The Annals of Applied Probability
1992, Vol. 2, No. 3, 610–628

OPTIMAL SPECTRAL STRUCTURE OF REVERSIBLE STOCHASTIC MATRICES, MONTE CARLO METHODS AND THE SIMULATION OF MARKOV RANDOM FIELDS[1]

BY ARNOLDO FRIGESSI, CHII-RUEY HWANG[2] AND LAURENT YOUNES

Istituto per le Applicazioni del Calcolo, CNR; Institute of Mathematics, Academia Sinica; and Istituto per le Applicazioni del Calcolo, CNR

In this paper we prove an optimal spectral structure theorem for stochastic matrices reversible with respect to a fixed probability measure π on a finite set and devise a new simulation algorithm for Markov random fields. We compute the minimum value for the second largest eigenvalue of all such matrices and characterize the class of matrices for which this minimum is attained. In fact, they share a common right eigenvector that can be written in terms of π. Furthermore, by iterating this procedure, we obtain a unique matrix which is minimal with respect to the lexicographic order of the eigenvalues. We give a probabilistic interpretation of the corresponding eigenvectors. Our results allow us to devise a dynamic Monte Carlo scheme which has an optimal worst-case performance. Regarding the simulation of lattice-based Gibbs distributions, we design a modified Gibbs sampler, whose performance is better in terms of both weak convergence at low temperatures and asymptotic variance of time averages at all temperatures.

1. Introduction. In many situations the evaluation of expectations in large sample spaces cannot be computed analytically or by enumerating all possible states. One has to resort to approximations. This is when the Monte Carlo methods come in. For example in image analysis and image synthesis, a direct sampling from the very big set of all possible images is impossible; instead, dynamic Monte Carlo methods are often used. This leads us to investigate the structure of the optimal Monte Carlo method with respect to certain criteria. The answer to this problem constitutes our first main theorem. This result gives us an idea on how to design a new simulation algorithm for Markov random fields which is our second main result. Since the results we obtained are quite general, we will describe our study in a general setting.

Let S be a finite set and consider the class of all stochastic matrices P on S, which are reversible with respect to a fixed positive probability measure π on S. Typically Monte Carlo methods are based on Markov chains with π as their equilibrium distribution.

Received August 1990; revised July 1991.

[1] This paper was written while C.-R. Hwang was visiting IAC. Partially supported by CNR Special Project "Statistics for random processes."

[2] Partially supported by NSC 80-0208-M001-13.

AMS 1980 *subject classifications*. Primary 60J10, 15A51; secondary 65C05, 60G60.

Key words and phrases. Reversible stochastic matrices, space average estimation, Monte Carlo methods, simulation of Markov random fields, Gibbs sampler, image analysis.

The set of all possible images mentioned above is a good example for such an S. For black and white images with 64×64 pixels, the set S has cardinality $2^{64\times 64}$.

We compute the minimum value for the *second largest eigenvalue* of all such matrices P and characterize the class of matrices for which this minimum is attained. In fact they share a common right eigenvector that can be given in terms of the invariant distribution. Furthermore, by iterating this procedure, we obtain a unique matrix which achieves the minimum of the lexicographic order of the eigenvalues. We give a probabilistic interpretation of the corresponding eigenvectors.

Our results allow us to devise a dynamic Monte Carlo scheme for the approximation of space averages relative to π, which has an optimal worst case performance in terms of *asymptotic variance of time averages*.

Regarding the simulation of lattice-based Gibbs distributions, we design a modified Gibbs sampler dynamics, whose performance is better in terms of both weak convergence at low temperatures and asymptotic variance of time averages at all temperatures.

Monte Carlo methods are developed in order to approximate

$$\langle f\rangle = \sum_{s\in S} f(s)\pi(s), \tag{1}$$

where f is a real valued function on S. A standard way to do this is by sampling independently $X(0), X(1), \dots, X(n-1)$ from π and evaluating

$$\frac{1}{n}\sum_{k=0}^{n-1} f[X(k)].$$

Think of S as a huge set (all possible images) so that $\sum_{s\in S} f(s)\pi(s)$ is numerically impossible and we need to resort to the ergodicity to estimate (1).

More generally, this independent sampling can be replaced by a Markov chain on S with transition probability P whose equilibrium measure is π. One may call these methods *dynamic Monte Carlo* schemes as compared to the *static* ones based on independent sampling [Sokal (1989)]. Among all these methods it is convenient to choose those for which

$$nE_{\mu_0}\left[\left\{\frac{1}{n}\sum_{k=0}^{n-1} f[X(k)] - \langle f\rangle\right\}^2\right]$$

is small as $n \uparrow \infty$. This expectation is computed according to the Markov chain with initial distribution μ_0 on S and transition probability P.

Throughout this paper we will only consider matrices $P = \{p_{st}\}$ that are reversible relative to π, that is,

$$\pi(s)p_{s,t} = \pi(t)p_{t,s}$$

for all s, t in S, $s \neq t$. It is well known [e.g., Keilson (1979)], that for irreducible transition matrices P,

$$\lim_{n\to\infty} nE_{\mu_0}\left[\left\{\frac{1}{n}\sum_{k=0}^{n-1} f[X(k)] - \langle f\rangle\right\}^2\right] = \lim_{n\to\infty} n\,\mathrm{var}_{\mu_0}\left(\frac{1}{n}\sum_{k=0}^{n-1} f[X(k)]\right)$$

exists and does not depend on μ_0. We denote this limit by $v(f, P, \pi)$. For the computation of this *asymptotic variance*, we therefore can and will assume that $\mu_0 = \pi$.

Contrary to classical variance reduction techniques [e.g., Rubinstein (1983)], it is our purpose to find P, given π, which reduces $v(f, P, \pi)$ *without exploiting any prior knowledge on the specific* f. In this sense we are using a minimax criterion: Our results will be optimal (within a certain class) from the worst-case point of view w.r.t. any possible f.

Here is the outline of the paper. In the next section we state a general result in terms of the second largest eigenvalue. We present a theorem that might be regarded as a Frobenius-type theorem. It allows us to describe the structure of the matrices P that have the smallest possible second largest eigenvalue and are reversible w.r.t. some specified measure π. This value turns out to be negative.

We then describe the relation between asymptotic variance of time averages and eigenvalues. This permits us to identify the class of matrices P which minimize $v(f, P, \pi)$ for all possible observables f. It is important to underline that given a specific f, one can sometimes exploit its properties to construct another P with lower asymptotic variance.

The procedure can be further repeated and we describe how to build a matrix with the lowest third eigenvalue, given that the second is already the smallest possible and so on. This matrix gives rise to a Monte Carlo method with smaller asymptotic variance compared with static independent sampling, since all eigenvalues (except one) are negative.

Another important quantity is the bias of the time averages given by $(1/n)\Sigma_{k=0}^{n-1}E_{\mu_0}[f(X(k))] - \langle f\rangle = (1/n)\Sigma_{k=0}^{n-1}(\mu_0 P^k f - \langle f\rangle)$. Note that the mean square errors of time averages and their variances coincide up to an order $1/n$ for large times n. Their difference, which is equal to the square of the bias, is of order $1/n^2$. The bias itself is of order $1/n$. We consider one particular case: f is the indicator function of the set $\{s\}$ and the Markov chain starts at $s \in S$. The bias is then given by $(1/n)(\Sigma_{k=0}^{n-1}p_{ss}^{(k)} - \pi(s))$, where $p_{ss}^{(k)}$ is the sth diagonal entry of the power matrix P^k. At the end of Section 2, we shall find the minimum of the potential $\Sigma_k(p_{ss}^{(k)} - \pi(s))$, over all possible reversible matrices w.r.t. π.

Originally we were led to look at these problems by our interest in fast computational methods for image reconstruction, inspired by the fundamental work of Grenander and D. and S. Geman on the Markov random field approach to imaging. Here one has to *sample* from a finite lattice based Gibbs distribution π, in order to synthesize or restore an image [Grenander (1984), Geman and Geman (1984), Geman (1990)]. The prohibitive dimension of the state space S makes any direct sampling from π impossible in practice and one has to rely on *local updating methods*. A simple version reads: At every step only one site i of the current configuration x is updated, according to a sparse transition matrix P_i. Standard assumptions on the site-visitation schedule and on reversibility of P_i w.r.t. π easily imply that the corresponding Markov chain will converge weakly to π.

Here, too, we want to compare different updating rules in terms of their rate of weak convergence, that is, in terms of *second largest eigenvalues in absolute value* of their transition matrices. This value does not always come from a positive eigenvalue [see, e.g., Frigessi, Hwang, Sheu and di Stefano (1990)]. Hence, for sampling, one has to consider also the smallest eigenvalue, even for a first order estimate of the convergence rate.

In Section 3 we use our results from Section 2, to propose a modified version of the local Gibbs sampler updating scheme of Geman and Geman (1984). Our algorithm reduces the asymptotic variance and at low temperature converges weakly faster than the Gibbs sampler, yet it requires only negligible extra computational effort.

It is important to distinguish between the two objectives: (i) fast weak convergence to equilibrium; and (ii) small asymptotic variance of time averages. A stochastic matrix P for which weak convergence is fast may have large asymptotic variance and conversely, so it is better to use different dynamics for the two different purposes. Note, however, that our method in Section 3 improves on the Gibbs sampler (at low temperature) in both senses.

2. Main results.

Without loss of generality and for simplicity of exposition, assume that the states in S are ordered according to increasing π-measure:

$$0 < \pi(1) \le \pi(2) \le \cdots \le \pi(N), \quad \text{where } N = |S|.$$

Denote by $\mathbf{1}$ the constant function equal to 1 [in vector notation, $\mathbf{1} = (1,\dots,1)^T$] and by δ_k the Dirac function at state k [in vector notation $\delta_k = (0,\dots,1,\dots,0)^T$] with one as the only nonzero element at the kth coordinate.

THEOREM 1. (a) *The second largest eigenvalue of any stochastic matrix P, reversible w.r.t. π, is greater than or equal to*

$$-\frac{\pi(1)}{1-\pi(1)}.$$

For all matrices whose second largest eigenvalue attains this lower bound, the corresponding eigenvector is

$$e_2 = \delta_1 - \langle\delta_1\rangle\mathbf{1} = (1-\pi(1), -\pi(1),\dots,-\pi(1))^T.$$

Furthermore, their first column has a zero as first entry and all other elements *are equal to*

$$\frac{\pi(1)}{1-\pi(1)}.$$

(b) *The construction above can be iterated to finally obtain a matrix with the following properties*: (i) *all the elements along the diagonal are zero, except possibly the last one*; (ii) *its eigenvalues are* $1 = \lambda_1 > 0 > \lambda_2 \geq \cdots \geq \lambda_N$ *and satisfy the property that* λ_{i+1} *attains the smallest possible value among all matrices* (*reversible w.r.t.* π) *that already possess the eigenvalues* $1, \lambda_2, \dots, \lambda_i$; (iii) *its columns have constant entries under the diagonal, which are, respectively,* $-\lambda_2, \dots, -\lambda_N$; (iv) *its eigenvectors are*

$$e_1 = \mathbf{1}$$

and

$$e_{k+1} = \delta_k - \langle \delta_k | 1, 2, \dots, k-1 \rangle_\pi, \qquad k = 1, \dots, N-1,$$

where $\langle f | 1, 2, \dots, k-1 \rangle_\pi$ *is the conditional expectation of f given the σ-algebra generated by the sets* $\{1\}, \dots, \{k-1\}$ *under the probability* π; *in vector notation*,

$$e_{k+1} = \Biggl(\underbrace{0, \dots, 0}_{k-1 \text{ terms}}, 1 - \frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}, -\frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}, \dots, -\frac{\pi(k)}{\pi(k) + \cdots + \pi(N)} \Biggr)^T.$$

(c) *Moreover, this matrix is the unique one which satisfies the previous condition* (ii).

To prove the theorem, we need a lemma. Denote

$$\langle f, g \rangle_\pi = \sum_{s \in S} f(s) g(s) \pi(s).$$

LEMMA 1. *For every stochastic matrix P, reversible w.r.t. π,*

$$\langle Pe_2, e_2 \rangle_\pi \geq -\pi(1)^2,$$

where $e_2 = \delta_1 - \langle \delta_1 \rangle \mathbf{1}$ *with equality if and only if* $p_{11} = 0$.

PROOF. From the reversibility and stochasticity of P,

$$\begin{aligned} Pe_2 &= P[1 - \pi(1), -\pi(1), \dots, -\pi(1)]^T \\ &= \left[p_{11} - \pi(1), \frac{\pi(1)}{\pi(2)} p_{12} - \pi(1), \dots, \frac{\pi(1)}{\pi(N)} p_{1N} - \pi(1) \right]^T. \end{aligned}$$

Furthermore,

$$\begin{aligned}\langle Pe_2, e_2\rangle_\pi &= (p_{11} - \pi(1))\pi(1)(1 - \pi(1)) \\ &\qquad - \pi(1)\sum_{k=2}^{N}\left(\frac{\pi(1)}{\pi(k)}p_{1k} - \pi(1)\right)\pi(k) \\ &= (p_{11} - \pi(1))\pi(1) \geq -\pi(1)^2.\end{aligned}$$

This proves the lemma. □

In the preceding proof, we never used the fact that $\pi(1)$ was the lowest probability. The statement of the lemma is therefore true for all $s \in S$:

$$\langle P[\delta_s - \langle\delta_s\rangle\mathbf{1}], [\delta_s - \langle\delta_s\rangle\mathbf{1}]\rangle_\pi \geq -\pi(s)^2,$$

with equality iff $p_{ss} = 0$.

Proof of Theorem 1. The second largest eigenvalue $\lambda_2(P)$ of any stochastic matrix P, reversible w.r.t. to π, satisfies

$$\lambda_2(P) = \sup_{f \neq 0, \langle f\rangle = 0} \frac{\langle Pf, f\rangle_\pi}{\langle f, f\rangle_\pi}. \tag{2}$$

Using Lemma 1 and $\langle e_2, e_2\rangle_\pi = \pi(1)(1 - \pi(1))$, it follows that

$$\lambda_2(P) \geq -\frac{\pi(1)}{1 - \pi(1)}.$$

This establishes the first part of (a). Moreover, the supremum in (2) is attained by the eigenvectors of P corresponding to the eigenvalue $\lambda_2(P)$, since P is self-adjoint w.r.t. $\langle \cdot \rangle_\pi$. Assume $\lambda_2(P) = -(\pi(1)/(1 - \pi(1))$ so that from Lemma 1, $p_{11} = 0$. This implies $e_2 = \delta_1 - \langle\delta_1\rangle\mathbf{1}$ is a corresponding eigenvector.

By computing the equations involved in $Pe_2 = -(\pi(1)/(1 - \pi(1))e_2$, we can check that, for all $j > 1$, $p_{1j} = (\pi(j)/(1 - \pi(1)))$. This implies by detailed balance that the entries on the first column under the diagonal are constant and equal to $\pi(1)/(1 - \pi(1))$. We have therefore proved part (a). Also, if a matrix P has eigenvalue $\lambda_2(P) = -(\pi(1)/(1 - \pi(1)))$, then its first row and column are fixed, which is a first step towards the uniqueness stated in part (c). Once we complete the proof of part (b), we will have constructed a matrix whose eigenvalues are all smaller than or equal to $-(\pi(1)/(1 - \pi(1)))$; anticipating this, we see that this lower bound of the second eigenvalue is attained.

We now turn to part (b). Notice that, according to the above discussion, if a matrix has $-\pi(1)/(1 - \pi(1))$ as the second largest eigenvalue, then it must have the form

$$P = \begin{bmatrix} 0 & \frac{\pi(2)}{1-\pi(1)} & \cdots & \frac{\pi(N)}{1-\pi(1)} \\ \frac{\pi(1)}{1-\pi(1)} & & & \\ \vdots & & P_2 & \\ \frac{\pi(1)}{1-\pi(1)} & & & \end{bmatrix}. \tag{3}$$

Now, P_2 is in detailed balance with the row vector $(\pi(2), \dots, \pi(N))$, and has constant row sums. Therefore, we can repeat the same arguments as above; the fact that the row sum in this case is not 1 does not make much difference. Moreover, one can easily check that finding an $(N-1)$-dimensional eigenvector of P_2, orthogonal to $(1, \dots, 1)^T \in \mathbb{R}^{N-1}$, is equivalent to finding an N-dimensional eigenvector for P, orthogonal to e_1 and e_2: If $(x_2, \dots, x_N)$ are the coordinates of the $(N-1)$-dimensional eigenvector, then $(0, x_2, \dots, x_N)$ is an eigenvector of P and every N-dimensional vector orthogonal to e_1 and e_2 must have 0 as its first entry.

Therefore, the second eigenvalue of P_2 is the third eigenvalue of P and the first and second rows and columns of any $N \times N$ matrix, having these eigenvalues, are fixed. Clearly, this can be pushed further, thus obtaining at each step a new eigenvalue that fixes a new row and column in the matrix. In order to check that the corresponding eigenvectors are given by

$$e_k = \left[0 \cdots 0, 1 - \frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}, \right.$$
$$\left. -\frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}, \dots, -\frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}\right]^T,$$

one has to remark that the matrix P_k that plays the role of P_2 at step k is in detailed balance with

$$\left(\frac{\pi(k)}{\pi(k) + \cdots + \pi(N)}, \dots, \frac{\pi(N)}{\pi(k) + \cdots + \pi(N)}\right).$$

Concerning the probabilistic interpretation of e_{k+1} as $\delta_k - \langle \delta_k | 1, 2, \dots, k-1 \rangle_\pi$, one can check it by elementary computations with conditional expectations: just notice, for example, that functions that are measurable with respect to $\sigma(\{1\}, \dots, \{k-1\})$ are linear combinations of $1, \delta_1, \dots, \delta_{k-1}$.

Finally, the values of $\lambda_2, \dots, \lambda_N$ can be computed. They are given in Remark 4. From this, one easily verifies that $\lambda_2 \geq \cdots \geq \lambda_N$.

REMARK 1. Concerning part (a) of Theorem 1, it may be surprising that, by only fixing the second eigenvalue to be optimal, at the same time one of the corresponding principal axes is also automatically fixed (there may, of course, exist other eigenvectors with the same eigenvalue; see Remark 3 for an example). This is why we interpret our result as a structural result on stochastic matrices that reminds us of a Frobenius-type theorem.

REMARK 2. A final nonzero element on the diagonal may remain in the matrix of part (b). In fact this happens if and only if $\pi(N-1) \neq \pi(N)$.

REMARK 3. As a simple example, consider the uniform case, $\pi(1) = \cdots = \pi(N) = 1/N$. At the end of the above program, the matrix P will have the form

$$(4) \qquad \begin{bmatrix} 0 & \dfrac{1}{N-1} & \cdots & \dfrac{1}{N-1} \\ \dfrac{1}{N-1} & 0 & \cdots & \dfrac{1}{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{1}{N-1} & \dfrac{1}{N-1} & \cdots & 0 \end{bmatrix}$$

with eigenvalues $\lambda_2 = \lambda_3 = \cdots = \lambda_N = -(1/(N-1))$.

REMARK 4. On the basis of the iterative construction in the proof of Theorem 1, it is not difficult to write down the values of $\lambda_2, \dots, \lambda_N$. They are

$$\lambda_{k+1} = -\frac{\pi(k)}{\pi(k+1) + \cdots + \pi(N)} \cdot \prod_{l=1}^{k-1} \left(1 - \frac{\pi(l)}{\pi(l+1) + \cdots + \pi(N)}\right).$$

If $\pi(k-1) = \pi(k)$, then $\lambda_k = \lambda_{k+1}$.

REMARK 5. It is not surprising that $\delta_1 - \langle \delta_1 \rangle \mathbf{1}$ realizes the above bound. In other words, this means that the most difficult quantity to estimate by time averages is the probability of the least likely state, which is consistent with intuition.

The next part of this section is devoted to the statistical interpretation of Theorem 1. For completeness, we state a well-known result concerning the asymptotic variance of time averages already mentioned in the Introduction. □

THEOREM 2. *Let P be an irreducible and a reversible* (*w.r.t.* π) *stochastic matrix. Let $X(0), X(1), \dots$ be a Markov chain on S with transition matrix P*

and $f\colon S \to \mathbb{R}$. For any initial distribution,

$$v(f,P,\pi) = \lim_{n\to\infty} n \operatorname{var}\left(\frac{1}{n}\sum_{k=0}^{n-1} f[X(k)]\right)$$
$$= \langle (I-P)^{-1}(I+P)(f-\langle f\rangle\mathbf{1}), f-\langle f\rangle\mathbf{1}\rangle_\pi.$$

PROOF. See, for example, Keilson (1979) or Caracciolo, Pelisetto and Sokal (1989) and Peskun (1973). □

Easy computations show that the above limit is equal to

$$\sum_{k=2}^{N} \frac{1+\lambda_k}{1-\lambda_k}\langle f, \bar{e}_k\rangle_\pi^2, \tag{5}$$

where $P\bar{e}_k = \lambda_k \bar{e}_k$, $\|\bar{e}_k\|_\pi = 1$. This formula links the asymptotic variance $v(f,P,\pi)$ to the eigenvalues and eigenvectors of P.

Clearly, in order to make the asymptotic variance small, it is useful to use dynamics P with possibly *all* negative and small eigenvalues (except the largest one which is 1). In this sense, one should interpret the common remark that "negative eigenvalues help" in terms of asymptotic variance and not of weak convergence, for which the second largest eigenvalue in absolute value matters.

The next result states precisely what we anticipated in the Introduction concerning variance reduction.

COROLLARY 1. *Let P be a stochastic matrix, reversible w.r.t. π. Let $v(P,\pi)$ be the maximum asymptotic variance of $(1/n)\sum_{k=0}^{n-1} f[X(k)]$ for norm 1 functions f. Then*

$$v(P,\pi) \geq 1 - 2\pi(1),$$

and any matrix that realizes this equality must have the properties given in Theorem 1(a) and hence be of the form (3).

PROOF. By expression (5), $v(P,\pi) = (1+\lambda_2)/(1-\lambda_2)$; using $\lambda_2 = -(\pi(1)/(1-\pi(1)))$, that is, the best possible value, we get the desired result. □

REMARK 6. Any matrix of type (3) will lead to a worst-case asymptotic variance $v(M,\pi) = 1 - 2\pi(1)$. Whenever no specific knowledge of f is available, this is optimal. We like to stress that for *specific* functions f, Markov chains with smaller asymptotic variance can be constructed.

There is a general and very easy way to find a P which for a specific given f does better than, say, the static P_0 (with all rows equal to π). For this, assume that $\langle f\rangle = 0$; denote by A the orthogonal projection operator in the direction f whose entries are $a_{ij} = \pi(j) f_i f_j$. Then, the row sums of A are all 0 and A is reversible w.r.t. π. Consider now the matrix $P_\varepsilon = P_0 - \varepsilon A$, with $\varepsilon > 0$. This

P_ε will be a stochastic matrix, reversible w.r.t. π if all its entries are nonnegative, that is, for small enough ε. Now f is an eigenvector of P_ε, with eigenvalue $-\varepsilon\|f\|_\pi^2$ and

$$v(f, P_\varepsilon, \pi) = \frac{1 - \varepsilon\|f\|_\pi^2}{1 + \varepsilon\|f\|_\pi^2}\|f\|_\pi^2.$$

The largest ε that can be chosen is such that, for all i, j, $\varepsilon f_i f_j \leq 1$. This gives $\varepsilon = 1/\max_i f_i^2$. Thus we obtain a P_ε such that

$$v(f, P_\varepsilon, \pi) = \frac{\max f_i^2 - \|f\|_\pi^2}{\max f_i^2 + \|f\|_\pi^2}\|f\|_\pi^2.$$

Two essential remarks should be added at this point. First, P_ε is of no practical use, because one has to know $\langle f \rangle$ in order to compute it (we assumed $\langle f \rangle = 0$). Moreover, in general, this P_ε is not the best possible for f. However, it can be better than the matrix we described in Theorem 1, where no information on f was used.

For illustration, consider the following example: Take $N = 3$, $\pi = (\frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ and $f = (0, 3, -2)^T$. The matrix given in Theorem 1 is

$$M = \begin{bmatrix} 0 & \frac{2}{5} & \frac{3}{5} \\ \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{1}{5} & \frac{8}{15} & \frac{4}{15} \end{bmatrix}$$

and one has $Mf = -\frac{8}{15}f$, so that $v(f, M, \pi) = \frac{7}{23}\|f\|_\pi^2$. The ε given above is $\frac{1}{9}$ and the matrix P_ε is such that $v(f, P_\varepsilon, \pi) = \frac{2}{7}\|f\|_\pi^2 < v(f, M, \pi)$.

Remark 7. From a computational point of view, running the chain based on (3) or even the one based on the matrix described in Theorem 1(b) may be of no greater effort than some alternative methods. For example, consider the uniform case (4) again as compared to

$$\begin{bmatrix} \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & & \vdots \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}.$$

Notice that some off-line computing is necessary, in general, in order to compute the underdiagonal column values. Of course, one may decide to modify only part of the matrix. We just claim that, whenever a static Monte Carlo method is useful and computationally feasible, it is possible to adopt a simple dynamic Monte Carlo method, thus reducing the worst-case asymptotic variance at no significant computational extra cost. The interest lies in its general purpose, blind (w.r.t. special f's) behavior.

Remark 8. Variance reduction techniques are often proposed and applied in Monte Carlo methods. They generally turn out to be of a static type [Rubinstein (1983)]. One of the most interesting approaches is called importance sampling. The idea is to express $\langle f\rangle_\pi$ as $\sum_s(f(s)\pi(s)/g(s))g(s)$, where $g(s)$ is a suitably chosen probability measure, and to approximate this by

$$\frac{1}{n}\sum_{k=0}^{n-1}\frac{f(X(k))\pi(X(k))}{g(X(k))},$$

where $(X(0),\dots,X(n-1))$ is a sample from $g(\cdot)$. The optimal g which minimizes the asymptotic variance is given by

$$g(s)=\frac{|f(s)|\pi(s)}{\sum_t|f(t)|\pi(t)}.$$

Of course sampling from $g(s)$ is as hard as computing $\langle f\rangle$, hence one will design approximations $\hat{g}(s)$ that resemble $g(s)$ but are computationally feasible. Our approach here is quite different and difficult to compare in general, since $\hat{g}(s)$ has to be specified [for the optimal $g(s)$, the variance is zero].

We conclude this section with a theorem that gives the minimum value of the bias for a particular choice of f.

Proposition 1. *Let $\mathscr{R}_\pi$ be the set of all irreducible aperiodic stochastic matrices, in detailed balance w.r.t. π. If $P\in\mathscr{R}_\pi$, denote by $P_{kk}^{(n)}$ the kth diagonal term of P^n. Then*

$$\inf_{P\in\mathscr{R}_\pi}\sum_{n\geq 0}\left(P_{kk}^{(n)}-\pi(k)\right)=\begin{cases}(1-\pi(k))^2, & \textit{if } \pi(k)\leq\frac{1}{2},\\ \pi(k)(1-\pi(k)), & \textit{if } \pi(k)\geq\frac{1}{2}.\end{cases}$$

Proof. For any π-reversible stochastic matrix Q, we have that $\langle Qf_k, f_k\rangle=\pi(k)(q_{kk}-\pi(k))$, where $f_k=\delta_k-\pi(k)$ (see Lemma 1).

Therefore, if $P\in\mathscr{R}_\pi$,

$$\begin{aligned}\sum_{n\geq 0}\left(P_{kk}^{(n)}-\pi(k)\right)&=\frac{1}{\pi(k)}\left\langle\sum_{n\geq 0}P^nf_k, f_k\right\rangle_\pi\\&=\frac{1}{\pi(k)}\langle(I-P)^{-1}f_k, f_k\rangle_\pi.\end{aligned}$$

Let $\varphi(P)=\langle(I-P)^{-1}f_k, f_k\rangle_\pi$; the problem is thus to find the infinum of φ over $\mathscr{R}_\pi$. As $\varphi(P)$ is also defined if P is not aperiodic, let $\overline{\mathscr{R}}_\pi$ be the set of irreducible and π-reversible stochastic matrices. This is a convex set. Let φ be defined on $\overline{\mathscr{R}}_\pi$ and let

$$\psi(P)=\langle Pf_k, f_k\rangle_\pi.$$

It is not difficult to check that φ is a convex function on $\overline{\mathscr{R}}_\pi$ (it is twice differentiable); for $P \in \overline{\mathscr{R}}_\pi$, denote by $C(P)$ the set of matrices H for which $P + \lambda H \in \overline{\mathscr{R}}_\pi$, for small enough $\lambda > 0$. The directional derivative of φ at P in the direction H is equal to

$$\varphi'(P)(H) = \langle H(I-P)^{-1} f_k, (I-P)^{-1} f_k \rangle_\pi \quad \text{for all } H \in C(P).$$

The directional derivative of ψ is of course $\psi'(P)(H) = \langle Hf_k, f_k \rangle_\pi$.

A necessary and sufficient condition for P to be a minimum of φ (respectively, of ψ) over $\overline{\mathscr{R}}_\pi$ is that for all $H \in C(P)$,

$$\varphi'(P)(H) \geq 0, \qquad (\psi'(P)(H) \geq 0).$$

We shall prove the following fact: There exists a $P \in \overline{\mathscr{R}}_\pi$, such that ψ attains its minimum at P and $Pf_k = \lambda f_k$. This will imply that for $H \in C(P)$, $\varphi'(P)(H) = (1/(1-\lambda)^2)\psi'(P)(H)$, thus $\varphi'(P)(H) \geq 0$ and P will be a minimum also for φ.

In order to obtain the minimum of ψ, first remark that $\psi(P) = (p_{kk} - \pi(k))\pi(k)$, so that we have to find the least possible value of p_{kk} over $\overline{\mathscr{R}}_\pi$. This value, of course, has to be nonnegative, but it is not always possible to reach 0, as is shown by the following trivial estimate:

$$p_{kk} = 1 - \sum_{l \neq k} p_{kl} = 1 - \sum_{l \neq k} \frac{\pi(l)}{\pi(k)} p_{lk} \geq 1 - \frac{1}{\pi(k)} \sum_{l \neq k} \pi(k) = \frac{2\pi(k) - 1}{\pi(k)}.$$

This lower bound is positive as soon as $\pi(k) > \frac{1}{2}$. Notice that $\pi(k) > \frac{1}{2}$ is only possible for the most probable state, namely $k = N$, since we have ordered S according to increasing probabilities.

In fact, one has

$$\inf_{\overline{\mathscr{R}}_\pi} p_{kk} = \max\left(0, \frac{2\pi(k) - 1}{\pi(k)}\right). \tag{6}$$

We shall prove (6) by using the trick given in Remark 6. This will provide at the same time a P for which ψ is minimum and f_k is an eigenvector, so that we will also get the minimum of φ.

Denote by $f_{k1}, \dots, f_{kN}$ the coordinates of f_k: One has $f_{kk} = 1 - \pi(k)$ and $f_{kl} = -\pi(k)$ for $l \neq k$. Consider the projection matrix A over f_k; its entries are $a_{ij} = \pi(j) f_{ki} f_{kj}$. Denote by P_ε the matrix $P_0 - \varepsilon A$, where P_0 has all rows equal to $\pi(1), \dots, \pi(N)$. We know that P_ε is a stochastic matrix for $\varepsilon \leq (1/\max_j f_{kj}^2)$. Its eigenvalues are 1, $-\varepsilon \| f \|_\pi^2$ and 0 with multiplicity $N - 2$, so that $P \in \overline{\mathscr{R}}_\pi$. Denote by P_k the matrix P_ε for

$$\varepsilon = \frac{1}{\max_j f_{kj}^2} = \frac{1}{\max\left(\pi(k)^2, (1 - \pi(k))\right)^2}.$$

A simple computation implies that the kth diagonal term of P_k is precisely that one given in (6).

Moreover,

$$\frac{\|f_k\|_\pi^2}{\max f_{kj}^2} = \min\left(\frac{\pi(k)}{1-\pi(k)}, \frac{1-\pi(k)}{\pi(k)}\right),$$

so that, if $\pi(k) \neq \frac{1}{2}$, $P_k \in \mathscr{R}_\pi$. If $\pi(k) = \frac{1}{2}$, one can approximate P_k by P_ε for $\varepsilon < (1/\max f_{kj}^2)$. This proves that

$$\inf_{P \in \mathscr{R}_\pi} \varphi(P) = \pi(k) \inf_{P \in \mathscr{R}_\pi} \sum_{n \geq 0} (P_{kk}^n - \pi(k)) = \varphi(P_k).$$

Now,

$$\varphi(P_k) = \frac{1}{1 + \|f_k\|_\pi^2/\max f_{kj}^2}\|f_k\|_\pi^2$$

gives the result of the proposition.

Remark 9. The methods in the previous proof can be used for some other functions in order to obtain lower bounds on $\sum_n \langle P^n f, f\rangle_\pi$. Unfortunately, this does not work for all f, because it is not true in general that $\langle Pf, f\rangle_\pi$ attains its minimum for some P such that $Pf = \lambda f$.

However, here is an example of f for which it works. Let $f = e_k$, one of the eigenvectors of Theorem 1. If $P \in \mathscr{R}_\pi$, $\langle Pe_k, e_k\rangle_\pi$ only depends on p_{ij} for $i, j \geq k$ (the $k-1$ first coordinates of e_k are 0). Let M be the matrix constructed in part (b) of Theorem 1; it is not hard to see that it is constructed so that $\langle Me_k, e_k\rangle_\pi$ is minimum. As e_k is an eigenvector of M, $\langle \sum_n P^n e_k, e_k\rangle_\pi$ also attains its minimum at M.

Remark 10. Consider the asymptotic variance of the time average for a fixed f (with $\langle f\rangle = 0$), which is given by $\langle (I-P)^{-1}(I+P)f, f\rangle_\pi$. For this function of P (which is also convex), whenever the methods of Proposition 1 are applicable (see Remark 9), they also provide a P for which the asymptotic variance is minimum. For example, the matrix P_k obtained in the preceding proof is the best in $\mathscr{R}_\pi$ for estimating $\pi(k)$. Its serious drawback is that it already depends on $\pi(k)$.

3. Markov random field simulation.

We now specialize our discussion to the simulation of *Markov random fields* on a finite graph. For this, we need some new notation.

Let D be a finite lattice (say, $D \subset \mathbb{Z}^2$). Let S_0 be a finite set (e.g., the colors) and $S = (S_0)^D$ (the set of all possible pictures). An element x of S will thus be a $|D|$-tuple, $x = (x_\sigma)_{\sigma \in D}$, $x_\sigma \in S_0$. We consider the measure π on S given by $\pi(x) = \exp(-(U(x)/T))/Z_T$, where Z_T is a normalizing constant.

Typically, S is a very large set, so large that it is impossible to sample directly from π; therefore, in this case, the methods presented in the previous section are impracticable. In most applications the function U is chosen to have nice local properties:

$$U(x) = -\sum_{c} u_c(x),$$

where c runs over some family of (small) subsets of D (cliques) and $u_c(x)$ only depends on x_σ, $\sigma \in c$. U is often referred to as the energy function and u_c as the potential function, in statistical mechanics.

This implies that the conditional probabilities $\pi_\sigma(x_\sigma | x_\rho, \rho \neq \sigma)$ at site σ given the rest of the configuration are easy to compute; they depend only on a small number of coordinates x_ρ. Equivalently, we can say that it is easy to compute the differences $U(x) - U(x')$, when x and x' differ only at one site.

Hence, Monte Carlo methods for simulating Markov random fields are made up of elementary steps, each of them consisting of the random updating of the current configuration *at only one site*. Such methods are presented and studied widely in the literature; see, for example, Geman and Geman (1984), Geman (1990), Sokal (1989). Frigessi, Hwang, Sheu and di Stefano (1990) compare some widely used algorithms in terms of rigorous rate of weak convergence; see also Gelfand and Smith (1990) for interesting applications of the Gibbs sampler outside image analysis.

One of the most popular algorithms is the *Gibbs sampler*, where each elementary step acts as follows: A site σ is chosen to be updated; the new value at this site is drawn at random according to the conditional distribution at site σ, $\pi_\sigma(\cdot | x_\rho, \rho \neq \sigma)$. In this paper we assume that the site to be updated is chosen uniformly at random. With this assumption, we can consider the following construction: Each elementary (local) update corresponds to a probability kernel P_σ; every elementary step of the algorithm consists of choosing a site σ at random and applying P_σ. This step therefore corresponds to the matrix

$$P = \frac{1}{|D|} \sum_{\sigma \in D} P_\sigma.$$

Our intent is to use the ideas of the preceding section to improve on P, *but only acting at the local level*. In other words, we try to improve the Gibbs sampler, staying within the class of *feasible* Monte Carlo algorithms. In fact, we will show how it is possible not only to reduce the asymptotic variance of time averages, but also to accelerate the weak convergence for small T. For this last point, we use methods from Frigessi, Hwang, Sheu and di Stefano (1990) for weak convergence comparisons.

Let us now focus on the updating at a site σ. The matrix $P_\sigma = (p_\sigma(x, y))$ is defined as

$$\begin{aligned} p_\sigma(x,y) &= \delta_{N_\sigma(x)}(y)\pi\big(y_\sigma | y_\rho, \rho \neq \sigma\big) \\ &= \delta_{N_\sigma(x)}(y) \frac{\exp[-U(y)/T]}{\sum_{z \in N_\sigma(x)} \exp[-U(z)/T]}, \end{aligned}$$

where $N_\sigma(x) = \{y \colon y_\rho = x_\rho, \rho \neq \sigma\}$.

Order all possible configurations as

$$S = \{x_1, \dots, x_l, x_{l+1}, \dots, x_{2l}, \dots, x_{(L-1)l+1}, \dots, x_{Ll}\},$$

with

$$l = |S_0|, \qquad L = l^{|D|-1} \text{ and } \{x_{pl+1}, \dots, x_{(p+1)l}\} = N_\sigma(x_{pl+1}).$$

With this ordering, P_σ is a block diagonal matrix, each block corresponding to a single $N_\sigma(x)$, that is, to a fixed configuration outside σ. We are going to modify the entries of these blocks as compared to those of the Gibbs sampler.

Indeed, each of these blocks has rank one, because every row is given by $\pi_\sigma(\cdot|x_\rho, \rho \neq \sigma)$. Here, we retrieve the structure of the static matrix of the previous section; we will use, on this matrix, the construction of Theorem 1(a).

For this, fix an $x \in S$ and the set $N_\sigma(x)$ (in fact, only the coordinates x_ρ, $\rho \neq \sigma$ are relevant). Let U_m be the maximal energy of configurations within $N_\sigma(x)$. Call H the set of those y for which $U(y) = U_m$ and $h = |H|$. The elements of H are therefore those for which $\pi(y_\sigma|x_\rho, \rho \neq \sigma)$ attains its minimum.

We iterate the construction of Theorem 1 over the set of all least likely states in each block. This gives:

1. For all $y \in H$, replace $p(y, y)$ with 0.
2. If $z \notin H$, then replace $p(z, z)$ by

$$p(z,z) - h\left(\frac{p(y,y)}{1 - p(y,y)} - p(y,y)\right).$$

(One can check that this is a positive number.)

3. If $y \in H$ and $z \neq y$, replace $p(y, z)$ by

$$\frac{p(y,z)}{1 - p(y,y)}$$

and $p(z, y)$ by

$$\frac{p(y,y)}{1 - p(y,y)}.$$

4. If $z \notin H$, $y \notin H$ and $z \neq y$, then leave $p(y, z)$ and $p(z, y)$ unchanged.

Steps (1) and (3) follow from Remark 4 of Theorem 1.

At this point, it is clear that ordering the configurations in order to make the above transformations is not necessary. This is fortunate, because this ordering would depend on the particular site $\sigma \in D$.

After performing the changes 1–4 for each different $N_\sigma(x)$, we obtain a new local updating transition probability $\tilde{P}_\sigma$. Thus, we get a new simulation method for Markov random fields, given by

$$\tilde{P} = \frac{1}{|D|} \sum_{\sigma \in D} \tilde{P}_\sigma.$$

It is easy to check that $\tilde{P}$ is still reversible w.r.t. π. In order to state a theorem on some properties of $\tilde{P}$, we need some definitions concerning the local minima of U as given in Frigessi, Hwang, Sheu and di Stefano (1990).

For $x \neq y$, a path from x to y is a sequence $x = x_0, x_1, \dots, x_k = y$ such that $x_{i+1} \in N_{\sigma_i}(x_i)$ for some σ_i and $x_i \neq x_j$ for $i \neq j$.

A configuration $x \in S$ is a local minimum of U if for any y with $U(y) < U(x)$ and any path from x to y, there exists z, belonging to this path such that $U(z) > U(x)$. Two local minima are equivalent if they can be linked by a path of constant energy. An equivalence class is called a *bottom*.

We can now state the theorem. Recall that the Gibbs sampler is based on P and that we denoted by $v(f, P, \pi)$ the asymptotic variance of the time averages of f for the Markov chain with transition P.

THEOREM 3. *The matrix $\tilde{P}$ constructed above is such that*:

(a) *For any nonconstant f*,

$$v(f, \tilde{P}, \pi) < v(f, P, \pi).$$

As a consequence, the second eigenvalue of $\tilde{P}$, $\lambda_2(\tilde{P})$, is strictly smaller than the second eigenvalue of P.

(b) *If T (the temperature) is small enough and U has at least two bottoms, then the second eigenvalue in absolute value of $\tilde{P}$, $\rho_2(\tilde{P})$, is strictly smaller than the second eigenvalue in absolute value of P.*

The theorem therefore states that this modified Gibbs sampler is always better than the original one in terms of both asymptotic variance of time averages and weak convergence, at least at low temperatures.

Notice that it is not necessarily true that $\rho_2 = |\lambda_2|$ (the largest absolute value might come from a negative eigenvalue); however, this equality holds for $\tilde{P}$ at least at low temperature.

PROOF OF THEOREM 3. According to Caracciolo, Pelisetto and Sokal (1989), it suffices to show that $\langle Pf, f \rangle_\pi > \langle \tilde{P}f, f \rangle_\pi$ for all f such that $\langle f \rangle = 0$ and $f \neq 0$.

We first prove the nonstrict inequality in (a) for each P_σ and $\tilde{P}_\sigma$, and because of the block-diagonal structure of these matrices, it suffices to prove it for each elementary block. With such blocks, things become simpler: Denoting the two corresponding blocks of P_σ and $\tilde{P}_\sigma$ by G and $\tilde{G}$, we want to show that $\langle G\tilde{f}, \tilde{f} \rangle_\pi \geq \langle \tilde{G}\tilde{f}, \tilde{f} \rangle_\pi$ for any l-dimensional vector $\tilde{f}$, with equality only if $\tilde{f}$ is constant. Indeed, we have [see, e.g., Caracciolo, Pelisetto and Sokal (1989)]

$$\langle (I - G)\tilde{f}, \tilde{f} \rangle_\pi = \sum_{i<j} \pi(j) g_{ij} \left(\tilde{f}_i - \tilde{f}_j \right)^2,$$

and a similar formula for $\tilde{G}$. The fact that the off diagonal terms of $\tilde{G}$ are not smaller than the off diagonal terms of G implies the desired inequality. In fact, notice that we constructed $\tilde{G}$ by strictly increasing the off diagonal terms of the first row; hence we get that equality is possible only if $\tilde{f}_i = \tilde{f}_1$ for all i. We therefore have proved that, for all f, $\langle Pf, f\rangle_\pi \geq \langle \tilde{P}f, f\rangle_\pi$.

If $f \neq 0$, then f is not constant because $\langle f \rangle = 0$. Thus there exist a site σ and two configurations x and y only differing at site σ, such that $f(x) \neq f(y)$, because any two configurations can be connected by a path. Therefore, if we take the blocks G and $\tilde{G}$ corresponding to $N_\sigma(x)$, then $\langle G\tilde{f}, \tilde{f}\rangle_\pi > \langle \tilde{G}\tilde{f}, \tilde{f}\rangle_\pi$, where $\tilde{f}$ is the restriction of f to $N_\sigma(x)$. This implies that $\langle Pf, f\rangle_\pi > \langle \tilde{P}f, f\rangle_\pi$.

We turn to (b). Since we know that $P - \tilde{P}$ is positive, we only have to show that the second eigenvalue in absolute value of $\tilde{P}$ comes from the second largest positive one, namely $\lambda_2(\tilde{P})$. This will be shown by proving that $\lambda_2(\tilde{P})$ tends to 1 if T tends to 0, while the other eigenvalues are bounded away from -1.

For this, the methods developed in Frigessi, Hwang, Sheu and di Stefano [(1990), Theorem 5] can be applied with minor modifications. We shall therefore only briefly sketch the proof of (b).

The assumption we made on the bottom of U implies that, when $T = 0$, the Markov chain associated with $\tilde{P}$ has more than one ergodic class; in fact, each bottom is an ergodic class. This implies that the second eigenvalue of $\tilde{P}$ tends to 1 when T tends to 0.

It remains to show that, if a state x is not a local minimum of U, then it is transient, and there is a positive probability at $T = 0$ that the chain starting from x reaches a local minimum: This will imply that there cannot be eigenvalues equal to -1 at $T = 0$. We refer once again to Frigessi, Hwang, Sheu and di Stefano (1990).

Let us now turn to a practical description of the algorithm associated to $\tilde{P}$. In some simple cases (e.g., Ising model), it is possible to tabulate the transition probabilities associated with $\tilde{P}$. However, this is in general not realistic, and the transformations that lead to $\tilde{P}$ must be made *on line* during the simulation. For convenience, we describe the algorithm in terms of energy.

Assume that the current configuration is x. (i) Take at random a site $\sigma \in D$. (ii) Within $N_\sigma(x)$, which contains $l = |S_0|$ elements, look for the maximum U_m of $U(\cdot)$ and compute

$$Z_x = \sum_{y \in N_\sigma(x)} \exp\left[-\frac{U(y)}{T}\right].$$

Let $h = |\{y \in N_\sigma(x)\colon U(y) = U_m\}|$ and $\alpha = \exp[-U_m/T]$. (iii) If $U(x) = U_m$, then choose $y \in N_\sigma(x) \setminus \{x\}$ at random with probability

$$\frac{\exp[-U(y)/T]}{Z_x - \alpha};$$

else if $U(x) < U_m$, choose $y \in N_\sigma(x)$ at random with probability

$$\begin{cases} \dfrac{\alpha}{Z_x - \alpha}, & \text{if } U(y) = U_m, \\[2ex] \dfrac{\exp(-U(x)/T)}{Z_x} - h\left[\dfrac{\alpha}{Z_x - \alpha} - \dfrac{\alpha}{Z_x}\right], & \text{if } y = x, \\[2ex] \dfrac{\exp(-U(y)/T)}{Z_x}, & \text{if } y \neq x \text{ and } U(y) < U_m. \end{cases}$$

The extra computational cost, when compared to the Gibbs sampler, is small. The number of exponentials to be computed is the same; moreover, depending on $U(x)$, up to four extra floating point additions and multiplications may be needed. Finally, we have to compute the local maximum of U. In practice, for the usual Gibbs sampler, it is recommended to find the minimum of the energy $U_{\min}$ before the local updating and to update by using the relative energy $U(x) - U_{\min}$, in order to avoid numerical overflows resulting from the computation of $\exp(A)$ for some large A. In our case, again, computing the maximum at the same time adds a very small extra cost.

REMARK 11. In the definition of the modified Gibbs sampler, we did not complete all the procedure described in part (b) of Theorem 1, for two reasons: The first is that we are not sure that, from any configuration which is not a local minimum of the energy, this new Markov chain would reach a bottom with positive probability. Our proof cannot be extended to show that this new stochastic matrix has no eigenvalue -1 at temperature 0. The second reason is practical: Each new step of the procedure of Theorem 1(b) would involve more and more computational cost. We therefore restrict ourselves to only one step, which is easy to implement.

REMARK 12. We do not know how to compare the algorithm of Theorem 3 with another commonly used Monte Carlo method for Markov random field simulation, namely the Metropolis algorithm [Geman (1990)].

Acknowledgments. The authors would like to express their gratitude to the referees for their helpful comments and suggestions. C.-R. Hwang and L. Younes would like to thank IAC for their hospitality during their stay.

REFERENCES

CARACCIOLO, S., PELISETTO, A. and SOKAL, A. (1989). Non-local Monte Carlo algorithms for self-avoiding walks with fixed endpoints. Dept. Physics, New York Univ.

FRIGESSI, A., HWANG, C.-R., SHEU, S.-J. and DI STEFANO, P. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. Roy. Statist. Soc. Ser. B*. To appear.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 395–409.

GEMAN, D. (1990). Random fields and inverse problems in imaging. *Ecole d'Eté de Probabilités de Saint-Flour XVIII. Lecture Notes in Math.* **1427**. Springer, New York.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI* **6** 721–741.

GRENANDER, U. (1984). Tutorial in pattern analysis. Lecture Notes, Div. Appl. Math., Brown Univ.

KEILSON, J. (1979). *Markov Chain Models–Rarity and Exponentiality*. Springer, New York.

PESKUN, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60** 607–612.

RUBINSTEIN, R. (1983). *Simulation and the Monte Carlo Method*. Wiley, New York.

SOKAL, A. (1989). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lecture Notes, Lausanne.

ISTITUTO PER LE APPLICAZIONI
DEL CALCOLO, CNR
VIALE DEL POLICLINICO 137
00161 ROMA
ITALY

INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
TAIPEI, TAIWAN, 11529
REPUBLIC OF CHINA

UNIVERSITÉ PARIS SUD
BÂTIMENT 425
DEPARTMENT DE MATHÉMATIQUES
F-91405 ORSAY
CEDEX
FRANCE