BLOCK UPDATING IN CONSTRAINED MARKOV CHAIN MONTE CARLO SAMPLING

MERRILEE A. HURN*
MATHEMATICAL SCIENCES
UNIVERSITY OF BATH, U. K.

HÅVARD RUE
MATHEMATICAL SCIENCES
NTNU, NORWAY

NUALA A. SHEEHAN
MATHEMATICAL SCIENCES
LOUGHBOROUGH UNIVERSITY, U. K.

SEPTEMBER 3, 1997

ABSTRACT

Markov chain Monte Carlo methods are widely used to study highly structured stochastic systems. However when the system is subject to constraints, it is difficult to find irreducible proposal distributions. We suggest a "block-wise" approach for constrained sampling and optimisation.

KEYWORDS: Constrained distributions; Importance sampling; Irreducibility; Markov chain Monte Carlo; Multiple-site updating; Stochastic simulation and optimisation.

*ADDRESS FOR CORRESPONDENCE: Merrilee Hurn, Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, U. K.
Email: M.A.Hurn@maths.bath.ac.uk, Havard.Rue@math.ntnu.no, and N.A.Sheehan@lut.ac.uk

1 Introduction

Markov chain Monte Carlo methods are now widely advocated for tackling many previously intractable Bayesian and non-Bayesian problems (see Gilks, Richardson and Spiegelhalter, 1995, for a recent overview of theory and applications). The basic idea is to explore the high-dimensional distribution of interest $\pi(\boldsymbol{x}),\ \boldsymbol{x} \in \Omega$, by generating an aperiodic and irreducible Markov chain with $\pi(\boldsymbol{x})$ as its equilibrium distribution. For simplicity we will describe the case when Ω is finite or countable. Then the transition matrix of the Markov chain $P(\boldsymbol{x}, \boldsymbol{x}')$ is often constructed as a combination of basic transition matrices $\{P_t(\boldsymbol{x}, \boldsymbol{x}')\}$, each describing a different way to move around the state space Ω. To update the Markov chain according to $P(\boldsymbol{x}, \boldsymbol{x}')$, we first select the basic transition type t either randomly or following some deterministic schedule, and then update the Markov chain according to P_t. Each P_t is constructed to satisfy detailed balance with respect to $\pi(\boldsymbol{x})$ as follows: when the current state is $\boldsymbol{x}$, we propose a possible new state $\boldsymbol{x}'$ according to the distribution $q_t(\boldsymbol{x}, \boldsymbol{x}')$. We then accept $\boldsymbol{x}'$ as the new state with probability

$$\alpha_t(\boldsymbol{x}, \boldsymbol{x}') = \min\left\{1, \frac{\pi(\boldsymbol{x}')q_t(\boldsymbol{x}', \boldsymbol{x})}{\pi(\boldsymbol{x})q_t(\boldsymbol{x}, \boldsymbol{x}')}\right\}, \quad \text{if } \pi(\boldsymbol{x})q_t(\boldsymbol{x}, \boldsymbol{x}') \neq 0. \tag{1}$$

Otherwise the Markov chain stays in the old state $\boldsymbol{x}$. Provided the transition matrix $P(\boldsymbol{x}, \boldsymbol{x}')$ is irreducible and aperiodic then, after an initial burn-in period, the states of the chain have $\pi(\boldsymbol{x})$ as their marginal distribution, albeit that values along the chain are dependent.

A vital part of MCMC is thus the formulation of this collection of proposal distributions $\{P_t\}$ which allow the sampler to visit all regions of the sample space Ω. For this, it is often sufficient to use a naïve approach with each proposal distribution changing at most a single component of $\boldsymbol{x}$; that is, $P_t(\boldsymbol{x}, \boldsymbol{x}')$ proposes a new value for x_t leaving the rest of $\boldsymbol{x}$, $\boldsymbol{x}_{-t}$, unchanged. As an example of such a component-wise proposal approach, the Gibbs sampler generates a new value for the t^{th} component of $\boldsymbol{x}$, holding all the other components $\boldsymbol{x}_{-t}$ fixed, taking $q_t(\boldsymbol{x}, \boldsymbol{x}')$ as the conditional distribution $\pi(x_t|\boldsymbol{x}_{-t})$. In this case, the new state $\boldsymbol{x}'$ is accepted with probability 1.

The problem on which we wish to focus occurs when the distribution of interest $\pi(\boldsymbol{x})$ is a conditional distribution subject to the constraint $C(\boldsymbol{x}) = 0$. Such distributions arise in a variety of fields and the constraints may result from either logical features of the application, additional exact external knowledge, or as a designed part of the model itself. One example of the first category is the problem of sampling genotypic frequencies on a pedigree where the genotype configuration $\boldsymbol{x}$ is required to satisfy the laws of Mendelian inheritance (e.g. Sheehan and Thomas, 1993). Two examples of the second type of constrained problem are the incorporation of ground truth into satellite image processing, and the lattice simulation of binary alloys when the proportion of the two atomic types is fixed. As an example of modelling-induced constraints, Geman *et. al.* (1990) construct models for image boundary detection which "forbid" certain undesirable features such as very small regions and dead-end boundaries by giving such features probability zero. In any such constrained situation, a combination of component-wise $P_t(\boldsymbol{x}, \boldsymbol{x}')$ may be insufficient to achieve irreducibility. Moreover, the complexity of $C(\boldsymbol{x})$ may make it difficult to construct alternative move-types which maintain the constraint and provide the required mixing of the Markov chain.

Geman (1990) and Geman *et. al.* (1990) describe constrained optimisation using simulated annealing for "forbidden state" problems. Here the quantity sought is the $\boldsymbol{x}$ which maximises $\pi(\boldsymbol{x})$ subject to $C(\boldsymbol{x}) = 0$. The constraint is initially relaxed to achieve irreducibility, but then as the annealing progresses, an increasing penalty is imposed on forbidden $\boldsymbol{x}$. Alternatively when it is functionals $E_\pi f(\boldsymbol{x})$ which are of interest, Hastings (1970) suggests using importance sampling ideas, generating samples from some $\pi^*(\boldsymbol{x})$ on a possibly larger state space $\Omega^* \supseteq \Omega$, where $\pi^*(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega$, and then reweighting them using the identity

$$E_{\pi} f(\boldsymbol{x}) = E_{\pi^*}[f(\boldsymbol{x})\frac{\pi(\boldsymbol{x})}{\pi^*(\boldsymbol{x})}]. \tag{2}$$

To avoid the irreducibility problem, it is common to choose $\pi^*(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega^*$ and Ω^* the product space of all the $\boldsymbol{x}$ components. Single component proposals can then be used to visit all the constrained and unconstrained $\boldsymbol{x}$. Although the idea of importance sampling is appealing, the efficiency can be low if the event $\{C(\boldsymbol{x}) = 0\}$ has low probability under $\pi^*(\boldsymbol{x})$. In addition, it may be necessary to estimate the ratio of the normalising constants in $\pi(\boldsymbol{x})/\pi^*(\boldsymbol{x})$. Neither of these approaches actually generates the samples themselves from $\pi(\boldsymbol{x})$.

As motivation for the algorithm we present in detail in Section 2, suppose we are interested only in the conditional distribution of a small block B of components, $\boldsymbol{x}_B$, conditional on the rest $\boldsymbol{x}_{-B}$. Assume the constraint is satisfied by $\boldsymbol{x} = \boldsymbol{x}_B \cup \boldsymbol{x}_{-B}$. We can learn about $\boldsymbol{x}_B$ under $\pi(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ by importance sampling from some suitable $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ and applying Equation (2). Can we then learn about the whole of $\boldsymbol{x}$ under $\pi(\boldsymbol{x})$ in this way, investigating "suitably sized" blocks B of the components in turn? The result in this note shows how to explore $\pi(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ using a Markov chain whose equilibrium distribution is $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$, merging the results from different blocks B together in such a way as to obtain a Markov chain with equilibrium distribution $\pi(\boldsymbol{x})$. Unlike importance sampling, this approach actually generates MCMC samples from $\pi(\boldsymbol{x})$.

The outline of this note is as follows: in Section 2 we detail our proposed procedure and make some general remarks. In Section 3 we discuss what is meant by "suitably sized" blocks to fulfill the irreducibility requirement, and also the rôle and specification of $\pi^*(\boldsymbol{x})$. We conclude by testing our methodology on an Ising model constrained to have a fixed number of connected components.

2 THE ALGORITHM

For descriptive purposes, the algorithm can be split into three parts; generating a block B of components of $\boldsymbol{x}$ to update from $\mathcal{B} = \{B\}$, proposing a possible new configuration for this set of components using $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$, and finally ensuring that the algorithm maintains the correct distribution $\pi(\boldsymbol{x})$.

To restate the goal, we want to sample from $\pi(\boldsymbol{x})$ under the constraint $C(\boldsymbol{x}) = 0$. A distribution $\pi^*(\boldsymbol{x})$ can be found for which an irreducible single-component sampler exists and which satisfies the condition $\pi^*(\boldsymbol{x}) > 0, \forall \boldsymbol{x} \in \Omega^* \supseteq \Omega$. When the current state is $\boldsymbol{x} \in \Omega$, a new state $\boldsymbol{x}'$ is proposed as follows: first a block B of components is chosen either randomly or deterministically from the available set of blocks $\mathcal{B}$. Holding the components $\boldsymbol{x}_{-B}$ outside the block fixed, an MCMC sampler satisfying detailed balance with respect to $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ is used to draw values for the block components. Let $\boldsymbol{x}_B, \boldsymbol{x}_B^{*1}, \boldsymbol{x}_B^{*2}, \ldots$ denote the successive states of this secondary chain. Note that these states are not required to satisfy the constraint $C(\boldsymbol{x}_B^{*k} \cup \boldsymbol{x}_{-B}) = 0$. The secondary chain is terminated when the first legal configuration is reached, that is the first k for which $C(\boldsymbol{x}_B^{*k} \cup \boldsymbol{x}_{-B}) = 0$. Denote this $\boldsymbol{x}_B^{*k}$ by $\boldsymbol{x}'_B$. This legal $\boldsymbol{x}' = \boldsymbol{x}'_B \cup \boldsymbol{x}_{-B}$ is then used as the proposed new state for the original chain. To maintain $\pi(\boldsymbol{x})$ as the equilibrium distribution for the main chain, we accept or reject this proposal according to Equation (1). Note that k is random, and it is not immediately clear how to evaluate the proposal distribution $q_B(\boldsymbol{x}, \boldsymbol{x}')$ underlying this mechanism. However, in the Appendix we prove that

$$\frac{q_B(\boldsymbol{x}', \boldsymbol{x})}{q_B(\boldsymbol{x}, \boldsymbol{x}')} = \frac{\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})}{\pi^*(\boldsymbol{x}'_B|\boldsymbol{x}_{-B})} = \frac{\pi^*(\boldsymbol{x})}{\pi^*(\boldsymbol{x}')}, \tag{3}$$

resulting in the simple expression for the acceptance ratio

$$\alpha_B(\boldsymbol{x}, \boldsymbol{x}') = \min\left\{1, \frac{\pi(\boldsymbol{x}')\pi^*(\boldsymbol{x})}{\pi(\boldsymbol{x})\pi^*(\boldsymbol{x}')}\right\} \tag{4}$$

which gives the required detailed balance. Irreducibility and aperiodicity of these transitions follow from the irreducibility and aperiodicity of the $\pi^*(\boldsymbol{x})$ sampler; this is discussed further in Section 3.2. We conclude this section with some general remarks.

REMARK 1 The stopping rule deciding when to terminate the secondary chain can be a more general rule satisfying the conditions stated in the Appendix. For example, $\boldsymbol{x}'_B$ could be the second legal configuration reached rather than the first, or the rule could be to sample either until a legal configuration is reached or for a fixed number m steps in the secondary chain, whichever is the smaller. Note in the latter case if the proposed $\boldsymbol{x}'$ is illegal, then Equation (4) will automatically reject it. The algorithm also generalises to continuous state spaces where the constraint is an event with non-zero probability.

REMARK 2 The MCMC sampler for $\pi^*(\boldsymbol{x})$ is not required to have reached equilibrium, nor do we place any restrictions on how the transitions are made for the secondary $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ chain other than that they are reversible, irreducible and aperiodic (for example, a random scan Gibbs sampler would be allowed here). The assertion is *not* that $q_B(\boldsymbol{x}', \boldsymbol{x}) = \pi^*(\boldsymbol{x})$. However, for certain choices of $\pi^*(\boldsymbol{x})$ it may actually be possible to draw exact samples thus avoiding the need for the secondary Hastings algorithm; in this case, the stronger assertion does hold, although Equation (3) remains unchanged.

REMARK 3 If we take $\pi^*(\boldsymbol{x}) \propto \pi(\boldsymbol{x}), \forall \boldsymbol{x} \in \Omega$, then the acceptance probability is always 1, giving the algorithm the feel of a block-wise Gibbs sampler. Note that this also demonstrates that importance sampling using this form for $\pi^*(\boldsymbol{x})$ is producing $\boldsymbol{x}$ samples from $\pi(\boldsymbol{x})$.

REMARK 4 In contrast with importance sampling, it is not necessary to estimate the relative normalising constants of $\pi(\boldsymbol{x})$ and $\pi^*(\boldsymbol{x})$ since these vanish in the acceptance ratio Equation (4). Hastings (1970) discusses this for importance sampling in §2.5 of his paper, pointing out that when it is possible to sample directly from $\pi^*(\boldsymbol{x})$, the MCMC sampler can be set up with $\pi(\boldsymbol{x})$ and $\pi^*(\boldsymbol{x})$ as the target and the proposal distributions respectively to avoid the estimation. The approach suggested here could be viewed as an extension of this point.

REMARK 5 Using MCMC samples as proposals within a second MCMC step is suggested by Neal (1996) for sampling multimodal distributions using simulated tempering. A sequence of linked distributions is set up, the first being the target distribution while the others are successively less modal and so easier to explore. Proposals $\boldsymbol{x}'$ are generated by passing the current $\boldsymbol{x}$ up through the sequence of distributions and then back down again, at each level applying a single MCMC update with that level's target distribution. Our suggestion differs in two ways. Firstly we are primarily concerned with sampling under constraints on $\boldsymbol{x}$ which may require us to enlarge the sample space Ω for irreducibility. Secondly, because of this, rather than always using a single update with $\pi^*(\boldsymbol{x})$ as the target distribution, we have a variable number of updates governed by a stopping rule on $\boldsymbol{x}$ which returns us to Ω.

3 IRREDUCIBILITY AND THE CHOICE OF π^*

In this Section, we consider the specification of the alternative distribution $\pi^*(\boldsymbol{x})$ and the choice of the set of blocks $\mathcal{B}$. Although the two are closely linked we divide the discussion, considering first the rôle $\pi^*(\boldsymbol{x})$ plays in allowing movement around Ω. We then examine how the blocking scheme can be related to the constraint $C(\boldsymbol{x})$. We concentrate on the case where Ω^* is the product space of the $\boldsymbol{x}$ components.

3.1 CHOICE OF RELAXED DISTRIBUTION

By specifying $\pi^*(\boldsymbol{x}) > 0, \forall \boldsymbol{x} \in \Omega^*$, irreducibility can be achieved using single-component updates such as the Gibbs sampler and the Metropolis algorithm. However there is more to the choice of $\pi^*(\boldsymbol{x})$: it should promote movement between legal $\boldsymbol{x} \in \Omega$ by creating paths of high probability through $\Omega^* \setminus \Omega$, but targeting where this probability is allocated to avoid over-long excursions. We also require the acceptance ratio, Equation (4), to be reasonably high. Although these properties are desirable in all situations, there may not be a general rule for formulating such a distribution. However we can suggest some guidelines applicable in a fairly wide range of problems.

As already noted, it is possible to make the acceptance ratio Equation (4) identically 1 by taking $\pi^*(\boldsymbol{x}) \propto \pi(\boldsymbol{x}), \forall \boldsymbol{x} \in \Omega$. But what about $\pi^*(\boldsymbol{x})$ for illegal $\boldsymbol{x}$? Here the choice may depend on whether there is a natural extension of $\pi(\boldsymbol{x})$ to Ω^*, the obvious case being if $\pi(\boldsymbol{x})$ is actually a restriction of some distribution $\phi(\boldsymbol{x})$ defined for $\boldsymbol{x} \in \Omega^*$,

$$\pi(\boldsymbol{x}) \propto \phi(\boldsymbol{x}) I_{[C(\boldsymbol{x})=0]}, \qquad \boldsymbol{x} \in \Omega. \tag{5}$$

If this is the case, then we might wish to maintain the $\phi(\boldsymbol{x})$ structure in $\pi^*(\boldsymbol{x})$. If not, then the choice may be influenced solely by the second goal, that of restricting the excursion lengths through $\Omega^* \setminus \Omega$. Suppose we are dealing with a constraint on $\boldsymbol{x}$ which is not single-valued for illegal $\boldsymbol{x}$, for example $C(\boldsymbol{x})$ might involve the number of components taking a particular value. Some $\boldsymbol{x} \in \Omega^*$ are then "more illegal" than others, and we could follow the idea of Green (1986) and weight this degree of illegality by including a factor $\exp(-\delta\, |C(\boldsymbol{x})|)$, $\delta > 0$, in $\pi^*(\boldsymbol{x})$; note that this factor will not affect the acceptance rate of legal $\boldsymbol{x}$. So, for $\boldsymbol{x} \in \Omega^* \setminus \Omega$, we suggest taking

$$\pi^*(\boldsymbol{x}) \propto \begin{cases} \phi(\boldsymbol{x})\ \exp(-\delta\, |C(\boldsymbol{x})|), & \text{if Equation (5) holds,} \\ \epsilon\ \exp(-\delta\, |C(\boldsymbol{x})|), & \text{otherwise,} \end{cases} \tag{6}$$

for some constant $\epsilon > 0$ reflecting the magnitude of $\phi(\boldsymbol{x})$.

3.2 CHOICE OF BLOCKING SCHEME $\mathcal{B}$

We now discuss the choice of the blocking scheme $\mathcal{B}$ controlling which components are considered jointly when sampling from $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$. We are interested in the connections between this choice and the constraint $C(\boldsymbol{x})$ in order to achieve irreducibility. When Ω^* is finite, we can expand the constraint in a way which helps to show why some blocking schemes may *not* be adequate and suggests which others might work. Irreducibility does still need to be checked on a case-by-case basis; even quite local constraints may require updating a large proportion of $\boldsymbol{x}$!

To illustrate our discussion, we will consider a form of Ising model where $\boldsymbol{x} = (x_1, \ldots, x_n)$ represents the binary values taken on a $p \times p$ toroidal lattice whose components (or sites) are indexed $1, \ldots, n = p^2$. The unconstrained Ising model has distribution

$$\phi(\boldsymbol{x}) \propto \exp(-\beta \sum_{s \sim t} I_{[x_s \neq x_t]}), \qquad \boldsymbol{x} \in \Omega^* = \{0,1\}^n \tag{7}$$

where $s \sim t$ indicates all four nearest neighbouring sites on the toroidal lattice, β is a positive parameter, and I is an indicator variable.

Suppose we are interested in the Ising model under some constraint $C(\boldsymbol{x}) = 0$. Since Ω^* is finite, we can expand the constraint uniquely as

$$\begin{aligned} C(\boldsymbol{x}) &= C_0 + \sum_{1 \leq i \leq n} x_i V_i(x_i) + \sum_{1 \leq i < j \leq n} x_i x_j V_{ij}(x_i, x_j) + \cdots \\ &\quad + x_1 x_2 \cdots x_n V_{1,2,\ldots,n}(x_1, \ldots, x_n), \end{aligned} \tag{8}$$

where $x_i V_i(x_i) = C((0, 0, \ldots, 0, x_i, 0, \ldots, 0)) - C((0, \ldots, 0))$ and so on (Besag, 1974). Many of these V terms will be identically zero. Knowledge of which terms are non-zero carries information about possible blocking schemes, in some sense indicating the "clique structure" of the constraint (*cf* the discussion of the Hammersley-Clifford theorem in Besag, 1974).

As a first example, consider the Ising model under the constraint of a fixed number, N_0, of sites taking the value 0. Written in the form of Equation (8), $C(\boldsymbol{x}) = n - N_0 - \sum_{1 \leq i \leq n} x_i$, with higher order V terms being zero. The set $\mathcal{B}$ must contain at least enough blocks B to guarantee movement around Ω using only B-wise changes. Here this can be achieved by considering all pairs of components for joint updating (this form being known as a "spin-exchange" algorithm).

Next consider instead constraining the edge length to be L_{edge}, where edge length is defined as the number of pairs of neighbouring sites taking different values. Here the constraint may be written as $C(\boldsymbol{x}) = L_{\text{edge}} - \sum_{i \sim j} I_{[x_i \neq x_j]}$ with all other V terms being zero. By analogy with the previous example, is it enough to ensure irreducibility if we choose $\mathcal{B}$ to be blocks which are one component larger than the order of the V terms, in this case all the three components sets? Unfortunately, the answer is no; consider the highly constrained system where the edge length on the toroidal $p \times p$ lattice must equal $2p$. This strict condition is met by only a few configurations such as those consisting of two horizontal bands, the first of 0's, the second of 1's. Except for very small values of n, it is not possible to move between legal configurations by making changes to just three components at a time, the blocks must be larger here.

What these examples demonstrate is that even constraints built up of local terms can give severe irreducibility problems. The form of Equation (8) can point to suitable blocking schemes, but unless irreducibility can be seen to be achieved for a certain blocking scheme, it may be necessary to update the whole of $\boldsymbol{x}$ (where irreducibility is guaranteed by the use of $\pi^*(\boldsymbol{x})$). Fortunately it is still possible to use a mixture of different smaller blocks in $\mathcal{B}$, possibly associated with V terms at different scales, provided there is a positive probability in this mixture of selecting a block which is large enough to lead to irreducible transitions.

4 An example

To illustrate the ideas presented so far, we consider sampling an Ising model, Equation (7), constrained to have a fixed number of clusters. Two components are said to be connected if there exists a route of neighbouring sites $s \sim t$ between them, each pair of which take the same value $x_s = x_t$, and a cluster is then a maximal set of connected components. For this example, we consider an Ising model with parameter $\beta = 0.6$ on a 64×64 toroidal lattice, constrained to have exactly 256 clusters. Such a high value of β suggests fairly strong component interactions and so 256 is an unusually large number of clusters. In addition, single-site samplers will not mix particularly well for the unconstrained model. The form of $C(\boldsymbol{x})$ here does not immediately suggest a minimal blocking scheme, and so we will employ a mixture of different blocks for $\mathcal{B}$. These blocks are chosen to be the components lying in square subwindows of the toroidal lattice, with these randomly positioned squares having as side length a Poisson random variable with parameter 8 conditioned to be in the range 2 to 64. The motivation for a spatial blocking scheme is that the connection pattern may be altered quite significantly by allowing a small region to evolve through illegal states. Irreducibility follows from the possibility of updating the whole lattice.

Since the constrained distribution is the restriction of a distribution defined over Ω^*, we could preserve the Ising structure by choosing $\pi^*(\boldsymbol{x}) \propto \exp(-0.6 \sum_{s \sim t} I_{[x_s \neq x_t]})$. This would lead to an acceptance rate of 1 for the resulting $\boldsymbol{x}'$, but the potential drawback is that excursions through the illegal $\Omega^* \setminus \Omega$ may tend to be very short or very long. The former case occurs because the "stickiness" of the distribution resists single-component changes when the component in question already matches the majority of its neighbours, and so the unchanged configuration is returned as the next legal state. On the other hand, once a move is made away from the initial legal $\boldsymbol{x}$, the distribution's desire to have fewer than 256 clusters leads to a long wait before the next legal state. Our first option to avoid this problem is to maintain the Ising structure on $\Omega^* \setminus \Omega$ but with a smaller interaction parameter, $\beta_2 \leq 0.6$ (smaller values of β_2 mean a greater chance of making an initial move out of Ω). Not all the proposed $\boldsymbol{x}'$ will now be accepted in the main chain. The length of the excursion through $\Omega^* \setminus \Omega$ can then be controlled in the manner of Equation (6) by taking

$$\pi^*(\boldsymbol{x}) \propto \exp(-\beta_2 \sum_{s \sim t} I_{[x_s \neq x_t]} - \delta\, |\text{No. clusters in } \boldsymbol{x} - 256|), \qquad \boldsymbol{x} \in \Omega^*. \tag{9}$$

A second cap on the excursion length is to modify the stopping rule to be the sooner of the two events, (a) the first legal $\boldsymbol{x}'$ reached and (b) a fixed number n_{max} steps in the secondary $\pi^*(\boldsymbol{x})$ chain. Stopping using the second of these events will almost always lead to a rejection of the resulting $\boldsymbol{x}'$ since it will only be a legal state if the two events occur simultaneously.

Table 1 presents the results of simulations using seven combinations of β_2 and δ values: β_2 is taken to be 0.6, 0.3 or 0 (successively increasing the secondary chain mixing but lowering the acceptance rate of the resulting $\boldsymbol{x}'$); δ is assigned the values 0, 0.3 or 0.6. By starting with a random labelling of the lattice with its corresponding high number of clusters, an initial legal configuration can be found quickly by single-site updating using $\beta_2 = 0.6$ and $\delta = 0$ since this distribution prefers to decrease the number of clusters to below the target of 256. To assess the relative efficiency of the different parameter settings, each scheme is run for approximately $64 \times 64 \times 500$ updates, where an update is counted as a single component step in the $\pi^*(\boldsymbol{x})$ sampler, the total number of updates permitted is equivalent to 500 complete sweeps of the entire lattice, and the block change in progress when this $2\,048\,000^{th}$ update is made is allowed to run until the stopping rule is satisfied. The differing behaviour in $\Omega^* \setminus \Omega$ means that the seven chains will generate different numbers of proposed $\boldsymbol{x}'$. Table 1 shows the total number of updates, together with the average excursion length. Since the Ising model exhibits symmetry in the two states 0 and 1, the frequency with which each sampler moves from being predominantly 0 or predominantly 1 is indicative of the chain's mixing. Defining "predominantly 0" here to be more than $2\,058$ 0-components, and "predominantly 1" to be more than $2\,058$ 1-components, the number of these "0/1 swaps" is monitored. The 0/1 symmetry also means that in the long term the ergodic average of the number of 1 components should be $2\,048$; the observed average is calculated as another mixing indicator, along with the minimum and maximum number of 1 components. The observed edge length in each realisation is related to the probability of that realisation via Equation (7), and so as a final diagnostic the minimum, average and maximum edge length is recorded.

Table 1 demonstrates some of the effects of different $\pi^*(\boldsymbol{x})$, in particular the benefits of using a secondary distribution other than the unconstrained $\pi(\boldsymbol{x})$, and the subsequent rôle played by δ in controlling the excursion length. Obviously over-restricting the excursion length with too high a value of

δ can be counter-productive if the effect is to impede movement around $\Omega^* \setminus \Omega$. The degree of coercion required may depend on how atypical the condition $C(\boldsymbol{x}) = 0$ is for a particular $\pi^*(\boldsymbol{x})$, for example when $\delta = 0$ the two chains taking $\beta = 0.6$ and $\beta = 0.3$ have quite noticeably different mean excursion lengths. The average numbers of 1 components are not all close to $2\,048$, but this is not so surprising since successive $\boldsymbol{x}$ states can be quite highly correlated when only a few components change between them, and the total number of updates is only equivalent to 500 sweeps of the entire lattice. Note though the significance of these results is that all seven combinations of β_2 and δ values result in MCMC samples from the constrained distribution $\pi(\boldsymbol{x})$ without even the need to estimate relative normalising constants.

5 CONCLUSIONS

In this note we have demonstrated how a block updating approach can be used to avoid irreducibility problems in MCMC sampling from constrained distributions. We have discussed the choice of the unconstrained secondary distribution and the blocking scheme with regard to the constraint $C(\boldsymbol{x}) = 0$. In contrast to importance sampling, no normalising constants need be estimated, and the $\boldsymbol{x}$ values generated do form an MCMC sample from the constrained distribution of interest.

ACKNOWLEDGEMENTS

We are grateful to Peter Green for comments and helpful suggestions for neatening the proof of Equation (3), as given in the Appendix. We would also like to thank Chris Jennison for useful discussions. MAH acknowledges a Nuffield Foundation Award, MAH and HR acknowledge a grant for collaborative research from the British Council and the Research Council of Norway.

APPENDIX: PROOF OF EQUATION (3)

We give the proof of detailed balance based on the stopping rule described in Section 2. We then give conditions under which our proof will hold for more general stopping rules.

States $\boldsymbol{x} \in \Omega$ will be referred to as legal configurations and states $\boldsymbol{x} \in \Omega^* \setminus \Omega$ as illegal. The proposal distribution of $\boldsymbol{x}$ from $\boldsymbol{x}'$ is the probability that after leaving $\boldsymbol{x}'$ the first legal configuration the reversible MCMC sampler for $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ reaches is $\boldsymbol{x}$. Denote this probability by $P^*(\boldsymbol{x}' \to \boldsymbol{x})$. Let $\Gamma_B = \{\gamma\}$ be the set of paths in Ω^* which begin at $\boldsymbol{x}'$ and finish at $\boldsymbol{x}$ without visiting any legal configuration in between (and recalling that only components in B may change). Then

$$q_B(\boldsymbol{x}', \boldsymbol{x}) = \sum_{\gamma \in \Gamma_B} P^*(\boldsymbol{x}' \to \boldsymbol{x} \text{ along } \gamma). \tag{10}$$

Let Γ_B^- be the set of paths which begin at $\boldsymbol{x}$ and finish at $\boldsymbol{x}'$ without visiting any legal configuration en route. By tracing each path backwards there is a one-to-one correspondence between Γ_B and Γ_B^-. It is therefore sufficient to consider a particular path and its reverse, $\gamma = \{\boldsymbol{x}' \to \boldsymbol{x}^1 \to \boldsymbol{x}^2 \to \ldots \to \boldsymbol{x}^n \to \boldsymbol{x}\}$ and $\gamma^- = \{\boldsymbol{x} \to \boldsymbol{x}^n \to \ldots \boldsymbol{x}^1 \to \boldsymbol{x}'\}$. Now

$$P^*(\boldsymbol{x}' \to \boldsymbol{x} \text{ along } \gamma) = P(\boldsymbol{x}', \boldsymbol{x}^1) P(\boldsymbol{x}^1, \boldsymbol{x}^2) \cdots P(\boldsymbol{x}^n, \boldsymbol{x})$$

where $P(\boldsymbol{x}^i, \boldsymbol{x}^j)$ is the one-step transition probability of the MCMC sampler satisfying detailed balance with respect to $\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$, that is

$$P(\boldsymbol{x}^i, \boldsymbol{x}^j)\ \pi^*(\boldsymbol{x}_B^i|\boldsymbol{x}_{-B}^i) = P(\boldsymbol{x}^j, \boldsymbol{x}^i)\ \pi^*(\boldsymbol{x}_B^j|\boldsymbol{x}_{-B}^j)$$

where $\boldsymbol{x}^i_{-B} = \boldsymbol{x}^j_{-B}$. Thus,

$$\begin{aligned} P^*(\boldsymbol{x}' \to \boldsymbol{x} \text{ along } \gamma) &= P(\boldsymbol{x}^1, \boldsymbol{x}') \frac{\pi^*(\boldsymbol{x}_B^1|\boldsymbol{x}_{-B}^1)}{\pi^*(\boldsymbol{x}_B'|\boldsymbol{x}_{-B}')}\ P(\boldsymbol{x}^2, \boldsymbol{x}^1) \frac{\pi^*(\boldsymbol{x}_B^2|\boldsymbol{x}_{-B}^2)}{\pi^*(\boldsymbol{x}_B^1|\boldsymbol{x}_{-B}^1)} \cdots \\ &\qquad \cdots\ P(\boldsymbol{x}, \boldsymbol{x}^n) \frac{\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})}{\pi^*(\boldsymbol{x}_B^n|\boldsymbol{x}_{-B}^n)} \\ &= \frac{\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})}{\pi^*(\boldsymbol{x}_B'|\boldsymbol{x}_{-B}')}\ P(\boldsymbol{x}, \boldsymbol{x}^n) \cdots P(\boldsymbol{x}^1, \boldsymbol{x}'). \end{aligned} \tag{11}$$

Using (11) in (10) gives the desired result

$$q_B(\boldsymbol{x}', \boldsymbol{x}) = \sum_{\gamma \in \Gamma_B} \frac{\pi^*(\boldsymbol{x}_B|\boldsymbol{x}_{-B})}{\pi^*(\boldsymbol{x}_B'|\boldsymbol{x}_{-B}')} P^*(\boldsymbol{x} \to \boldsymbol{x}' \text{ along } \gamma^-) = \frac{\pi^*(\boldsymbol{x})}{\pi^*(\boldsymbol{x}')}\ q_B(\boldsymbol{x}, \boldsymbol{x}')$$

since $\boldsymbol{x}_{-B} = \boldsymbol{x}'_{-B}$.

We can use a more general rule to terminate the secondary chain by defining some stopping time $T(\boldsymbol{x}', \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n, \boldsymbol{x})$. This stopping time has to maintain the one-to-one correspondence between the paths from $\boldsymbol{x}'$ to $\boldsymbol{x}$ and the reverse route $\boldsymbol{x}$ to $\boldsymbol{x}'$, Γ_B and Γ_B^- with

$$T(\boldsymbol{x}', \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n, \boldsymbol{x}) = TRUE = T(\boldsymbol{x}, \boldsymbol{x}^n, \boldsymbol{x}^{n-1}, \ldots, \boldsymbol{x}^1, \boldsymbol{x}'),$$

but with

$$T(\boldsymbol{x}', \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^k) = FALSE = T(\boldsymbol{x}, \boldsymbol{x}^n, \boldsymbol{x}^{n-1}, \ldots, \boldsymbol{x}^k),$$

for all intermediate states $\boldsymbol{x}^k$ in the path.

An example of a valid stopping rule would be to wait for the m'th legal state or a fixed number of steps n_{max}, whichever occurs sooner. An example of an invalid stopping rule would be starting from $\boldsymbol{x}$ to wait until the first state $\boldsymbol{x}'$ of higher probability under $\pi(\boldsymbol{x})$, $\pi(\boldsymbol{x}') \geq \pi(\boldsymbol{x})$; this rule is invalid because chains starting at $\boldsymbol{x}'$ will not terminate on reaching $\boldsymbol{x}$.

REFERENCES

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, Series B*, **36**, 192-236.

Geman, D. (1990), *Random Fields and Inverse Problems in Imaging*, (Springer, Berlin).

Geman, S., Geman, D., Graffigne, C. and Dong, P. (1990), Boundary detection by constrained optimization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 609-628.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1995), *Practical Markov Chain Monte Carlo*, (Chapman and Hall, London).

Green, P. (1986), Contribution to the Discussion of Besag (1986), *Journal of the Royal Statistical Society, Series B*, **48**, 284-285.

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

Neal, R. (1996), Sampling from multimodal distributions using tempered transitions, *Statistics and Computing*, **6**, 353-366.

Sheehan, N. and Thomas, A. (1993), On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme, *Biometrics*, **49**, 163-175.

TABLE 1: Monitored statistics for various choices of $\pi^*(\boldsymbol{x})$: the total number of single component updates used; the average excursion length in terms of updates per block; the total number of swaps between predominantly 0 and predominantly 1 configurations; the average number of updates required per swap; the minimum, average and maximum number of 1-components; the minimum, average and maximum edge lengths.

	$\beta_2 = 0.6$ $\delta = 0.0$	$\beta_2 = 0.6$ $\delta = 0.3$	$\beta_2 = 0.6$ $\delta = 0.6$	$\beta_2 = 0.3$ $\delta = 0.0$	$\beta_2 = 0.3$ $\delta = 0.3$	$\beta_2 = 0.3$ $\delta = 0.6$	$\beta_2 = 0.0$ $\delta = 0.6$
No. of updates	2 049 546	2 048 170	2 048 012	2 048 032	2 048 020	2 048 000	2 048 003
Mean no. of updates per block	49.57	13.18	4.90	12.38	4.97	2.98	7.57
No. of 0/1 swaps	38	50	119	45	78	98	33
Mean no. of updates per swap	53 935	40 963	17 210	45 512	26 257	20 898	62 061
Min no. of 1's	1 841	1 740	1 694	1 823	1 820	1 727	1 755
Mean no. of 1's	2 068.1	2 092.7	2 016.4	2 064.2	2 051.3	2 008.1	2 010.5
Max no. of 1's	2 291	2 343	2 313	2 348	2 378	2 269	2 240
Min edge length	2 674	2 640	2 584	2 656	2 636	2 630	2 628
Mean edge length	2 820.9	2 817.2	2 795.5	2 831.2	2 800.8	2 817.5	2 791.4
Max edge length	3 332	3 338	3 360	3 328	3 328	3 338	3 330