J. Appl. Prob. **46,** 309–329 (2009)
Printed in England
© *Applied Probability Trust* 2009

FIXED PRECISION MCMC ESTIMATION BY MEDIAN OF PRODUCTS OF AVERAGES

WOJCIECH NIEMIRO,* ** *Nicolaus Copernicus University*
PIOTR POKAROWSKI,* *** *University of Warsaw*

Abstract

The standard Markov chain Monte Carlo method of estimating an expected value is to generate a Markov chain which converges to the target distribution and then compute correlated sample averages. In many applications the quantity of interest θ is represented as a product of expected values, $\theta = \mu_1 \cdots \mu_k$, and a natural estimator is a product of averages. To increase the confidence level, we can compute a median of independent runs. The goal of this paper is to analyze such an estimator $\hat{\theta}$, i.e. an estimator which is a 'median of products of averages' (MPA). Sufficient conditions are given for $\hat{\theta}$ to have fixed relative precision at a given level of confidence, that is, to satisfy $\mathrm{P}(|\hat{\theta} - \theta| \le \theta\varepsilon) \ge 1 - \alpha$. Our main tool is a new bound on the mean-square error, valid also for nonreversible Markov chains on a finite state space.

Keywords: Markov chain Monte Carlo; rare-event simulation; mean-square error; bias; confidence estimation

2000 Mathematics Subject Classification: Primary 60J10; 65C05
Secondary 68W20; 82B80

1. Introduction

This paper is about constructing exact, nonasymptotic, confidence bounds in the course of Monte Carlo (MC) simulations. In many applications in 'rare-event simulations', statistical physics, chemistry, or biology, the quantity of interest, denoted henceforth by θ, is a positive number of unknown order of magnitude. For this reason, we focus on bounding the *relative* error. The goal is to obtain an MC estimator $\hat{\theta}$ such that

$$\mathrm{P}(|\hat{\theta} - \theta| \le \theta\varepsilon) \ge 1 - \alpha.$$

This requirement means that the estimator should have fixed relative precision $\varepsilon > 0$ at a given level of confidence $1 - \alpha < 1$.

In this paper much attention is given to the case where the parameter of interest can be expressed as a product,

$$\theta = \prod_{j=1}^{k} \mu_j,$$

where each μ_j is computed using the Markov chain Monte Carlo (MCMC) method. Thus, we assume that μ_j is the stationary mean of a functional of some Markov chain (in general,

Received 30 July 2007; revision received 19 December 2008.

* Work partially supported by the Polish Ministry of Science and Higher Education (grant number N N201 387234).

** Postal address: Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Poland. Email address: wniem@mat.uni.torun.pl

*** Postal address: Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland. Email address: pokar@mimuw.edu.pl

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

we have k chains defined on different state spaces). This product representation is the basis of many efficient computational algorithms. We give several examples in Section 5. Standard MCMC algorithms estimate stationary means μ_j by sample averages. An estimate of θ is then the product of averages. Finally, a 'median trick' is used to enhance the confidence level. In this way we arrive at an estimator which is a *median of products of averages* (MPA). Although such an estimator is the main object of our investigations, some auxiliary results used in the analysis of MPA estimators are presumably of independent interest.

The paper is organized as follows. We begin in Section 2 with a careful analysis of a 'median trick'. We explain how it can be used to obtain exponential inequalities for unbounded variables and discuss applications to rare-event simulations [4, Chapter 6]. This is illustrated by one specific example, namely estimation of the tail probability of a random sum; see [5] and [16]. We suggest that an estimator which can be used here is a *median of averages* (MA) of independent and identically distributed (i.i.d.) variables.

In Section 3 we consider inequalities for the relative *mean-square error* (MSE) for products of estimators. The results are tailored for the application to MPA estimators in Section 5.

Section 4 is devoted to MCMC algorithms based on ergodic averages along a trajectory of a Markov chain. This scheme of computations is widely used in practice, but the nonasymptotic analysis is difficult because it involves dependent variables. Our basic tool is a new bound for the MSE of Markov chain averages; see Theorem 4.2. In contrast with the inequality of [1] used in [15], our bound holds for chains which are not necessarily reversible and is in some instances much tighter. We also obtain an inequality for the bias (see Theorem 4.1), which is a generalized version of the results of [8], [14], and [30].

In Section 5 we collate the results of the earlier sections. We use them in the analysis of the MPA scheme, based on the product representation. Theorem 5.1 gives lower bounds on the number of samples necessary to guarantee the fixed relative precision of the MPA estimate. The bounds depend on a few key quantities, assumed to be known *a priori*. Our result is of similar form to that in [15]. However, we work in a more general setting and pay more attention to optimizing constants in our inequalities. Some examples illustrate the range of applications of MPA estimators and our bounds. We also mention other MC estimators and bounds on their cost. Comparison of these bounds shows that, for several important problems, an MPA estimator is the most efficient.

2. The 'median trick'

In this section we discuss the problem of constructing confidence bounds based on inequalities for the MSE. Well-known and classical ways of doing this are via Chebyshev or exponential inequalities, such as the Bernstein inequality [6]. Less known is a 'median trick' introduced in [23]. The most popular approach to confidence estimation which uses the central limit theorem will not be discussed, because we are interested in exact bounds. The setup considered below will be needed in Section 5.

Assume that $\hat{\theta}_1, \dots, \hat{\theta}_m$ are estimators of a parameter $\theta > 0$, each of them computed using an independent sample of size n (or proportional to n). Thus, $\hat{\theta}_1, \dots, \hat{\theta}_m$ are i.i.d. random variables. Suppose that a bound on the relative MSE is available and that it is of the form

$$\mathrm{E}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2 \leq \frac{B}{n}(1 + r(n)), \tag{2.1}$$

where B is an explicitly known constant and $r(\cdot)$ is a nonnegative function (also explicitly known) such that $r(n) \to 0$ as $n \to \infty$. Note that (2.1) is quite natural, because the variance of

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

standard estimators usually decreases as $1/n$ and the remainder $r(n)$ can absorb bias. Conditions similar to (2.1) may come up in various problems of applied probability, in particular those related to MC algorithms; see, e.g. [25]. As in the introduction, we look for an estimator $\hat{\theta}$ of $\theta > 0$ such that

$$\mathrm{P}(|\hat{\theta} - \theta| \le \varepsilon\theta) \ge 1 - \alpha. \tag{2.2}$$

Let $\hat{\theta} = \operatorname{med}(\hat{\theta}_1, \dots, \hat{\theta}_m)$. We choose n and m large enough to guarantee (2.2). It is reasonable to require that the total cost of sampling, nm, is minimum.

Proposition 2.1. *There are universal constants $C_1 \approx 8.305$ and $C_2 \approx 2.315$ with the following properties. If (2.1),*

$$\frac{n}{1 + r(n)} \ge C_1 \frac{B}{\varepsilon^2}, \tag{2.3}$$

$$m \ge C_2 \ln(2\alpha)^{-1} \quad \text{and} \quad m \text{ is odd}, \tag{2.4}$$

are satisfied, then (2.2) holds.

Proof. The idea is to fix an initial moderate level of confidence $1 - a < 1 - \alpha$ and choose n such that $\mathrm{P}(|\hat{\theta}_i - \theta| \le \varepsilon\theta) \ge 1 - a$ for all i. Then we boost the level of confidence from $1 - a$ to $1 - \alpha$ by computing a median. If n satisfies

$$\frac{B}{n}(1 + r(n)) \le a\varepsilon^2 \tag{2.5}$$

then the Chebyshev inequality yields $\mathrm{P}(|\hat{\theta}_i - \theta| > \varepsilon\theta) \le a$. Suppose that $a < \frac{1}{2}$. Consider the Bernoulli scheme in which $|\hat{\theta}_i - \theta| \le \varepsilon\theta$ is interpreted as the 'success' in the ith trial. The event $|\hat{\theta} - \theta| > \varepsilon\theta$ can occur only if the number of successes is less than $m/2$. Therefore, we obtain

$$\begin{aligned} \mathrm{P}(|\hat{\theta} - \theta| > \varepsilon\theta) &\le \sum_{i=(m+1)/2}^{m} \binom{m}{i} a^i (1 - a)^{m-i} \\ &\le a^{m/2}(1 - a)^{m/2} \sum_{i=(m+1)/2}^{m} \binom{m}{i} \\ &= a^{m/2}(1 - a)^{m/2} 2^{m-1} x \\ &= \tfrac{1}{2}[4a(1 - a)]^{m/2} \\ &= \frac{1}{2} \exp\biggl\{\frac{m}{2} \ln[4a(1 - a)]\biggr\}. \end{aligned} \tag{2.6}$$

The above derivation is based on [20]. A similar result without the $\frac{1}{2}$ factor can be deduced from the well-known Chernoff bound, which is a special case of Hoeffding's first inequality [17, Theorem 1]. The right-hand side of (2.6) is less than α if m satisfies

$$m \ge \frac{2 \ln(2\alpha)^{-1}}{\ln[4a(1 - a)]^{-1}}. \tag{2.7}$$

Therefore, (2.5) and (2.7) together imply (2.2).

It remains to optimize the choice of a. The goal is to minimize nm subject to (2.5) and (2.7). An exact solution of this minimization problem depends on the actual form of $r(n)$ and may

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

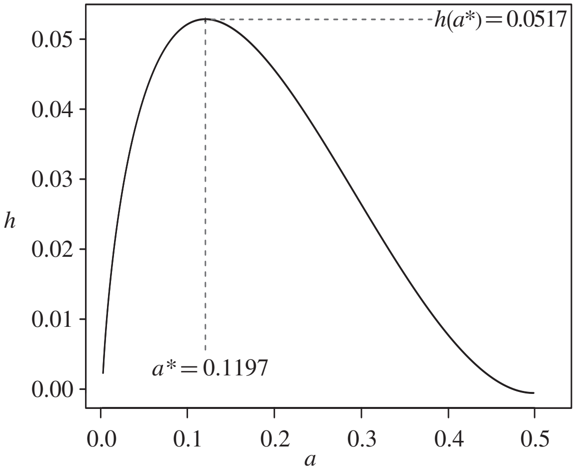

FIGURE 1: Graph of $h(a)$.

be complicated. There exists however a solution which is quite universal and *nearly* optimal under (2.1). Note that, by (2.5), the lower bound on n behaves roughly as $B/(a\varepsilon^2)$ for $\varepsilon \to 0$. Therefore, the lower bound on nm is approximately

$$\frac{2}{a \ln[4a(1-a)]^{-1}} \ln(2\alpha)^{-1} \frac{B}{\varepsilon^2}.$$

To minimize this expression, it is enough to find the maximum of the function $h(a) = -(a/2)\ln[4a(1-a)]$ on the interval $(0, \frac{1}{2})$. There is exactly one maximum at $a^* \approx 0.119\,69$ and $h(a^*) \approx 0.051\,708$ (see Figure 1). Let $C_1 = 1/a^*$ and $C_2 = a^*/h(a^*)$. Inequalities (2.3) and (2.4) are just (2.5) and (2.7) with $a = a^*$.

2.1. Rare-event simulation

We are interested in the behavior of estimators of θ as $\theta \to 0$. We say that an estimator Z of θ has *bounded relative error* if it is unbiased, $\mathrm{E}\, Z = \theta$, and

$$\frac{\operatorname{var} Z}{\theta^2} \le B, \tag{2.8}$$

where B is a constant independent of θ. This concept plays an important role; see [4, Chapter 6] or [29]. Note that (2.8) is often deduced for nonnegative Z from a stronger condition, namely

$$\frac{Z}{\theta} \le B. \tag{2.9}$$

Indeed, it follows from (2.9) that $\mathrm{E}\, Z^2 \le \mathrm{E}\, Z\theta B = \theta^2 B$.

There is a problem which seems to be mostly overlooked in the literature. How exactly can (2.8) be used to construct fixed relative precision estimates at a given level of confidence? Below we discuss three possible methods.

Method 1: Chebyshev inequality. In MC simulations we can generate n independent copies of the random variable Z, denoted by $Z_1, \dots, Z_n$. The obvious candidate for a good estimator

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

of θ is the sample average, $\bar{Z} = (1/n)\sum_{i=1}^{n} Z_i$. The Chebyshev inequality and (2.8) give

$$\mathrm{P}(|\bar{Z} - \theta| > \theta\varepsilon) \leq \frac{B}{n\varepsilon^2}.$$

The right-hand side is less than or equal to α if

$$n \geq \frac{B}{\varepsilon^2 \alpha}. \tag{2.10}$$

Method 2: Bernstein inequality. Consider the case when (2.9) holds. For simplicity, additionally assume that $\varepsilon \leq 1$. Then the Bernstein inequality [6] yields

$$\mathrm{P}(|\bar{Z} - \theta| > \theta\varepsilon) \leq 2\exp\biggl\{-\frac{n\varepsilon^2}{2B + 2B\varepsilon/3}\biggr\} \leq 2\exp\biggl\{-\frac{3\varepsilon^2}{8B}n\biggr\}.$$

To make the right-hand side less than or equal to α, we need n to satisfy

$$n \geq \frac{8B}{3\varepsilon^2}\ln\biggl[\frac{\alpha}{2}\biggr]^{-1}. \tag{2.11}$$

Method 3: the median trick. Let us now consider nm independent copies of the random variable Z, denoted by Z_{il} and arranged in m blocks, each of length n. Let

$$\hat{\theta}_i = \frac{1}{n}\sum_{l=1}^{n} Z_{il} \quad \text{and} \quad \hat{\theta} = \mathrm{med}(\hat{\theta}_1, \dots, \hat{\theta}_m).$$

Estimator $\hat{\theta}$ is thus an MA. Note that (2.8) implies that $\mathrm{E}(\hat{\theta}_i - \theta)^2/\theta^2 \leq B/n$; so here (2.1) holds with $r(n) = 0$. Condition (2.3) simplifies to $n \geq C_1 B/\varepsilon^2$. Combining this with (2.4) we see that the number of samples sufficient for (2.2) is approximately

$$nm \sim C_1 C_2 \frac{B}{\varepsilon^2}\ln[2\alpha]^{-1}, \qquad \varepsilon, \alpha \to 0. \tag{2.12}$$

In fact, (2.2) holds if $n \geq C_1 B/\varepsilon^2$ and $m \geq C_2 \ln[2\alpha]^{-1}$, where n is an integer and m is an odd integer. For small α, (2.11) and (2.12) are much better than (2.10). The right-hand side of (2.12) is of the same form as (2.11) but with a larger constant, $C_1 C_2 \approx 19.34 > \frac{8}{3}$. On the other hand, MA uses only (2.8), whilst, for the Bernstein inequality, we need (2.9).

Example. (*Asmussen–Kroese estimator.*) One of the typical problems in the field of rare-event simulations is to estimate the tail probability of a random sum; see [4, Chapter 6]. This is needed, e.g. for computing the probability of ruin via the Khinchine–Pollaczek formula; see [2, pp. 285–287]. The classical MC algorithm introduced by Siegmund uses importance sampling and exponential change of measure. This method requires that the summands have light tails. In a series of papers [3], [5], [16], *conditional MC* algorithms have been developed for the case of heavy tails. Below we briefly describe one of the algorithms, focusing attention on the facts relevant to the subject of this paper.

We compute $\theta = \theta(u) = \mathrm{P}(S_N > u)$, where $S_N = X_1 + \cdots + X_N$ with i.i.d. summands having the tail function $\bar{F}(u) = 1 - F(u) = \mathrm{P}(X_1 > u)$ and N is an independent random variable. As a rule, $u > 0$ is large and θ is very small. Let us consider the estimator

$$Z = Z(u) = N\bar{F}(M_{N-1} \vee (u - S_{N-1})), \tag{2.13}$$

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

where $M_N = \max(X_1, \dots, X_N)$. Obviously, Z is an unbiased estimator of θ, because $Z = \mathrm{E}(S_N > u, M_N = X_N \mid X_1, \dots, X_{N-1})$. This estimator is denoted by Z_1 in [16] and by Z_4 in [4, Section 6.3]. For several classes of heavy-tailed distributions F, estimator (2.13) has bounded relative error provided, e.g. that N has moments of sufficiently high order; see [16, Theorem 4.2]. Although (2.8) is satisfied, (2.9) fails to hold if N is unbounded; see [16, Lemma 4.1]. Of the three methods of constructing exact confidence bounds, the Chebyshev inequality and the median trick can in principle be applied (provided that the actual constant B is extracted from the proofs in [16]). Bernstein's inequality breaks down.

To the authors' knowledge, in this example the MA estimator is the only known estimator for which an exponential inequality for large relative deviations holds uniformly for $u \to \infty$.

3. Product estimators

Assume that the quantity of interest is represented as a product of positive factors, $\theta = \mu_1 \cdots \mu_k$. Let $\hat{\mu}_1, \dots, \hat{\mu}_k$ be independent nonnegative random variables, where $\hat{\mu}_j$ is interpreted as an estimate of μ_j, possibly biased. Consider the estimator $\hat{\theta} = \hat{\mu}_1 \cdots \hat{\mu}_k$. Similarly as in [11], we will bound the relative MSE of $\hat{\theta}$ in terms of the relative MSE and the relative bias of $\hat{\mu}_j$. Let

$$\nu_j^2 = \frac{\mathrm{E}(\hat{\mu}_j - \mu_j)^2}{\mu_j^2}, \qquad b_j = \frac{\mathrm{E}\,\hat{\mu}_j - \mu_j}{\mu_j}, \qquad \nu^2 = \frac{\mathrm{E}(\hat{\theta} - \theta)^2}{\theta^2}. \tag{3.1}$$

Proposition 3.1. *Let $D > 0$ be a constant. If $\nu_j^2 \le D/k$ and $|b_j| \le D/2k$ for $j = 1, \dots, k$, then*

$$\nu^2 \le D + \tfrac{9}{4} D^2 \mathrm{e}^{2D}.$$

Let us begin with the following lemma.

Lemma 3.1. *Assume that $|x_j| \le D/k$ for $j = 1, \dots, k$. Then*

$$\prod_{j=1}^{k} (1 + x_j) = 1 + \sum_{j=1}^{k} x_j + r,$$

where $|r| \le D^2 \mathrm{e}^D / 2$.

Proof. Since

$$r = \sum_{j_1 < j_2} x_{j_1} x_{j_2} + \sum_{j_1 < j_2 < j_3} x_{j_1} x_{j_2} x_{j_3} + \cdots + x_1 \cdots x_k,$$

by our assumption we have

$$\begin{aligned} |r| &\le \binom{k}{2} \frac{D^2}{k^2} + \binom{k}{3} \frac{D^3}{k^3} + \cdots + \binom{k}{k} \frac{D^k}{k^k} \\ &\le D^2 \biggl(\frac{1}{2!} + \frac{1}{3!} D + \cdots + \frac{1}{k!} D^{k-2} \biggr) \\ &\le \frac{D^2}{2} \biggl(1 + \frac{1}{3} D + \frac{1}{3 \cdot 4} D^2 + \cdots \biggr) \\ &\le \frac{D^2}{2} \mathrm{e}^D. \end{aligned}$$

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

Lemma 3.2. *If ν_j^2, ν^2, and b_j are defined as in (3.1), then*

$$\nu^2 = \prod_{j=1}^{k}(1+\nu_j^2+2b_j) - 2\prod_{j=1}^{k}(1+b_j)+1.$$

Proof. Since the $\hat{\mu}_j$ are independent,

$$\frac{\mathrm{E}(\hat{\theta}-\theta)^2}{\theta^2} = \frac{\mathrm{E}\,\hat{\theta}^2 - 2\theta\,\mathrm{E}\,\hat{\theta}+\theta^2}{\theta^2} = \frac{\mathrm{E}\,\hat{\mu}_1^2\cdots\mathrm{E}\,\hat{\mu}_k^2}{\mu_1^2\cdots\mu_k^2} - 2\frac{\mathrm{E}\,\hat{\mu}_1\cdots\mathrm{E}\,\hat{\mu}_k}{\mu_1\cdots\mu_k}+1.$$

To conclude the proof, it is enough to note that

$$\frac{\mathrm{E}\,\hat{\mu}_j^2}{\mu_j^2} = \frac{\mathrm{E}(\hat{\mu}_j-\mu_j)^2 + 2\mu_j\,\mathrm{E}(\hat{\mu}_j-\mu_j)+\mu_j^2}{\mu_j^2} = \nu_j^2+2b_j+1,$$

$$\frac{\mathrm{E}\,\hat{\mu}_j}{\mu_j} = \frac{\mathrm{E}(\hat{\mu}_j-\mu_j)+\mu_j}{\mu_j} = b_j+1.$$

From the preceding lemmas we immediately obtain the proof of our basic result in this section.

Proof of Proposition 3.1. The formula for ν^2 given in Lemma 3.2 can be rewritten as follows, using Lemma 3.1:

$$\begin{aligned}\nu^2 &= 1+\sum\nu_j^2+2\sum b_j+r'-2\Big(1+\sum b_j+r''\Big)+1\\ &= \sum\nu_j^2+r'-2r'',\end{aligned}$$

where $|r'| \le \frac{1}{2}(2D)^2\mathrm{e}^{2D}$, because $\nu_j^2+2|b_j| \le 2D/k$, and $|r''| \le \frac{1}{2}(D/2)^2\mathrm{e}^{D/2}$, because $|b_j| \le D/2k$. Thus, $r'-2r'' \le \frac{9}{4}D^2\mathrm{e}^{2D}$. Of course, $\sum\nu_j^2 \le D$, and the result follows.

4. Bias and the MSE of Markov chain estimators

In this section we consider a Markov chain $X_0, X_1, \ldots$ on a finite state space $\mathcal{X}$. Assume that the chain is irreducible and aperiodic, but not necessarily reversible. Let P be the one-step transition matrix. The stationary distribution is denoted by π. Assume that f is a function defined on $\mathcal{X}$. We focus attention on estimating the stationary mean,

$$\mu = \mathrm{E}_\pi f = \sum_{x\in\mathcal{X}} f(x)\pi(x).$$

Many computational problems in physics, chemistry, and biology are of this form. If the space $\mathcal{X}$ is large and π is exponentially concentrated, it is impossible to sample directly from π, and MCMC methods have to be applied. The standard practice is to estimate μ by a sample average. To reduce bias, an initial part of the trajectory (the so-called burn-in time t) is usually discarded; cf. [31]. Thus, we consider

$$\bar{f}_{t,n} = \frac{1}{n}\sum_{i=t}^{t+n-1} f(X_i)$$

as an estimator of μ. If $t=0$ then we write $\bar{f}_n = (1/n)\sum_{i=0}^{n-1} f(X_i)$.

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

The following results involve the second largest eigenvalue of the multiplicative reversibilization of P, defined in [14]. Let us consider the Hilbert space L^2_π of functions $f\colon \mathcal{X} \to \mathbb{R}$ endowed with the scalar product $\langle f, g\rangle = \sum_{x\in\mathcal{X}} f(x)g(x)\pi(x)$. The norm is defined by $\|f\|^2 = \langle f, f\rangle$. We will freely identify functions and probability distributions with column vectors in $\mathbb{R}^s$, where $s = |\mathcal{X}|$. For example, the scalar product we work with can be rewritten as $\langle f, g\rangle = f^\top \Pi g$, where $\Pi = \operatorname{diag}[\pi(x)]_{x=1,\dots,s}$. We identify P with an operator on L^2_π given by $Pf(x) = \sum_y P(x, y)f(y)$. The adjoint operator is $P^* = \Pi^{-1}P^\top\Pi$. Indeed, $\langle f, Pg\rangle = f^\top\Pi Pg = f^\top\Pi P\Pi^{-1}\Pi g = \langle P^* f, g\rangle$. We say that P^*P is the *multiplicative reversibilization* of P. Operator P^*P is self-adjoint and nonnegative definite. Let us denote its eigenvalues by $1 = \lambda_1^2 > \lambda_2^2 \ge \cdots \ge \lambda_s^2 \ge 0$. We can assume that P^*P is irreducible, so the largest eigenvalue 1 is single. The corresponding (right) eigenspace is one-dimensional and it is spanned by 1, the constant function equal to 1. For simplicity, write $\lambda = \lambda_2$ and say that it is the *second largest singular value* of P.

Now we are in a position to prove our basic results about the bias and MSE of MCMC estimates.

4.1. Bias

Let $\pi_t(x) = \mathrm{P}(X_t = x)$. The initial distribution is thus π_0. Define a chi-squared 'distance from stationarity' as

$$\chi_t^2 = \sum_{x\in\mathcal{X}} \frac{(\pi_t(x) - \pi(x))^2}{\pi(x)} = \|\Pi^{-1}(\pi_t - \pi)\|^2.$$

The stationary variance of f is, by definition, $\sigma^2 = \|f - \mu\|^2$. Throughout this section, we will write $g = f - \mu$.

Theorem 4.1. *We have*

$$|\mathrm{E}(f(X_t)) - \mu| \le \sigma\chi_t \le \lambda^t\sigma\chi_0.$$

Applying Theorem 4.1 to $f(x) = \mathbf{1}(\pi_t(x) > \pi(x))$ and $\pi_0(x) = \mathbf{1}(x = x_0)$ we obtain $\|\pi_t - \pi\|_{\mathrm{TV}} := \frac{1}{2}\sum_x |\pi_t(x) - \pi(x)| \le \lambda^t\sigma/\sqrt{\pi(x_0)}$. This inequality implies the result of [14, Theorem 2.1] upon noting that $\sigma \le \frac{1}{2}$. For reversible chains, the same inequality is given in [8]. Letting $f(x) = \mathbf{1}(x = x_j)$ and $\pi_0(x) = \mathbf{1}(x = x_i)$ in Theorem 4.1, we obtain

$$\frac{|\mathrm{P}(X_t = x_j \mid X_0 = x_i) - \pi(x_j)|}{\pi(x_j)} \le \frac{\lambda^t}{\sqrt{\pi(x_j)\pi(x_i)}},$$

which is the inequality proved for reversible chains in [30, Proposition 3.1].

To prove Theorem 4.1, we need the following lemma, which we believe belongs to the folklore.

Lemma 4.1. *If* $\langle g, 1\rangle = 0$ *then* $\|P^t g\| \le \lambda^t\|g\|$ *for* $t = 0, 1, \dots$.

Proof. We have $\|Pg\|^2 = \langle Pg, Pg\rangle = \langle g, P^*Pg\rangle$. Now, use the well-known minimax characterization of eigenvalues of a self-adjoint operator (cf., e.g. [18, p. 176]). Since $v_1 \equiv 1$, the second largest eigenvalue of P^*P is

$$\lambda^2 = \max_{g\colon \langle g,1\rangle=0} \frac{\langle g, P^*Pg\rangle}{\langle g, g\rangle}.$$

Thus, for $\langle g, 1\rangle = 0$, we have $\|Pg\|^2 \le \lambda^2\|g\|^2$. To obtain the conclusion by induction, it is enough to note that $0 = \langle g, 1\rangle = \pi^\top g = \pi^\top Pg = \langle Pg, 1\rangle$.

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

Since the eigenvalues of P^*P and PP^* are the same, we have the following result.

Corollary 4.1. *If* $\langle g, 1\rangle = 0$ *then* $\|(P^*)^t g\| \le \lambda^t \|g\|$ *for* $t = 0, 1, \ldots$.

Proof of Theorem 4.1. To obtain the first inequality in the conclusion of the theorem, we proceed as follows. By the Cauchy–Schwarz inequality,

$$\begin{aligned}|\mathrm{E} f(X_t) - \mu| &= |(\pi_t^\top - \pi^\top)(f - \mu)|\\ &= |\langle \Pi^{-1}(\pi_t - \pi), g\rangle|\\ &\le \|\Pi^{-1}(\pi_t - \pi)\| \|g\|\\ &= \chi_t \sigma.\end{aligned}$$

The second of the claimed inequalities follows from Corollary 4.1. Indeed, $\Pi^{-1}(P^\top)^t = (P^*)^t \Pi^{-1}$ and $\langle \Pi^{-1}(\pi_t - \pi), 1\rangle = 0$. Therefore,

$$\begin{aligned}\chi_t &= \|\Pi^{-1}(\pi_t - \pi)\|\\ &= \|\Pi^{-1}(P^\top)^t(\pi_0 - \pi)\|\\ &= \|(P^*)^t \Pi^{-1}(\pi_0 - \pi)\|\\ &\le \lambda^t \|\Pi^{-1}(\pi_0 - \pi)\|\\ &= \lambda^t \chi_0.\end{aligned}$$

4.2. Mean-square error

It is well known that, for an arbitrary initial distribution,

$$\lim_{n\to\infty} n\,\mathrm{E}(\bar{f}_n - \mu)^2 = \lim_{n\to\infty} n \operatorname{var} \bar{f}_n = \tau^2, \tag{4.1}$$

where τ^2 is called the asymptotic variance, to avoid confusion with the stationary variance σ^2. The following result replaces asymptotics with respect to n in (4.1) (which is useless for our purposes) with a useful inequality.

Theorem 4.2. *Under our standing assumptions, we have*

$$\left|\mathrm{E}(\bar{f}_n - \mu)^2 - \frac{1}{n}\tau^2\right| \le \frac{1}{n^2}\frac{2\lambda\sigma^2 + (1+\lambda)\|f - \mu\|_\infty \sigma \chi_0}{(1-\lambda)^2},$$

where $\|f - \mu\|_\infty = \max_x(|f(x) - \mu|)$.

Proof. Set $Q_i = \operatorname{diag}[\pi_i(x)]_{x=1,\ldots,s}$, and write $R_i = Q_i - \Pi$. Now,

$$\begin{aligned}n^2\mathrm{MSE} &= n^2\,\mathrm{E}(\bar{f}_n - \mu)^2\\ &= \mathrm{E}\biggl(\sum_{i=0}^{n-1} g(X_i)\biggr)^2\\ &= 2\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1} \mathrm{E}\, g(X_i)g(X_j) + \sum_{i=0}^{n-1} \mathrm{E}\, g(X_i)^2\\ &= 2\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1} g^\top Q_i P^{j-i} g + \sum_{i=0}^{n-1} g^\top Q_i g\\ &= n^2\mathrm{BIAS} + n^2\mathrm{MSE}^*,\end{aligned}$$

where

$$n^2\mathrm{BIAS} = 2\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1} g^\top R_i P^{j-i} g + \sum_{i=0}^{n-1} g^\top R_i g,$$
$$n^2\mathrm{MSE}^* = 2\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1} g^\top \Pi P^{j-i} g + \sum_{i=0}^{n-1} g^\top \Pi g.$$

It is enough to show that

(i) $n^2|\mathrm{BIAS}| \le \|g\|_\infty \sigma \chi_0 (1+\lambda)/(1-\lambda)^2$;

(ii) $n^2|\mathrm{MSE}^* - \tau^2/n| \le 2\sigma^2\lambda/(1-\lambda)^2$.

First, we consider the bias term and prove (i). For $j \ge i$, in view of the Cauchy–Schwarz inequality, Lemma 4.1, and Theorem 4.1, we have

$$\begin{aligned}
|g^\top R_i P^{j-i} g| &\le |g^\top R_i \Pi^{-1}\Pi P^{j-i} g| \\
&= |\langle \Pi^{-1} R_i g, P^{j-i} g\rangle| \\
&\le \|\Pi^{-1} R_i g\| \|P^{j-i} g\| \\
&\le \|g\|_\infty \chi_i \|g\| \lambda^{j-i} \\
&\le \|g\|_\infty \chi_0 \|g\| \lambda^{j} \\
&= \|g\|_\infty \chi_0 \sigma \lambda^{j},
\end{aligned}$$

because $\|\Pi^{-1} R_i g\|^2 = \sum_x g(x)^2 (\pi_i(x) - \pi(x))^2/\pi(x) \le \|g\|_\infty^2 \chi_i^2$.

Now, setting $C = \|g\|_\infty \sigma \chi_0$, we obtain

$$\begin{aligned}
n^2|\mathrm{BIAS}| &\le 2C \sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1} \lambda^j + C\sum_{i=0}^{n-1}\lambda^i \\
&\le C\biggl(2\sum_{i=0}^{\infty} \frac{\lambda^{i+1}}{1-\lambda} + \frac{1}{1-\lambda}\biggr) \\
&= C\frac{1+\lambda}{(1-\lambda)^2}.
\end{aligned}$$

We have shown (i).

Now we turn to (ii). The asymptotic variance can be expressed in terms of P and f via the so-called fundamental matrix of the Markov chain. Let $T = P - 1\pi^\top$. For $i > 0$, we have $T^i = P^i - 1\pi^\top$ and $P^i g = T^i g$, because $\pi^\top g = 0$. The fundamental matrix is

$$Z = \sum_{i=0}^{\infty} T^i = (I - T)^{-1}. \tag{4.2}$$

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

We will make use of the following formula for the asymptotic variance:

$$\begin{aligned}\tau^2 &= \operatorname{var}_\pi f(X_0) + 2\sum_{i=1}^{\infty} \operatorname{cov}_\pi(f(X_0), f(X_i)) \\ &= \|g\|^2 + 2\sum_{i=1}^{\infty} \langle g, P^i g\rangle \\ &= g^\top \Pi\biggl(I + 2\sum_{i=1}^{\infty} T^i\biggr)g \\ &= g^\top[2\Pi Z - \Pi]g. \end{aligned} \tag{4.3}$$

Formula (4.3) is classical and can be found, e.g. in [7, p. 232].

Of course, $\sum_{i=1}^{n} T^i = ZT(I - T^n)$. Hence,

$$\begin{aligned}\sum_{i=0}^{n-1}\sum_{j=1}^{n-i-1} T^j &= \sum_{i=0}^{n-1} ZT(I - T^{n-i-1}) \\ &= nZT - Z\sum_{i=1}^{n} T^j \\ &= nZT - Z^2T(I - T^n) \\ &= n(Z - I) + Z^2T(I - T^n). \end{aligned}$$

Therefore,

$$\begin{aligned} n^2\mathrm{MSE}^* &= 2g^\top \Pi \sum_{i=0}^{n-1}\sum_{j=1}^{n-i-1} T^j g + ng^\top \Pi g \\ &= ng^\top[2\Pi(Z - I) + \Pi]g + 2g^\top \Pi Z^2 T(I - T^n)g. \end{aligned}$$

By (4.2), the first term on the right-hand side is equal to $n\tau^2$.

It remains to bound the second term. To this end, we use the simple observation that $g^\top \Pi T = g^\top \Pi P = g^\top (P^*)^\top \Pi$. The Cauchy–Schwarz inequality, Lemma 4.1, and Corollary 4.1 imply that

$$\begin{aligned} |g^\top \Pi Z Z T(I - T^n)g| &= \biggl|g^\top \Pi\biggl(\sum_{j=0}^{\infty} T^j\biggr)\biggl(\sum_{i=1}^{n} T^i\biggr)g\biggr| \\ &= \biggl|g^\top\biggl(\sum_{j=0}^{\infty}((P^*)^\top)^j\biggr)\Pi\biggl(\sum_{i=1}^{n} P^i\biggr)g\biggr| \\ &= \sum_{j=0}^{\infty}\sum_{i=1}^{n} |\langle (P^*)^j g, P^i g\rangle| \\ &\le \sum_{j=0}^{\infty}\sum_{i=1}^{n} \|P^j g\|\|P^i g\| \\ &\le \|g\|^2 \sum_{j=0}^{\infty}\lambda^j \sum_{i=1}^{n}\lambda^i \end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 \frac{\lambda(1-\lambda^n)}{(1-\lambda)^2} \\
&\le \sigma^2 \frac{\lambda}{(1-\lambda)^2}.
\end{aligned}
$$

Consequently, $|n^2\mathrm{MSE}^* - n\tau^2| \le 2\sigma^2\lambda/(1-\lambda)^2$. This completes the proof.

4.3. Corollaries

In this subsection we state some simple consequences of the preceding results in a form 'ready to use' in the analysis of MPA estimators in Section 5.

Corollary 4.2. *We have*

$$\mathrm{E}(\bar{f}_{t,n} - \mu)^2 \le \frac{1}{n}\frac{1+\lambda}{1-\lambda}\sigma^2 + \frac{1}{n^2}\frac{2\lambda\sigma^2 + (1+\lambda)\|f-\mu\|_\infty \sigma\chi_t}{(1-\lambda)^2}.$$

Proof. In view of Theorem 4.2 it is enough to show that

$$\tau^2 \le \frac{1+\lambda}{1-\lambda}\sigma^2.$$

This inequality is well known for reversible chains, but it also holds in the nonreversible case (λ^2 denotes the second largest eigenvalue of P^*P). Indeed, by (4.3), the Cauchy–Schwarz inequality, and Lemma 4.1,

$$\tau^2 = \|g\|^2 + 2\sum_{i=1}^{\infty}\langle g, P^i g\rangle \le \|g\|^2\bigg(1 + 2\sum_{i=1}^{\infty}\lambda^i\bigg) = \frac{1+\lambda}{1-\lambda}\sigma^2.$$

This completes the proof.

Corollary 4.2 plays a role analogous to Proposition 4.2 of Aldous [1] and Proposition 3.2 of Gillman [15]. Let us point out the main differences. Aldous's inequality involves $1/\min_x \pi(x)$. This quantity is of moderate order of magnitude for uniform distributions, but it is disastrously large in the problems considered in Section 5; cf. Examples 5.1 and 5.2. Gillman's bound on the MSE of $\bar{f}_{t,n}$ (in our notation) is implicit in the proof of his Proposition 3.2. This bound is not dependent on $1/\min_x \pi(x)$, but it does not go to 0 as $n \to \infty$ with t fixed. Both the cited results are derived only for reversible chains.

In the next corollaries we assume that $f \ge 0$ and write $B = \|f\|_\infty/\mu$. Note that $\sigma^2/\mu^2 \le B$ (cf. (2.8) and (2.9)). We also use the notation $\varrho = 1/(1-\lambda)$.

Corollary 4.3. *If $f \ge 0$ then*

$$\frac{|\mathrm{E}\bar{f}_{t,n} - \mu|}{\mu} \le \frac{\varrho\sqrt{B}\chi_t}{n}.$$

Indeed, by Theorem 4.1 we have

$$|\mathrm{E}\bar{f}_{t,n} - \mu| \le \frac{1}{n}\sum_{i=0}^{n-1}|\mathrm{E}f(X_{t+i}) - \mu| \le \frac{1}{n}\sum_{i=0}^{n-1}\lambda^i\chi_t\sigma \le \frac{1}{n}\frac{\chi_t\sigma}{1-\lambda}.$$

Corollary 4.4. *If $f \ge 0$ then*

$$\frac{\mathrm{E}(\bar{f}_{t,n} - \mu)^2}{\mu^2} \le \frac{2\varrho B}{n}\bigg(1 + \frac{\varrho}{n} + \frac{\varrho\sqrt{B}\chi_t}{n}\bigg).$$

Indeed, by Corollary 4.2 we have

$$\begin{aligned}\mathrm{E}(\bar f_{t,n}-\mu)^2 &\le \frac{1+\lambda}{n(1-\lambda)}\sigma^2+\frac{2\lambda}{n^2(1-\lambda)^2}\sigma^2+\frac{1+\lambda}{n^2(1-\lambda)^2}B\mu\sigma\chi_t\\ &\le \frac{2\varrho\sigma^2}{n}+\frac{2\varrho^2\sigma^2}{n^2}+\frac{2\varrho^2B\mu\sigma\chi_t}{n^2}\\ &\le \frac{2\varrho B\mu^2}{n}\biggl(1+\frac{\varrho}{n}+\frac{\varrho\sqrt{B}\chi_t}{n}\biggr).\end{aligned}$$

Theorem 4.1 immediately entails the following result.

Corollary 4.5. *For a deterministic initial distribution,* $\mathrm{P}(X_0=x)=1$,

$$\chi_t\le \mathrm{e}^{-t/\varrho}\pi(x)^{-1/2}.$$

Indeed, it is easy to see that $\chi_0\le\pi(x)^{-1/2}$ and $\lambda^t=(1-1/\varrho)^t\le \mathrm{e}^{-t/\varrho}$.

Finally, from Corollaries 4.3, 4.4, and 4.5, we derive the following tidy bounds.

Corollary 4.6. *Assume that* $f\ge 0$ *and* $\mathrm{P}(X_0=x)=1$. *If*

$$t\ge\varrho\ln(\pi(x)^{-1/2}B^{-1/2})$$

then

$$\frac{\mathrm{E}(\bar f_{t,n}-\mu)^2}{\mu^2}\le\frac{2\varrho B}{n}\biggl(1+\frac{2\varrho B}{n}\biggr)\quad\textit{and}\quad\frac{|\,\mathrm{E}\,\bar f_{t,n}-\mu|}{\mu}\le\frac{\varrho B}{n}.$$

Indeed, it is enough to note that $\chi_t\le\sqrt{B}$.

5. Median of products of averages

As mentioned in the introduction, we consider the problem of computing a quantity which is expressed in the form $\theta=\mu_1\cdots\mu_k$. Each μ_j is the expectation of some function f_j with respect to a probability distribution π_j on some finite space, $\mu_j=\mathrm{E}_{\pi_j}f_j$. Assume that we can generate a Markov chain with transition matrix P_j such that π_j is its stationary distribution. The sampling procedure under consideration starts from x_j and makes $t+n$ steps. The first t steps are discarded and the remaining n steps are used to compute averages. The resulting estimates of the μ_js are multiplied. Finally, the whole procedure is repeated m times and the median is taken as an estimate of θ. The basic parameters of the algorithm are t, n, k, and m. Clearly, the number of samples is $(n+t)km$. A formal description of the algorithm is given in Algorithm 5.1, below. Note that in this section we have to modify earlier notation to accommodate different 'building blocks' in one algorithm.

Algorithm 5.1. (*Median of products of averages.*)

```
Inputs: $t, n, k, m$, and $(P_j, x_j)$ for $j = 1, \dots, k$.
for $i = 1$ to $m$ do
    for $j = 1$ to $k$ do
        generate the trajectory $X^0_{ij}, X^1_{ij}, \dots, X^t_{ij}, \dots, X^{t+n-1}_{ij}$ of the Markov chain with transition rule $P_j$ and initial state $X^0_{ij} = x_j$
        compute $\hat\mu_{ij} = (1/n)\sum_{l=0}^{n-1} f_j(X^{t+l}_{ij})$
    end for
```

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

```
        compute $\hat{\theta}_i = \prod_{j=1}^{k} \hat{\mu}_{ij}$
    end for
    compute $\hat{\theta} = \mathrm{med}(\hat{\theta}_1, \dots, \hat{\theta}_m)$
    Output: $\hat{\theta}$.
```

Assume that $f_j \geq 0$ for every $j = 1, \dots, k$. Suppose that we know *a priori* $B_\bullet > 0$ and $\pi_\bullet > 0$ such that $f_j/\mu_j \leq B_\bullet$ and $\pi_j(x_j) \geq \pi_\bullet$. Moreover, we denote the *second largest singular value* of P_j by λ_j, write $\varrho_j = (1 - \lambda_j)^{-1}$, and assume that $\varrho_j \leq \varrho_\bullet$ for every $j = 1, \dots, k$. Our main result shows how the quantities $B_\bullet$, $\pi_\bullet$, $\varrho_\bullet$, and k determine t, n, and m, and, consequently, the cost of the algorithm.

Theorem 5.1. *Let* $0 < \varepsilon < 1$ *and* $0 < \alpha < \frac{1}{2}$. *Assume that*

(i) $t \geq \varrho_\bullet \ln(\pi_\bullet^{-1/2} B_\bullet^{-1/2})$;

(ii) $n \geq 2C_1 \varrho_\bullet B_\bullet k \varepsilon^{-2}(1 + \varepsilon^2)$;

(iii) $m \geq C_2 \ln(2\alpha)^{-1}$ *and* m *is odd,*

where C_1 *and* C_2 *are the universal constants defined in Section 2. Then the final estimate* $\hat{\theta}$ *satisfies* $\mathrm{P}(|\hat{\theta} - \theta| \leq \theta\varepsilon) \geq 1 - \alpha$.

Bounds on the cost of MCMC algorithms occur in many papers devoted to the computational complexity of counting problems [11], [15], [19, Chapter 3]–[22]. In these papers, to prove that a given algorithm in a given problem has the required relative precision, the authors derived *ad hoc* bounds which correspond to our Propositions 2.1 and 3.1, and Corollaries 4.5 and 4.6. The conditions of Theorem 5.1 highlight distinct roles played by the problem-dependent parameters $B_\bullet$, $\pi_\bullet$, $\varrho_\bullet$, and k. The conclusion is applicable to general MPA algorithms. We have optimized the constants C_1 and C_2 so that in selected examples (see Examples 5.1 and 5.2, below) the cost of the algorithms has been reduced at least several times compared to earlier results [15], [20], [21]. Moreover, Theorem 5.1 does not require reversibility and covers, e.g. the 'systematic sweep' or 'systematic scan' schemes [9], [12]. Analysis of such schemes is very difficult and first bounds have been obtained recently for spin systems in [12].

Proof of Theorem 5.1. First we are going to apply Corollary 4.6 to the averages $\hat{\mu}_{ij}$, then Proposition 3.1 to the products $\hat{\theta}_i$, and finally Proposition 2.1 to the median $\hat{\theta}$.

Let

$$v_j^2 = \frac{\mathrm{E}(\hat{\mu}_{ij} - \mu_j)^2}{\mu_j^2}, \qquad b_j = \frac{\mathrm{E}\,\hat{\mu}_{ij} - \mu_j}{\mu_j}, \qquad v^2 = \frac{\mathrm{E}(\hat{\theta}_i - \theta)^2}{\theta^2}.$$

Corollary 4.6, when translated to our new notation, asserts that assumption (i) implies that

$$v_j^2 \leq \frac{2\varrho_\bullet B_\bullet}{n}\biggl(1 + \frac{2\varrho_\bullet B_\bullet}{n}\biggr), \qquad |b_j| \leq \frac{\varrho_\bullet B_\bullet}{n}.$$

If we write

$$D = \frac{2\varrho_\bullet B_\bullet k}{n}\biggl(1 + \frac{2\varrho_\bullet B_\bullet}{n}\biggr), \qquad B = 2\varrho_\bullet B_\bullet k,$$

then $\nu_j^2 \le D/k$ and $|b_j| \le D/(2k)$, so Proposition 3.1 implies that

$$\begin{aligned}\nu^2 &\le D + \frac{9}{4}D^2 \mathrm{e}^{2D} \\ &\le \frac{B}{n}\left(1+\frac{B}{n}+\frac{9B}{4n}\left(1+\frac{B}{n}\right)^2 \exp\left\{\frac{2B}{n}\left(1+\frac{B}{n}\right)\right\}\right), \end{aligned} \tag{5.1}$$

because $D \le (B/n)(1+B/n)$. This is clearly an expression of the form (2.1) and we are in a position to apply Proposition 2.1. It remains to verify that assumption (ii) implies (2.3) or, equivalently, that the right-hand side of (5.1) is less than or equal to ε^2/C_1. The following elementary computation shows this is indeed true. Set $H = B/n$. Since

$$H \le \frac{\varepsilon^2}{(1+\varepsilon^2)C_1} < \frac{\varepsilon^2}{C_1} < \frac{\varepsilon^2}{8} < \frac{1}{8},$$

it follows that

$$\begin{aligned} H\left(1+H+\frac{9}{4}H(1+H)^2\mathrm{e}^{2H(1+H)}\right) &< H\left(1+\frac{\varepsilon^2}{8}\left[1+\frac{9}{4}\left(\frac{9}{8}\right)^2\mathrm{e}^{9/32}\right]\right) \\ &< H(1+\varepsilon^2) \\ &\le \frac{\varepsilon^2}{C_1}, \end{aligned}$$

and the proof is complete.

5.1. Remarks on alternative approaches

Let us compare the computational complexity of Algorithm 1 with alternative schemes known in the literature. The criterion, as before, is the number of samples necessary to guarantee that $\mathrm{P}(|\hat{\theta} - \theta| \le \theta\varepsilon) \ge 1 - \alpha$. We use the $O(\cdot)$ notation, thus neglecting constants.

The cost of Algorithm 5.1 is, by Theorem 5.1,

$$(n+t)km = O(\varrho_\bullet k(kB_\bullet\varepsilon^{-2} + \ln\pi_\bullet^{-1})\ln\alpha^{-1}). \tag{5.2}$$

Algorithm 5.1 is classical and close to computational practice, but many important theoretical results have been obtained for algorithms which use averages over final states of *multiple* independent runs of the chain; see [11], [19, Chapter 3]–[22], and [30]. This scheme, given by Algorithm 5.2, below, is easier to analyze because it involves averages of independent variables. It can be shown that the total number of samples in Algorithm 5.2 is

$$tnkm = O(\varrho_\bullet k^2 B_\bullet \varepsilon^{-2}\ln(\pi_\bullet^{-1}B_\bullet\varepsilon^{-1}k)\ln\alpha^{-1}). \tag{5.3}$$

Algorithm 5.2. (*Median of products of averages with multiple runs.*)

```
**Inputs:** $t, n, k, m$, and $(P_j, x_j)$ for $j = 1, \dots, k$.
**for** $i = 1$ to $m$ **do**
    **for** $j = 1$ to $k$ **do**
        **for** $l = 1$ to $n$ **do**
            generate the trajectory $X_{ijl}^0, X_{ijl}^1, \dots, X_{ijl}^t$ of the Markov chain with transition rule
            $P_j$ and initial state $X_{ijl}^0 = x_j$
        **end for**
```

```
            compute $\hat{\mu}_{ij} = (1/n)\sum_{l=1}^{n} f_j(X_{ijl}^t)$
        end for
        compute $\hat{\theta}_i = \prod_{j=1}^{k} \hat{\mu}_{ij}$
    end for
    compute $\hat{\theta} = \mathrm{med}(\hat{\theta}_1, \ldots, \hat{\theta}_m)$
    Output: $\hat{\theta}$.
```

Finally, let us consider Algorithm 5.3, below, which is seemingly simpler because it uses *products of averages along trajectories* (without medians).

Algorithm 5.3. (*Products of averages.*)

```
    Inputs: $t, n, k$, and $(P_j, x_j)$ for $j = 1, \ldots, k$.
    for $j = 1$ to $k$ do
        generate the trajectory $X_j^0, X_j^1, \ldots, X_j^t, \ldots, X_j^{t+n-1}$ of the Markov chain with transition
        rule $P_j$ and initial state $X_j^0 = x_j$
        compute $\hat{\mu}_j = (1/n)\sum_{l=0}^{n-1} f_j(X_j^{t+l})$
    end for
    compute $\hat{\theta} = \prod_{j=1}^{k} \hat{\mu}_j$
    Output: $\hat{\theta}$.
```

The length n of a single trajectory in Algorithm 5.3 must of course be greater than in Algorithm 5.1 to achieve the same relative accuracy and level of confidence. To derive bounds analogous to those in Theorem 5.1, we can use some exponential inequality for the deviations of $\hat{\mu}_j$ from μ_j and then the Bonferroni inequality to obtain a confidence bound for products. Exponential inequalities for Markov chain averages [10], [15], [24] allow us to obtain $\mathrm{P}(|\hat{\mu}_j - \mu_j| > \eta\mu_j) \le A\exp\{-Rn\eta^2/(B_j\varrho_j)\}$ for some absolute constants $A, R > 0$, where B_j bounds f_j/μ_j. To infer that $\mathrm{P}(|\prod\hat{\mu}_j - \prod\mu_j| > \prod\mu_j\varepsilon) \le \alpha$ via the Bonferroni inequality, we have to ensure that, say, $\mathrm{P}(|\hat{\mu}_j - \mu_j| > \mu_j\varepsilon/(2k)) \le \alpha/k$. We omit the details, because the best we can hope to obtain in this way is

$$(t+n)k = O(\varrho_\bullet k(k^2 B_\bullet \varepsilon^{-2}\ln(\alpha^{-1}k) + \ln\pi_\bullet^{-1})). \tag{5.4}$$

Bound (5.3) is clearly worse than (5.2). In most examples of practical relevance, (5.4) is also worse than (5.2).

5.2. Examples

Many models of statistical physics describe equilibrium properties of configurations of particles. Let $\mathcal{X}$ be a finite space of configurations. The *Gibbs distribution* at inverse temperature $\beta > 0$ is given by $\pi_\beta(x) = Z_\beta^{-1}\mathrm{e}^{-\beta V(x)}$, $x \in \mathcal{X}$, where V is a potential (energy) function and Z_β is a normalizing constant called the *partition function*. The *Boltzmann distribution* on the space of possible energy levels is induced by the Gibbs distribution:

$$\rho_\beta(v) = \sum_{x\in\mathcal{X}\colon V(x)=v} \pi_\beta(x) = \frac{w(v)\mathrm{e}^{-\beta v}}{Z_\beta}, \tag{5.5}$$

where the function $w(v) = |V^{-1}(w)|$ is called the *density* of states and

$$Z_\beta = \sum_v w(v)\mathrm{e}^{-\beta v}. \tag{5.6}$$

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

The summation in (5.6) is computationally feasible, because the set of energy levels v is typically of moderate size, say approximately 10^4, in contrast with the size of $\mathcal{X}$, which is usually exponentially large. Therefore, knowledge of the density of states, even up to proportionality, is sufficient for computing the Boltzmann distribution, which is of primary interest.

Modern algorithms, based on the idea of the multihistogram due to Ferrenberg and Swendsen [13], usually involve several series of MCMC simulations performed at different temperatures $\beta_1 < \beta_2 < \cdots < \beta_k$. A Markov chain at temperature β_j converges to π_{β_j} and is used to compute estimates $\hat{\rho}_{\beta_j}(v)$ of Boltzmann probabilities. If the length of simulation at different temperatures is equal then the *multihistogram* estimator of $w(v)$ is given by

$$\hat{w}(v) = \frac{\sum_{j=1}^{k} \hat{\rho}_{\beta_j}(v)}{\sum_{j=1}^{k} \exp\{-\beta_j v\}/Z_{\beta_j}}. \tag{5.7}$$

A simple rationale behind (5.7) is the equation $\sum_j \rho_{\beta_j}(v) = w(v) \sum_j \exp\{-\beta_j v\}/Z_{\beta_j}$, which immediately follows from (5.5). Let us refer to [13] and [26, Chapter 8] for a more general version and different derivations of (5.7). Common practice, also recommended by the cited authors, is to iteratively approximate $w(v)$ and unknown Z_{β_j} intermittently using (5.6) and (5.7). Both sets of values are in this way estimated up to a multiplicative constant.

An alternative way is to use the expression

$$\begin{aligned}\frac{Z_{\beta_{j+1}}}{Z_{\beta_j}} &= \sum_{x \in \mathcal{X}} \exp\{-(\beta_{j+1} - \beta_j)V(x)\}\pi_{\beta_j}(x) \\ &= \mathrm{E}_{\beta_j} \exp\{-(\beta_{j+1} - \beta_j)V\} \end{aligned} \tag{5.8}$$

and the ‘telescopic product’

$$Z_{\beta_j} = Z_{\beta_1} \frac{Z_{\beta_2}}{Z_{\beta_1}} \cdots \frac{Z_{\beta_j}}{Z_{\beta_{j-1}}}. \tag{5.9}$$

Note that even if the value of Z_{β_1} is unknown, we can estimate the ratio in (5.8) by MCMC methods and, thus, estimate the collection of Z_{β_j} $(j = 1, \dots, k)$ up to proportionality. Then we can use (5.7) directly and, thus, avoid the iterative procedure mentioned before. The computational scheme based on (5.8) and (5.9) fits in the setup considered in Section 5: Z_{β_k} is expressed as a product of the expected values $\mu_j = Z_{\beta_{j+1}}/Z_{\beta_j}$. This scheme is applied in [27] and [28] to the analysis of a model of protein folding. It is interesting that algorithms based on (5.8) and (5.9) were proposed earlier in theoretical papers on the computational complexity of counting problems [11], [20], [21], [30].

Below we consider two classical examples from statistical physics, apply our theorem, and compare the complexity of the three algorithms described in the previous subsections. For a detailed presentation of the analyzed models and background of the considered problems, we refer the reader to [20] and [21].

Example 5.1. (*The Ising model.*) The state space $\mathcal{X}$ consists of all spanning subgraphs of a given graph $(\mathcal{V}, \mathcal{E})$. The problem is to compute the partition function

$$Z = \sum_{x \in \mathcal{X}} \gamma^{|x|} \tau^{|\mathrm{ODD}(x)|},$$

where $\gamma, \tau > 0$ and $\mathrm{ODD}(x)$ stands for the set of all odd-degree vertices of graph x. Jerrum and Sinclair [20] gave an instance of Algorithm 5.2 and they proved the following bounds.

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

Let $N = |\mathcal{V}|$ and $M = |\mathcal{E}|$. Then, in our notation, $k \le N$, $B_\bullet \le 10$, $\varrho_\bullet \le 2N^4M^2$, and $\pi_\bullet^{-1} \le 2^M$. Gillman [15] applied these bounds in his analysis of Algorithm 5.1 and (implicitly) of Algorithm 5.3. The cost is the following:

$$\begin{aligned} &O(N^6M^2\varepsilon^{-2}\ln\alpha^{-1}) \quad \text{for Algorithm 5.1,}\\ &O(N^6M^3\varepsilon^{-2}\ln\alpha^{-1}) \quad \text{for Algorithm 5.2,}\\ &O(N^7M^2\varepsilon^{-2}\ln(\alpha^{-1}N)) \quad \text{for Algorithm 5.3.}\end{aligned}$$

To simplify the expressions for Algorithms 5.1 and 5.2, we have used rather unrestrictive assumptions that $\ln[N\varepsilon^{-1}] \ll M \ll N\varepsilon^{-2}$.

Let us note that from our Theorem 5.1 it follows that the cost of Algorithm 5.1 is asymptotically equivalent to $40C_1C_2N^6M^2\varepsilon^{-2}\ln\alpha^{-1}$ for $\varepsilon, \alpha \to 0$. The constant $40C_1C_2 \approx 769$ is about three times less than that in [15].

Example 5.2. (*The monomer-dimer model.*) Consider the state space $\mathcal{X}$ consisting of all matchings in a graph $(\mathcal{V}, \mathcal{E})$. The problem is to compute the partition function

$$Z = \sum_{x\in\mathcal{X}} \tau^{|x|},$$

where $\tau > 0$. Jerrum and Sinclair [21] gave an instance of Algorithm 5.2. In this case the following bounds are shown. Let $N = |\mathcal{V}|$ and $M = |\mathcal{E}|$. Then $k \le 2N\ln[\tau' M] + 2$, $B_\bullet \le \mathrm{e} \approx 2.718$, $\varrho_\bullet \le 4MN\tau'$, where $\tau' = \max(\tau, 1)$, and $\pi_\bullet^{-1} \le (2N)!\tau'^N$. From (5.2), (5.3), and (5.4), we obtain the following bounds:

$$\begin{aligned} &O(N^3M\tau'(\ln(M\tau'))^2\varepsilon^{-2}\ln\alpha^{-1}) \quad \text{for Algorithm 5.1,}\\ &O(N^4M\tau'(\ln(N\tau'))^3\varepsilon^{-2}\ln\alpha^{-1}) \quad \text{for Algorithm 5.2,}\\ &O(N^4M\tau'(\ln(N\tau'))^3\varepsilon^{-2}\ln(\alpha^{-1}N)) \quad \text{for Algorithm 5.3.}\end{aligned}$$

Again, we have used some simplifying assumptions to make the bounds more readable. In the expressions for Algorithms 5.1, 5.2, and 5.3 we assume respectively that $\ln N \ll \ln[M\tau']\varepsilon^{-2}$, $\ln\varepsilon^{-1} \ll N\ln[N\tau']$, and $\ln M \ll N$. None of these assumptions seems to be restrictive.

We have to admit however that the results in the spirit of computational complexity theory, with precise nonasymptotic bounds, so far remain too pessimistic to be applied in practice.

Appendix A. Proofs of the complexity bounds for Algorithms 5.2 and 5.3

Proof of (5.3). We derive bounds on t, n, and m. Let $f_j/\mu_j \le B_j$, and note that $\sigma_j^2 = \operatorname{var}_{\pi_j} f_j \le B_j\mu_j^2$.

Choice of t. By Theorem 4.1,

$$|\mathrm{E} f_j(X_{ijl}^t) - \mu_j| \le \exp\biggl\{-\frac{t}{\varrho_j}\biggr\}\frac{\sigma_j}{\sqrt{\pi_j(x_j)}} \le \exp\biggl\{-\frac{t}{\varrho_j}\biggr\}\sqrt{\frac{B_j}{\pi_j(x_j)}}\mu_j.$$

Let us divide both sides by μ_j, replace B_j, ϱ_j, and $\pi_j(x_j)$ by $B_\bullet$, $\varrho_\bullet$, and $\pi_\bullet$, and then choose t sufficiently large to make the right-hand side less than or equal to $a\varepsilon^2/(4k)$. Here we can

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

choose any fixed $a < \frac{1}{2}$, but for definiteness, let us take $a = a_* = 1/C_1$ as in Proposition 2.1. We see that

$$\frac{|\mathrm{E} f_j(X^t_{ijl}) - \mu_j|}{\mu_j} \le \sqrt{\frac{B_\bullet}{\pi_\bullet}} \mathrm{e}^{-t/\varrho_\bullet} \le \frac{a\varepsilon^2}{4k} \tag{A.1}$$

is satisfied if we choose

$$t \ge \varrho_\bullet \ln \frac{4k\sqrt{B_\bullet}}{a\varepsilon^2\sqrt{\pi_\bullet}} = O(\varrho_\bullet \ln(\pi_\bullet^{-1} B_\bullet \varepsilon^{-1} k)). \tag{A.2}$$

Choice of n. Since $\hat{\mu}_{ij}$ is an average of i.i.d. variables $\operatorname{var} \hat{\mu}_{ij} = (1/n) \operatorname{var} f_j(X^t_{ijl})$. But it follows from (A.1) that $\mathrm{E}\, f_j(X^t_{ijl}) \le 2\mu_j$, so

$$\operatorname{var} f_j(X^t_{ijl}) \le \mathrm{E}\, f_j(X^t_{ijl})^2 \le \mathrm{E}\, f_j(X^t_{ijl})\mu_j B_j \le 2\mu_j^2 B_j,$$

because $0 \le f_j(X^t_{ijl}) \le \mu_j B_j$. We can see that

$$\frac{\operatorname{var} \hat{\mu}_{ij}}{\mu_j^2} \le \frac{2B_j}{n} \le \frac{2B_\bullet}{n} \le \frac{a\varepsilon^2}{2k} \tag{A.3}$$

is satisfied if we choose

$$n \ge \frac{4kB_\bullet}{a} = O(B_\bullet \varepsilon^{-2} k). \tag{A.4}$$

Now, (A.1) together with (A.3) implies that

$$\nu_j^2 := \frac{\mathrm{E}(\hat{\mu}_{ij} - \mu_j)^2}{\mu_j^2} \le \left(\frac{a\varepsilon^2}{4k}\right)^2 + \frac{a\varepsilon^2}{2k} \le \frac{2a\varepsilon^2}{3k}.$$

We are in a position to apply Proposition 3.1 with $D = 2a\varepsilon^2/3$ and obtain

$$\nu^2 := \frac{\mathrm{E}(\hat{\theta}_i - \theta)^2}{\theta^2} \le D + \frac{9}{4} D^2 \mathrm{e}^{2D} \le a\varepsilon^2.$$

By the Chebyshev inequality, it follows that $\mathrm{P}(|\hat{\theta}_i - \theta| > \varepsilon) \le a$.

Choice of m. Exactly as in the proof of Theorem 5.1,

$$m \ge C_2 \ln(2\alpha)^{-1} = O(\ln \alpha^{-1}). \tag{A.5}$$

It remains to take the product of (A.2), (A.4), and (A.5), and multiply by k.

Proof of (5.4). We will use a simplified version of an inequality due to Lezaud [24, Theorem 1.1]. Without loss of generality, we can assume that $0 \le f_j \le 1$ and set $B_j = 1/\mu_j$. The stationary variance of f_j is bounded by $1/B_j$. Since in Algorithm 5.3 we have $\hat{\mu}_j = (1/n) \sum_{l=0}^{n-1} f_j(X_j^{t+l})$, the Lezaud inequality yields

$$\begin{aligned} \mathrm{P}\left(|\hat{\mu}_j - \mu_j| > \frac{\varepsilon\mu_j}{2k}\right) &\le 3\left(1 + \frac{\exp\{-t/\varrho_j\}}{\sqrt{\pi_j(x_j)}}\right) \exp\left\{-\frac{n(\varepsilon_j\mu_j/2k)^2}{5\varrho_j/B_j}\right\} \\ &\le 3\left(1 + \frac{\mathrm{e}^{-t/\varrho_\bullet}}{\sqrt{\pi_\bullet}}\right) \exp\left\{-\frac{n\varepsilon^2}{20k^2\varrho_\bullet B_\bullet}\right\}. \end{aligned} \tag{A.6}$$

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

Choice of t. If

$$t \geq \varrho_{\bullet} \ln \frac{1}{\sqrt{\pi_{\bullet}}} = O(\varrho_{\bullet} \ln \pi_{\bullet}^{-1})$$

then (A.6) simplifies to

$$\mathrm{P}\bigg(|\hat{\mu}_j - \mu_j| > \frac{\varepsilon \mu_j}{2k}\bigg) \leq 6 \exp\bigg\{-\frac{n\varepsilon^2}{20k^2 \varrho_{\bullet} B_{\bullet}}\bigg\}. \tag{A.7}$$

Choice of n. If

$$n \geq \frac{\varrho_{\bullet} k^2 B_{\bullet}}{40\varepsilon^2} \ln \frac{6k}{\alpha} = O\bigg(\frac{\varrho_{\bullet} k^2 B_{\bullet}}{\varepsilon^2} \ln[\alpha^{-1} k]\bigg)$$

then it follows from (A.7) that

$$\mathrm{P}\bigg(|\hat{\mu}_j - \mu_j| > \frac{\varepsilon \mu_j}{2k}\bigg) \leq \frac{\alpha}{k}. \tag{A.8}$$

By Lemma 2.3 with $D = \varepsilon/2$,

$$\mu_j\bigg(1 - \frac{\varepsilon}{2k}\bigg) \leq \hat{\mu}_j \leq \mu_j\bigg(1 + \frac{\varepsilon}{2k}\bigg) \quad \text{for } j = 1, \dots, k$$

implies that

$$\bigg(\prod_j \mu_j\bigg)(1 - \varepsilon) \leq \prod_j \hat{\mu}_j \leq \bigg(\prod_j \mu_j\bigg)(1 + \varepsilon),$$

because

$$\prod_j \bigg(1 \pm \frac{\varepsilon}{2k}\bigg) = 1 \pm \frac{\varepsilon}{2} + r \quad \text{where} \quad |r| \leq \frac{\varepsilon^2}{8} \mathrm{e}^{\varepsilon/2} < \frac{\varepsilon}{2}.$$

Now we can use the Bonferroni inequality and (A.8) to obtain $\mathrm{P}(|\hat{\theta} - \theta| > \varepsilon\theta) \leq \alpha$, which completes the proof.

References

[1] Aldous. D. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Prob. Eng. Inf. Sci.* **1,** 33–46.

[2] Asmussen, S. (2000). *Ruin Probabilities* (Adv. Ser. Statist. Sci. Appl. Prob. **2**). World Scientific, River Edge, NJ.

[3] Asmussen, S. and Binswanger, K. (1997). Simulation of ruin probabilities for subexponential claims. *ASTIN Bull.* **27,** 297–318.

[4] Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis* (Stoch. Modelling Appl. Prob. **57**). Springer, New York.

[5] Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Prob.* **38,** 545–558.

[6] Bernstein, S. N. (1924). On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Savantes Ukraine, Sect. Math.* **1,** 38–49.

[7] Bremaud, P. (1999). *Markov Chains*. Springer, New York.

[8] Diaconis, P. and Strook, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.* **1,** 36–61.

[9] Diaconis, P., Holmes, S. and Neal, R. M. (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Prob.* **10,** 726–752.

[10] Dinwoodie, I. H. (1995). A probability inequality for the occupation measure of a reversible Markov chain. *Ann. Appl. Prob.* **5,** 37–43.

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press

[11] Dyer, M. and Frieze, A. (1991). Computing the volume of convex bodies: a case where randomness provably helps. In *Probabilistic Combinatorics and its Applications* (Proc. AMS Symp. Appl. Math. **44**), American Mathematical Society, Providence, RI, pp. 123–169.

[12] Dyer, M., Goldberg, L. A. and Jerrum, M. (2006). Systematic scan for sampling colorings. *Ann. Appl. Prob.* **16,** 185–230.

[13] Ferrenberg, A. M. and Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63,** 1195–1198.

[14] Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Prob.* **1,** 62–87.

[15] Gillman, D. (1998). A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.* **27,** 1203–1220.

[16] Hartinger, J. and Kortschak, D. (2006). On the efficiency of Asmussen–Kroese-estimator and its applications to stop-loss transforms. In *6th Internat. Workshop on Rare Event Simulation* (Bamberg, October 2006), pp. 162–171.

[17] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58,** 13–30.

[18] Horn, R. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

[19] Jerrum, M. (2003). *Counting, Sampling and Integrating: Algorithms and Complexity*, Birkhäuser, Basel.

[20] Jerrum, M. and Sinclair, A. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM. J. Comput.* **22,** 1087–1116.

[21] Jerrum, M. and Sinclair, A. (1996). The Markov chain Monte Carlo method: an approach to approximate counting and integration, In *Approximation Algorithms for NP-hard Problems*, ed. D. Hochbaum, PWS, Boston, pp. 482–520.

[22] Jerrum, M., Sinclair, A. and Vigoda, E. (2004). A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. Assoc. Comput. Mach.* **51,** 671–697.

[23] Jerrum, M. R., Valiant, L. G. and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43,** 169–188.

[24] Lezaud, P. (1998). Chernoff-type bound for finite Markov chains. *Ann. Appl. Prob.* **8,** 849–867.

[25] Niemiro, W. (2008). Nonasymptotic bounds on the estimation error for regenerative MCMC algorithm under a drift condition. Submitted.

[26] Newman, M. E. J. and Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics.* Clarendon Press, New York.

[27] Pokarowski, P., Droste, K. and Kolinski, A. (2005). A minimal protein-like lattice model: an alpha-helix motif. *J. Chem. Phys.* **122,** 214915.

[28] Pokarowski, P., Kolinski, A. and Skolnick, J. (2003). A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. *Biophysical J.* **84,** 1518–1526.

[29] Sandman, W. (ed.) (2006). *Proceedings of the 6th International Workshop on Rare Event Simulation* (Bamberg, October 2006), University of Bamberg, Germany.

[30] Sinclair, A. J. and Jerrum, M. R. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inf. Comput.* **82,** 93–133.

[31] Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: foundations and new algorithm. Lecture Notes: Cours de Troisieme Cycle de la Physique en Suisse Romande (Lausanne, June 1989). Unpublished manuscript.

https://doi.org/10.1239/jap/1245676089 Published online by Cambridge University Press