are good rules of thumb for general practice, in terms of choosing k as well as other choices (e.g., the generating methods) presented in this paper, one should not adopt them blindly without examining the special structures of the problem at hand.

APPENDIX

A.1. *Proof for the unit circle example.* We first note that (3.4) implies, for any $t \geq 1$,

$$(A.1) \qquad (\theta_t^X + \theta_t^Y) \bmod (2\pi) = -(\theta_{t-1}^X + \theta_{t-1}^Y) \bmod (2\pi).$$

From the right panel of Figure 7, $\widehat{MX_0} = \widehat{QN} = \widehat{QY_0} + \widehat{Y_0N}$, where $\widehat{AB}$ denotes the counter-clockwise arc between points A and B on the unit circle. Furthermore, $\widehat{MX_0} = \theta_0^X - \pi$, $\widehat{QY_0} = 2\tau$ and $\widehat{Y_0N} = 2\pi - \theta_0^Y$. Combining these identities, we obtain $\theta_0^X + \theta_0^Y = 2\tau + 3\pi$. Consequently, (A.1) implies (3.5) by noting that $\theta_t^X + \theta_t^Y \in (0, 4\pi)$ and the "alternating" nature of (A.1), which is also clear from Figure 7, and thus the four possible lines in (3.5) are reachable from each other by (X_t, Y_t). Similar arguments apply to other possible initial configurations of X_0 and Y_0 (e.g., when X_0 and Y_0 are on different sides of the horizontal axes).

To prove $\rho_t(\tau) = -\cos(2\tau)$, we observe that (3.5) implies $\sin^2(\theta_t^X + \theta_t^Y) = \sin^2(2\tau)$ for any $t \geq 0$. Using the identity $\sin^2(\alpha + \beta) = \cos^2(\alpha) + \sin^2(\beta) - 2\cos(\alpha)\cos(\beta)\cos(\alpha + \beta)$, we then obtain

$$(A.2) \qquad \cos^2(\theta_t^Y) + \cos^2(\theta_t^X) + 2\cos(\theta_t^Y)\cos(\theta_t^X)\cos(2\tau) = \sin^2(2\tau).$$

Thus the orbit of $(X_t^{(1)}, Y_t^{(1)}) \equiv (\cos(\theta_t^X), \cos(\theta_t^Y))$ is an ellipse. Taking expectations on both sides of (A.2) and using $E[\cos^2(\theta_t^Y)] = E[\cos^2(\theta_t^X)] = 1/2$, we obtain $2E[\cos(\theta_t^Y)\cos(\theta_t^X)]\cos(2\tau) = -\cos^2(2\tau)$. Together with $\mathrm{Var}(\cos^2(\theta_t^X)) = \mathrm{Var}(\cos^2(\theta_t^Y)) = 1/2$, this yields $\rho_t(\tau) = -\cos(2\tau)$ when $\tau \neq \pi/4$. For $\tau = \pi/4$, (3.5) gives $\theta_t^Y = (2k + 1)\pi \pm \pi/2 - \theta_t^X$, and thus $\cos(\theta_t^Y) = \pm\sin(\theta_t^X)$. Consequently, $2E[\cos(\theta_t^Y)\cos(\theta_t^X)] = \pm E[\sin(2\theta_t^X)] = 0$, so $\rho_t(\tau) = -\cos(2\tau)$ still holds.

A.2. *Proof of Theorem* 5. Since the (marginal) distribution of $U_T^{(j)}$ is Uniform$(0, 1)$ and thus does not depend on T, by the Hoeffding identity (1.1), inequality (4.4) becomes immediate if we can establish

$$(A.3) \qquad F_{T+1}^{(h)}(u, v) \leq F_T^{(h)}(u, v) \qquad \forall (u, v),$$

where $F_t^{(h)}(u, v)$ is the joint CDF of $(h(U_t^{(1)}), h(U_t^{(2)}))$.

We first show that (A.3) holds for $h(x) = x$ by mathematical induction. Given the exchangeability, we can assume $u \leq v$, and thus $[ku] \leq [kv]$, where $[x]$ denotes the integer part of x. Let $p_{ij} = P(U_t^{(1)} \leq ku - i, U_t^{(2)} \leq kv - j)$; then by the recursion (4.3),

$$F_{t+1}(u, v) = P\big(U_t^{(1)} \leq ku - \sigma_t^{(1)}, U_t^{(2)} \leq kv - \sigma_t^{(2)}\big)$$

(A.4)
$$= \frac{1}{k(k-1)} \sum_{i=0}^{[ku]} \sum_{j=0}^{[kv]} p_{ij} \mathbb{1}_{\{i \neq j\}}.$$

To evaluate this expression, we consider the following possibilities according to the value of (i, j):

(A) When $i \leq [ku] - 1$ and $j \leq [kv] - 1$, $p_{ij} = P(U_t^{(1)} \leq 1, U_t^{(2)} \leq 1) = 1$; there are $[ku]([kv] - 1)^+$ such pairs of (i, j)'s within $0 \leq i \neq j \leq k$. Recall $(x)^+ = \max\{x, 0\}$.

(B) When $i \leq [ku] - 1$ and $j = [kv]$, $p_{ij} = P(U_t^{(2)} \leq kv - [kv]) = \{kv\}$, where $\{x\} = x - [x]$; there are $[ku]$ such pairs of (i, j)'s.

(C) When $i = [ku]$ and $j \leq [kv] - 1$, $p_{ij} = P(U_t^{(1)} \leq ku - [ku]) = \{ku\}$; there are $[kv]$ such pairs of (i, j)'s when $[ku] = [kv]$, but only $[kv] - 1$ such pairs when $[ku] < [kv]$ because of the requirement that $i \neq j$.

(D) When $i = [ku]$ and $j = [kv]$, $p_{ij} = P(U_t^{(1)} \leq ku - [ku], U_t^{(2)} \leq kv - [kv]) = F_t(\{ku\}, \{kv\})$. There is no such pair when $[ku] = [kv]$ and one such pair when $[ku] < [kv]$.

Putting these four possibilities together, we have

$$F_{t+1}(u, v)$$

(A.5)
$$= \begin{cases} 0, & \text{if } [kv] = 0, \\ \dfrac{ku(kv-1) - \{ku\}(\{kv\} - 1)}{k(k-1)}, & \text{if } 0 < [kv] = [ku], \\ \dfrac{ku(kv-1) - \{ku\}\{kv\} + F_t(\{ku\}, \{kv\})}{k(k-1)}, & \text{if } [kv] > [ku]. \end{cases}$$

From (A.5), if (A.3) holds for $T = 0$ the rest follows immediately by induction. For $t = 1$, since $F_0(\{ku\}, \{kv\}) = \{ku\}\{kv\}$, we have from (A.5)

(A.6) $$F_1(u, v) = \begin{cases} 0, & \text{if } [kv] = 0, \\ \dfrac{ku(kv-1) - \{ku\}(\{kv\} - 1)}{k(k-1)}, & \text{if } 0 < [kv] = [ku], \\ \dfrac{u(kv-1)}{k-1}, & \text{if } [kv] > [ku]. \end{cases}$$

Among the three expressions above, the second one is the largest as a function of u and v. It is easy to check that this second function is less than or equal to

$F_0(u, v) = uv$ if and only if

$$(A.7) \qquad\qquad \{ku\}(1 - \{kv\}) \leq ku(1 - v).$$

Using $[ku] = [kv]$, (A.7) is also equivalent to $\{kv\}(u - \{ku\}) \leq [ku](1 - u)$, which is obvious when $u \leq \{ku\}$. When $0 \leq \{ku\} < u$, $\{ku\}(1 - \{kv\}) \leq u(1 - \{kv\})$. But then $u(1 - \{kv\}) \leq ku(1 - v)$ is equivalent to $[kv] \leq k - 1$, which is obviously true except when $v = 1$. But the $v = 1$ case is trivial because then $u = 1$ in order to maintain $[ku] = [kv] = k$, and thus $\{ku\} = 0$, and hence (A.7) holds for all $0 \leq u \leq v \leq 1$. This proves $F_1(u, v) \leq F_0(u, v)$ for all (u, v) and hence (A.3) by induction.

To show that (A.3) holds for any nondecreasing h, let $x_w = \sup\{x : h(x) \leq w\}$ for any given w. Then $\{x : h(x) \leq w\} = A_w^{(h)}(x)$, where $A_w^{(h)}(x) = \{x : x \leq x_w\}$ if $h(x_w) \leq w$ and $A_w^{(h)}(x) = \{x : x < x_w\}$ if $h(x_w) > w$. This means that the probability calculations of event $\{U : h(U) \leq w\}$ are the same as those for either $\{U : U \leq x_w\}$ or $\{U : U < x_w\}$. This allows us to go from $F_T^{(h)}(u, v)$ to $F_T(x_u, x_v)$, for which we already have proved the desired result. A technical complication here is that, depending on the continuity properties of h at x_u and x_v, one or two "$\leq$" operations in the definition of $F_T(x_u, x_v) = P(U_T^{(1)} \leq x_u, U_T^{(2)} \leq x_v)$ may need to be replaced by "$<$". However, as far as (A.3) is concerned, such modifications are immaterial because they do not affect any part of (A)–(D). Or mathematically, we have, for any given (u, v),

$$\begin{aligned}
F_{T+1}^{(h)}(u, v) &= P\big(h(U_{T+1}^{(1)}) \leq u, h(U_{T+1}^{(2)}) \leq v\big) \\
&= E\big[A_u^{(h)}(U_{T+1}^{(1)}) \cap A_v^{(h)}(U_{T+1}^{(2)})\big] \leq E\big[A_u^{(h)}(U_T^{(1)}) \cap A_v^{(h)}(U_T^{(2)})\big] \\
&= P\big(h(U_T^{(1)}) \leq u, h(U_T^{(2)}) \leq v\big) = F_T^{(h)}(u, v),
\end{aligned}$$

where the middle inequality holds because (A.3) holds with $h(x) = x$, possibly with the aforementioned modifications from "$\leq$" to "$<$" in the definition of the bivariate CDF.

A.3. *Proof of Theorem* 6. For any $h \in L^2[0, 1]$, Stein (1987) showed that Theorem 6 is true for the original LHS, that is, when $T = 1$. Therefore, for any monotone $h \in L^2[0, 1]$, Theorem 6 is a consequence of Theorem 5, because the smallest possible value of $\mathrm{Corr}(h(U_T^{(1)}), h(U_T^{(2)}))$ is $-1/(k - 1)$, as seen in (3.9). More specifically, for any $T > 1$, any possible limit of $-(k - 1)\,\mathrm{Corr}(h(U_T^{(1)}), h(U_T^{(2)}))$, as $k \to \infty$, must be bounded below and above by 1, and hence can only be 1. For nonmonotone $h \in L^2[0, 1]$, the proof turns out to be much more technical. It essentially requires going through Stein's (1987) original arguments but using the more complicated bivariate CDF via (A.5); details are given in Craiu (2001).

A.4. *Proof of Theorem* 7. First we prove (4.6) when both h_1 and h_2 are identity functions. By (A.5),

$$(A.8) \qquad D(t, t+1) = \sup_{[ku] \neq [kv]} \left\{ \left| \frac{F_t(\{ku\}, \{kv\})}{k(k-1)} - \frac{F_{t-1}(\{ku\}, \{kv\})}{k(k-1)} \right| \right\}$$

$$(A.9) \qquad \qquad \leq \sup_{u,v} \left\{ \left| \frac{F_t(u, v)}{k(k-1)} - \frac{F_{t-1}(u, v)}{k(k-1)} \right| \right\} = \frac{D(t-1, t)}{k(k-1)}.$$

This allows us to conclude that

$$(A.10) \qquad \qquad D(t, t+1) \leq \frac{D(0, 1)}{[k(k-1)]^t},$$

where $D(0, 1) = \sup_{u,v} \{|F^{(1)}(u, v) - F^{(0)}(u, v)|\} = \sup_{u,v} \{uv - F^{(1)}(u, v)\}$. By (A.6), $D(0, 1)$ is the maximum value of the following three suprema:

(i) $\quad \displaystyle\sup_{[ku]=[kv]=0} uv = \frac{1}{k^2};$

(ii) $\quad \displaystyle\sup_{0 < [kv]=[ku]} \left\{ uv - \frac{ku(kv-1) - \{ku\}(\{kv\} - 1)}{k(k-1)} \right\}$

$\qquad = \displaystyle\sup_{0 < [kv]=[ku]} \left\{ \frac{ku(1-v) - \{ku\}(1-\{kv\})}{k(k-1)} \right\} \leq \frac{1}{k-1};$

(iii) $\quad \displaystyle\sup_{[kv] \neq [ku]} \left\{ uv - \frac{ku(kv-1)}{k(k-1)} \right\} = \displaystyle\sup_{[kv] \neq [ku]} \left\{ \frac{u(1-v)}{k-1} \right\} = \frac{1}{k-1}.$

Putting all these facts together, we have

$$D(t, t+m) \leq \sum_{i=t}^{t+m-1} D(i, i+1)$$

$$\leq \sum_{i=t}^{t+m-1} \frac{1}{k^i (k-1)^{i+1}} = \frac{1}{k^t (k-1)^{t+1}} \frac{1 - [k(k-1)]^{-m}}{1 - [k(k-1)]^{-1}},$$

which is less than $k^{-(t-1)}(k-1)^{-(t+2)}$ when $k \geq 2$. The extension to nondecreasing functions h_1, h_2 is immediate by defining $h^{-1}(u) = \sup\{s : h(s) \leq u\}$ and proceeding in the same fashion as in Appendix A.2. That is, for any nondecreasing functions h_1 and h_2,

$$D_{h_1, h_2}(t, t+m) \leq \sup_{u,v} \left| F_{t+m}\left(h_1^{-1}(u), h_2^{-1}(v)\right) - F_t\left(h_1^{-1}(u), h_2^{-1}(v)\right) \right|$$

$$\leq D(t, t+m),$$

where the definition of $F_{t+m}(x_u, x_v)$ and $F_t(x_u, x_v)$ may need to be modified as in Appendix A.2.

A.5. *Maximum reduction factor in the uniform case.* To maximize (4.7) subject to $1 \le [kc] \le k - 1$, let $i_c = [kc]$ and $f_c = \{kc\}$. By symmetry, the maximum of (4.7) occurs at either $i_c = 1$ or $i_c = k - 2$. When $i_c = 1$, (4.7) can be expressed as

$$(A.11) \qquad S_k^{(f)}(f_c) = \frac{k(1 - f_c)f_c}{(1 + f_c)(k - 1 - f_c)}.$$

Straightforward differentiation then shows that the maximizer must satisfy $(k - 3)f_c^2 + 2(k - 1)f_c - (k - 1) = 0$, which only has one acceptable solution $f_c = \frac{\sqrt{k-1}}{\sqrt{2k-4}+\sqrt{k-1}}$. Therefore the maximizer of (4.7) is $c_1^* = \frac{1}{k}(1 + \frac{\sqrt{k-1}}{\sqrt{2k-4}+\sqrt{k-1}})$ with the corresponding maximal value of $S_k^{(f)}$ given in (4.8). The maximum from $i_c = k - 2$ is the same, with the maximizer $c_2^* = \frac{1}{k}(k - 2 + \frac{\sqrt{k-1}}{\sqrt{2k-4}+\sqrt{k-1}})$.

REFERENCES

ALBERT, J. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679.

ARJAS, E. and GASBARRA, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *J. Amer. Statist. Assoc.* **91** 1101–1109.

ARVIDSEN, N. I. and JOHNSSON, T. (1982). Variance reduction through negative correlation—a simulation study. *J. Statist. Comput. Simulation* **15** 119–127.

BAILEY, B. J. R. (1981). Alternatives to Hastings' approximation to the inverse of the normal cumulative distribution function. *Appl. Statist.* **30** 275–276.

BILLINGSLEY, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

BONDESSON, L. (1983). Problem 128. *Statist. Neerlandica* **37** 149.

BOREL, E. (1924). *Traité du Calcul des Probabilités et de ses Applications*. Gauthier-Villars, Paris.

CASELLA, G., LAVINE, M. and ROBERT, C. P. (2001). Explaining the perfect sampler. *Amer. Statist.* **55** 299–305.

CASELLA, G., MENGERSEN, K. L., ROBERT, C. P. and TITTERINGTON, D. M. (2002). Perfect samplers for mixtures of distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 777–790.

CRAIU, R. V. (2001). Multivalent framework for approximate and exact sampling and resampling. Ph.D. dissertation, Univ. Chicago.

CRAIU, R. V. (2004). Antithetic acceleration of the multiple-try Metropolis. Technical Report 0407, Dept. Statistics, Univ. Toronto.

CRAIU, R. V. and MENG, X.-L. (2001). Chance and fractals. *Chance* **14**(2) 47–52.

CRAIU, R. V. and MENG, X.-L. (2002). Multi-process parallel antithetic coupling for backward and forward MCMC. Technical Report 0205, Dept. Statistics, Univ. Toronto.

DAMIEN, P., WAKEFIELD, J. and WALKER, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 331–344.

ESARY, J. D., PROSCHAN, F. and WALKUP, D. W. (1967). Association of random variables, with applications. *Ann. Math. Statist.* **38** 1466–1474.

FILL, J. A. (1998). An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.* **8** 131–162.

FILL, J. A., MACHIDA, M., MURDOCH, D. J. and ROSENTHAL, J. S. (2000). Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures Algorithms* **17** 229–316.

FISHMAN, G. S. (1972). Variance reduction in simulation studies. *J. Stat. Comput. Simul.* **1** 173–182.

FOSS, S. G. and TWEEDIE, R. L. (1998). Perfect simulation and backward coupling. *Comm. Statist. Stochastic Models* **14** 187–203.

FRÉCHET, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon. Sect. A*(*3*) **14** 53–77.

FRIGESSI, A., GÅSEMYR, J. and RUE, H. (2000). Antithetic coupling of two Gibbs sampler chains. *Ann. Statist.* **28** 1128–1149.

GELFAND, A. E., SMITH, A. F. M. and LEE, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87** 523–532.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

GENTLE, J. (1998). *Random Number Generation and Monte Carlo Methods*. Springer, New York.

GEROW, K. and HOLBROOK, J. (1996). Statistical sampling and fractal distributions. *Math. Intelligencer* **18**(2) 12–22.

HÄGGSTRÖM, O. and NELANDER, K. (1998). Exact sampling from anti-monotone systems. *Statist. Neerlandica* **52** 360–380.

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1965). *Monte Carlo Methods*. Methuen, London.

HAMMERSLEY, J. M. and MORTON, K. V. (1956). A new Monte Carlo technique: Antithetic variates. *Proc. Cambridge Philos. Soc.* **52** 449–475.

HELTON, J. C. and DAVIS, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety* **81** 23–69.

HOBERT, J. P., ROBERT, C. P. and TITTERINGTON, D. M. (1999). On perfect simulation for some mixtures of distributions. *Statist. Comput.* **9** 287–298.

HOEFFDING, W. (1940). Masstabinvariante Korrelationtheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universitat Berlin* **5** 181–233.

IMAN, R. L. (1999). Latin hypercube sampling. *Encyclopedia of Statistical Sciences Update* **3** 408–411. Wiley, New York.

JOAG-DEV, K. and PROSCHAN, F. (1983). Negative association of random variables, with applications. *Ann. Statist.* **11** 286–295.

JOE, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

KENDALL, W. S. (1998). Perfect simulation for the area-interaction point process. *Probability Towards 2000. Lecture Notes in Statist.* **128** 218–234. Springer, New York.

KENNEDY, W. J. and GENTLE, J. E. (1980). *Statistical Computing*. Dekker, New York.

KISH, L. (1965). *Survey Sampling*. Wiley, New York.

LANCASTER, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **44** 289–292.

LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.

LEW, R. A. (1981). An approximation to the cumulative normal distribution with simple coefficients. *Appl. Statist.* **30** 299–301.

LOH, W.-L. (1996). On Latin hypercube sampling. *Ann. Statist.* **24** 2058–2080.

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245.

MENG, X.-L. (2000). Towards a more general Propp–Wilson algorithm: Multistage backward coupling. In *Monte Carlo Methods* (N. Madras, ed.) 85–93. Amer. Math. Soc., Providence, RI.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

MØLLER, J. (1999). Perfect simulation of conditionally specified models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 251–264.

MØLLER, J. and NICHOLLS, G. K. (1999). Perfect simulation for sample-based inference. Preprint, Dept. Mathematical Statistics, Chalmers Institute of Technology.

MORAN, P. A. P. (1967). Testing for correlation between non-negative variates. *Biometrika* **54** 385–394.

MURDOCH, D. J. (2000). Exact sampling for Bayesian inference: Unbounded state spaces. In *Monte Carlo Methods* (N. Madras, ed.) 111–121. Amer. Math. Soc., Providence, RI.

MURDOCH, D. J. and GREEN, P. J. (1998). Exact sampling from a continuous state space. *Scand. J. Statist.* **25** 483–502.

MURDOCH, D. J. and MENG, X.-L. (2001). Towards perfect sampling for Bayesian mixture priors. In *Bayesian Methods, with Applications to Science, Policy and Official Statistics* (E. George, ed.) 381–390. Eurostat.

NEAL, R. M. (2003). Slice sampling (with discussion). *Ann. Statist.* **31** 705–767.

OWEN, A. B. (1992). A central limit theorem for Latin hypercube sampling. *J. Roy. Statist. Soc. Ser. B* **54** 541–551.

OWEN, A. B. (1997). Monte Carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.* **34** 1884–1910.

OWEN, A. B. (2002). Quasi Monte Carlo. Presented at the First Cape Cod Workshop on Monte Carlo Methods.

PAGE, E. S. (1965). On Monte Carlo methods in congestion problems. II. Simulation of queuing systems. *Oper. Res.* **13** 300–305.

PEMANTLE, R. (2000). Towards a theory of negative dependence. *J. Math. Phys.* **41** 1371–1390.

PHILIPPE, A. and ROBERT, C. (2003). Perfect simulation of positive Gaussian distributions. *Statist. Comput.* **13** 179–186.

PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.

RUBINSTEIN, R. Y. and SAMORODNITSKY, G. (1985). Variance reduction by the use of common and antithetic random variables. *J. Statist. Comput. Simulation* **22** 161–180.

RÜSCHENDORF, L. and UCKELMANN, L. (2000). Variance minimization and random variables with constant sum. Technical report, Institüt für Mathematische Stochastik, Univ. Freiburg.

SIMON, G. (1976). Computer simulation swindles, with applications to estimates of location and dispersion. *Appl. Statist.* **25** 266–274.

SNIJDERS, T. A. B. (1984). Antithetic variates for Monte Carlo estimation of probabilities. *Statist. Neerlandica* **38** 55–73.

STEIN, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29** 143–151. [Corrigendum *Technometrics* **32** (1990) 367.]

THORISSON, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer, New York.

VAN DYK, D. and MENG, X.-L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10** 1–111.

WILSON, D. B. (1998). Annotated bibliography of perfectly random sampling with Markov chains. In *Microsurveys in Discrete Probability* (D. Aldous and J. Propp,

eds.) 209–220. Amer. Math. Soc., Providence, RI. Updated versions appear at http://dimacs.rutgers.edu/˜dbwilson/exact.

WILSON, D. B. (2000a). How to couple from the past using a read-once source of randomness. *Random Structures Algorithms* **16** 85–113.

WILSON, D. B. (2000b). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). In *Monte Carlo Methods* (N. Madras, ed.) 143–176. Amer. Math. Soc., Providence, RI.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO
CANADA M5S 3G3
E-MAIL: craiu@ustat.toronto.edu

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: meng@stat.harvard.edu