Stochastic Processes and their Applications 49 (1994) 207-216
North-Holland

Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms

G.O. Roberts

University of Cambridge, UK

A.F.M. Smith

Imperial College London, UK

Received 15 July 1992
Revised 11 February 1993

Markov chain Monte Carlo (MCMC) simulation methods are being used increasingly in statistical computation to explore and estimate features of likelihood surfaces and Bayesian posterior distributions. This paper presents simple conditions which ensure the convergence of two widely used versions of MCMC, the Gibbs sampler and Metropolis–Hastings algorithms.

Markov chain Monte Carlo * Gibbs sampler * Metropolis–Hastings algorithm * statistical computation * ergodicity * lower semicontinuity

1. Introduction

The basic idea of Markov chain Monte Carlo (MCMC) is very straightforward. We wish to generate random variates from a distribution with density $\pi(x)$ for $x \subseteq \mathscr{X} \subset \mathbb{R}^n$ but cannot do this directly, typically because n is large or the form of $\pi(x)$ is intractable. Instead, we construct a Markov chain with state space $\mathscr{X}$ and equilibrium distribution $\pi(x)$ and simulate a long run of the chain. Following an initial, transient phase, realized values of the chain can be used as a basis for summarizing features of $\pi(x)$ of interest. All we require are algorithms for constructing appropriately behaved chains.

An early version of such a MCMC algorithm was given by Metropolis et al. (1953) in a statistical physics context, with subsequent generalization by Hastings (1970), who focused on statistical problems. A version particularly suited to certain problems in spatial statistics and Bayesian image analysis (the so-called Gibbs sampler) was introduced by Geman and Geman (1984), and subsequently shown by Gelfand and Smith (1990) to have great potential for general Bayesian computation. A number of recent review papers provide

Correspondence to: Dr. G.O. Roberts, Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, UK.

SSDI 0304-4149(93)00016-8

surveys of developments and the rapidly growing literature of the subject: for statistical physics, see Sokal (1989) and Gidas (1992); for spatial statistics and Bayesian computation, see Besag and Green (1993); for general Bayesian computation, see Smith and Roberts (1993); for applications to medical statistics, see Gilks et al. (1993).

However, study of this large and growing literature reveals a major deficiency. Whereas the structural properties required for a constructed chain to converge appropriately are well understood (essentially irreducibility and aperiodicity), simple conditions are not available which relate these to the analytic features of the target distribution, $\pi(x)$. The aim of this paper is to provide such simple conditions for the Gibbs sampler and Metropolis–Hastings algorithms, concentrating on conditions sufficient for typical applications rather than abstract generality.

A number of important theoretical questions concerning MCMC remain unresolved. In particular, useful analytical bounds on convergence rates remain elusive in general; however, conditions ensuring geometric ergodicity can sometimes be given (see, for example, Chan, 1993, or Roberts and Polson, 1994). Ergodic central limit theorems do exist (see, for example, Geyer, 1993), however, they involve regularity conditions which are usually extremely difficult to check. In summary, there has hitherto been a large gulf between the general results provided by probability theory, and their ready applicability in our context. While confining our attention to the irreducibility issue, we attempt to bring together the two areas.

We proceed as follows. In Section 2, we provide an informal description of the algorithms and, in Section 3, we set them in an appropriate formal framework. Convergence of the Gibbs sampler is then examined in detail in Section 4 and convergence of the Metropolis–Hastings algorithm in Section 5.

2. Informal description of the algorithms

The following is intended as an informal description in order to give the 'flavour' of the basic algorithms. A formal mathematical discussion is given in Section 3. Here, it suffices for the reader to think in terms of finite, discrete $\mathscr{X}$.

2.1. Gibbs sampler algorithm

Let $\pi(x) = \pi(x_1, \ldots, x_k)$, $x \in \mathbb{R}^n$, $1 < k \leqslant n$, denote the target density, where, for $i = 1, \ldots, k$, $x_i = (x_{i1}, \ldots, x_{in(i)})$, $n(i) \geqslant 1$, $n(1) + \cdots + n(k) = n$, and the x_{ij} are scalar components of x, and let $\pi(x_i | x_{-i})$ denote the induced conditional densities for each of the component subvectors x_i, given values of the other components $x_{-i} = (x_j; j \neq i)$, $i = 1, \ldots, k$.

A systematic form of the Gibbs sampler algorithm proceeds as follows. First, pick arbitrary starting values $x^0 = (x_1^0, \ldots, x_k^0)$. Then, successively make random variate drawings from each of $\pi(x_i | x_{-i})$ in turn, as follows:

$$\begin{array}{ll} x_1^1 & \text{from} \quad \pi(x_1 \,|\, x_{-1}^0)\ ; \\ x_2^1 & \text{from} \quad \pi(x_2 \,|\, x_1^1, x_3^0, \ldots, x_k^0)\ ; \\ x_3^1 & \text{from} \quad \pi(x_3 \,|\, x_1^1, x_2^1, x_4^0, \ldots, x_k^0)\ ; \\ & \vdots \\ x_k^1 & \text{from} \quad \pi(x_k \,|\, x_{-k}^1)\ . \end{array}$$

This defines a transition mechanism from x^0 to $x^1 = (x_1^1, \ldots, x_k^1)$. Iteration of this process generates a sequence $x^0, x^1, \ldots, x^t, \ldots$ which is a realization of a Markov chain with transition probability from x to y given by

$$\prod_{l=1}^{k} \pi(y_l \,|\, x_j, j>l, y_j, j<l)\ .$$

We note that the algorithm is defined by the choice of blocking $(x_1, \ldots, x_k)$ of $x \in \mathbb{R}^n$ and the forms of $\pi(x_i | x_{-i})$ induced by $\pi(x)$.

2.2. Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm constructs a transition probability from $X^t = x$ to the next realized state X^{t+1} as follows. First, a (for the moment arbitrary) transition probability function $q(x, y)$ is chosen and, if $X^t = x$, y generated from $q(x, y)$ is considered as a candidate value for X^{t+1}. Secondly, an additional probability function

$$\alpha(x, y) = \begin{cases} \min\left\{ \dfrac{\pi(y) q(y, x)}{\pi(x) q(x, y)}, 1 \right\} & \text{if } \pi(x) q(x, y) > 0\,, \\ 1 & \text{if } \pi(x) q(x, y) = 0\,, \end{cases}$$

is defined and we set $X^{t+1} = y$ with probability $\alpha(x, y)$, otherwise, with probability $1 - \alpha(x, y)$, setting $X^{t+1} = y$. This procedure defines a Markov chain with transition probability from x to y given by

$$\begin{cases} q(x, y)\alpha(x, y) & \text{if } y \neq x\,, \\ 1 - \sum\limits_{Z} q(x, z)\alpha(x, z) & \text{if } y = x\,. \end{cases}$$

The Metropolis et al. (1953) algorithm corresponds to forms of the above with $q(x, y) = q(y, x)$. For discussion of other choices of $q(x, y)$ forms, see the references given in Smith and Roberts (1993).

3. Mathematical framework

Let $X = (X^0, X^1, \ldots, X^t, \ldots)$, $X^t \in E \subseteq \mathbb{R}^n$, be a Markov chain (MC) with transition kernel $K : E \times E \to \mathbb{R}^+$ such that, with respect to a σ-finite measure ν on the Borel σ-field of $\mathbb{R}^n$, for ν-measurable A,

$$P(X^t \in A \mid X^{t-1} = x) = \int_A K(x, y)\, \mathrm{d}\nu(y) + r(x) I[x \in A] ,$$

where

$$r(x) = 1 - \int_E K(x, y)\, \mathrm{d}\nu(y) .$$

Note that K is substochastic, describing only the accepted iterations. We assume that K is non-degenerate, so that $r(x) < 1$ for all $x \in E$. The iterative form

$$K^{(t)}(x, y) = \int_{\mathbb{R}^n} K^{(t-1)}(x, z) K(z, y)\, \mathrm{d}\nu(z)$$
$$+ K^{(t-1)}(x, y) r(y) + [1 - r(x)]^{t-1} K(x, y) , \qquad (1)$$

is then a (possibly substochastic) kernel describing t-step transitions which involve at least one accepted move and $K_x^{(t)}(\cdot) = K^t(x, \cdot)$ is the density (with respect to ν) of X^t, given $X^0 = x$, excluding realizations with $X^j = x, j = 1, \ldots, t$.

Assuming the existence of a density (with respect to ν), and using notation for distribution and density interchangeably, we recall that an invariant distribution π for this MC satisfies

$$\pi(A) = \int P(X^1 \in A \mid X^0 = x) \pi(x)\, \mathrm{d}\nu(x) ,$$

for all ν-measurable A. Defining $D = \{x \in E;\ \pi(x) > 0\}$, we also recall that K is called π-irreducible if, for all $x \in D$, $\pi(A) > 0$ implies that we have $P(X^t \in A \mid X^0 = x) > 0$ for some $t \geqslant 1$, and is called aperiodic if there does not exist a ν-measurable partition $E = (B_0, \ldots, B_{r-1})$, for some $r \geqslant 2$, such that $P(X^t \in B_{t \bmod (r)} \mid X^0 = x^{(0)} \in B_0) = 1$ for all t. Writing $|g| = \int_E |g(x)|\, \mathrm{d}\nu(x)$ for all ν-measurable functions g defined on E, a key result regarding the ergodic behaviour of the MC is the following.

Theorem 1. *If K is π-irreducible and aperiodic then, for all $x \in D$,*
(i) $|K_x^{(t)} - \pi| \to 0$ *as* $t \to \infty$;
(ii) *for real-valued, π-integrable f,*

$$t^{-1}\{f(X^1) + \cdots + f(X^t)\} \to \int_E f(x) \pi(x)\, \mathrm{d}\nu(x) \quad a.s. \quad as\ t \to \infty .$$

Proof. See Tierney (1991), based on Nummelin (1984). □

In the next two sections, we shall re-examine, against this formal background, the two MCMC algorithms introduced informally in Section 2. We shall see that, by construction, in each case the target distribution π is an invariant distribution of the chain. The key issue, therefore, will be to identify conditions which ensure π-irreducibility and aperiodicity. This we accomplish by establishing simple (and very weak) sufficient conditions.

4. Convergence conditions for the Gibbs sampler algorithm

Assuming an underlying product measure $\mathrm{d}\nu(x) = \mathrm{d}\nu_1(x_1)\cdots\mathrm{d}\nu_k(x_k)$, for the blocking $x = (x_1, \ldots, x_k)$, $1 < k \leqslant n$, the Gibbs sampler construction requires that the conditionals

$$\pi(x_i \mid x_{-i}) = \pi(x) \bigg/ \int \pi(x)\ \mathrm{d}\nu_i(x_i)\ ,$$

$i = 1, \ldots, k$, be well-defined over the appropriate support regions. With $D = \{x \in E;\ \pi(x) > 0\}$, we seek to define $K_{\mathrm{G}} : D \times D \to \mathbb{R}^n$ by

$$K_{\mathrm{G}}(x, y) = \prod_{l=1}^{k} \pi(y_l \mid x_j, j > l, y_j, j < l)\ ,$$

provided $\int \pi(y_1, \ldots, y_i, x_{i+1}, \ldots, x_k)\ \mathrm{d}\nu_i(y_i) > 0$, and $K_{\mathrm{G}}(x, y) = 0$ otherwise. We therefore assume that $x^{(0)} \in D$. It is then straightforward to check that, when K_{G} is a well-defined kernel, π is an invariant distribution of the chain applied by K_{G}. In what follows, we shall focus on the two important special cases where ν is either a discrete measure (so that D is a lattice in $\mathbb{R}^n$) or n-dimensional Lebesgue measure. Other cases can be treated similarly.

Lemma 1. *If ν is discrete, then K_{G} is well-defined and π-irreducibility of K_{G} is a sufficient condition for the results of Theorem* 1.

Proof. Well-definedness of K_{G} is trivial in the discrete case and the fact that $K_{\mathrm{G}}(x, x) > 0$ for all $x \in D$ implies that K_{G} is aperiodic. □

This concludes the discussion of the discrete case. All that is required is to check that the structure of the discrete transition probability matrix ensures π-irreducibility.

Discussion of the continuous (Lebesgue measure) case is necessarily more technical, but the resulting sufficient conditions will be seen to be very simple. As a preliminary, we define a function $h : \mathbb{R}^n \to \mathbb{R}^+$ to be lower semicontinuous at 0 (l.s.c. at 0) if, for all x with $h(x) > 0$, there exists an open neighbourhood $N_x \ni x$ and $\varepsilon > 0$ such that, for all $y \in N_x$, $h(y) \geqslant \varepsilon > 0$.

Lemma 2. *If ν is n-dimensional Lebesgue measure, and π is l.s.c. at* 0, *then K_{G} is well-defined.*

Proof. For all $x = (x_1, \ldots, x_k) \in D$, l.s.c. at 0 implies that $\int \pi(x)\ \mathrm{d}x_i > 0$ for $i = 1, \ldots, k$. □

We note immediately that l.s.c. at 0 is an extremely weak condition (weaker than lower semicontinuity), but makes intuitive sense in the Gibbs kernel context by ensuring both probability mass around component points (required for the conditionals) and around points in $\mathbb{R}^n$ (required for π).

The next result establishes technical conditions for K_{G} to be aperiodic. As a preliminary, we define, for $x \in D$,

$$B_x^{(t)} = \{y \in D;\ K_{\mathrm{G}}^{(t)}(x, y) > 0\}\ ,$$

where $K_{\mathrm{G}}^{(t)}$ is the iterative form of Section 3 derived from K_{G}.

Lemma 3. *Suppose that π is l.s.c. at 0 and that, for $i = 1, \ldots, k$, $\int \pi(x)\ \mathrm{d}x_i$ is locally bounded on D. Then,*

(i) *$K_{\mathrm{G}}(x, \cdot)$ and $K_{\mathrm{G}}(\cdot, x)$ are both l.s.c. at 0 for all $x \in D$;*

(ii) *$K_{\mathrm{G}}^{(t)}(x, \cdot)$ is l.s.c. at 0 for all $x \in D$ and for all $t \geqslant 1$;*

(iii) *$B_x^{(t)}$ is open in D for all $x \in D$ and for all $t \geqslant 1$;*

(iv) *$B_x^{(t)} \subset B_x^{(t+1)}$ for all $t \geqslant 1$, so that K_{G} is aperiodic.*

Proof. (i) By definition,

$$K_{\mathrm{G}}(x, y) = \frac{\prod_{l=1}^{k} \pi(y_j, j \leqslant l, x_j, j > l)}{\prod_{l=1}^{k} \pi(y_j, j < l, x_j, j > l)}\ .$$

It is straightforward to check that the product of functions l.s.c. at 0 is also a function l.s.c. at 0, so that there exists open neighbourhoods N_x of x and N_y of y and $\varepsilon > 0$ such that

$$\prod_{l=1}^{k} \pi(w_j, j \leqslant l, z_j, j > l) \geqslant \varepsilon > 0 \quad \text{for all } w \in N_y,\ z \in N_x\ .$$

But, by the local boundedness of $\int \pi(x)\ \mathrm{d}x_i$ for $i = 1, \ldots, k$, there exist open neighbourhoods M_x of x and M_y of y and constant $C > 0$ such that

$$\prod_{l=1}^{k} \pi(w_j, j < l, z_j, j > l) \leqslant C \quad \text{for all } w \in M_y,\ z \in M_x\ .$$

The results now follows since

$$K_{\mathrm{G}}(w, z) \geqslant \varepsilon C^{-1} > 0 \quad \text{for all } w \in M_y \cap N_y,\ z \in M_x \cap N_x\ .$$

(ii) We proceed by contradiction and assume that for some x, $y \in D$ and $t \geqslant 1$ with $K_{\mathrm{G}}^{(t)}(x, y) > 0$, there exists a sequence $\{y^{(j)} \in D,\ j \geqslant 1\}$ converging to y, but with $\limsup_{j \to \infty} K_{\mathrm{G}}^{(t)}(x, y^{(j)}) = 0$. We note that

$$K_{\mathrm{G}}^{(t)}(x, y^{(j)}) = \int K_{\mathrm{G}}^{(t-1)}(x, z) K_{\mathrm{G}}(z, y^{(j)})\ \mathrm{d}z\ ,$$

and since, by (i), $K_{\mathrm{G}}(z, \cdot)$ is l.s.c. at 0, we note there exists $\varepsilon(z) > 0$ such that, for all z with $K_{\mathrm{G}}(z, y) > 0$, $\liminf_{j \to \infty} K_{\mathrm{G}}(z, y^{(j)}) \geqslant \varepsilon(z) > 0$. By Fatou's lemma we then have

$$\begin{aligned} \limsup_{j \to \infty} K_{\mathrm{G}}^{(t)}(x, y^{(j)}) &\geqslant \int \liminf_{j \to \infty} K_{\mathrm{G}}^{(t-1)}(x, z) K_{\mathrm{G}}(z, y^{(j)})\ \mathrm{d}z \\ &\geqslant \int_S K_{\mathrm{G}}^{(t-1)}(x, z) \varepsilon(z)\ \mathrm{d}z > 0\ , \end{aligned}$$

since $S = \{z;\ K_{\mathrm{G}}^{(t-1)}(x, z) K_{\mathrm{G}}(z, y) > 0\}$ is non-null by virtue of the assumption $K_{\mathrm{G}}^{(t)}(x, y) > 0$.

(iii) This is immediate from (ii) and the definition of $B_x^{(t)}$.

(iv) For any $y \in B_x^{(t)}$, $x \in D$, $t \geqslant 1$, $K_{\mathrm{G}}(\cdot, y)$ is l.s.c. at 0, by (i), and since $K_{\mathrm{G}}(y, y) > 0$ there exists an open neighbourhood, $N_y \subseteq B_x^{(t)}$, of y and $\varepsilon(y) > 0$ such that, for all $z \in N_y$, $K_{\mathrm{G}}(z, y) \geqslant \varepsilon(y) > 0$. It follows that

$$K_{\mathrm{G}}^{(t+1)}(x, y) = \int K_{\mathrm{G}}^{(t)}(x, z) K_{\mathrm{G}}(z, y)\, \mathrm{d}z \geqslant \varepsilon(y) \int_{N_y} K_{\mathrm{G}}^{(t)}(x, z)\, \mathrm{d}z > 0\,,$$

by virtue of the l.s.c. at 0 of $K_{\mathrm{G}}^{(t)}(x, \cdot)$, established in (ii). It follows from the definition that $y \in B_x^{(t+1)}$. □

As a preliminary to establishing conditions for the π-irreducibility of K_{G}, we note that the latter corresponds to $B_x^{(\infty)} = D$, for all $x \in D$, where $B_x^{(\infty)} = \lim_{t \to \infty} B_x^{(t)}$. This motivates the technical result of Lemma 4, which in turn requires the following preliminary definition. Given the Gibbs sampler blocking $x = (x_1, \ldots, x_k)$, we define

$$\begin{aligned} \mathscr{H} = \{ & \text{closed hyper-rectangles } H \subseteq D, \text{ with hypersurfaces} \\ & \text{defined by } x_{ij} = \text{constant}, j = 1, \ldots, n(i), i = 1, \ldots, k\}\,. \end{aligned}$$

Lemma 4. *Under the conditions of Lemma 3, if $H \in \mathscr{H}$ then, for all $x \in D$, either $H \subseteq B_x^{(\infty)}$ or $H \subseteq B_x^{(\infty)c}$.*

Proof. Suppose, for an $H \in \mathscr{H}$, and $x \in D$, that $H \cap B_x^{(\infty)} \neq \emptyset$. If $w \in H \cap B_x^{(\infty)}$ then $w \in B_x^{(t)}$ for some $t \geqslant 1$ and, since $K_{\mathrm{G}}^{(t)}(x, \cdot)$ is l.s.c. at 0, there exists an open neighbourhood $N_w \subseteq B_x^{(t)} \cap H$ and $\varepsilon_1 > 0$ such that, for all $z \in N_w$, $K_{\mathrm{G}}^{(t)}(x, z) \geqslant \varepsilon_1 > 0$. Now consider any $y \in H$. Since $K_{\mathrm{G}}(z, y) > 0$, by compactness there exists $\varepsilon_2 > 0$ such that $K_{\mathrm{G}}(z, y) \geqslant \varepsilon_2 > 0$. It follows that

$$K_{\mathrm{G}}^{(t+1)}(x, y) = \int K_{\mathrm{G}}^{(t)}(x, z) K_{\mathrm{G}}(z, y)\, \mathrm{d}z \geqslant \varepsilon_1 \varepsilon_2 \int_{N_w} \mathrm{d}z > 0\,,$$

so that $y \in B_x^{(t+1)} \subset B^{(\infty)}$. □

Sufficient conditions for convergence of the Gibbs sampler with dominating n-dimensional Lebesgue measure are then given by the following.

Theorem 2. *If ν is n-dimensional Lebesgue measure, π is l.s.c. at 0, $\int \pi(x)\, \mathrm{d}x_i$ is locally bounded for $i = 1, \ldots, k$, and D is connected, the results of Theorem 1 apply to K_{G}.*

Proof. The aperiodicity of K_{G} is established by (iv) of Lemma 3.

To establish π-irreducibility, i.e. $B_x^{(\infty)} = D$ for all $x \in D$, we proceed by contradiction by supposing that, given $x \in D$, there exists $z \in D$ with $z \notin B_x^{(\infty)}$. Since D is connected, there exists a continuous function $g: [0, 1] \to D$ such that $g(0) = x$, $g(1) = z$. Let $t_x = \inf\{t \in [0, 1];\ g(t) \notin B_x^{(\infty)}\}$. Since $K_{\mathrm{G}}(x, x) > 0$ and $K_{\mathrm{G}}(x, \cdot)$ is l.s.c. at 0, we must have $t_x > 0$.

Also, since $B_x^{(\infty)}$ is open, $g(t_x)\notin B_x^{\infty}$. Let N be an open neighbourhood of $g(t_x)$ and let $\varepsilon=\inf_{y\in N^c}\{|t_x-y|\}$. It follows that there exists a hypercube $H\in\mathscr{H}$, $H\subset N$ of side $<(4\varepsilon^2/n)^{1/2}$ and containing $g(t_x)$. This is a contradiction, by Lemma 4, since $H\not\subseteq B^{(\infty)}$ and $H\not\subseteq B^{(\infty)c}$. □

Connectedness of D is an intuitively obvious sufficient condition for π-irreducibility but is clearly not a necessary condition. However, it has the advantage of being sufficient whatever the form of blocking or continuous one-to-one transformation of x adopted in defining K_G (as is easily seen by consideration of the proof of Theorem 2). The following result, essentially a necessary and sufficient condition, provides a significant generalization given a fixed parametrization and blocking of x.

Corollary 1. *Suppose* $D=\bigcup_{i=1}^{\infty} D_i$, *where each* D_i *is connected and suppose that, for any two components* D_{i_1} *and* D_{i_2}, *there exists an integer*, $m(i_1, i_2)$, *and a sequence of integers*, $\{j(0), j(1), \ldots, j(m(i_1, i_2))\}$ *such that* $j(0)=i_1$, $j(m(i_1, i_2))=i_2$, *and for each* $0\leqslant r\leqslant m(i_1, i_2)-1$, *there exist* $x^{(r)}\in D_{j(r)}$, $y^{(r)}\in D_{j(r+1)}$, *such that* $x^{(r)}$ *and* $y^{(r)}$ *differ by only one component in the chosen Gibbs sampler blocking. Then* K_G *is* π-*irreducible.*

Proof. Since Lemma 4 implies, for $i=1, 2, \ldots$, that, for any $x\in D$, either $D_i\subset B^{(\infty)}$ or $D_i\subset B^{(\infty)c}$, the result follows by induction on $m(i_1, i_2)$ (the number of moves required to link the two components, D_{i_1} and D_{i_2}). □

It should be noted that in many applications, the level of generality given here is not needed. In most cases, $D=\{x\,|\,\pi(x)>0\}$ is a product set $D=\prod_{i=1}^{k} D_i$, and then everything simplifies:

– $\pi(x_i|x_{-i})$ is well-defined for all $x\in D$ and $i=1, \ldots, k$, and so the kernel K_G is clearly well-defined.

– Using Fubini's theorem, it is easily seen that

$$P(X^1\in A\,|\,X^0=x)>0 \quad \text{whenever } x\in D \text{ and } \pi(A)>0\,,$$

and, hence, it follows that K_G is π-irreducible and aperiodic.

On the other hand, there exist important special cases where D is not a product set, e.g., hardcore-point processes as used in spatial statistics. In such cases, Corollary 1 becomes important since D may not be connected but will typically satisfy the conditions of the corollary.

5. Convergence conditions for the Metropolis–Hastings algorithm

Recalling the informal introduction of the algorithm given in Section 2.2, we now formally take $q: D\times D\to\mathbb{R}^+$ to be a Markov chain kernel (with respect to ν) with $D=\{x\in\mathbb{R}^n;\ \pi(x)>0\}$. With $\alpha: D\times D\to[0, 1]$ as defined in Section 2.2, we define $K_H: D\times D\to\mathbb{R}^+$ by

$$K_H(x, y)=q(x, y)\alpha(x, y)\ .$$

This is a (substochastic) kernel governing moves of the chain $X^0, X^1, \ldots, X, \ldots$ from x to

y which are 'accepted' according to the probability $\alpha(x, y)$. It is straightforward to check that π is an invariant distribution of the chain defined by K_{H}. For general measures ν, the convergence properties of the Metropolis–Hastings algorithm are inherited from the properties of q as follows.

Theorem 3. (i) *If q is aperiodic; or $P(X^t = X^{t-1}) > 0$ for some $t \geqslant 1$, then the Metropolis–Hastings algorithm is aperiodic.*

(ii) *If q is π-irreducible and $q(x, y) = 0$ if, and only if, $q(y, x) = 0$, then the Metropolis–Hastings algorithm is π-irreducible.*

Proof. (i) If the Metropolis–Hastings chain is periodic then we clearly cannot have $P(X^t = X^{t-1}) > 0$ for any $t \geqslant 1$. The chain is therefore driven by q and is thus aperiodic.

(ii) The condition $q(x, y) = 0$, if, and only if, $q(y, x) = 0$ implies that $\alpha(x, y) > 0$ for all $x, y \in D$. Let $K_{\mathrm{H}}^{(t)}$, $q^{(t)}$ denote the iterated kernels obtained by setting K equal to K_{H} and q, respectively, in (1), and define $U_x^{(t)} = \{y, K_{\mathrm{H}}^{(t)}(x, y) > 0\}$, $V_x^{(t)} = \{y; q^{(t)}(x, y) > 0\}$. We show by induction that $U_x^{(t)} \supseteq V_x^{(t)}$ for all $t \geqslant 1$. Suppose, therefore, that $U_x^{(t)} \supseteq V_x^{(t)}$ and consider $z \in V_x^{(t+1)}$, which implies that

$$\int_{U_x^{(t)}} q^{(t)}(x, y) q(y, z) \; \mathrm{d}\nu(y) > 0 \, .$$

However, if $z \notin U_x^{(t+1)}$ the support of the function

$$q^{(t)}(x, \cdot) q(\cdot, z) \alpha(\cdot, z) \, ,$$

has ν-measure 0, which implies that the support of the function

$$K_{\mathrm{H}}^{(t)}(x, \cdot) q(\cdot, z) \, ,$$

also has ν-measure 0, contradicting the above inequality. Since $U_x^{(1)} \supseteq V_x^{(1)}$, the result follows. □

6. Discussion

We begin with the Gibbs sampler algorithm and simply comment on the extremely weak nature of the conditions we have shown to be sufficient for convergence. Indeed, it is a challenging problem in itself to think of any functions arising in conventional statistical modelling for which the lower semi-continuity at 0 and locally bounded conditions do not hold.

To illustrate the kind of problem that could arise, consider $x = (x_1, x_2)$ with $\pi(x)$ uniform on the region

$$S = \{(x_1, x_2); \; -1 \leqslant x_1 \leqslant 1, \, 0 \leqslant x_2 \leqslant x_1^{-1/2}\} \, .$$

Clearly, problems arise when x_1 is close to 0. However, simply defining $D = S \cap \{x; x_1 \neq 0\}$ solves the problem and amounts in practice to avoiding initializing the Gibbs sampler at

$x_1=0$. Other 'problem examples' that come to mind are similarly dealt with and suggest that the notion of restricting the starting values is essentially equivalent to finding a (ν a.e. equivalent) version of the density which is l.s.c. at 0 and choosing a starting value with positive density. This will preclude starting points on the boundary of D, which can lead to the Gibbs sampler 'getting stuck'.

In the case of the Metropolis–Hastings algorithm, a key issue is the relationship of the domain of definition of q to the support domain D of π. Our sufficient condition is that they coincide. To illustrate what can go wrong otherwise, consider, for example, π uniform on the region $\{(x, y);\ (x, y) \in [0, 1] \times [0, 1] \cup [1, 2] \times [1, 2]\}$ and suppose q is defined on $[0, 2] \times [0, 2]$ by

$$q\{(x_1, y_1), (x_1, y_2)\} = \tfrac{1}{4}, \quad 0 \leqslant y_2 \leqslant 2,$$

$$q\{(x_1, y_1), (x_2, y_1)\} = \tfrac{1}{4}, \quad 0 \leqslant x_2 \leqslant 2,$$

so that D is strictly contained in the domain of q. It is easily seen that the Metropolis–Hastings chain is reducible.

Acknowledgements

The second author acknowledges invaluable support from the SERC Complex Stochastic Systems Intitiative.

References

J. Besag and P.J. Green, Spatial statistics and Bayesian computation, J. Roy. Statist. Soc. Ser. B 55 (1993) 25–38.

K.S. Chan, Asymptotic behaviour of the Gibbs sampler, J. Amer. Statist. Assoc. 88 (1993) 320–326.

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

S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans. Pattn. Anal. Mach. Intell. 6 (1984) 721–741.

C.J. Geyer, Practical Markov Chain Monte Carlo (with discussion), Statist. Sci. 7(4) (1993) 457–511.

B. Gidas, Metropolis-type Monte Carlo simulation algorithms and simulated annealing, in: Doyle and Snell, eds., Trends in Contemporary Probability (1992).

W.R. Gilks, D.G. Clayton, D.J. Spiegelhalter, N.G. Best, A.J. McNeil, L.D. Sharples and A.J. Kirby, Modelling complexity: applications of Gibbs sampling in medicine, J. Roy. Statist. Soc. Ser. B 55 (1993) 39–52.

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

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equations of state calculations by fast computing machine, J. Chem. Phys. 21 (1953) 1087–1091.

E. Nummelin, General Irreducible Markov Chains and Non-negative Operators (Cambridge Univ. Press, Cambridge, 1984).

G.O. Roberts and N.G. Polson, On the geometric convergence of the Gibbs sampler, to appear in: J. Roy. Statist. Soc. Ser. B 56 (1994).

A.F.M. Smith and G.O. Roberts, Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo Methods, J. Roy. Statist. Soc. Ser. B 55 (1993) 3–24.

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

L. Tierney, Markov Chains for exploring posterior distributions, Tech. Rept., School of Statist., Univ. of Minnesota (Minneapolis, MN, 1991).