Ergodic averages for monotone functions using upper and lower dominating processes

Jesper Møller

Department of Mathematical Sciences, Aalborg University, Denmark.

Kerrie Mengersen

Department of Mathematical Sciences, Queensland University of Technology, Australia.

Summary. We show how the mean of a monotone function (defined on a state space equipped with a partial ordering) can be estimated, using ergodic averages calculated from upper and lower dominating processes of a stationary irreducible Markov chain. In particular, we do not need to simulate the stationary Markov chain and we eliminate the problem of whether an appropriate burn-in is determined or not. Moreover, when a central limit theorem applies, we show how confidence intervals for the mean can be estimated by bounding the asymptotic variance of the ergodic average based on the equilibrium chain. Our methods are studied in detail for three models using Markov chain Monte Carlo methods and we also discuss various types of other models for which our methods apply.

Keywords: Asymptotic variance; Bayesian models; Burn-in; Ergodic average; Ising model; Markov chain Monte Carlo; Mixture model; Monotonocity; Perfect simulation; Random walk; Spatial models; Upper and lower dominating processes

1. Introduction

Suppose that π is a given target distribution on a state space Ω and we wish to estimate the mean

$$\mu = \int \phi(x)\pi(\mathrm{d}x) \tag{1}$$

for a given real function ϕ. In many applications it is not known or at least not straightforward to generate a stationary chain, so instead a non-stationary chain $Y_1, Y_2 \ldots$ is generated by Markov chain Monte Carlo (MCMC) and μ is estimated by the ergodic average $\sum_{t=M+1}^{N} \phi(Y_t)/(N-M)$, where $M \geq 0$ is an "appropriate" burn-in and $N \gg M$ is "sufficiently" large, (see, for example, Robert and Casella 2004). This estimator is consistent provided the chain is irreducible and M is independent of the Y chain. The problem is to determine M and N so that the estimator is close to μ with a high degree of confidence.

Propp and Wilson (1996) showed how upper and lower dominating processes can be used for generating a perfect (or exact) simulation of a stationary Markov chain at a fixed time, provided the chain is monotone with respect to a partial ordering on Ω for which there exists unique maximal and minimal states. In this paper we introduce similar ideas but our aim is to obtain reliable estimates of mean values rather than perfect simulations.

More specifically, we consider irreducible Markov chains with π as the invariant distribution and make the following additional assumptions. Let $X = (X_t;\, t = 1, 2, \ldots)$ denote the possibly unknown equilibrium chain, i.e. $X_1 \sim \pi$ and hence $X_t \sim \pi$ for all $t \geq 1$, and

let

$$\bar{\phi}_t = \frac{1}{t}\sum_{s=1}^{t}\phi(X_s)$$

denote the ergodic average estimating μ. Assume there exist stochastic processes $U = (U_t;\, t = 1, 2, \ldots)$ and $L = (L_t;\, t = 1, 2, \ldots)$ so that

$$\bar{\phi}_t^L \le \bar{\phi}_t \le \bar{\phi}_t^U, \quad t = 1, 2, \ldots, \tag{2}$$

where the ergodic averages

$$\bar{\phi}_t^L = \frac{1}{t}\sum_{s=1}^{t}\phi(L_s), \quad \bar{\phi}_t^U = \frac{1}{t}\sum_{s=1}^{t}\phi(U_s) \tag{3}$$

are consistent estimators of μ. Though U and L will be Markov chains in most of our detailed examples, they do not need to be so as exemplified in Section 4.1 (explaining why we write "processes"). To ensure (2) we assume that with respect to a partial ordering $\prec$ on Ω, U and L are bounding X, i.e.

$$L_t \prec X_t \prec U_t, \quad t = 1, 2, \ldots, \tag{4}$$

and ϕ is monotone

$$x \prec y \implies \phi(x) \le \phi(y) \tag{5}$$

(or, as discussed later on, ϕ is a linear combination of monotone functions). Then (2) holds, and so it suffices to consider the processes $(\bar{\phi}_t^L;\, t = 1, 2, \ldots)$ and $(\bar{\phi}_t^U;\, t = 1, 2, \ldots)$. Consequently, we do not need to simulate the equilibrium chain and we eliminate the problem of whether an appropriate burn-in is determined or not. Assuming a central limit theorem applies, we show how confidence intervals for the mean can be estimated by bounding the asymptotic variance of $\bar{\phi}_t$. Note also that to assess if the process $(\phi(X_t);\, t = 1, 2, \ldots)$ has stabilised into equilibrium, it suffices to consider the processes $(\phi(L_t);\, t = 1, 2, \ldots)$ and $(\phi(U_t);\, t = 1, 2, \ldots)$. Our methods are studied in detail for three models using MCMC methods and we also discuss various types of other models for which our methods apply.

Note that in contrast to the Propp-Wilson algorithm we do not require that U_t and L_t coalesce for all sufficiently large t. Equivalently, we do not require that X is uniformly ergodic (Foss and Tweedie, 1998). For extensions of the Propp-Wilson algorithm which may be of relevance for our methods, see the references in Section 5.

The paper is organised as follows. Section 2 presents our ideas in a simple setting for a random walk, while Section 3 considers a general setting. Section 4 illustrates how our methods apply on the Ising model and a mixture model in which the weights are unknown. Finally, Section 5 discusses extensions and application areas of the methods.

2. A simple example

Despite its conceptual ease, the random walk example below is a challenging platform on which to evaluate the performance of our proposed methods in Section 3.

2.1. Upper and lower bounds for a random walk

Consider a stationary random walk $X = (X_t;\, t = 1, 2, \ldots)$ on a finite state space $\Omega = \{0, 1, \ldots, k\}$ with transition probabilities

$$p_i = P(X_{t+1} = \min\{i+1, k\} | X_t = i) > 0,$$

$$q_i = P(X_{t+1} = \max\{i-1, 0\} | X_t = i) = 1 - p_i > 0,$$

for $i = 0, 1, \ldots, k$, and invariant distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_k)$ given by

$$\pi_i = \pi_0 \prod_{j=0}^{i-1} p_j / q_{j+1}, \quad i = 1, \ldots, k.$$

We can construct this by a so-called stochastic recursive sequence (SRS). Let $X_1, R_1, R_2, \ldots$ be independent random variables with $X_1 \sim \pi$ and $R_t \sim$ Uniform $[0, 1]$, $t = 0, 1, \ldots$. Define a so-called updating function $\chi : \Omega \times [0, 1] \to \Omega$ by

$$\chi(i, r) = \begin{cases} \min\{i+1, k\} & \text{if } r \le p_i \\ \max\{i-1, 0\} & \text{otherwise.} \end{cases}$$

Then the SRS is given by

$$X_{t+1} = \chi(X_t, R_t), \quad t = 1, 2, \ldots.$$

This construction allows us to bound the equilibrium chain by an upper chain $U = (U_t; t = 1, 2, \ldots)$ and a lower chain $L = (L_t; t = 1, 2, \ldots)$ defined by

$$U_1 = k, \quad U_{t+1} = \chi(U_t, R_t), \quad t = 1, 2, \ldots,$$

$$L_1 = 0, \quad L_{t+1} = \chi(L_t, R_t), \quad t = 1, 2, \ldots.$$

Thereby

$$L_t \le X_t \le U_t, \quad t = 1, 2, \ldots, \tag{6}$$

and hence also for the ergodic averages

$$\bar{L}_t = \frac{1}{t} \sum_{s=1}^{t} L_s, \quad \bar{X}_t = \frac{1}{t} \sum_{s=0}^{t} X_s, \quad \bar{U}_t = \frac{1}{t} \sum_{s=0}^{t} U_s,$$

we have that

$$\bar{L}_t \le \bar{X}_t \le \bar{U}_t, \quad t = 1, 2, \ldots. \tag{7}$$

By irreducibility, as t grows, $\bar{L}_t$ and $\bar{U}_t$ converge to μ. Note that (4) and (5) are satisfied with $\prec$ given by $\le$ and ϕ the identity function. Indeed (4)-(7) are satisfied if we replace X by any Markov chain Y using the same coupling construction as above, i.e. when $Y_1 \in \Omega$ is an arbitrary initial state and $Y_{t+1} = \chi(Y_t, R_t),\ t = 1, 2, \ldots$.

2.2. Bounding the asymptotic variance for the ergodic average

In this simple example, the mean $\mu = \sum_{i=1}^{k} i\pi_i$ is easily determined, and so there is no need for estimating it by an ergodic average. Moreover, it is of course easy to generate X_1 from π, and hence to generate $\bar{X}_t$. However, in more complex situations as considered later in Sections 3-5, the mean value of interest is unknown and it is usually hard to make a draw from the invariant distribution. We can instead generate the upper and lower chains and use (7) (or the extensions considered in the following sections) together with the following considerations.

Since X is ergodic and Ω is finite, a central limit theorem (CLT) applies:

$$\sqrt{t}(\bar{X}_t - \mu) \text{ converges in distribution to Normal}(0, \sigma^2) \text{ as } t \to \infty \tag{8}$$

where

$$\sigma^2 = \sum_{t=-\infty}^{\infty} \gamma_{|t|} < \infty, \quad \gamma_t = \mathrm{Cov}(X_1, X_{t+1}). \tag{9}$$

We estimate σ^2 using for example a window type estimator (Geyer, 1992) or batch means (Ripley, 1987). For specificity, we consider in the sequel a window type estimator

$$\hat{\sigma}_N^2 = \sum_{t=-m}^{m} \hat{\gamma}_{|t|,N} \tag{10}$$

based on $X_1, \ldots, X_N$, but similar considerations will apply for batch means. Here

$$\hat{\gamma}_{t,N} = \frac{1}{N} \sum_{s=1}^{N-t} (X_{s+t} - \bar{X}_N)(X_s - \bar{X}_N), \tag{11}$$

see, for example, Priestly (1981, pp. 323-324). Geyer's initial series estimator is given by letting $m = 2l + 1$ where l is the largest integer so that the sequence $\hat{\gamma}_{2t,N} + \hat{\gamma}_{2t+1,N}$, $t = 0, \ldots, l$, is strictly positive, strictly decreasing and strictly convex. For an irreducible and reversible Markov chain this provides a consistent conservative estimator of σ^2, cf. Geyer (1992). By (6), (7) and (11), $\hat{\sigma}_N^2$ is bounded from above and below by

$$\hat{\sigma}^2_{\max,N} = \sum_{t=-m}^{m} a_{|t|,N}, \quad \hat{\sigma}^2_{\min,N} = \sum_{t=-m}^{m} b_{|t|,N}, \tag{12}$$

where for $t \geq 0$,

$$a_{t,N} = \frac{1}{N} \sum_{s=1}^{N-t} (U_{s+t}U_s - L_{s+t}\bar{L}_N - L_s\bar{L}_N + \bar{U}_N^2)$$

and

$$b_{t,N} = \frac{1}{N} \sum_{s=1}^{N-t} (L_{s+t}L_s - U_{s+t}\bar{U}_N - U_s\bar{U}_N + \bar{L}_N^2)$$

are upper and lower bounds on $\hat{\gamma}_{t,N}$. As illustrated below, though $\hat{\sigma}^2_{\max,N}$ is more conservative than $\hat{\sigma}_N^2$, it can still provide a useful bound.

2.3. Experimental results: Random walk

We illustrate the difficulties with using these ergodic averages and the bounds of asymptotic variances with a random walk when $p_i = p$ is constant. Further experimental results are given in Section 3.3. The running mean $\bar{X}_t$ and corresponding upper and lower bounds $\bar{U}_t$ and $\bar{L}_t,\ t = 1, \ldots, N$ are shown in the left plot of Figure 1 for a run length of $N = 10000$ iterations when $k = 5$ and $p = 1/2$, i.e. $\mu = 2.5$. Corresponding upper and lower bounds on the variance, $\hat{\sigma}^2_{\max,N}$ and $\hat{\sigma}^2_{\min,N}$ given by (12) are depicted in the right plot of Figure 1 for $N = 1, \ldots, 1000$. We have also obtained results (not included here) and compared values of $\bar{L}_N$, $\bar{X}_N$ and $\bar{U}_N$ together with values of $\hat{\sigma}^2_{\min,N}$, $\hat{\sigma}^2_N$ and $\hat{\sigma}_{\max,N}$ for various values of p, k and N. As expected, the bounds become wider as the range k of the random walk increases and become narrower as the value of p moves away from 0.5. In cases with $k \leq 10$, $0.2 \leq p \leq 0.5$ and N larger than 5000, $\hat{\sigma}^2_{\min,N}$ and $\hat{\sigma}^2_{\max,N}$ are close to $\hat{\sigma}^2_N$, and the running means seem to stabilise. However, this can be somewhat misleading because although $\hat{\sigma}_{\max,N}$ may be small, it may not follow that $\bar{L}_N \leq \mu \leq \bar{U}_N$. This is illustrated in Figure 1.

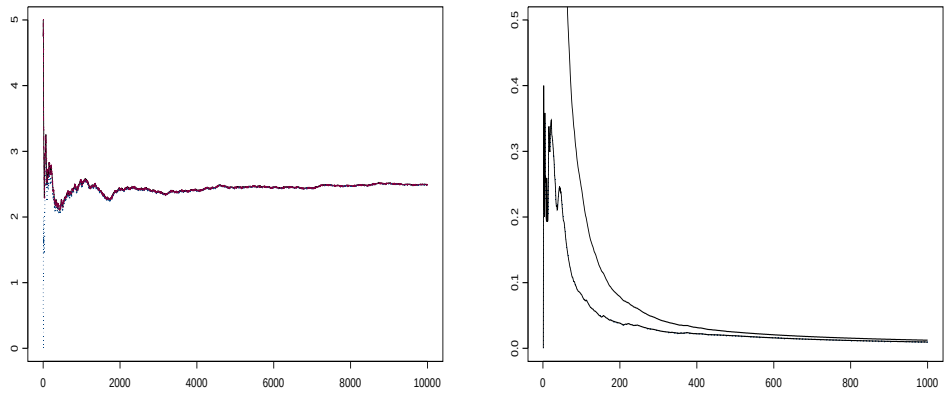

Fig. 1. Random walk with $k = 5$ and $p = 0.5$. *Left plot:* Running mean and upper and lower bounds over 10000 iterations. *Right plot:* Upper and lower bounds on the variance of the mean over the first 1000 iterations. The middle line is the variance based on the stationary chain. The lower bound is too close to zero to be easily seen on this plot.

The performance of the bounds should be evaluated in light of the very high autocorrelation in the chain which increases the conservativeness of Geyer's variance estimate, and the inherent variability of a random walk itself. The latter point is exemplified in Figure 2. The left plot of Figure 2 illustrates the behaviour of 100 independent replications of the running mean of a stationary random walk with $k = 5$ and $p = 0.5$ over 1000 iterations. Although the average of the 100 estimates is close to $k/2$ at each iteration, the individual estimates vary considerably: 95% confidence intervals for $k/2$ are (1.26, 3.70), (1.94, 3.02) and (2.12, 2.88) for $t = 100, 500$ and 1000, respectively. The right plot of Figure 2 illustrates the persistence of this variability for five replications of the running mean of the same random walk over a longer run length of 100000 iterations. As in the left plot, the estimates are quite unstable at $t = 1000$, ranging from 2.33 to 2.63, but noticeable differences persist even at $t = 100000$ with estimates ranging from 2.476 to 2.513.

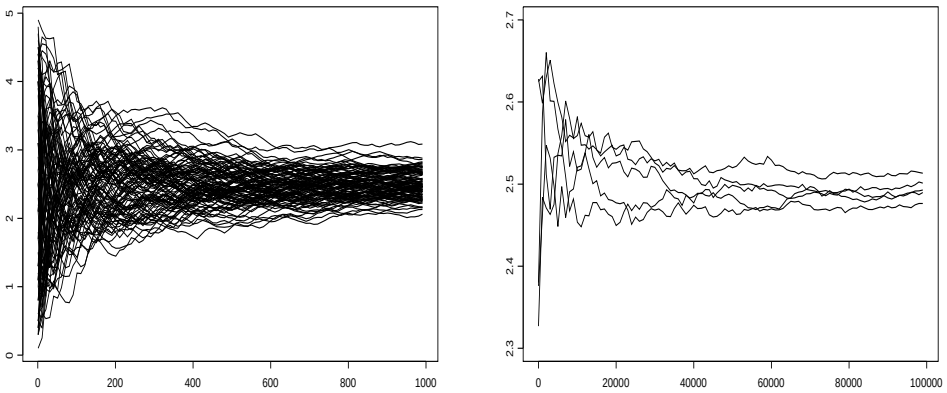

Fig. 2. Random walk with $k = 5$ and $p = 0.5$. *Left plot:* One hundred independent simulations of the running mean over 1000 iterations. *Right plot:* Five independent simulations of the running mean over 100000 iterations.

3. General setting and methods

In this section we consider the general setting in Section 1: Assume that (4)-(5) and hence (2) are satisfied, where the equilibrium chain X is irreducible and $\bar{\phi}_t^L$ and $\bar{\phi}_t^U$ are consistent estimators of μ given by (1). Moreover, as in (8) assume that a CLT applies:

$$\sqrt{t}(\bar{\phi}_t - \mu) \text{ converges in distribution to Normal}(0, \sigma^2) \text{ as } t \to \infty \tag{13}$$

where σ^2 is defined as in (9) but now $\gamma_t = \text{Cov}(\phi(X_1), \phi(X_{t+1}))$ for $t \geq 0$. Sufficient conditions for the CLT to hold can be found in Meyn and Tweedie (1993), Geyer (1996), Chan and Geyer (1994) and Roberts and Rosenthal (1998). For instance, it suffices to establish that X is geometrically ergodic and, if X is reversible, that $\text{E}\phi(X_t)^2 < \infty$.

Assuming that X is reversible, Geyer's initial series estimator applies (Section 2.2 with X_t replaced by $\phi(X_t)$): If we for the moment imagine that $X_1, \ldots, X_N$ are observed, then σ^2 is estimated by (10) where now for $0 \leq t < N$,

$$\hat{\gamma}_{t,N} = \frac{1}{N} \sum_{s=1}^{N-t} (\phi(X_{s+t}) - \bar{\phi}_N)(\phi(X_s) - \bar{\phi}_N).$$

For a real number or function f, write $f_+ = \max\{0, f\}$ for its positive part and $f_- = \max\{0, -f\}$ for its negative part, so $f = f_+ - f_-$. By (4)-(5) we have that

$$0 \leq \phi_+(L_t) \leq \phi_+(X_t) \leq \phi_+(U_t), \quad 0 \leq \phi_-(U_t) \leq \phi_-(X_t) \leq \phi_-(L_t),$$

$$0 \leq \bar{\phi}_{N+}^L \leq \bar{\phi}_{N+} \leq \bar{\phi}_N^U, \quad 0 \leq \bar{\phi}_{N-}^U \leq \bar{\phi}_{N-} \leq \bar{\phi}_{N-}^L.$$

Hence $\hat{\sigma}_N^2$ is bounded by $\hat{\sigma}_{\max,N}^2$ and $\hat{\sigma}_{\min,N}^2$ given by (12) where now for $t \geq 0$,

$$\begin{aligned} a_{t,N} =& \frac{1}{N}\sum_{s=1}^{N-t}\Big\{\phi_+(U_{s+t})\phi_+(U_s) - \phi_-(U_{s+t})\phi_+(L_s) - \phi_+(L_{s+t})\phi_-(U_s) + \phi_-(L_{s+t})\phi_-(L_s) \\ & - \phi_+(L_{s+t})\bar{\phi}_{N+}^L + \phi_+(U_{s+t})\bar{\phi}_{N-}^L + \phi_-(L_{s+t})\bar{\phi}_{N+}^U - \phi_-(U_{s+t})\bar{\phi}_{N-}^U - \phi_+(L_s)\bar{\phi}_{N+}^L \\ & + \phi_+(U_s)\bar{\phi}_{N-}^L + \phi_-(L_s)\bar{\phi}_{N+}^U - \phi_-(U_s)\bar{\phi}_{N-}^U + \bar{\phi}_N^{U\,2} - 2\bar{\phi}_{N+}^L\bar{\phi}_{N-}^U + \bar{\phi}_N^{L\,2}\Big\} \end{aligned} \tag{14}$$

and

$$\begin{aligned} b_{t,N} =& \frac{1}{N}\sum_{s=1}^{N-t}\Big\{\phi_+(L_{s+t})\phi_+(L_s) - \phi_-(L_{s+t})\phi_+(U_s) - \phi_+(U_{s+t})\phi_-(L_s) + \phi_-(U_{s+t})\phi_-(U_s) \\ & - \phi_+(U_{s+t})\bar{\phi}_{N+}^U + \phi_+(L_{s+t})\bar{\phi}_{N-}^U + \phi_-(U_{s+t})\bar{\phi}_{N+}^L - \phi_-(L_{s+t})\bar{\phi}_{N-}^L - \phi_+(U_s)\bar{\phi}_{N+}^U \\ & + \phi_+(L_s)\bar{\phi}_{N-}^U + \phi_-(U_s)\bar{\phi}_{N+}^L - \phi_-(L_s)\bar{\phi}_{N-}^L + \bar{\phi}_N^{L\,2} - 2\bar{\phi}_{N+}^U\bar{\phi}_{N-}^L + \bar{\phi}_N^{U\,2}\Big\}. \end{aligned} \tag{15}$$

These bounds depend entirely on the upper and lower processes and not on the equilibrium chain.

3.1. Method 1

Our first method is based on combining (2), (13) and the upper bound on $\hat{\sigma}_N^2$ to obtain a conservative confidence interval for μ: Asymptotically with at least probability $2(1-\alpha)$,

$$\bar{\phi}_N^L - q_\alpha\hat{\sigma}_{\max,N} \leq \mu \leq \bar{\phi}_N^U + q_\alpha\hat{\sigma}_{\max,N} \tag{16}$$

where $\hat{\sigma}_{\max,N} = \sqrt{\hat{\sigma}_{\max,N}^2}$.

3.2. Method 2

One potential problem with Method 1 is meta-stability: the processes $\bar{\phi}_N^L$ and $\bar{\phi}_N^U$ may appear to have converged at time N, but they have not yet done so, cf. Section 2.3. A more conservative alternative is to use i.i.d. blocks of upper and lower processes; details follow below. As illustrated in Sections 3.3, 4.2 and 4.4, the relative merit of one method over the other depends on the particular model.

Assume that there exist unique elements $\hat{0}, \hat{1} \in \Omega$ so that $\hat{0} \prec x \prec \hat{1}$ for all $x \in \Omega$. For example, for the random walk in Section 2.1, $\hat{0} = 0$ and $\hat{1} = k$. Further, suppose that $((U_t^{(1)}, L_t^{(1)})_{,\,t=1,\ldots,T_1}, T_1)$, $((U_t^{(2)}, L_t^{(2)})_{,\,t=1,\ldots,T_2}, T_2), \ldots$ are i.i.d. "blocks", where $T_1, T_2, \ldots$ are either equal fixed times or i.i.d. random times so that

$$U_1^{(i)} = \hat{1}, \quad L_1^{(i)} = \hat{0}, \quad i = 1, 2, \ldots,$$

$$U_t^{(1)} = U_t, \quad L_t^{(1)} = L_t, \quad t = 1, \ldots, T_1,$$

$$L_t^{(2)} \prec L_{t+T_1} \prec U_{t+T_1} \prec U_t^{(2)}, \quad t = 1, \ldots, T_2,$$

$$L_t^{(3)} \prec L_{t+T_1+T_2} \prec U_{t+T_1+T_2}, \prec U_t^{(3)}, \quad t = 1, \ldots, T_3,$$

and so on. For instance, in the case of the random walk in Section 2, we obtain such i.i.d. blocks when

$$U_{t+1}^{(2)} = \chi(U_t^{(2)}, R_{t+T_1}), \quad L_{t+1}^{(2)} = \chi(L_t^{(2)}, R_{t+T_1}), \quad t = 1, \ldots, T_2 - 1,$$

$$U_{t+1}^{(3)} = \chi(U_t^{(3)}, R_{t+T_1+T_2}), \quad L_{t+1}^{(3)} = \chi(L_t^{(3)}, R_{t+T_1+T_2}), \quad t = 1, \ldots, T_3 - 1,$$

etc. We may, for example, choose T_i as the first time n_i at which

$$\frac{1}{n_i} \sum_{t=1}^{n_i} \left(\phi(U_t^{(i)}) - \phi(L_t^{(i)}) \right) \leq \epsilon \tag{17}$$

where $\epsilon > 0$ is a user-specified parameter.

By (4)-(5), for $N = T_1 + \ldots + T_m$ and $m = 1, 2, \ldots$,

$$\tilde{\phi}_N^L \leq \bar{\phi}_N^L \leq \bar{\phi}_N \leq \bar{\phi}_N^U \leq \tilde{\phi}_N^U$$

where we set

$$\tilde{\phi}_N^U = \frac{1}{N} \sum_{i=1}^{m} W_i^U, \quad W_i^U = \sum_{s=1}^{T_i} \phi(U_{s+T_0+\ldots+T_{i-1}}^{(i)}),$$

$$\tilde{\phi}_N^L = \frac{1}{N} \sum_{i=1}^{m} W_i^L, \quad W_i^L = \sum_{s=1}^{T_i} \phi(L_{s+T_0+\ldots+T_{i-1}}^{(i)}),$$

and $T_0 = 0$. On one hand these new bounds are more conservative: in Method 1, $\bar{\phi}_N^U$ and $\bar{\phi}_N^L$ are consistent estimators of μ, whereas $\tilde{\phi}_N^U$ and $\tilde{\phi}_N^L$ almost surely converge to $\mathrm{E}W_1^U/\mathrm{E}T_1$ and $\mathrm{E}W_1^L/\mathrm{E}T_1$, respectively, which in general are different from μ. On the other hand, since the blocks are i.i.d., we may better "trust" the bounds $\tilde{\phi}_N^U$ and $\tilde{\phi}_N^L$: if these bounds are close, we may expect that $\bar{\phi}_N^U$ and $\bar{\phi}_N^L$ have been stabilised. If $T_i = n_i$ is specified by (17) then of course

$$\tilde{\phi}_N^U - \tilde{\phi}_N^L \leq \epsilon.$$

Finally, the classical CLT and strong law of large numbers apply for the i.i.d. blocks so that as $m \to \infty$, $\tilde{\phi}_N^U$ and $\tilde{\phi}_N^L$ are approximately normally distributed with variances $(\mathrm{Var}W_1^U)/(m(\mathrm{E}T_1)^2)$ and $(\mathrm{Var}W_1^L)/(m(\mathrm{E}T_1)^2)$ provided the moments exist. It is straightforward to estimate these moments from the i.i.d. blocks and thereby obtain consistent estimates $\tilde{\sigma}_{\max,N}$ and $\tilde{\sigma}_{\min,N}$ for the standard deviations. Thus asymptotically with at least probability $2(1-\alpha)$,

$$\tilde{\phi}_N^L - q_\alpha \tilde{\sigma}_{\min,N} \leq \mu \leq \tilde{\phi}_U^L + q_\alpha \tilde{\sigma}_{\max,N}. \tag{18}$$

3.3. Experimental results: Random walk continued

Consider again a random walk with $k = 5$ and all $p_i = p = 0.5$, and let ϕ be the identity function. Conservative 95% confidence bounds on the running mean based on (16) for Method 1 are shown in Figure 3 (left plot). Note that longer runs of least 10000 iterations seem needed, since many of the confidence intervals do not contain 2.5; see also Figure 2. The procedure for taking i.i.d. blocks under Method 2 is illustrated in Figure 3 (right plot). Blocks were identified using the criterion given in (17), with ϵ arbitrarily chosen to be equal

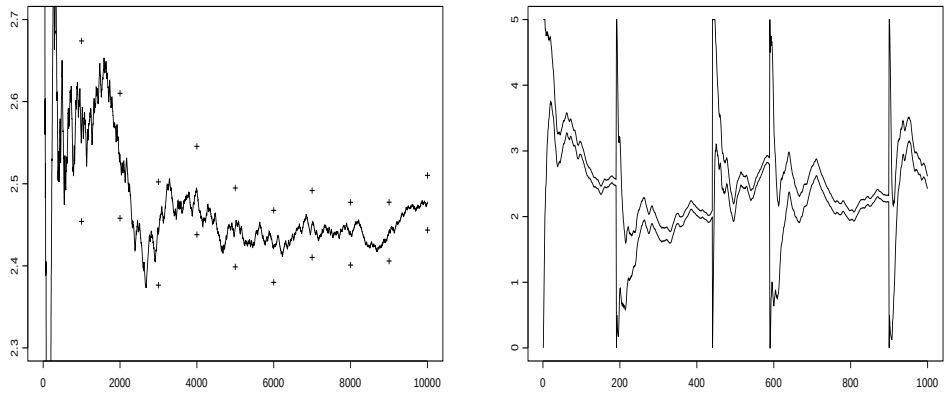

Fig. 3. Random walk with $k = 5$ and $p = 0.5$. *Left plot:* Conservative 95% confidence bounds (indicated by crosses) on the running mean when $t = 1, \ldots, 10000$ and Method 1 is used. *Right plot:* Method 2 for obtaining $L_t^{(i)}$ and $U_t^{(i)}$ when $t = 1, \ldots, 1000$.

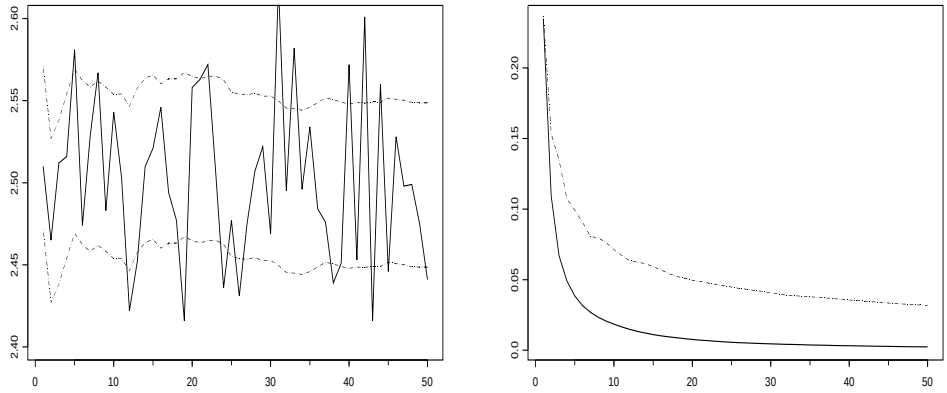

Fig. 4. Random walk with $k = 5$, $p = 0.5$, $\epsilon = 0.1$ and $N = 1, .., 500000$, depicted at every 10000th iteration. *Left plot:* $\bar{\phi}_N^L$ and $\bar{\phi}_N^U$ under Method 1 (solid line) and $\tilde{\phi}_N^L$ and $\tilde{\phi}_N^U$ under Method 2 (dashed lines). Note that $\bar{\phi}_N^L$ and $\bar{\phi}_N^U$ are effectively equal for these values of N. *Right plot:* $\hat{\sigma}_{\max,N}$ under Method 1 (solid line) and $\tilde{\sigma}_{\max,N}$ under Method 2 (dashed line).

to 0.1. For this example, $m = 280$ such blocks were identified from $N = 100000$ iterations. The solid lines in the figure represent $L_t^{(i)}$ and $U_t^{(i)}$ when $t = 1, \ldots, 1000$.

Runs of varying length N were simulated for other random walks with different ranges $k = 5, 10$, values of $p = 0.2, 0.5$ and values of $\epsilon = 0.1, 0.01$. Comparison of Methods 1 and 2 under these conditions confirmed the greater meta-stability of Method 1, gained at the expense of a larger variance, for the same number of iterations. The relative merit of one method over the other depends in particular on the values of k and ϵ. For example, for the same N and ϵ under Method 2, as k increases the value of m decreases and $\mathrm{Var}W_1^L$ and $\mathrm{Var}W_1^U$ increase because the lower and upper processes are re-initiated at 0 and k, respectively, for each i.i.d. block. In comparison, under Method 1 the processes are initiated at 0 and k only at time $t = 0$ and the variances are computed using all N iterations. The same behaviour is observed for fixed k and N as ϵ decreases. The comparative performance of the means and the upper bound on the variances under the two methods is illustrated in Figure 4 for $k = 5$, $p = 0.5$, $\epsilon = 0.1$ and N ranging from 10000 to 500000.

4. Other examples

In this section we consider two examples of more complicated models where the methods in Section 3 are helpful.

4.1. Ising model

Consider an Ising model defined on a square lattice $V = \{1, ..., M\}^2$ and with the set of first order edges

$$E = \{\{(i_1, i_2), (j_1, j_2)\} \subseteq V : (i_1 - j_1)^2 + (i_2 - j_2)^2 = 1\}$$

defining the neighbourhood relation. The state space is $\Omega = \{\pm 1\}^V$ and

$$\pi(x) \propto \exp\left(\beta \sum_{\{i,j\} \in E} x_i x_j\right), \quad x = (x_i)_{i \in V} \in \Omega,$$

where β is a real parameter.

For simplicity we consider first a Gibbs sampler with a simple random updating scheme. The updating function is $\chi : \Omega \times V \times [0, 1] \to \Omega$ with

$$\chi(x, i, r) = \begin{cases} (x_{(1,1)}, \ldots, 1, \ldots, x_{(M,M)}) & \text{if } (1 + \exp(-2\beta \sum_{j:\{i,j\} \in E} x_j))^{-1} \leq r \\ (x_{(1,1)}, \ldots, -1, \ldots, x_{(M,M)}) & \text{else} \end{cases}$$

where the 1 or -1 is placed at the ith coordinate. The Gibbs sampler is the SRS

$$X_{t+1} = \chi(X_t, I_t, R_t), \quad t = 0, 1, \ldots,$$

where $I_0, R_0, I_1, R_1, \ldots$ are mutually independent, $I_t \sim \mathrm{Uniform}(V)$ and $R_t \sim \mathrm{Uniform}[0, 1]$.

Define a partial ordering on Ω by

$$x \prec y \iff x_i \leq y_i \text{ for all } i \in V \tag{19}$$

with $x = (x_i)_{i\in V}$ and $y = (y_i)_{i\in V}$. Then $\hat{1} = (1,\dots,1)$ and $\hat{0} = (-1,\dots,-1)$ are the unique maximal and minimal elements. Suppose first that $\beta \geq 0$. Then the Gibbs sampler is monotone in the sense that

$$x \prec y \implies \chi(x,\cdot,\cdot) \leq \chi(y,\cdot,\cdot).$$

Hence we can define upper and lower chains in a similar way as in Section 2.1:

$$U_0 = \hat{1},\ L_0 = \hat{0},\ U_{t+1} = \chi(U_t, I_t, R_t),\ L_{t+1} = \chi(L_t, I_t, R_t), \quad t = 0,1,\dots. \tag{20}$$

If instead $\beta < 0$, the Gibbs sampler becomes anti-monotone, and we can use the cross-over trick of Kendall (1998) (see also Häggström and Nelander, 1998 and Møller, 1999):

$$U_0 = \hat{1},\ L_0 = \hat{0},\ U_{t+1} = \chi(L_t, I_t, R_t),\ L_{t+1} = \chi(U_t, I_t, R_t), \quad t = 0,1,\dots.$$

Then (4) is still satisfied, but U and L are not individual Markov chains.

Since the Gibbs sampler is ergodic and Ω is finite, we obtain the CLT (13). As required, $\bar{\phi}_t^L$ and $\bar{\phi}_t^U$ are consistent estimators of μ (this is obvious when $\beta \geq 0$ and not so hard to verify in the anti-monote case $\beta < 0$). The reason for using the Gibbs sampler instead of the more efficient Swendson and Wang (1987) algorithm is because the latter algorithm has a lack of monotonicity (Propp and Wilson, 1996; Mira et al., 2001).

4.2. Experimental results: Ising model

The Gibbs sampler described above is clearly reversible. For the experiments in this section we used a slightly different algorithm with a systematic updating scheme in which one iteration consists of $2M^2 - 1$ Gibbs updates scanning through the elements of V and back again in reverse order. This double scan Gibbs sampler is also reversible, monotone and ergodic. The autocorrelation is much smaller under this approach.

Let $\phi(x) = \sum_{i\in V} x_i$ which is monotone with respect to (19). By symmetry in the density, $\mu = 0$ is known. Bounds based on (16) were constructed for Ising models with $M = 5, 10, 64$ and parameters $\beta = 0.1, 0.5$. The results are illustrated for $M = 64, \beta = 0.1$ in Figure 5; the first 100 iterations of the running mean and corresponding upper and lower bounds under Method 1 are depicted in the left panel, and the right panel shows the conservative 95% confidence intervals based on (16) for $N = 1000$.

Method 2 was employed for an Ising model with $M = 5$ and $\beta = 0.5$. With $\epsilon = 5$, $m = 33$ blocks were sampled from $N = 500$ iterations. The estimated mean and standard deviation for the lower bound were $\tilde{\phi}_N^L = -0.0248$ and $\tilde{\sigma}_{\min,N} = 0.00289$. The corresponding figures for the upper bound were $\tilde{\phi}_N^U = 0.0321$ and $\tilde{\sigma}_{\max,N} = 0.00386$. For comparison, with the same values of M, β and N, the lower and upper bounds on the standard deviation under Method 1 were $\hat{\sigma}_{\min,N} = 0.0635$ and $\hat{\sigma}_{\max,N} = 0.0785$. In this case the 95%-confidence interval (16) for Method 1 is about four times wider than the 95%-confidence interval (18) for Method 2.

4.3. Mixture model

In this section we consider a Bayesian model for a simple mixture distribution, following similar ideas as in Robert and Casella (2004).

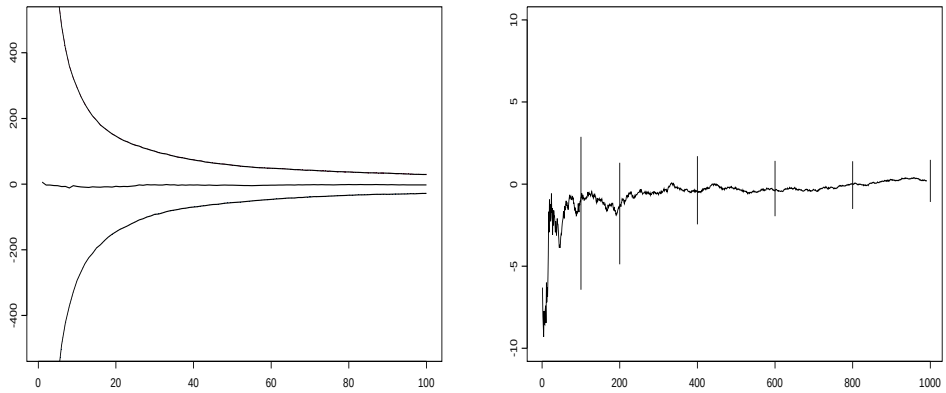

Fig. 5. Method 1 for an Ising model with $M = 64$ and $\beta = 0.1$. *Left plot:* Running mean $\bar{\phi}_t$ and upper and lower bounds $\bar{\phi}_t^U$ and $\bar{\phi}_t^L$. *Right plot:* Time series plot of $\bar{\phi}_t$ and corresponding 95% confidence intervals.

We assume that we have i.i.d. observations $Y_1 = y_1, \ldots, Y_n = y_n$ from a two-component mixture given by the density

$$f(x|p) = pf_1(x) + (1-p)f_2(x)$$

where the densities f_1 and f_2 are known and the parameter p follows a conjugate prior $\text{Beta}(\lambda_1, \lambda_2)$. Consider latent variables Z_i, $i = 1, .., n$ that allocate observation Y_i to component $j = 1$ or 2. Specifically, the Z_i given p are i.i.d. with $P(Z_i = 1|p) = p$ and $P(Z_i = 2|p) = 1 - p$, and the Y_i given the Z_i and p are independent with Y_i following f_j if $Z_i = j$. Thus a posteriori we obtain the full conditionals

$$p|\cdots \sim \text{Beta}(\lambda_1 + n_1, \lambda_2 + n_2),$$

$$P(Z_i = j|\cdots) \propto \omega_j f_j(y_i), \quad j = 1, 2,\ i = 1, \ldots, n,$$

where $\omega_1 = p$, $\omega_2 = 1 - p$ and n_j is the number of times $Z_i = j,\ i = 1, \ldots, n$.

For ease of exposition, consider first a Gibbs sampler with a random updating scheme, using inversion at each type of update from the full conditionals: The SRS for the chain $X_t = (p_t, Z_{t,1}, \ldots, Z_{t,n})$ is given by

$$X_{t+1} = \varphi(X_t, I_t, R_t), \quad t = 0, 1, \ldots,$$

where $I_t \sim \text{Uniform}\{0, 1, \ldots, n\}$, $R_t \sim \text{Uniform}[0, 1]$, the $I_0, R_0, I_1, R_1, \ldots$ are mutually independent, and the updating function is specified as follows. In case $I_t = 0$ then $X_{t+1} = (p_{t+1}, Z_{t,1}, \ldots, Z_{t,n})$ and $p_{t+1} = F^{-1}(R_t|n_{t,1})$, the inverse distribution function of $\text{Beta}(\lambda_1 + n_{t,1}, \lambda_2 + n - n_{t,1})$ (with $n_{t,1}$ equal to the number of times $Z_{t,i} = 1,\ i = 1, \ldots, n$). If $I_t = i \in \{1, \ldots, n\}$ then $X_{t+1} = (p_t, Z_{t,1}, \ldots, Z_{t,i-1}, Z_{t+1,i}, Z_{t,i+1}, \ldots, Z_{t,n})$ where $Z_{t+1,i} = 1$ if $R_t \leq p_t f_1(y_i)/[p_t f_1(y_i) + (1 - p_t)f_2(y_i)]$ and $Z_{t+1,i} = 2$ otherwise.

This Gibbs sampler is obviously monotone with respect to the following partial ordering on $[0,1] \times \{1,2\}^n$:

$$(p, z_1, \ldots, z_n) \prec (p', z_1', \ldots, z_n') \Longleftrightarrow p \leq p',\ z_i \geq z_i',\ i = 1, \ldots, n.$$

Furthermore, $\hat{1} = (1, 1, \ldots, 1)$ and $\hat{0} = (0, 0, \ldots, 0)$ are the unique maximal and minimal elements. Hence we define upper and lower chains in the same way as in (20).

Note that p_t is the chain of the interest. Since its state space $[0,1]$ is compact, it can be shown that the chain is uniformly ergodic. Consequently, for real functions $\phi(p)$, the requirements that $\bar{\phi}_t^L$ and $\bar{\phi}_t^U$ are consistent estimators of μ, and the CLT (13) holds, are satisfied.

4.4. Experimental results: Mixture model

The Gibbs sampler studied above is obviously reversible. For similar reasons as in Section 4.2, we used a systematic Gibbs sampler for the experiments in this section (where one iteration is one scan of the $n+1$ components). Reversibility of the target chain p_t is automatically ensured.

We illustrate Method 1 using a two-component normal mixture in which $f_1 \sim \mathrm{N}(0,1)$ and $f_2 \sim \mathrm{N}(2,1)$. Observations $y_1, .., y_{200}$ were simulated from this mixture with $p = 0.3$. Figure 6 depicts the running mean of p and upper and lower bounds for $N = 1000$ (left plot) and the corresponding 95% upper confidence bound on p (right plot) computed for every 1000th iteration to $N = 10000$. After $N = 100, 1000$ and 10000 iterations, the respective values of $(\bar{\phi}_N^L,\ \bar{\phi}_N^U)$ were (0.2743,0.3294), (0.2964,0.3019) and (0.2987,0.2992). The corresponding values of $(\hat{\sigma}_{\min,N}, \hat{\sigma}_{\max,N})$ at these values of N were (0,0.0286), (0,0.00313) and (0.0363,0.000517). Conservative 95% confidence bounds for p, computed using $\hat{\sigma}_{\max,N}$, were (0.2182,0.3855), (0.2903,0.3080) and (0.2977,0.3002), respectively. For comparison, under Method 2 with $\epsilon = 0.001$, the values of $(\tilde{\phi}_N^L,\ \tilde{\phi}_N^U)$ after $N = 100, 1000$ and 10000 iterations were (0.2374,0.5326), (0.2411,0.5148) and (0.2523,0.5115). The corresponding values of $\hat{\sigma}_{\max,N}$ were 0.0175, 0.00742 and 0.000992, leading to much more conservative 95% confidence bounds for p than those for Method 1 above. In particular, the asymmetry of the process is reflected in the quite conservative upper bounds $\bar{\phi}_U^N$ and $\tilde{\sigma}_{\max,N}$. More generally, the accuracy and precision of the bounds is dependent on the sample composition and sample size, as well as the values of p and other parameters in the mixture model.

5. Extensions and applications

Our methods also apply when ϕ is anti-monotone, i.e. when

$$x \prec y \implies \phi(x) \geq \phi(y).$$

We simply exploit the fact that $-\phi$ is monotone. Similarly, our methods easily apply when ϕ is a linear combination of monotone functions. In fact many lattice and point process models are of an exponential family type where the canonical sufficient statistic $t(x)$ is a linear combination of monotone functions (considering here for simplicity the one dimensional case of $t(x)$; in the higher dimensional case each coordinate function is often a linear combination of monotone functions). For the Ising model, for example,

$$t(x) = \sum_{\{i,j\} \in E} x_i x_j = \frac{1}{2} \sum_{\{i,j\} \in E} I[X_i = X_j = 1] + \frac{1}{2} \sum_{\{i,j\} \in E} I[X_i = X_j = -1] - |E|$$

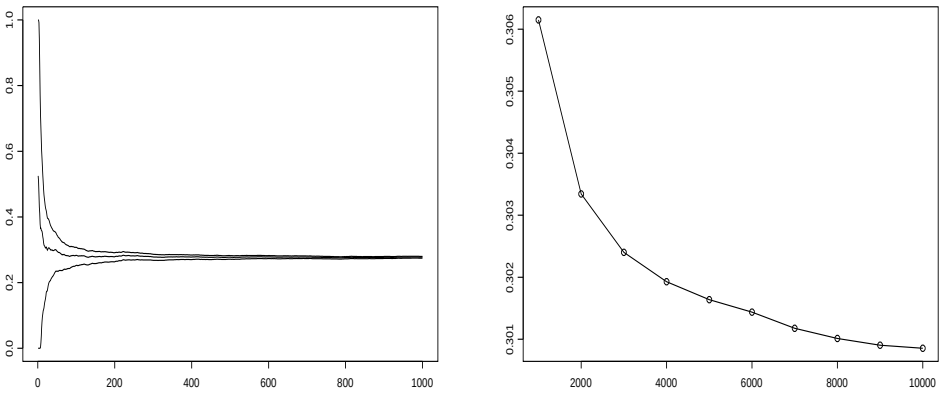

Fig. 6. Mixture model with $p = 0.3$ and normal distributions $\mathrm{N}(0,1)$ and $\mathrm{N}(2,1)$. *Left plot:* Running mean and upper and lower bounds for $N = 1000$. *Right plot:* 95% upper confidence bounds computed at $t = 1000, 2000, ..., 10000$.

where the first term is monotone, the second is anti-monotone and the third is constant; here $I[\cdot]$ denotes the indicator function and $|E|$ the cardinality of E.

Method 1 easily extends to a time-continuous setting. For example, spatial birth-death processes have been successfully used for perfect simulation of spatial point processes (Kendall, 1998; Kendall and Møller, 2000), and Method 1 can straightforwardly be modified to this case. However, Method 2 does not easily apply in that case, since there is no maximal element (or more generally, since the dominating Poisson-birth-death process in Kendall and Møller (2000) is used in a way for obtaining the upper and lower processes which makes it difficult to obtain i.i.d. blocks). Instead the ideas in Wilson (2000) may be exploited.

In particular, our methods apply for many stochastic models used in statistical physics and spatial statistics. Examples include Ising and hard-core models, and many of Besag's auto-models (Besag, 1974): the auto-logistic, the auto-binomial, the auto-Poisson and the auto-gamma model; for coupling constructions, see Møller (1999). Moreover, many spatial point process models, including the Strauss process and other repulsive pairwise interaction point process models (Møller and Waagepetersen, 2003) can be handled, using the modification of Method 1 discussed above. For the area-interaction point process (or mixture Widow-Rowlinson model, see Widom and Rowlinson, 1970 and Baddeley and van Lieshout, 1995) it is easier to use the coupling construction in Häggström et al. (1999).

On the other hand, it seems that our methods so far are of rather limited importance for general Bayesian problems, since it is usually not known how to construct the upper and lower dominating processes, or since the functions ϕ of interest are often not linear combinations of monotone functions. Some exceptions are the mixture model in Section 4.3 and the following models.

The Ising model with an external field: this model is equivalent to an auto-logistic model, and it appears, for example as posterior distributions used for reconstruction problems in image analysis (Geman and Geman, 1984).

The auto-gamma model has been used in the Bayesian literature, see Møller (1999) and the references therein. Møller (1999) and Wilson (2000) show how the U and L processes can be constructed.

Our methods are suitable for posterior distributions associated with mixtures of exponential families and conjugate priors (Casella et al., 2002) using the upper and lower chains introduced in Mira et al. (2001), where other examples of applications also are given.

Our methods also apply when using the upper and lower processes for the perfect simulated tempering algorithms and the Bayesian models considered in Møller and Nicholls (1999) and Brooks et al. (2002).

In conclusion, our methods apply whenever the Propp and Wilson (1996) algorithm does or when modifications such as those in Kendall and Møller (2000) do. Moreover, they may also apply in situations where almost sure coalescence of the upper and lower processes are not required (see, e.g. Møller, 1999), and it would be interesting to explore such cases, but we shall refrain from this in the present paper.

Acknowledgements

This research was supported by The Danish Natural Science Research Council and the Australian Research Centre for Dynamic Systems and Control.

References

Baddeley, A. J. and M. N. M. van Lieshout (1995). Area-interaction point processes. *Annals of the Institute of Statistical Mathematics 46*, 601–619.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B 36*, 192–326.

Brooks, S., Y. Fan, and J. Rosenthal (2002). Perfect forward simulation via simulation tempering. Technical report, Department of Statistics, University of Cambridge.

Casella, G., M. K., C. Robert, and D. Titterington (2002). Perfect slice samplers for mixtures of distributions. *Journal of the Royal Statistical Soceity Series B 64*(4), 777–790.

Chan, K. and C. Geyer (1994). Discussion of "Markov chains for exploring posterior distributions". *Annals of Statistics 22*, 1747–1758.

Foss, S. and R. Tweedie (1998). Perfect simulation and backward coupling. *Stochastic Models 14*, 187–203.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

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

Geyer, C. (1996). Estimation and optimization of functions. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, London, pp. 241–258. Chapman and Hall.

Häggström, O. and K. Nelander (1998). Exact sampling from anti-monotone systems. *Statistica Neerlandia 52*, 360–380.

Häggström, O., M. N. M. van Lieshout, and J. Møller (1999). Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli 5*, 641–659.

Kendall, W. (1998). Perfect simulation for the area-interaction point process. In C. Heyde and L. Accardi (Eds.), *Proability Towards 2000*, New York, pp. 218–234. Springer-Verlag.

Kendall, W. and J. Møller (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability 32*, 844–85.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Mira, A., J. Møller, and G. Roberts (2001). Perfect slice samplers. *Journal of the Royal Statistical Society Series B 63*, 583–606.

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

Møller, J. and G. Nicholls (1999). Perfect simulation for sample-based inference. Technical Report R-99-2011, Department of Mathematical Sciences, Aalborg University.

Møller, J. and R. Waagepetersen (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton: Chapman and Hall/CRC.

Priestly, M. (1981). *Spectral Aanlaysis and Time Series*. London: Academic Press.

Propp, J. and D. Wilson (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms 9*, 223–252.

Ripley, B. (1987). *Stochastic Simulation*. New York: John Wiley.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (2 ed.). New York: Springer.

Roberts, G. and J. Rosenthal (1998). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canadian Journal of Statistics 26*, 5–32.

Swendson, R. and J. Wang (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters 58*, 86–88.

Widom, B. and J. S. Rowlinson (1970). A new model for the study of liquid-vapor phase transitions. *Journal of Chemical Physics 52*, 1670–1684.

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