Probability in the Engineering and Informational Sciences, 1, 1987, 33–46. Printed in the U.S.A.

ON THE MARKOV CHAIN SIMULATION METHOD FOR UNIFORM COMBINATORIAL DISTRIBUTIONS AND SIMULATED ANNEALING

DAVID ALDOUS
Department of Statistics
University of California
Berkeley, California

Uniform distributions on complicated combinatorial sets can be simulated by the Markov chain method. A condition is given for the simulations to be accurate in polynomial time. Similar analysis of the simulated annealing algorithm remains an open problem. The argument relies on a recent eigenvalue estimate of Alon [4]; the only new mathematical ingredient is a careful analysis of how the accuracy of sample averages of a Markov chain is related to the second-largest eigenvalue.

1. INTRODUCTION

For an irreducible Markov chain (X_n) on a finite state space V with stationary distribution π, it is elementary that

$$H_K \equiv K^{-1}\sum_{n=1}^{K} h(X_n) \to \bar{h} \equiv \sum_{v \in V} \pi(v)h(v) \text{ a.s. as } K \to \infty \qquad (1.1)$$

for any functional h: $V \to R$ and any initial distribution X_0. This result is the basis of *the Markov chain simulation method*: if you want to simulate a dis-

Research supported by National Science Foundation Grant MCS84-03239.

 0269-9648/87 $5.00 + .00

tribution π on a set V and cannot see how to do so directly, then try to construct a Markov chain whose stationary distribution is π and simulate the Markov chain. The classic applications of this method, as discussed in the classic book of Hammersley and Handscomb [9], are to Ising models and other physics problems. More recently, variants of this idea have arisen in two combinatorial, computer science inspired, contexts described below. A fundamental theoretical issue is understanding how large must K be in Eq. (1.1) to ensure that H_K is a good estimator of $\bar{h}$. This is a hard question. We shall discuss what is known in the two contexts. Our main thesis is that the right way to study the issue is via the exponential time/polynomial time distinction familiar in computer science, rather than studying asymptotics in a fixed problem.

1.1. Simulating Uniform Combinatorial Distributions

Let V_N be a finite "combinatorial" set such as

$$\begin{array}{l}\text{(i) } \{0,1\}^N \\ \text{(ii) all permutations of } \{1,2,\ldots,N\} \\ \text{(iii) all trees on } N \text{ labeled vertices.}\end{array} \tag{1.2}$$

Let h: $V_N \to R$ be a functional. If we are interested in the "typical" values of h, it is natural to study the average value $\bar{h} = |V_N|^{-1} \sum_{v \in V_N} h(v)$. Knuth [12] contains many concrete examples arising in computer science. In special cases the exact or asymptotic value of $\bar{h}$ may be obtainable; and for N small an exhaustive pass through V_N may enable us to calculate $\bar{h}$ on a computer. But in general one must resort to simulation. If one can simulate a uniform random element U of V_N, then the obvious empirical average of repeated simulations

$$K^{-1} \sum_{n=1}^{K} h(U_n)$$

gives an estimator of $\bar{h}$ with error $O(K^{-1/2})$. For sufficiently simple sets V_N, such as examples in Eq. (1.2), it is indeed possible to simulate uniform random elements. For

$$\begin{array}{l}\text{(i) all permutations } \pi \text{ of } \{1,\ldots,N\} \text{ with a specified upper bound on } \Sigma i\pi(i), \\ \text{(ii) all } N \times N \text{ matrices with positive integer entries and specified row and column sums,} \\ \text{(iii) all trees which are subgraphs of a specific graph,}\end{array} \tag{1.3}$$

direct simulation is harder. But the Markov chain method is easy to implement. One defines some notion of a "minimal change" which takes one element v of V_N to another element v'. In the examples of Eq. (1.3) above, we could define the minimal changes as follows.

i. Given π, pick i_1, i_2 and let π' be the permutation

$$\pi'(i_1) = \pi(i_2), \qquad \pi'(i_2) = \pi(i_1), \qquad \pi'(i) = \pi(i) \text{ else,}$$

provided π' satisfies the bound.

ii. Given $A = (a_{i,j})$, pick i_1, i_2, j_1, j_2 and let A' be the matrix

$$a'_{i_1 j_1} = a_{i_1 j_1} + 1 \qquad a'_{i_1 j_2} = a_{i_1 j_2} - 1$$

$$a'_{i_2 j_1} = a_{i_2 j_1} - 1 \qquad a'_{i_2 j_2} = a_{i_2 j_2} + 1;\ a'_{ij} = a_{ij} \text{ else}$$

provided A' is nonnegative.

iii. Given a tree T, add an edge to produce "a tree with one circuit," then delete one edge from the circuit to produce a tree T'.

We require symmetry: if there is a minimal change taking v to v' then there is a minimal change taking v' to v. Given an arbitrary element v_0, we can define and simulate the process

$$v_0 = X_0, X_1, X_2, \ldots$$

where X_{n+1} is a random minimal change of X_n.

More formally, one specifies a graph $G = (V_N, \mathbb{E})$ with vertex-set V_N and edge-set $\mathbb{E}$, where (v, w) is an edge iff there is a minimal change taking v to w. Then (X_n) is the natural random walk on G, that is the Markov chain with transition matrix

$$\begin{aligned} P(v,w) &= 1/r_v \qquad \text{if } (v,w) \text{ is an edge} \\ &= 0 \qquad \text{if not,}\end{aligned} \tag{1.4}$$

where r_v is the degree of v. Assume G is connected. Then the chain is irreducible and it is well known (e.g., Ross [14]) that its stationary distribution is

$$\pi(v) = r_v/(2|\mathbb{E}|). \tag{1.5}$$

In particular, if the graph G is *regular* ($r_v \equiv r$ for all v) then the stationary distribution is uniform, so Eq. (1.1) gives

$$H_K \equiv K^{-1} \sum_{n=1}^{K} h(X_n) \to \bar{h} \equiv |V|^{-1} \sum_{v \in V} h(v). \tag{1.6}$$

In the nonregular case the limit in Eq. (1.6) is the weighted average $\Sigma\pi(v)h(v)$. But we can get an estimator of $\bar{h}$ by evaluating the degree $r(X_n)$ at each step of the simulation:

$$\hat{H}_K \equiv \frac{\sum_{n=1}^{K} h(X_n)/r(X_n)}{\sum_{n=1}^{K} 1/r(X_n)} \to \bar{h}. \tag{1.7}$$

For our theoretical analysis we will treat the regular case; the complications of the nonregular case do not affect our main arguments.

1.2. Combinatorial Optimization by Simulated Annealing

As above, let V_N be a combinatorial set and let h: $V_N \to R$ be a function. Suppose h has a unique minimum h_* attained at v_*; the problem is to find v_*. A general algorithm proposed by Kirkpatrick *et al.* [11], called *simulated annealing*, has attracted much interest. See Aragon *et al.* [5] for a good survey and references. As above, introduce a notion of "minimal change" and a graph structure. For a parameter $0 < c < \infty$ ("temperature") consider the distribution on V_N

$$\pi^c(v) = A_c r_v \exp(-h(v)/c) \tag{1.8}$$

where A_c is the normalization constant. This is the stationary distribution for the Markov transition matrix

$$\begin{aligned} P^c(v,w) &= r_v^{-1}\exp(-(h(w)-h(v))/c) \qquad \text{if } (v,w) \text{ is an edge and } h(w) > h(v); \\ &= r_v^{-1} \qquad \text{if } (v,w) \text{ is an edge and } h(w) \le h(v); \\ &= 0 \qquad \text{if } v \ne w \text{ and } (v,w) \text{ is not an edge;} \\ P^c(v,v) &= 1 - \sum_{w \ne v} P^c(v,w). \end{aligned} \tag{1.9}$$

It is easy to simulate the chain with transitions P^c; this is essentially Metropolis' method for estimating means of functionals under the distribution of Eq. (1.8) (Hammersley and Handscomb [9]). Now given (c_n) consider the nonhomogenous Markov chain where $X_0 = v_0$, arbitrary, and

$$P(X_{n+1} = w | X_n = v) = P^{c_n}(v,w). \tag{1.10}$$

Consider the facts

i. For $c_n \equiv c$, $P(X_n = v) \to \pi^c(v)$ as $n \to \infty$;
ii. As $c \to 0$, $\pi^c(v) \to 1_{(v=v_*)}$.

It is easy to deduce:

$$\text{If } c_n \to 0 \text{ sufficiently slowly then the chain } (X_n) \text{ defined by Eq. (1.10) satisfies } P(X_n = v) \to 1_{(v=v_*)} \text{ as } n \to \infty. \tag{1.11}$$

Thus given a *cooling schedule* (c_n),

$$X_K^* = \text{the } X_i \text{ which achieves } \min(h(X_1), \ldots, h(X_K)) \tag{1.12}$$

is a natural estimator of v_*.

Of course there are many practical questions about implementation of these algorithms which we shall not discuss. We want to focus on one issue: how large must K be in order that the estimators H_K in Eq. (1.6) and X_K^* in Eq. (1.12) be good? The reader may turn to Section 5 to see our answer for H_K.

2. NAIVE ASYMPTOTICS

By "naive asymptotics" we mean the study of what happens as the number K of steps of the simulation tends to infinity in a fixed problem. Results of Eqs. (1.6) and (1.11) say that the estimators H_K and X_K^* are eventually good, without any indication of how large K needs to be. It is fairly easy to improve these results to get ultimate rates of convergence:

$$E|H_K - \bar{h}|^2 \sim AK^{-1} \text{ as } K \to \infty \tag{2.1}$$

$$\text{if } c_n \ge A'/\log(n) \text{ then } P(X_K^* = v^*) \to 1 \text{ as } n \to \infty, \text{ for } (X_n) \text{ defined by Eq. (1.10),} \tag{2.2}$$

where A, A' are constants depending on the graph structure $(V, \mathbb{E})$ and on h. Indeed, to derive Eq. (2.1) we fix v_0, let T_m be the time of the mth visit to v_0, let $D_m = h(X_{T_m}) + h(X_{T_m+1}) + \ldots + h(X_{T_{m+1}-1})$, observe that (D_m) are i.i.d., quote the usual i.i.d. central limit theorem, and a simple time change argument yields a central limit theorem for H_K:

$$K^{1/2}(H_K - \bar{h}) \xrightarrow{D} \text{Normal } (0, \sigma^2) \text{ as } K \to \infty\text{: some } \sigma^2 \ge 0. \tag{2.3}$$

See e.g., Freedman [7], Chapter 3.

For Eq. (2.2) the details are not elementary, but the outline is. For fixed, very small, $c > 0$ the chain Eq. (1.9) is very unlikely to move so as to increase h, and therefore spends most of its time on the set V^* of local minima of h. The rate of escape from a particular local minimum v_1^* to some other is of order $\exp(-\Delta(v_1^*)/c)$, where $\Delta(v_1^*)$ is the smallest number such that there exists a path $v_1^* = w_0, w_1, w_2, \ldots, w_j = v_2^*$ to some other local minimum with $h(w_i) - h(v_1^*) \le \Delta$ along the path. To avoid being permanently trapped in a local minimum v_1^* which is not the global minimum, we want the cooling schedule (c_n) to satisfy

$$\sum_n \exp(-\Delta(v_1^*)/c_n) = \infty,$$

and this is how Eq. (2.2) arises. For the details see Hajek [8], Mitra *et al.* [13], Tsitsiklis [17].

But results Eqs. (2.1) and (2.2) have serious drawbacks. After all, we can solve the problems in $|V|$ steps by exhaustive search. So as a bare minimum we would like theoretical results saying that, under some circumstances, the simulations give good estimates in $o(|V|)$ steps. Results Eqs. (2.1) and (2.2) fail to do this. To the author it seems more sensible to pose the question differently. In the examples of Eqs. (1.2) and (1.3), we have sets V_N for which $|V_N|$ grows

exponentially fast with N. We are studying specific algorithms for finding $\bar{h}$, v^*, and the natural question is: under what conditions do the algorithms work in a polynomial (in N) number of steps? The arguments above are much too crude to be useful here. For the problem in Section 1.1 we shall give more sophisticated arguments which lead to reasonable conditions of Eq. (5.1) for polynomial time; the essential condition is that the graphs $(V_N, \mathbb{E}_N)$ be highly connected. Finding similar results for Section 1.2, simulated annealing, seems a much harder problem about which nothing is known.

3. NONASYMPTOTIC BOUNDS IN SIMULATION OF UNIFORM COMBINATORIAL DISTRIBUTIONS

The nonasymptotic bound on the error of H_K involves two ideas.

Idea 1. If τ is sufficiently large then

$$X_\tau \text{ has approximately the stationary distribution } \pi, \text{ independent of } X_0. \tag{3.1}$$

So the sequence $X_\tau, X_{2\tau}, \ldots, X_{M\tau}$ is approximately independent, and so the mean-square error of the average $M^{-1}(h(X_\tau) + \ldots + h(X_{M\tau}))$ is about $M^{-1}\mathrm{var}(h)$. More observations cannot hurt, so for $K = M\tau$ the mean-square error of H_K is at most $M^{-1}\mathrm{var}(h)$.

This leaves the problem of estimating how large τ has to be for Eq. (3.1) to hold.

Idea 2. Perron–Frobenius theory (Seneta [15]) says the ultimate rate of convergence of the distribution of X_n is exponential:

$$P(X_n = v) - \pi(v) \sim A_v \exp(-n/\tau_e) \text{ as } n \to \infty,$$

where τ_e has an eigenvalue interpretation. Now τ_e can in principle be estimated by analytic methods, and then $\tau = 0(\tau_e)$ should work for Eq. (3.1).

We now start to formalize these ideas. Being a random walk on a graph, the chain (X_n) is *reversible*:

$$\pi(v)P(v,w) = \pi(w)P(w,v) \qquad \text{all } (v,w). \tag{3.2}$$

The transition matrix P has real eigenvalues, of which the largest equals 1. Let $\lambda < 1$ be the second largest eigenvalue, and let

$$\tau_e = -1/\log(\lambda), \tag{3.3}$$

so that $\lambda^n = \exp(-n/\tau_e)$.

To formalize idea 2 we can quote a recent result of Alon [4], derived in the study of a different computer science problem (expanders). For the graph $G = (V, \mathbb{E})$ let $\epsilon > 0$ be the largest number such that

$$\text{for each } A \subset V \text{ with } |A| \le \frac{1}{2}|V|, \text{ the set } \{w \notin A: (w,v) \text{ is an edge for some } v \in A\} \text{ has cardinality} \ge \epsilon|A|. \tag{3.4}$$

Adapting Lemma 2.4 of Alon [4] of our setting gives

Proposition 3.1: If G is regular, degree r, then

$$\tau_e \le r(2 + 4\epsilon^{-2}). \tag{3.5}$$

In the nonregular case a similar bound involving $r^* = \max r_v$ can be obtained.

The next task, and the only novel (albeit easy) mathematics in this article, is to make clear the connection between τ_e and the error in sample averages, in order to formalize idea 1. This is carried out in the next section. The results depend only on the reversibility property, and so apply to chains more general than random walks on graphs.

We end this section with a digression. The parameter ϵ defined as in Eq. (3.4) is a measure of the "connectedness" of the graph G. It is not obvious how to estimate ϵ for a given graph. But for one special type of graph, there is a simple lower bound. Let $\mathbf{G}$ be a finite group; let H be a subset of $\mathbf{G}$ satisfying

$$H = H^{-1};\ H \text{ does not contain the identity;}$$

let G be the graph with vertices $\mathbf{G}$ and edges $\{(g, gh): g \in \mathbf{G}, h \in H\}$. Suppose G is connected: then its diameter is

$$\Delta = \min\{n: \text{for each } g \in \mathbf{G} \text{ there exist } h_1, \ldots, h_n \in H \text{ such that } g = h_1 h_2 \ldots h_n\}.$$

LEMMA 3.1: *For a graph of the special type described above,*

$$\epsilon \ge \frac{1}{2\Delta}. \tag{3.6}$$

Proposition 3.1 then yields

COROLLARY 3.1: *For a graph of the special type described above,*

$$\tau_e \le 17\Delta^2|H|. \tag{3.7}$$

Proof: Fix $A \subset \mathbf{G}$ with $|A| \le \frac{1}{2}|\mathbf{G}|$. Write $Ag = \{ag: a \in A\}$. Then $|\mathbf{G}|^{-1} \sum_{g \in \mathbf{G}} |A \cap Ag| = |A|^2/|\mathbf{G}|$, and so there exists some $\hat{g} \in G$ such that $|A \cap A\hat{g}| \le |A|^2/|\mathbf{G}|$. This implies

$$|A\hat{g} \backslash A| \ge |A| - |A|^2/|\mathbf{G}| \ge \frac{1}{2}|A|. \tag{a}$$

Now $g = h_1 h_2 \ldots h_d$ for some $d \le \Delta$, $h_i \in H$. Write $g_i = h_1 h_2 \ldots h_i$. Then

$$|Ag \backslash A| \le \sum_{i=1}^{d} |Ag_i \backslash Ag_{i-1}| = \sum_{i=1}^{d} |Ah_i \backslash A|. \tag{b}$$

By (a) and (b), there exists some $h_0 \in H$ such that $|Ah_0 \backslash A| \geq \frac{1}{2}|A|/\Delta$. The lemma follows by considering the set of edges (a, ah_0), $a \in A$.

4. SAMPLE AVERAGES FOR REVERSIBLE MARKOV CHAINS

Let $(X_n;\ n \geq 0)$ be an irreducible Markov chain on a finite set V, and suppose X is reversible, that is the stationary distribution π satisfies Eq. (3.2). Let h: $V \to R$ be an arbitrary function. Let h_1 and h_2 be the expectation and variance of h under the stationary distribution π:

$$h_1 = \Sigma \pi_i h(i); \qquad h_2 = \Sigma \pi_i (h(i) - h_1)^2.$$

Consider the partial sums and averages

$$S_N = \sum_{n=1}^{N} h(X_n); \qquad A_N = N^{-1} S_N.$$

We need nonasymptotic bounds for the error in using A_N as an estimate of h_1. Let λ and τ_e be as defined by Eq. (3.3), and suppose $\tau_e \geq 1$.

Proposition 4.1: Suppose X_0 has the stationary distribution π. Then

i. $EA_N = h_1$;
ii. $\operatorname{var}(A_N) \leq \alpha(\tau_e/N)h_2$, where $\alpha(x) = 2x(1 + x\exp(-1/x))$. **(4.1)**

Remark: Note $\alpha(x) \approx 2x$ for small x. Proposition 4.1 has a nice intuitive interpretation, as follows. Suppose $N = K(2\tau_e)$. Then $\operatorname{var}(A_N)$ is essentially bounded by $K^{-1/2}h_2$, which is exactly the variance of the average of K independent observations. Thus Proposition 4.1 says: for the purpose of estimating a mean $Eh(X)$, observing a stationary reversible Markov chain continuously is at least as good as getting independent observations at times $2\tau_e$ apart. No analogous results hold for general nonreversible chains (see Aldous [2]).

PROOF: Assertion (i) is obvious. For (ii), we may suppose $h_1 = 0$. Let

$$M_{i,j} = \pi_i^{1/2} P_{i,j} \pi_j^{-1/2} \tag{4.2}$$

be the symmetrization of P. Then

$$P_{i,j}^n = \pi_i^{-1/2} M_{i,j}^n \pi_j^{1/2}. \tag{4.3}$$

By diagonalizing M we can write

$$M_{i,j}^n = \sum_u \lambda_u^n d_{i,u} d_{j,u} \tag{4.4}$$

where D is orthonormal and $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \ldots \geq -1$ are the common eigenvalues of M and P. For the stationary chain (X_n),

$$Eh(X_0)h(X_n) = \sum_i \sum_j \pi_i h(i) P_{i,j}^n h(j).$$

A few lines of algebra put this into the form

$$Eh(X_0)h(X_n) = \sum_u \lambda_u^n v_u^2; \tag{4.5}$$

where $v_1 = h_1 = 0$, and where $\sum v_u^2 = h_2$. Now

$$\begin{aligned} \operatorname{var}(S_N) &= ES_N^2 \\ &= NEh^2(X_0) + \sum_{n=1}^{N-1} 2(N-n)Eh(X_0)h(X_n) \end{aligned}$$

by expanding the sum and using stationarity. From Eq. (4.5) we deduce that

$$\operatorname{var}(S_N) \leq C(N, \lambda_2)h_2, \text{ where}$$

$$C(N, \lambda_2) = \max_{-1 \leq \lambda \leq \lambda_2} \left\{ N + \sum_{n=1}^{N-1} 2(N-n)\lambda^n \right\}.$$

For $\lambda < 0$ the series is alternating and the right side is bounded by N. For $\lambda > 0$ the right side is increasing in λ, and hence the maximum is attained at λ_2. Then algebra gives

$$C(N, \lambda_2) = 2(1 - \lambda_2)^{-2}\{\lambda_2^{N+1} - (N+1)\lambda_2 + N\} - N.$$

Recalling that $\tau_e = (1 - \lambda_2)^{-1}$,

$$\begin{aligned} C(N, \lambda_2) &\leq 2\tau_e^2\{(1 - 1/\tau_e)^{N+1} + N(1 - \lambda_2) - \lambda_2\} \\ &\leq 2\tau_e^2\{\exp(-N/\tau_e) + N/\tau_e\}. \end{aligned}$$

Since $\operatorname{var}(A_N) = N^{-2}\operatorname{var}(S_N)$, we obtain assertion (ii).

To get similar results for chains with arbitrary initial distribution, it is necessary to consider averages of *delayed* sums (see Example 4.2 below). For N_0, $N_1 \geq 1$ let

$$A_{N_0, N_1}^* = N_1^{-1} \sum_{n=T+1}^{T+N_1} h(X_n) \tag{4.6}$$

where T has Poisson (N_0) distribution, independent of (X_n). Note $T = N_0 \pm 0(\sqrt{N_0})$; having a random rather than deterministic delay is technically convenient (see Remark at the end of this Section).

Proposition 4.2: For any initial distribution X_0,

$$E(A_{N_0, N_1}^* - h_1)^2 \leq \{1 + \pi_*^{-1}\exp(-N_0/\tau_e)\}\alpha(\tau_e/N_1)h_2, \tag{4.7}$$

for $\alpha(x)$ as in Proposition 4.1 and $\pi_* = \min \pi_i$.

Proof: Write

$$Q^N_{i,j} = EP^T_{i,j}, \text{ where } T \stackrel{D}{=} \text{Poisson } (N);$$
$$\rho(N) = \max_{i,j} Q^N_{i,j}/\pi_j;$$
$$b_j = E\left(\left(N_1^{-1}\sum_{n=1}^{N_1} H(X_n) - h_1\right)^2 | X_0 = j\right).$$

Then

$$E((A^*_{N_0,N_1} - h_1)^2 | X_0 = i) = \sum_j Q^{N_0}_{i,j} b_j \qquad \text{by conditioning on } X_T$$
$$\le \rho(N_0)\sum_j \pi_j b_j \qquad \text{by definition of } \rho(N)$$
$$\le \rho(N_0)\alpha(\tau_e/N_1)h_2 \qquad \text{by Proposition 4.1.}$$

Thus it suffices to prove

$$\rho(N) \le 1 + \pi_*^{-1}\exp(-N/\tau_e). \tag{4.8}$$

In the notation of Eqs. (4.3) and (4.4),

$$Q^N_{i,j} = EP^T_{i,j}$$
$$= \pi_i^{-1/2}\pi_j^{1/2} EM^T_{i,j} \qquad \text{by (4.3)}$$
$$= \pi_i^{-1/2}\pi_j^{1/2}\sum_u E\lambda_u^T d_{i,u}d_{j,u} \qquad \text{by (4.4)}$$
$$= \pi_i^{-1/2}\pi_j^{1/2}\sum_u d_{i,u}d_{j,u}\exp((\lambda_u - 1)N). \tag{4.9}$$

The contribution from the $u = 1$ term works out as π_j. For $u > 1$ we have $\lambda_u - 1 \le \lambda_2 - 1 = -1/\tau_e$, so

$$Q^N_{i,j} \le \pi_j + \pi_i^{-1/2}\pi_j^{1/2}\exp(-N/\tau_e)\sum_u |d_{i,u}d_{j,u}|.$$

But by orthonormality $\sum_u |d_{i,u}d_{j,u}| \le 1$, so

$$Q^N_{i,j}/\pi_j \le 1 + \pi_i^{-1/2}\pi_j^{-1/2}\exp(-N/\tau_e),$$

which gives (4.8).

Remarks: To make the bound in Proposition 4.2 small we need

$$N_0 \gg \tau_e \log(1/\pi_*)$$
$$N_1 \gg \tau_e.$$

Using statistical language, one might say that N_0 must be large enough to eliminate "bias" from the initial position X_0, and N_1 must be large enough to make the "sampling variance" small. It is perhaps counterintuitive that the requirement for N_0 is more stringent than for N_1; the following example illustrates the reason.

Example 4.1: Let $V = \{0,1\}^d$, regarded as the vertices of the d-dimensional cube $i = (i_1, \ldots, i_d)$, and let (X_n) be the natural random walk on V, i.e., on the graph with edges $(i,\hat{i})$ such that $\sum |i_r - \hat{i}_r| = 1$. Take d odd. Let

$$D = \left\{i \in V: \Sigma i_r < \frac{1}{2}d\right\} \tag{4.10}$$
$$h(i) = 1 \text{ for } i \in D; \qquad h(i) = -1 \text{ for } i \notin D.$$

So $h_1 = 0$, $h_2 = 1$. It can be shown that $\tau_e \approx d/2$ (see e.g., Aldous and Diaconis [3] for this and other properties of this special process). Suppose first that X_0 has the stationary (that is, uniform) distribution. Then Proposition 4.1 implies $\text{var}(A_N) \le (d/N)(1 + 0(d/N))$, and so $0(d)$ steps suffice to make A_N a good estimator of h_1. But now consider the case where $X_0 = 0 = (0, \ldots, 0)$. Then

$$h(X_n) = 1 \text{ for all } n < \zeta = \min\{n: X_n \notin D\}.$$

For large d it can be shown that $\zeta \approx \frac{1}{4}d\log(d)$, and that at time $\approx \frac{1}{4}d\log(d)$ the distribution of X is close to the stationary distribution. So in this example the "bias" of the initial position does indeed persist for a longer time than is required, once the bias is eliminated, for the subsequent sample averages to approach the true mean.

Example 4.2: Continuing with the previous process, consider now the function

$$h(i) = d^{-1/2}2^{d/2} \qquad \text{if } \Sigma i_r = 1$$
$$= 0 \qquad \text{otherwise.} \tag{4.11}$$

For large d we have $h_1 \approx 0$, $h_2 \approx 1$. For the process started at $X_0 = 0$ we have $h(X_1) = d^{-1/2}2^{d/2}$ certainly, and hence $A_N \ge N^{-1}d^{-1/2}2^{d/2}$ certainly. Thus N has to be, at best, almost $0(2^{d/2})$ before A_N can be a reasonable estimate of h_1. This example shows that for nondelayed averages one cannot hope to bound the error of A_N in terms of τ_e and h_2 for arbitrary initial positions.

Remark: Taking the delay to have Poisson distribution is tantamount to passing from the discrete-time to the continuous-time Markov chain. Indeed, $Q^N_{i,j}$ is the transition probability $P(X_N = j | X_0 = i)$ for the continuous-time chain, and Eq. (4.9) is the usual spectral expansion in continuous time. The point of the switch is to avoid discussion of negative eigenvalues $\lambda \approx -1$ of P which complicate convergence $P(X_n = j) \to \pi_j$ of the discrete-time chain but which are irrelevant for convergence of the continuous-time chain, and for the behavior of our sample averages.

5. POLYNOMIAL-TIME SIMULATION

Propositions 3.1 and 4.2 fit together to give Theorem 5.1 below. Let us first collect the hypotheses. Let $G = (V, \mathbb{E})$ be a connected regular graph of degree r. Let $0 < \epsilon \leq 1$ be defined by Eq. (3.3). For a function h: $V \to R$ let

$$\bar{h} = |V|^{-1} \Sigma h(v) \qquad h_2 = |V|^{-1} \Sigma (h(v) - \bar{h})^2.$$

Let (X_n) be the random walk Eq. (1.4) on G. Given integers K_0, K_1, take T to have Poisson (K_0) distribution independent of (X_n), and consider the delayed average

$$A = K_1^{-1} \sum_{n=T+1}^{T+K_1} h(X_n).$$

THEOREM 5.1: *If* $K_0 = K_1 = Lr(2 + 4\epsilon^{-2}) \log |V|$ *then* $E(A - \bar{h})^2 \leq \beta(L) h_2$ *for all* h *and all initial distributions* X_0; *where* $\beta(L)$ *depends only on* L, *and* $\beta(L) \to 0$ *as* $L \to \infty$.

In the context of simulating uniform combinatorial distributions, the point is that we have sets V_N for which $|V_N|$ typically grows exponentially fast in N. Theorem 5.1 implies the Markov chain simulation method gives arbitrarily accurate estimates in a polynomial (in N) number of steps, provided the graphs $G_N = (V_N, \mathbb{E}_N)$ are such that

i. $\log |V_N|$
ii. ϵ_N^{-1}
iii. r_N **(5.1)**

each grow only polynomially fast.

In examples like Eq. (1.4), conditions (i) and (iii) are immediate. The only issue is whether the graph is sufficiently "highly-connected" for (ii) to hold. It is unfortunately hard to bound ϵ in particular examples; in this sense, our results are unsatisfactory. From the theoretical viewpoint one definitely needs some conditions of the connectivity of G, otherwise (e.g., for the circuit graph, the discrete circle) $0(|V|^2)$ steps may be required instead of $0(\log |V|)$ steps.

6. MISCELLANEOUS REMARKS

i. Informally, knowing how to write an algorithm to generate a uniform random element of a set V in a definite number of steps is equivalent to knowing a formula for $|V|$; Jerrum *et al.* [10] give a formal treatment of this relation. Thus the Markov chain method is a natural candidate when no such formula is known.

The idea of using Markov chains in this particular context has been "in the air" for many years, though there seems no good reference. The author learned it from Andre Broder via Persi Diaconis.

ii. Random walks on graphs arise in other areas—as models in their own right, as analogies for electrical networks, in nonprobabilistic graph theory, as tools in interacting particle systems, in recreational mathematics. A survey will be given elsewhere.

iii. The random walk corresponds to the $c = \infty$ case of simulated annealing. For $c < \infty$ the random process depends on the function h as well as the graph structure, and so is complicated to study. Indeed, for $c < \infty$ it is not apparent how to formalize either of the ideas of Section 3.

iv. In place of τ_e, one can formulate the notion of "the time τ until X_τ has approximately the stationary distribution" in terms of total variation distance. See Aldous [1] for definitions. For this τ the analysis of Section 4 becomes trivial: delayed sample averages based on $K\tau$ steps have $O(K^{-1/2})$ error. The difficulty is that there are no known general results like Proposition 3.1 which give bounds on this τ. Aldous and Diaconis [3] describe techniques for bounding τ in special cases. Broder [6] treats a nice example, and has unpublished work on other examples mentioned in Section 1.

v. One can also consider simulating uniform distributions on continuous sets, such as the group of orthogonal matrices (see Sloane [16]). Similarly, simulated annealing can be employed on continuous problems (see Vanderbilt and Louis [18]).

References

1. Aldous, D. (1983). Random walks on finite groups and rapidly mixing Markov chains. In: *Séminaire de Probabilités* XVII. Springer Lecture Notes in Mathematics 986, p. 243–297.
2. Aldous, D. (1986). Implications of relaxation times for 1-dimensional Markov chains. Technical Report No. 65, Statistics Dept., Berkeley, Calif.
3. Aldous, D. and Diaconis, P. (1986). Strong uniform times and finite random walks. Technical Report #59, Statistics Department, Berkeley, California. (To appear in *Adv. Appl. Math.*)
4. Alon, N. (1986). Eigenvalues and expanders. *Combinatorica*, to appear.
5. Aragon, C. R., Johnson, D. S., McGeogh, L. A., and Schevon, C. (1986). Optimization by simulated annealing: an experimental evaluation. Preprint, Computer Science Dept., Berkeley, Calif.
6. Broder, A. Z. (1986). How hard is it to marry at random? (On the approximation of the permanent). *STOC 86*, Assoc. Computing Machinery, Baltimore, Md., p. 50–58.
7. Freedman, D. (1971). *Markov Chains.* Holden-Day.
8. Hajek, B. (1986). Cooling schedules for optimal annealing. Coordinated Sci. Lab., University of Illinois, Urbana, Ill.
9. Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* Methuen, London.
10. Jerrum, M., Valiant, L., and Vazirani, V. (1985). Random generation of combinatorial structures from a uniform distribution. Department of Computer Science, University of Edinburgh.

11. Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220: 671–680.
12. Knuth, D. (1968,69,73). *The Art of Computer Programming*, Vols. 1–3. Addison-Wesley, Reading, Mass.
13. Mitra, D., Romeo, F., and Sangiovanni-Vincentelli, A. (1986). Convergence and finite-time behavior of simulated annealing. *Adv. Appl. Prob.*, 18: 747–771.
14. Ross, S.M. (1980). *Introduction to Probability Models*. Academic Press, New York.
15. Seneta, E. (1980). *Non-negative Matrices and Markov Chains*. Springer-Verlag, New York.
16. Sloane, N. J. A. (1983). Encrypting by random rotrations. In: *Cryptography*, ed. T. Beth. Springer Lecture Notes in Computer Science 149.
17. Tsitsiklis, J. N. (1986). Markov chains with rare transitions and simulated annealing. Preprint, Electrical Eng. Dept., M.I.T.
18. Vanderbilt, D. and Louis, S. G. (1984). A Monte Carlo simulated annealing approach to optimization over continuous variables. *J. Computational Phys.* 56: 259–271.

Probability in the Engineering and Informational Sciences, 1, 1987, 47–59. Printed in the U.S.A.

GIBBS' MEASURES ON COMBINATORIAL OBJECTS AND THE CENTRAL LIMIT THEOREM FOR AN EXPONENTIAL FAMILY OF RANDOM TREES

J. MICHAEL STEELE
Program in Engineering Statistics
Princeton University
Princeton, New Jersey

A model for random trees is given which provides an embedding of the uniform model into an exponential family whose natural parameter is the expected number of leaves. The model is proved to be analytically and computationally tractable. In particular, a central limit theorem (CLT) for the number of leaves of a random tree is given which extends and sharpens Rényi's CLT for the uniform model. The method used is general and is shown to provide tractable exponential families for a variety of combinatorial objects.

1. INTRODUCTION

The theory of random combinatorial objects has shown a considerable devo tion to the uniform distribution. This is a natural development, and a gre number of interesting and useful results have been produced. Still, for the pu pose of modeling and for technical applications, it seems valuable to be abl to go beyond uniformity and the elementary extensions amounting to cond tional uniformity.

Research supported in part by National Science Foundation Grant No. DMS-8414069.

 0269-9648/87 $5.00 + .00