A NOTE ON THE SWENDSEN-WANG ALGORITHM AND ORDERED COLOURS

Peter J. Green

University of Bristol,
Department of Mathematics,
Bristol, BS8 1TW, UK.

ABSTRACT

The approach of Swendsen and Wang to simulating from the Potts model is extended to more general Markov random fields, obtained by dropping symmetry over sites and colours, and allowing general pairwise-difference interactions. The key idea is a more general notion of bond variable, taking integer values, leading to a more complicated cluster structure. Convergence is proved under quite general conditions. Only informal speculations are made about implementing the resulting algorithms.

Preliminary version: comments welcome.

1. Introduction

There is currently much interest in the use of dynamic Monte Carlo methods such as the Metropolis and Gibbs samplers to evaluate distributions and expectations in complex stochastic systems. Such methods are being used in applications as diverse as statistical image analysis (Geman and Geman, 1984), general multi-parameter Bayesian inference (Gelfand and Smith, 1990), spatial epidemiology (Besag, York and Mollié, 1991), and pedigrees in genetics (Sheehan and Thomas, 1991).

Given a (high-dimensional) joint distribution of interest, $p(\mathbf{x})$ say, the general approach is to evaluate the required quantities by averaging statistics calculated from samples drawn from the distribution; this simulation is not performed directly, however, but by constructing a Markov chain whose limiting distribution is $p(\mathbf{x})$. There is considerable freedom of choice in constructing such chains (Hastings, 1970, Green and Han, 1991).

An obvious concern in using such methods is whether convergence is sufficiently fast, and the samples sufficiently independent, for the approach to be practically viable. Loosely speaking, when there is a high degree of dependence between the variables in $p(\mathbf{x})$, it can take a very long time to visit all corners of the state space adequately if one uses the conventional strategy of changing one, or a few, of the components of $\mathbf{x}$ at each transition.

This difficulty is well-documented in the computational physics literature, where the phenomenon of chief concern is known as critical slowing-down (see Sokal, 1990). Near a critical point, the autocorrelation time of such Markov chains typically becomes infinite. The finite-lattice version of this phenomenon is that convergence to equilibrium becomes increasingly slow, and successive samples increasingly dependent, as

interaction parameters increase.

One way of overcoming this difficulty is to devise a Markov chain with the same limiting distribution $p(\mathbf{x})$, but with transitions of completely different character, changing many variables at once. Such chains offer the prospect of generating sample paths that explore the sample space much more quickly, and thus avoiding critical slowing-down. But such chains are also difficult to construct, or rather, it is difficult to construct chains that do not involve rejection sampling with extremely low acceptance rates and which are therefore impractical.

A notable success in this direction, however, is the algorithm of Swendsen and Wang (1987) for sampling from the Ising model, and other Potts models. This algorithm exploits a duality between the Potts model and the "random-cluster" model, a variant of bond percolation, to give dramatically better performance in terms of convergence speed and autocorrelation time as the critical point is approached. Although there have been studies in the statistical literature assessing the performance of the Swendsen-Wang algorithm against that of conventional "single-flip" methods (e.g. Kirkland (1989)), the algorithm has had no impact on practical applications because of the lack of relevance of the symmetrical interaction structure in the Potts model to the applied science contexts listed above.

In this note, we discuss an extension of the idea behind the Swendsen-Wang algorithm to address models with more general pairwise interactions, depending monotonically on differences between neighbouring variables. At the same time, we provide a cleaner derivation of the standard Swendsen-Wang construction than is usually given, and prove convergence under very general conditions. Thus we considerably extend the relevance and applicability of the method. We do not, however, discuss any practical applications in this note, and neither do we make more than speculative remarks about implementing the algorithms.

2. Bond variables

Suppose that $\mathbf{x}$ is a random vector, with components x_i indexed by $i \in S = \{1,2,\ldots,n\}$, a finite set of sites or pixels. Each x_i takes values in a finite set of colours $C = \{0,1,2,\ldots,L-1\}$, and we denote the whole sample space C^S by Ω. We are interested in distributions on Ω with probability functions of the form

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i \in S} a_i(x_i) \prod_{i \sim j} b_{ij}(|x_i - x_j|), \tag{1}$$

where $\{a_i()\}$ and $\{b_{ij}()\}$ are prescribed finite, positive, functions, each $b_{ij}()$ is *non-increasing*, and Z is the appropriate normalising constant. Thus there are at most pairwise interactions, and these involve only differences between the variables. The notation $i \sim j$ indicates that i and j are neighbours in the conditional independence graph of the variables $\{x_i\}$: that is, we will take $i \sim j$ if and only if $b_{ij}()$ is not identically constant. The second product in (1) is thus a product over all pairs of neighbours.

From the positivity of $a_i()$ and $b_{ij}()$, it is immediate that $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \Omega$. Note that model (1) might more usually be written in the form

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i \in S} \alpha_i(x_i) - \sum_{i \sim j} \beta_{ij}(|x_i - x_j|)\right),$$

but the multiplicative form is more useful in what follows.

Our approach to manipulating such distributions is to introduce auxiliary random variables, which we term *bond variables*. These are denoted by e_{ij}, taking values in C, and defined for each i,j such that $i \sim j$. The vector with components (e_{ij}) is denoted by **e**.

The joint distribution of (**x** , **e**) will be defined through the conditional distribution of **e** given **x**. Let

$$c_{ij}(L-1) = b_{ij}(L-1)$$

and

$$c_{ij}(e) = b_{ij}(e) - b_{ij}(e+1)$$

for $e = 0,1,2,\ldots L-2$. Then $\{c_{ij}()\}$ are non-negative functions (since b_{ij} is non-increasing), and

$$b_{ij}(d) = \sum_{e=0}^{L-1} c_{ij}(e)\, I[d \le e],$$

for all $d = 0,1,\ldots,L-1$, where $I[\ ]$ is the indicator function. Given **x**, the $\{e_{ij}\}$ are defined to be conditionally independent, with

$$\begin{aligned} p(e_{ij}|\mathbf{x}) &= 0 \quad \text{if } e_{ij} < |x_i - x_j| \\ &= \frac{c_{ij}(e_{ij})}{b_{ij}(|x_i - x_j|)} \quad \text{if } e_{ij} \ge |x_i - x_j|. \end{aligned} \tag{2}$$

It is clear from our assumptions that this is a proper distribution. Combining (1) and (2), we have the joint distribution for **x** and **e**:

$$p(\mathbf{x},\mathbf{e}) = \frac{1}{Z} \prod_{i \in S} a_i(x_i) \prod_{i \sim j} c_{ij}(e_{ij})\, I[|x_i - x_j| \le e_{ij}]. \tag{3}$$

Note that, by construction, the normalising constant Z is the same in (1) and (3). This joint distribution is *not* everywhere positive: for any **e**, $p(\mathbf{x},\mathbf{e})$ is only positive for

$$\mathbf{x} \in \Omega(\mathbf{e}) = \{\mathbf{x} \in \Omega : |x_i - x_j| \le e_{ij} \text{ for all } i \sim j\}. \tag{4}$$

The conditional distribution of **x** given **e** can now be read off from (3): clearly

$$p(\mathbf{x}|\mathbf{e}) \propto \prod_{i \in S} a_i(x_i) \quad \text{for } \mathbf{x} \in \Omega(\mathbf{e}). \tag{5}$$

The normalising constant would not be easy to obtain. The bond variables **e** induce an equivalence relation on S, where i and j are equivalent if $i = j$ or there is a path $i = i_0 \sim i_1 \sim i_2 \sim \ldots \sim i_k = j$ between pairs of neighbours, with each $e_{i_{r-1}i_r} < L-1$, strictly. We call the equivalence classes *clusters*: it is evident that components of **x** indexed by sites in different clusters are conditionally independent given **e**.

3. Special cases

The Potts model (Potts, 1952) is the special case of model (1) that arises when $a_i(x_i) \equiv 1$ for all $x_i \in C$, all $i \in S$, and

$$b_{ij}(|x_i - x_j|) = \begin{cases} 1 & x_i = x_j \\ = e^{-\beta} & x_i \neq x_j . \end{cases} \tag{6}$$

The probability $p(\mathbf{x})$ is then just the exponential of the negative of the number of unlike-coloured neighbours, suitably normalised. It is a modest generalisation of this to allow β to depend on (i,j). In our representation above, this corresponds to $c_{ij}(L-1) = \exp(-\beta_{ij})$, $c_{ij}(0) = 1 - \exp(-\beta_{ij})$, and $c_{ij}(e) = 0$ otherwise, so that the bond variables e_{ij} taken only two values, 0 and $(L-1)$. These have usually been described as on and off, true and false, or presence and absence of a bond between the sites i and j. In this case, the clusters induced by $\mathbf{e}$ will each be all of one colour.

The extra generality here in allowing a different function $a_i(\,)$ at each site permits some additional heterogeneity, most importantly that arising when we observe a degraded version of $\mathbf{x}$, say $\mathbf{y}$, and wish to study the posterior distribution $p(\mathbf{x}|\mathbf{y})$. If the degradation is independent, pixel-wise, so that

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i \in S} p(y_i|x_i)$$

and the prior distribution of $\mathbf{x}$ has the form (1), then the posterior will also have this form, with an additional factor of $p(y_i|x_i)$ in $a_i(x_i)$.

The major extension incorporated here is that of relaxing (6), however. By allowing bond variables taking integer, rather than just boolean, values, we can thus sensibly handle pairwise interaction Markov random fields on ordered colours (usually grey levels).

4. Metropolis and Gibbs samplers

The purpose of the constructions described in Section 2 is to assist in devising dynamic Monte Carlo procedures for simulating from $p(\mathbf{x})$ in the model (1). Since direct simulation is impractical, we aim to construct a Markov chain, with transition function $P(\mathbf{x},\mathbf{x}')$, say, on $\Omega \times \Omega$, which satisfies the detailed balance equation

$$p(\mathbf{x})\, P(\mathbf{x},\mathbf{x}') = p(\mathbf{x}')\, P(\mathbf{x}',\mathbf{x}) \tag{7}$$

for all $\mathbf{x},\mathbf{x}' \in \Omega$, and thus for which $\{p(\mathbf{x}), \mathbf{x} \in \Omega\}$ is an equilibrium distribution.

The Swendsen-Wang procedure is effectively the Gibbs sampler applied to $p(\mathbf{x},\mathbf{e})$, with block updating of all of $\mathbf{e}$, then all of $\mathbf{x}$, alternately. That is, given $\mathbf{x}^{(t)} = \mathbf{x}$, we first draw $\mathbf{e}$ from $p(\mathbf{e}|\mathbf{x})$ using (2), then draw $\mathbf{x}^{(t+1)} = \mathbf{x}$ from $p(\mathbf{x}|\mathbf{e})$ using (5). This amounts to using the transition function

$$P(\mathbf{x},\mathbf{x}') = \sum_{\mathbf{e}} p(\mathbf{e}|\mathbf{x})\, p(\mathbf{x}'|\mathbf{e}), \tag{8}$$

and (7) holds since

$$p(\mathbf{x})P(\mathbf{x},\mathbf{x}') = \sum_{\mathbf{e}} p(\mathbf{x})\, \frac{p(\mathbf{x},\mathbf{e})}{p(\mathbf{x})}\, \frac{p(\mathbf{x}',\mathbf{e})}{p(\mathbf{e})},$$

which is clearly symmetric in $\mathbf{x}$ and $\mathbf{x}'$.

Whilst simulating from (2) is always trivial, it may be much less straightforward to simulate from $p(\mathbf{x}|\mathbf{e})$, given by (5), particularly in the present generality. (In the case of the Potts model, (5) just involves an independent uniform choice of colour for

each cluster, so this case presents no difficulty.) More flexibility is available if we use the more general Metropolis/Hastings approach (see Metropolis, et al., (1953), Hastings (1970), and Green and Han (1991)). Having drawn $\mathbf{e}$ from $p(\mathbf{e}|\mathbf{x})$, suppose that we now draw a proposal $\mathbf{x}'$ for the new state $\mathbf{x}^{(t+1)}$, by sampling from an arbitrary transition function $q(\mathbf{x},\mathbf{x}';\mathbf{e})$, indexed by $\mathbf{e}$. This proposal is not immediately taken as the new state of the chain. Rather, it is only accepted, and $\mathbf{x}^{(t+1)}$ set equal to $\mathbf{x}'$, with probability $\alpha(\mathbf{x},\mathbf{x}';\mathbf{e})$; otherwise it is rejected, and no move is made, so that $\mathbf{x}^{(t+1)} = \mathbf{x}$.

The corresponding transition function is

$$\begin{aligned} P(\mathbf{x},\mathbf{x}') &= \sum_{\mathbf{e}} p(\mathbf{e}|\mathbf{x})\, q(\mathbf{x},\mathbf{x}';\mathbf{e})\ \alpha(\mathbf{x},\mathbf{x}';\mathbf{e}) \quad \text{if } \mathbf{x} \neq \mathbf{x}' \\ &= 1 - \sum_{\mathbf{x}' \neq \mathbf{x}} P(\mathbf{x},\mathbf{x}') \qquad\qquad \text{if } \mathbf{x} = \mathbf{x}', \end{aligned} \tag{9}$$

and if we take

$$\alpha(\mathbf{x},\mathbf{x}';\mathbf{e}) = \min\ \{1,\ \frac{p(\mathbf{x}',\mathbf{e})q(\mathbf{x}',\mathbf{x};\mathbf{e})}{p(\mathbf{x},\mathbf{e})q(\mathbf{x},\mathbf{x}';\mathbf{e})}\} \tag{10}$$

then it is readily shown that detailed balance (7) holds.

The ordinary Swendsen-Wang procedure (Gibbs sampler) is obtained by taking

$$q(\mathbf{x},\mathbf{x}';\mathbf{e}) = p(\mathbf{x}'|\mathbf{e})$$

whence from (10),

$$\alpha(\mathbf{x},\mathbf{x}';\mathbf{e}) \equiv 1,$$

so that the proposal is never rejected.

In many contexts, including most image analysis applications, the number of sites n will be very large. The ratio term in (10) then involves highly multivariate probability functions, and so with substantial probability the acceptance probability α will be very small. In this situation, such Metropolis methods will be unacceptably inefficient. Fortunately, the generation and acceptance of proposals can be performed on a cluster-by-cluster basis, thus effectively reducing this dimensional problem.

Let $Cl(\mathbf{e}) \subset 2^S$ be the collection of clusters induced by the particular configuration of bond variables $\mathbf{e}$. For $A \in Cl(\mathbf{e})$, suppose that a proposal $\mathbf{x}_A{}'$ for recolouring $\mathbf{x}_A$ is generated with probability $q(\mathbf{x}_A,\mathbf{x}_A{}';\mathbf{e})$ for each $\mathbf{x}_A{}' \in C^A$, and accepted with probability $\alpha(\mathbf{x}_A,\mathbf{x}_A{}';\mathbf{e})$. These choices are made independently for each cluster. The resulting Markov transition matrix is

$$\begin{aligned} P(\mathbf{x},\mathbf{x}') = \sum_{\mathbf{e}} p(\mathbf{e}|\mathbf{x}) \prod_{A \in Cl(\mathbf{e})} \{ & q(\mathbf{x}_A,\mathbf{x}_A{}';\mathbf{e})\, \alpha(\mathbf{x}_A,\mathbf{x}_A{}';\mathbf{e}) + \\ & \delta_{\mathbf{x}_A \mathbf{x}_A{}'} \sum_{\mathbf{x}_A^\circ} q(\mathbf{x}_A,\mathbf{x}_A^\circ;\mathbf{e})\,(1-\alpha(\mathbf{x}_A,\mathbf{x}_A^\circ;\mathbf{e}))\} \end{aligned} \tag{11}$$

But by the conditional independence of $\mathbf{x}$ given $\mathbf{e}$ in different clusters, we have

$$p(\mathbf{x})p(\mathbf{e}|\mathbf{x}) = p(\mathbf{e})p(\mathbf{x}|\mathbf{e}) = p(\mathbf{e}) \prod_{A \in Cl(\mathbf{e})} p(\mathbf{x}_A|\mathbf{e}). \tag{12}$$

Combining (11) with (12), it is then easy to show that $p(\mathbf{x})P(\mathbf{x},\mathbf{x}')$ is symmetric in $\mathbf{x}$ and $\mathbf{x}'$, if α is defined, by analogy with (10), as

$$\alpha(\mathbf{x}_A, \mathbf{x}_A'; \mathbf{e}) = \min\{1, \frac{p(\mathbf{x}_A'|\mathbf{e})q(\mathbf{x}_A', \mathbf{x}_A; \mathbf{e})}{p(\mathbf{x}_A|\mathbf{e})q(\mathbf{x}_A, \mathbf{x}_A'; \mathbf{e})}\}. \tag{13}$$

Since clusters will typically contain far fewer than n vertices, these cluster recolouring acceptance probabilities will rarely be very small.

A particular instance of this cluster-wise Metropolis procedure that is simple to implement, is obtained by taking

$$q(\mathbf{x}_A, \mathbf{x}_A'; \mathbf{e}) = \theta p(\mathbf{x}_A'|\mathbf{e}) + (1-\theta)\delta_{\mathbf{x}_A \mathbf{x}_A'},$$

where $\theta \in (0,1]$ can depend on A and $\mathbf{e}$. The acceptance probability (13) is then identically 1. The motivation for this proposal distribution is that it arises if we attempt to sample from $p(\mathbf{x}_A'|\mathbf{e})$ using a rejection method, and abandon the attempt after a specified number of trials, retaining the colouring $\mathbf{x}_A$ instead. See section 6 for discussion of such a rejection method.

5. Convergence of the chain

To establish that the Markov chains defined in the previous section converge in distribution to the equilibrium $p(\mathbf{x})$, $\mathbf{x} \in \Omega$, we have to check that aperiodicity and irreducibility hold. The state space Ω is finite, so no other regularity conditions are needed.

Let $\mathbf{e}^*$ denote the configuration of bond variables in which $e_{ij} = L - 1$ for all $i \sim j$ (so that there are n clusters each containing one site). Since $b_{ij}(\,)$ is positive, $c_{ij}(L-1) > 0$ for all $i \sim j$. Hence, using (2), $p(\mathbf{e}^*|\mathbf{x}) > 0$ for all $\mathbf{x} \in \Omega$.

Suppose we only use proposal distributions $q(\,)$ satisfying

$$q(\mathbf{x}, \mathbf{x}'; \mathbf{e}^*) > 0 \quad \text{for all } \mathbf{x}, \mathbf{x}' \in \Omega$$

(it is clear from (5) that $p(\mathbf{x}'|\mathbf{e}^*) > 0$ for all $\mathbf{x}' \in \Omega$, so the Gibbs sampler case is included). Then it is immediate from (9) that $P(\mathbf{x}, \mathbf{x}') > 0$ for all $\mathbf{x}, \mathbf{x}' \in \Omega$, so that all such chains are both aperiodic and irreducible.

6. Uniform colourings conditional on bond variables

Although efficient implementation of algorithms using these ideas will not be attempted here, we do sketch an approach to generating $\mathbf{x}$ uniformly from $\Omega(\mathbf{e})$ for any prescribed set of bond variable values $\mathbf{e}$. When the functions $\{a_i(\,)\}$ are identically 1, then the conditional distribution of the site variables given in (5) is precisely this uniform distribution. In other cases, this distribution may be a useful proposal distribution for the Metropolis/Hastings approach.

We wish to generate $\mathbf{x}$ uniformly over C^S subject to the constraint $|x_i - x_j| \leq e_{ij}$ for all $i \sim j$. Values assigned in different clusters will be independent, so we have only to consider a single cluster $A \subseteq S$, say. A rejection method will be necessary, but it is clearly likely to be very inefficient to sample each $x_i, i \in A$, uniformly from C, and then reject the whole realisation and repeat, if $|x_i - x_j| > e_{ij}$ for any $i \sim j, i, j \in A$. Instead, we propose a random walk approach. First construct a spanning tree for the cluster - that is a subgraph containing all sites $i \in A$ and having no loops. Any spanning tree is valid, but for efficiency, one should be chosen that picks edges (i,j) for

which e_{ij} is as small as possible. A formal construction of a *minimum* spanning tree is probably not worthwhile. Given the tree, assign an arbitrary "root" site the integer value 0, and then proceed outwards from the root along the edges of the tree until all sites are numbered. In traversing an edge (i,j) from a site i numbered z_i, assign j a uniform random integer z_j from $\{z_i - e_{ij}, z_i - e_{ij} + 1, \ldots, z_i + e_{ij}\}$. Having assigned integers to all sites in the cluster, test all edges of the original graph not in the tree, and reject the whole assignment if $|z_i - z_j| > e_{ij}$ for any $i \sim j$. Reject also if $\max_{i \in A} z_i - \min_{i \in A} z_i > L - 1$. Finally draw an integer w at random from $0, 1, \ldots, L-1$, and set $x_i = z_i + w - \min_{i \in A} z_i$ for each $i \in A$. Reject if any $x_i > L - 1$, otherwise the assignment of colours is complete. It is evident that all $\mathbf{x} \in \Omega(\mathbf{e})$ are equally likely to be generated.

Many of the tests, for example monitoring the range $\max z_i - \min z_i$, may be performed as the tree is traversed, so that rejection can take place before incurring the expense of visiting all sites in the cluster. Various other devices might be employed in implementing this scheme efficiently, including the use of linked lists in constructing the tree, and a stack of visited sites, so that rejection and shifting the assigned colours have only a small housekeeping cost.

In the Potts model, all e_{ij} are 0 or $(L-1)$, so the procedure described simply chooses a random colour for the cluster, and never rejects. Intuitively, the method might be expected to work well if most e_{ij} are 0 or $(L-1)$, and the remainder (or at least those associated with edges of the spanning tree) are small.

Acknowledgements

I am grateful to Alan Sokal for some helpful correspondence, including electronic mail copies of his articles.

References

Besag, J., York, J. C., and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math., 43, 1-59.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. J. Amer. Statist. Assoc., 85, 398-409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell., 12, 609-628.

Green, P. J. and Han, X-L. (1991) Metropolis methods, gaussian proposals, and antithetic variables. In Stochastic models, statistical methods and algorithms in image analysis. edited by P. Barone, A. Frigessi and M. Piccioni, Lecture Notes in Statistics, Springer, Berlin (to appear).

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains, and their applications. Biometrika, 57, 97-109.

Kirkland, M. (1989) Simulating Markov random fields. Ph. D. thesis, University of Strathclyde.

Potts, R. B. (1952) Some generalised order-disorder transformations. Proc. Camb. Phil. Soc., 48, 106-109.

Sheehan, N. and Thomas, A. W. (1991) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Preprint, University of Bath.

Sokal, A. D. (1989) Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne.

Sokal, A. D. (1990) How to beat critical slowing-down: 1990 update. Lattice '90 conference, Tallahassee, Florida, October 1990.

Swendsen, R. H. and Wang, J-S. (1987) Nonuniversal critical dynamics in Monte Carlo simulations. Phys. Review Lett., 58, 86-88.