GEOMETRIC CONVERGENCE OF THE METROPOLIS–HASTINGS SIMULATION ALGORITHM

LARS HOLDEN
NORWEGIAN COMPUTING CENTER AND UNIVERSITY OF OSLO

ABSTRACT. Necessary and sufficient conditions for geometric convergence of the Metropolis–Hastings simulation algorithm with a general generation function are established. If these conditions are violated, then the algorithm does not in general converge. An explicit expression for the convergence rate is found. The convergence rate depends heavily on the size of the domain where the generation function is positive, a lower bound of the ratio between the generation function and the limiting function in this domain and the number of jumps necessary to jump between two arbitrary states. The results in the paper also give a qualitative understanding of the convergence rate.

1. Introduction. This paper discusses the convergence rate for the Metropolis–Hastings simulation algorithm proposed in Hastings (1970). The algorithm is a generalization of the Metropolis–algorithm, see Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953). Ripley (1987) gives a good overview of this and similar simulation algorithms. Meyn & Tweedie (1993) is a more technical and updated description of Markov chain theory. Diaconis & Saloff-Coste (1995) present some recent results on the Metropolis–Hastings algorithm.

The Metropolis–Hastings simulation algorithm is used for sampling from a distribution $f(x)$. It is only necessary to know $f(x)$ up to a constant, i.e. $f(x) = c \cdot h(x)$ where $h(x)$ is known and c is unknown. The Metropolis–Hastings simulation algorithm is a Markov chain Monte Carlo (MCMC) method. One starts with any initial value x^0 and then a sequence of values $x^{i+1} = F(x^i)$ is generated. Let $p^i(x)$ be the probability distribution after i iterations. There are several different proofs that $p^i(x) \to f(x)$ as $i \to \infty$: See for example Corollary 1 of Theorem 2.2 in Billingsley (1986) and Theorem 18.5.1 in Meyn & Tweedie (1993) under different regularity conditions. Under stronger assumptions the convergence rate is geometric with ratio equal to the next largest eigenvalue of $F(\cdot)$. This convergence has been studied in Frigessi, Hwang, Stefano & Sheu (1993), for example. However, the

Date: December 17, 1996.
1991 *Mathematics Subject Classification.* 60J27.
Key words and phrases. Markov chain Monte Carlo, Metropolis–Hastings algorithm, convergence.
Research supported by Research Council of Norway.

size of this eigenvalue for a particular Markov chain is difficult to quantify. Geometric convergence is proved in Mengersen & Tweedie (1994) when the generation function satisfies $q(x \mid y) = q(x) \geq a \cdot f(x)$. In Roberts & Tweedie (1996) it is proved geometric convergence in the total variation norm if the tails of the limiting distribution is sufficient light. In this paper necessary and sufficient conditions for geometric convergence in the relative supremum norm for a general generation function are established. An explicit formula for the convergence rate which is not too conservative, and for some cases optimal, is proved.

There is currently a lot of interest in the theory and applications of MCMC: See for instance Geyer (1992) and Geyer & Thompson (1995). The Metropolis–Hastings algorithm is used in a large number of applications. For several years the present author has applied the algorithm in the simulation of marked point processes including variable number of points: See for example Skare, Skorstad, Hauge & Holden (1996), where a complex marked point process model is described. The Metropolis–Hastings algorithm is used for simulation from the model by changing one point at a time. Traditionally, the position of a new point is drawn uniformly. The position of the points in the posterior distribution, given the seismic data, is far from uniform. Intuition told us that the convergence would improve considerably if the generation function was close to the limiting distribution. Hence, much effort was spent in finding an ad hoc generation function which satisfied the following three criteria:

- it is possible to simulate from the generation function
- it is possible to calculate the probability for generating a particular realization
- the generation function generates realizations with high probability.

The first two assumptions are needed, and the last was believed to be critical for the convergence. The results in this paper show that the convergence rate depends critically on how close the generation function is to the limiting distribution.

The relative supremum norm is used in the convergence results. The theorem and the results show that this is a natural norm for proving convergence. The algorithm also converges in other norms but this may be much more difficult to prove and the convergence rate may not be as good. In some cases the algorithm converges in other norms, such as $T.V.$, L_1 or L_∞, and not in the relative supremum norm because of properties in a domain $A \subset \Omega$ where $\int_A f(x)\,dx$ is small. If the tails of the distribution are not important, other norms may be better.

In Hektoen & Holden (1996) there is a similar theorem on the convergence rate for the sequential importance resampling (SIR) algorithm.

2. The Metropolis–Hastings simulation algorithm. Let $\Omega \subset \mathbb{R}^n$ be a Borel measurable state space and $f(x)$ a probability density which is positive in Ω. The densities $p^0(x)$ and $q(x \mid y)$, $x, y \in \Omega$ are positive in Ω or a subset of Ω. All the densities are assumed absolutely continuous.

METROPOLIS-HASTINGS ALGORITHM. To generate a sample from the probability density $f(x)$:

1. Generate an initial state $x^0 \in \Omega$ from the density $p^0(x)$.
2. For $i = 1, \ldots, n$:
   (a) Generate an alternative state y from the density $q(y \mid x^i)$.
   (b) Calculate
$$\alpha(y, x^i) = \min\left\{1, \frac{f(y)q(x^i \mid y)}{f(x^i)q(y \mid x^i)}\right\}.$$
   (c) Set
$$x^{i+1} = \begin{cases} y & \text{with probability } \alpha(y, x^i) \\ x^i & \text{with probability } 1 - \alpha(y, x^i). \end{cases}$$

In this paper it is assumed that $q(y \mid x) > 0$ implies that $q(x \mid y) > 0$ for all $x, y \in \Omega$ since states proposed by $q(y \mid x) > 0$ will not be accepted if $q(x \mid y) = 0$. The following definitions are needed:

$$\begin{aligned} \Omega(y) &= \{x \in \Omega;\ q(x \mid y) > 0\}, \\ h(x, y) &= \min\big\{f(x)\, q(y \mid x), f(y)\, q(x \mid y)\big\}, \\ Q(x, y) &= \frac{h(x, y)}{f(y)} = \min\left\{q(x \mid y), q(y \mid x)\frac{f(x)}{f(y)}\right\}. \end{aligned}$$

Notice that $\Omega(y)$ may have lower dimension than Ω. Integration over $\Omega(y)$ or a subset of $\Omega(y)$ is with respect to the Lebesgue measure in this dimension.

3. An expression for the probability density. Let $p^i(x)$ be the probability density after i iterations of the Metropolis–Hastings simulation algorithm. The following lemma is crucial for the later theorem since it formulates the probability density for $p^{i+1}(x)$ as a function of $p^i(x)$ in a compact formula.

LEMMA. *The probability density of the Metropolis–Hastings simulation algorithm satisfies*

$$p^{i+1}(y) = p^i(y) + \int_{\Omega(y)} \left(\frac{p^i(x)}{f(x)} - \frac{p^i(y)}{f(y)}\right) h(x, y)\, dx$$

and

$$\begin{aligned} \frac{p^{i+1}(y)}{f(y)} - 1 = &\left(\frac{p^i(y)}{f(y)} - 1\right)\left(1 - \int_{\Omega(y)} Q(x, y)\, dx\right) \\ &+ \int_{\Omega(y)} \left(\frac{p^i(x)}{f(x)} - 1\right) Q(x, y)\, dx, \end{aligned}$$

where $\int_{\Omega(y)} Q(x,y)\,dx \le 1$.

PROOF. The definition of the Metropolis–Hastings algorithm gives

$$p^{i+1}(y) = \int_{\Omega(y)} p^i(x)\,q(y \mid x)\,\alpha(y,x)\,dx + \int_{\Omega(y)} p^i(y)\,q(z \mid y)\big(1-\alpha(z,y)\big)\,dz.$$

Using that $h(x,y)$ is symmetric and that $\alpha(x,y) = h(x,y)/(f(y)q(x \mid y))$ gives

$$\begin{aligned} p^{i+1}(y) &= p^i(y) + \int_{\Omega(y)} \Big(p^i(x)\,q(y \mid x)\,\alpha(y,x) \\ &\qquad - p^i(y)\,q(x \mid y)\,\alpha(x,y)\Big)\,dx \\ &= p^i(y) + \int_{\Omega(y)} \left(p^i(x)\,q(y \mid x)\frac{h(x,y)}{f(x)q(y \mid x)} \right. \\ &\qquad \left. - p^i(y)\,q(x \mid y)\frac{h(x,y)}{f(y)\,q(x \mid y)}\right)dx \\ &= p^i(y) + \int_{\Omega(y)} \left(\frac{p^i(x)}{f(x)} - \frac{p^i(y)}{f(y)}\right) h(x,y)\,dx. \end{aligned}$$

This proves the first part of the lemma. Further,

$$\begin{aligned} \frac{p^{i+1}(y)}{f(y)} - 1 &= \frac{p^i(y)}{f(y)} + \int_{\Omega(y)} \left(\frac{p^i(x)}{f(x)} - \frac{p^i(y)}{f(y)}\right) Q(x,y)\,dx - 1 \\ &= \left(\frac{p^i(y)}{f(y)} - 1\right)\left(1 - \int_{\Omega(y)} Q(x,y)\,dx\right) \\ &\quad + \int_{\Omega(y)} \left(\frac{p^i(x)}{f(x)} - 1\right) Q(x,y)\,dx. \end{aligned}$$

Finally, $\int_{\Omega(y)} Q(x,y)\,dx \le \int_{\Omega(y)} q(x \mid y)\,dx \le 1$, which proves the rest of the lemma. □

4. Convergence for positive generation function. If the generation function is positive, it is possible to move between any two states in one jump. This makes the convergence faster and the result less technical.

PROPOSITION. *Assume that*

$$q(x \mid y) \ge a f(x)$$

is satisfied for all $x, y \in \Omega$ *where the constant* $a \in [0,1]$. *Then the probability density of the Metropolis–Hastings simulation algorithm satisfies for* $y \in \Omega$*:*

$$\left|\frac{p^{i+1}(y)}{f(y)} - 1\right| \le (1-a) \sup_{x \in \Omega} \left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}.$$

The proposition states that $|p^{i+1}(y)/f(y) - 1|$ does not increase and that the Metropolis–Hastings simulation algorithm converges if $a > 0$. The convergence is fast if $q(x \mid y) \approx f(x)$ and immediate if $q(x \mid y) = f(x)$. An example which shows that there may be equality instead of the $\leq$ sign is given after the proof.

PROOF. The assumptions in the proposition imply

$$Q(x,y) = \min\left\{q(x \mid y), q(y \mid x)\frac{f(x)}{f(y)}\right\} \geq af(x).$$

In order to simplify the notation we introduce

$$R^j(x) = \frac{p^j(x)}{f(x)} - 1 \quad \text{and} \quad R^j_M = \sup_{x\in\Omega}\left\{\left|\frac{p^j(x)}{f(x)} - 1\right|\right\}.$$

The lemma gives

$$\begin{aligned}
R^{i+1}(y) &= R^i(y)\left(1 - \int_\Omega Q(x,y)\,dx\right) + \int_\Omega R^i(x)\,Q(x,y)\,dx \\
&\leq R^i_M - \int_\Omega R^i_M\,Q(x,y)\,dx + \int_\Omega R^i(x)\,Q(x,y)\,dx \\
&= R^i_M - \int_\Omega \Big(R^i_M - R^i(x)\Big)Q(x,y)\,dx \\
&\leq R^i_M - a\int_\Omega \Big(R^i_M - R^i(x)\Big)f(x)\,dx \\
&= R^i_M\left(1 - a\int_\Omega f(x)\,dx\right) + a\int_\Omega \Big(p^i(x) - f(x)\Big)dx \\
&= R^i_M(1-a).
\end{aligned}$$

Define $\tilde{p}^i(x)$ and the corresponding $\tilde{R}^i(x)$ such that $\tilde{R}^i(x) = -R^i(x)$. The above calculation is also valid for $\tilde{R}^i(x)$. Then $|R^{i+1}(y)| \leq R^i_M(1-a)$. We have $a \in [0,1]$ since both $f(\cdot)$ and $q(\cdot \mid y)$ are densities. This proves the proposition. □

The following example shows that there may be equality instead of the $\leq$ sign in the proposition. The important properties in order to get equality are that for the particular y chosen

$$\left|\frac{p^i(y)}{f(y)} - 1\right| = \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}$$

and that $Q(x,y) = af(x) \quad \text{for} \quad x \in \Omega$.

EXAMPLE 1. Let $\Omega = (0,1)$, $f(x) = 1$,

$$q(x \mid y) = \begin{cases} a & \text{for } x \leq 1/2 \\ 2-a & \text{for } x > 1/2, \end{cases}$$

and

$$p^i(x) = \begin{cases} (1+\epsilon)\,f(x) & \text{for } x \leq 1/2, \\ (1-\epsilon)\,f(x) & \text{for } x > 1/2. \end{cases}$$

Then

$$\left|\frac{p^{i+1}(y)}{f(y)}-1\right| = (1-a)\sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)}-1\right|\right\}.$$

5. Vanishing generation function. In this section the proposition is generalized in order to show convergence also for vanishing generation function. The assumptions made are necessary in order for the algorithm to converge. However, first an example demonstrates why several iterations may be necessary in order to get a reduction for a relevant norm.

The proposition shows that $\sup_{x\in\Omega}\{|p^i(x)/f(x)-1|\}$ does not increase as the number of iterations increases. If $q(x \mid y) = 0$ for some values, then $a = 0$ and the proposition may not be used for proving convergence. This is the case when Metropolis–Hastings simulates a marked point process where only one or a few points are changed in each iteration. It is easy to give examples where for some value of y that $p^{i+1}(y)/f(y) = p^i(x)/f(x)$ for x such that $Q(x,y) > 0$. Then $p^{i+1}(y)/f(y) = p^i(y)/f(y)$, and the relative supremum norm does not decrease.

The norm

$$\int_{x\in\Omega}\left|\frac{p^i(x)}{f(x)}-1\right|dx$$

may increase at least for some values of i as the number of iterations increases. This is shown in the following example. One important property in the example is that

$$\frac{p^i(y)}{f(y)}-1 = \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)}-1\right|\right\}$$

in areas where $Q(x,y) > 0$.

EXAMPLE 2. Let $\Omega = (0,1)$, $f(x) = 2x$,

$$q(x \mid y) = \begin{cases} 0 & \text{for } |x-y| \geq \beta, \\ 1/2\beta & \text{for } |x-y| < \beta \text{ and } \beta < y < 1-\beta, \\ 1/(y+\beta) & \text{for } |x-y| < \beta \text{ and } y \leq \beta, \\ 1/(1-y+\beta) & \text{for } |x-y| < \beta \text{ and } y \geq 1-\beta, \end{cases}$$

where $0 < \beta < 1/6$, and

$$p^i(x) = \begin{cases} f(x) & \text{for } x \leq 2\beta, \\ (1+\epsilon)f(x) & \text{for } 2\beta < x \leq \gamma-\beta, \\ f(x) & \text{for } \gamma-\beta < x \leq \gamma+\beta, \\ (1-\epsilon)f(x) & \text{for } \gamma+\beta < x \leq 1-2\beta, \\ f(x) & \text{for } x > 1-2\beta, \end{cases}$$

where γ is determined such that $\int p^i(x)\,dx = 1$. This gives for $\beta < y \leq 1-\beta$

$$Q(x,y) = \begin{cases} 0 & \text{for } |x-y| \geq \beta, \\ 1/2\beta & \text{for } |x-y| \leq \beta \text{ and } \beta < y < x < 1-\beta, \\ x/2y\beta & \text{for } |x-y| \leq \beta \text{ and } \beta < x \leq y \leq 1-\beta. \end{cases}$$

The example is constructed such that the absolute value may be moved and we may perform the following calculation:

$$\begin{aligned} \int_{y\in\Omega} \left| \frac{p^{i+1}(y)}{f(y)} - 1 \right| dy &= \int_{y\in\Omega} \left| \left(\frac{p^i(y)}{f(y)} - 1 \right) \left(1 - \int_{x\in\Omega} Q(x,y)\,dx \right) \right. \\ &\qquad \left. + \int_{x\in\Omega} \left(\frac{p^i(x)}{f(x)} - 1 \right) Q(x,y)\,dx \right| dy \\ &= \int_{y\in\Omega} \left| \frac{p^i(y)}{f(y)} - 1 \right| \left(1 - \int_{x\in\Omega} Q(x,y)\,dx \right) dy \\ &\qquad + \int_{y\in\Omega} \int_{x\in\Omega} \left| \frac{p^i(x)}{f(x)} - 1 \right| Q(x,y)\,dx\,dy \\ &= \int_{y\in\Omega} \left| \frac{p^i(y)}{f(y)} - 1 \right| \left(1 - \int_{x\in\Omega} Q(x,y)\,dx \right) dy \\ &\qquad + \int_{x\in\Omega} \left| \frac{p^i(x)}{f(x)} - 1 \right| \int_{y\in\Omega} Q(x,y)\,dy\,dx \\ &= \int_{y\in\Omega} \left| \frac{p^i(y)}{f(y)} - 1 \right| \left(1 - \int_{x\in\Omega} Q(x,y)\,dx \right. \\ &\qquad \left. + \int_{z\in\Omega} Q(y,z)\,dz \right) dy \\ &> \int_{y\in\Omega} \left| \frac{p^i(y)}{f(y)} - 1 \right| dy, \end{aligned}$$

since $\int_{z\in\Omega} Q(y,z)\,dz > \int_{x\in\Omega} Q(x,y)\,dx$ where $|p^i(y)/f(y) - 1| > 0$.

When the generation function vanishes, several jumps $\{x^j\}_{j=0,s}$ where $x^0 = x$ and $x^s = y$, are necessary in order to jump between any states $x, y \in \Omega$. Let $D_j(x^{j+1})$ be the domain of x^j which is passed in the jumps from x to y. The larger the domains $D_j(x^{j+1})$ are, the more probable it is to jump from x to y. Hence, the integral $\int_{D_j(x^{j+1})} f(x^j)dx^j$ is critical for the convergence rate. If the space spaned by $\left\{D_i(x^{i+1})\right\}_{i=0}^{j}$ has less dimension than the space spaned by $\left\{D_i(x^{i+1})\right\}_{i=0}^{j+1}$, then $D_j(x)$ consists of one or a limited number of points. Then the integral $\int_{D_j(x^{j+1})} f(x^j)dx^j$ is interpretted as $\sum_{x^j\in D_j(x^{j+1})} f(x^j)$. This is illustrated in Example 8. In the following theorem a lower bound is necessary both on the size of $D_j(x^{j+1})$ and on the generation function in $D_j(x^{j+1})$. The properties of $D_j(x^{j+1})$ are formalized in the following definition: Define $S = \left\{S_y^x\right\}_{x,y\in\Omega}$ as a set of sequences $S_y^{x^0} = \{D_j(x^{j+1})\}_{j=0}^{s-1}$, where $x^s = y$, $x_j \in D_j(x^{j+1})$ for all $x^{j+1} \in D_{j+1}(x^{j+2})$,

$D_0(x^1) = \{x^0\}$ and $D_j(x^{j+1}) \subseteq \Omega(x^{j+1})$ for $j = 0, \ldots, s-1$. Let S_j be the set which consists of element j in all the sequences in $S_y^{x^0}$. Then we have the following theorem.

THEOREM. *Let the state space Ω be an open subset of $\mathbb{R}^n$ and assume that*

$$\inf_{z\in\Omega} \int_{\Omega(z)} f(x)\,dx > 0 \tag{1}$$

and that

$$\sup_{z\in\Omega} \int_{\Omega(z)} f(x)\,dx$$

is bounded. Assume the set of sequences $S_y^x = \left\{D_j(x^{j+1})\right\}_{j=0}^{s-1}$, for all $x, y \in \Omega$ satisfies

1.

$$q(x^j|x^{j+1}) \geq a_j f(x^j) \quad \textit{and} \quad q(x^{j+1}|x^j) \geq a_j f(x^{j+1}) \tag{2}$$

for $x^j \in D_j(x^{j+1}), j = 0, \ldots, s-1$.

2.

$$b_j = \inf_{D_j(x^{j+1})\in S_j} \int_{D_j(x^{j+1})} f(x^j)\,dx^j > 0 \quad \textit{for } j = 1, ..., s-1.$$

If $s = 1$, set $c = a_0$, and if $s > 1$, set $c = a_0 \prod_{j=1}^{s-1}(a_j b_j)$. Then

$$\left|\frac{p^{i+s}(y)}{f(y)} - 1\right| \leq (1-c) \sup_{x\in\Omega} \left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}, \tag{3}$$

where $c \in (0, 1]$. If such a set S_y^x does not exist for all $x, y \in \Omega$, then there exists $\epsilon > 0$ and $p^0(x)$ such that

$$\sup_{x\in\Omega} \left|\frac{p^j(x)}{f(x)} - 1\right| = \epsilon \tag{4}$$

for all $j \geq 0$.

In Examples 7 and 8 possible sets S are illustrated. This theorem generalizes the proposition since if $q(x \mid y) \geq af(x)$ for $x, y \in \Omega$, $s = 1$ and $D_1(y) = \Omega$ for all $x, y \in \Omega$. The critical assumptions in the theorem is that it is possible to jump between any two states $x, y \in \Omega$ and (2). The constants b_j may always be made positive by assumption (1).

Normally one will choose $D_j(x^{j+1})$ as large as possible. This makes b_j larger, but possibly a_j smaller. The size of $D_j(x^{j+1})$ is therefore a trade off between the size of a_j and b_j.

An often asked question is: How independent are two states generated by the Metropolis–Hastings simulation algorithm as a function of the number of iterations between the two states? The theorem states that the error in the relative supremum norm for the last state given the first state decreases at least by a factor $(1-c)$ per s iterations. In

addition, it is necessary to bound the probability of staying in the first state in several iterations.

This theorem may be used for comparison between different generation functions. This is also possible if these generation functions have different computational cost such that number of iterations varies.

PROOF OF THE MAIN THEOREM. First the following lower bounds on $Q(\cdot,\cdot)$ are needed. Equation (2) implies that

$$(5)\quad Q(x^j, x^{j+1}) = \min\left\{q(x^j \mid x^{j+1}), q(x^{j+1} \mid x^j)\frac{f(x^j)}{f(x^{j+1})}\right\} \geq a_j\, f(x^j)$$

for $x^j \in D_j(x^{j+1})$. Then the integral is bounded:

$$(6)\quad 1 \geq \int_{D_j(x^{j+1})} Q(x^j, x^{j+1})\, dx^j \geq a_j \int_{D_j(x^{j+1})} f(x^j)\, dx^j \geq a_j b_j$$

for $j = 1, \ldots, s-1$. This also shows that $c = a_0 \prod_{j=1}^{s-1}(a_j b_j) \in (0, 1]$. In order to simplify the notation introduce

$$R^j(x) = \frac{p^{i+j}(x)}{f(x)} - 1 \qquad \text{and} \qquad R_M = \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}.$$

The proposition implies that $R^j(x) \leq R_M$ for $j = 0, \ldots, s-1$ and $x \in \Omega$. The lemma gives

$$\begin{aligned} R^{j+1}(x^{j+1}) &= R^j(x^{j+1})\left(1 - \int_{\Omega(x^{j+1})} Q(x^j, x^{j+1})\, dx^j\right) \\ &\quad + \int_{\Omega(x^{j+1})} R^j(x^j)\, Q(x^j, x^{j+1})\, dx^j \\ &\leq R_M - \int_{\Omega(x^{j+1})} \Big(R_M - R^j(x^j)\Big)\, Q(x^j, x^{j+1})\, dx^j \end{aligned}$$

Notice that the integration is with respect to $\Omega(x^{j+1})$ which may have a lower dimension than Ω. If this equation is used for $j = 0, \ldots, s-1$

iteratively, this gives

$$\begin{aligned}
R^{i+s}(y) &\le R_M - \int_{\Omega(x^s)} \cdots \int_{\Omega(x^1)} \Big(R_M - R^0(x^0)\Big) \\
&\qquad \times Q(x^0,x^1)\,dx^0 Q(x^1,x^2)\,dx^1 \cdots Q(x^{s-1},x^s)\,dx^{s-1} \\
&\le R_M - \int_\Omega \int_{D_{s-1}(x^s)} \cdots \int_{D_1(x^2)} Q(x^0,x^1)Q(x^1,x^2)\,dx^1 \times \cdots \\
&\qquad \cdots \times Q(x^{s-1},x^s)\,dx^{s-1}\Big(R_M - R^0(x^0)\Big)\,dx^0 \\
&\le R_M - a_0 \int_\Omega \int_{D_{s-1}(x^s)} \cdots \int_{D_1(x^2)} Q(x^1,x^2)\,dx^1 \times \cdots \\
&\qquad \cdots \times Q(x^{s-1},x^s)\,dx^{s-1}\Big(R_M - R^0(x^0)\Big)\,f(x^0)\,dx^0 \\
&\le R_M - a_0 \prod_{j=1}^{s-1}(a_j b_j) \int_\Omega \Big(R_M - R^0(x^0)\Big)\,f(x^0)\,dx^0 \\
&= R_M - c\int_\Omega \Big(R_M - R^0(x^0)\Big)\,f(x^0)\,dx^0 \\
&= R_M\left(1 - c\int_\Omega f(x^0)\,dx^0\right) + \int_\Omega R^0(x^0)\,f(x^0)\,dx^0 \\
&= R_M(1-c).
\end{aligned}$$

In the calculation we have used the lower bound on $Q(\cdot,\cdot)$ from (5), changed the order of integration using the fact that S spans Ω and the properties of the sequence $D_j(x^{j+1})$. Before the order is shifted it is integrated over all possible sequences $\{x^j\}_{j=0}^{j=s}$ fixing only $x^s = y$. After the integration order is shifted it is only integrated over the sets $D_j(x^{j+1})$ with both $x^s = y$ and x^0 fixed. Notice that the integration domain $D_j(x^{j+1})$ depends on both $x^s = y$ and x^0. Then (6) is used. Define $\tilde{p}^i(x)$ and the corresponding $\tilde{R}^i(x)$ such that $\tilde{R}^i(x) = -R^i(x)$. The above calculation is also valid for $\tilde{R}^i(x)$ so that $\big|R^{i+s}(y)\big| \le R_M(1-c)$. This proves that the existence of a set of sequences S_y^x for all $x, y \in \Omega$, implies (3). It is left to prove the implication in the other direction.

Choose $a \in (0,1)$ and $s > 0$. Define $S_y^{a,s}$ such that each $D_j(x^{j+1})$ is as large as possible satisfying the first half of (2) and for any $x^0 \in \Omega$, i.e.

$$D_j(x^{j+1}) = \Big\{x^j \in \Omega(x^{j+1}); \quad q(x^j \mid x^{j+1}) \ge a f(x^j)\Big\}.$$

Using only half of (2) makes $D_j(x^{j+1})$ larger. Define

$$\begin{aligned}
A_y^{a,s} &= \operatorname{span}(S_y^{a,s}) \\
&= \big\{x \in \Omega; \quad x \in D_0(x^1) \text{where } D_0(x^1) \text{ is in a sequence in } S_y^{a,s}\big\}.
\end{aligned}$$

Assume $A_y^{a,s}$ has positive measure in $\mathbb{R}^n$. If this is not the case, y is replaced by another state in Ω. Let $A_y = \sup_{a,s}\big\{\operatorname{span}(S_y^{a,s})\big\}$.

Assume first $A_y \neq \Omega$. Let

$$p^0(x) = \begin{cases} (1+\epsilon)\, f(x) & \text{for } x \in \Omega \setminus A_y, \\ (1-\beta\epsilon)\, f(x) & \text{else,} \end{cases}$$

where β is determined such that $\int_\Omega p^0(x)\, dx = 1$. A chain with $x^0 \in \Omega \setminus A_y$ does not join A_y, and a chain with $x^0 \in A_y$ does not join $\Omega \setminus A_y$ for any s. Then according to the lemma $p^j(x) = (1+\epsilon)\, f(x)$ for $x \in \Omega \setminus A_y$ and all $j \geq 0$, which implies (4).

Assume then $A_y = \Omega$. For a given $\delta > 0$, there exist a and s such that the probability of a chain with $x^0 \in \Omega \setminus A_y^{a,s}$ entering $A_y^{a,s}$ is less than δ. This is proved as follows: In order to enter $A_y^{a,s}$ it is necessary to pass a domain $D_{j+1} \subseteq \Omega(x^j)$ for any x^j such that

$$q(x^j \mid x^{j+1}) < a\, f(x^j) \qquad \text{for } x^{j+1} \in D_{j+1}.$$

The probability for both generating and accepting a point $x^{j+1} \in D_{j+1}$ is bounded by

$$\int_{D_{j+1}} \frac{f(x^{j+1})\, q(x^j | x^{j+1})}{f(x^j)\, q(x^{j+1} | x^j)} q(x^{j+1} | x^j)\, dx^{j+1} \leq a \int_{D_{j+1}} f(x^{j+1})\, dx^{j+1},$$

which may be made arbitrarily small by choosing a small since the integral is bounded by the assumptions in the theorem. Let

$$p^0(x) = \begin{cases} (1+\epsilon)\, f(x) & \text{for } x \in \Omega \setminus A_y^{a,s}, \\ (1-\beta\epsilon)\, f(x) & \text{else,} \end{cases}$$

where β is determined such that $\int_\Omega p^0(x)\, dx = 1$. Then $p^j(x) \geq (1+\epsilon)\,(1-\delta)\, f(x)$ for $x \in \Omega \setminus A_y^{a,s}$ and $j \leq s$. This implies that

$$\sup_{x \in \Omega} \left| \frac{p^j(x)}{f(x)} - 1 \right| = \epsilon$$

for all $j > 0$, which proves the theorem. □

6. Some examples. It is recognized in practical applications that the Metropolis–Hastings simulation algorithm does not work satisfactory without the assumptions in the theorem. The two assumptions regarding the construction of the set S are critical. The following two examples demonstrate why convergence is not achieved if either of these two assumptions is violated. Two additional examples are given. These further examples show why the additional assumptions in the theorem, not connected to the construction of S, are necessary. In the first example there is not a finite number of jumps between any states $x, y \in \Omega$ independent of x, y.

EXAMPLE 3. Let $\Omega = \mathbb{R}$, let f be a normal distribution with expectation μ, and let $q(x \mid y) = 0$ for $|x - y| > c$ for a constant c.

If

$$p^0(x) = \begin{cases} (1+\epsilon)f(x) & \text{for } x \le \mu, \\ (1-\epsilon)f(x) & \text{for } x > \mu, \end{cases}$$

then $\sup_{x\in\mathbb{R}} |p^s(x)/f(x) - 1| = \epsilon$ for all $s > 0$, since $p^i(x)$ will not be changed for sufficiently large values of $|x|$ in a finite number of jumps. Hence, the algorithm does not converge in this norm. However, the algorithm converges in both L_1 and L_∞. This example also violates the assumption $\inf_y \int_{\Omega(y)} f(x)\,dx > 0$ by setting $|y|$ sufficiently large.

The other critical assumption is (2). The following example demonstrates that we may not in general get convergence if this assumption is violated.

EXAMPLE 4. Let $\Omega = (0,1)$, $f(x) = 1$, $q(x \mid y) = 2x$, and

$$p^0(x) = \begin{cases} (1+\epsilon)f(x) & \text{for } x \le 1/2, \\ (1-\epsilon)f(x) & \text{for } x > 1/2. \end{cases}$$

Then $\sup_{x\in\Omega} |p^s(x)/f(x) - 1| = \epsilon$ for all $s > 0$, since the algorithm never leaves a state with arbitrarily small values of x. Hence, also in this example the algorithm does not converge in the relative supremum norm. The algorithm converges in L_1 but not in L_∞.

The following example demonstrates why $\sup_y \int_{\Omega(y)} f(x)\,dx$ is assumed bounded.

EXAMPLE 5. Let $x = (x_1, x_2) \in \mathbb{R}^2$,

$$\Omega = \left\{(x_1, x_2) \in \mathbb{R}^2; \quad x_1 \ge 1 \text{ and } 0 < x_2 \le x_1^{-2}\right\}.$$

and $f(x) = 3$ for all $x \in \Omega$. Further let

$$q\big((x_1, x_2) \mid (y_1, y_2)\big) = \begin{cases} 1/2x_1^2 & \text{if } x_1 = y_1, \\ 1/2(x_2^2 - 1) & \text{if } x_2 = y_2, \\ 0 & \text{else,} \end{cases}$$

and

$$p^0(x) = \begin{cases} (1+\epsilon)f(x) & \text{if } x_1 > 2, \\ (1-\beta\epsilon)f(x) & \text{else,} \end{cases}$$

where β is chosen such that $\int_\Omega p^0(x)\,dx = 1$. A chain starting with sufficiently large values of x_1 has arbitrarily small probability of entering the region with $x_1 < 2$. Hence, $\sup_{x\in\Omega} |p^j(x)/f(x) - 1| = \epsilon$ for all $j \ge 0$.

The following example shows why Ω must be open. The theorem is also valid for closed Ω, but this necessitates additional technical assumptions which exclude the following and similar examples.

EXAMPLE 6. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\Omega = [0,1]^n \cup [1,2]^n$, and $f(x) = 1/2$. Further let

$$q(x \mid y) = \begin{cases} 1/n & \text{if } x_i = y_i \text{ for at least } n-1 \text{ values of } i = 1, \ldots, n, \\ 0 & \text{else,} \end{cases}$$

and

$$p^i(x) = \begin{cases} (1+\epsilon)f(x) & \text{if } x_1 < 1, \\ (1-\epsilon)f(x) & \text{else.} \end{cases}$$

Then $\sup_{x \in \mathbb{R}} |p^{i+s}(x)/f(x) - 1| = \epsilon$ for all $s > 0$, since it is necessary to pass through a subspace with $n-1$ coordinates equal to 1 in order to jump from one half of the state space to the other half. This subspace has measure 0 in $\mathbb{R}^n$ and, more important, also has measure 0 in $\Omega(x)$ for points x which are not already in the subspace.

The following examples illustrate the theorem in two cases where convergence is obtained. In the first case the same integration measure is used for both $f(x)$ and $q(x \mid y)$ and in the second case different measures are used.

EXAMPLE 7. Let $\Omega = (0,1)$, $f(x) = 1$ and

$$q(x \mid y) = \begin{cases} 0 & \text{for } |x-y| \geq \beta, \\ 1/2\beta & \text{for } |x-y| < \beta \text{ and } \beta < y < 1-\beta, \\ 1/(y+\beta) & \text{for } |x-y| < \beta \text{ and } y \leq \beta, \\ 1(1-y+\beta) & \text{for } |x-y| < \beta \text{ and } y \geq 1-\beta, \end{cases}$$

where $0 < \beta < 1/2$. This gives

$$Q(x,y) = \begin{cases} 0 & \text{for } |x-y| \geq \beta, \\ 1/(\max\{x,y\} + \beta) & \text{for } x, y \leq \beta, \\ 1/(\max\{1-x, 1-y\} + \beta) & \text{for } x, y \geq 1-\beta, \\ 1/2\beta & \text{else.} \end{cases}$$

Set

$$D_j(x^{j+1}) = \left(j\frac{y-x}{s} + x - \gamma, j\frac{y-x}{s} + x + \gamma \right)$$

where γ is chosen as large as possible, but such that it is possible to jump from any position in $D_j(x^{j+1})$ to any position in $D_{j+1}(x^{j+2})$ for all states $x, y \in \Omega$. This is not an optimal choice of D_j, since that is more technical. This choice gives $\gamma = \big(\beta - (1/s)\big)/2$ where $s > 1/\beta$. Then we may calculate

$$b_j = \int_{D(x^{j+1})} f(x^j)\, dx^j = 2\gamma = \beta - \frac{1}{s}.$$

It is easily seen that $a_j = 1/2\beta$. The bound on the error is then

$$\left|\frac{p^{i+s}(y)}{f(y)} - 1\right| \leq \left(1 - \left(\frac{1}{2\beta}\right)^s \left(\beta - \frac{1}{s}\right)^{s-1}\right) \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}$$
$$= \left(1 - \frac{1}{2\beta}\left(\frac{1}{2} - \frac{1}{2\beta s}\right)^{s-1}\right) \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}.$$

Notice that choosing s as small as possible but larger than $1/\beta$ gives an upper bound on when the reduction in the relative supremum norm starts. However, choosing s larger will give a better bound on the convergence when the number of iterations is larger.

EXAMPLE 8. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\Omega = (0,1)^n$, and

$$f(x) = \begin{cases} \nu & \text{if } x_i < \beta \text{ for } i = 1, \ldots, n, \\ \mu & \text{if } x_i \geq 1 - \beta \text{ for } i = 1, ..., n, \\ (1 - (\nu + \mu)\beta^n)/(1 - 2\beta^n) & \text{else,} \end{cases}$$

where $\beta < 1/2$ and $\nu \geq \mu > 1$. Further

$$q(x \mid y) = \begin{cases} 1/n & \text{if } x_i = y_i \text{ for at least } n-1 \text{ values of } i = 1, \ldots, n, \\ 0 & \text{else.} \end{cases}$$

For μ large this example is similar to a Strauss process with strong attraction; see for example Ripley (1987). The movement of a chain between domains with high density is only possible by passing through domains with low density. The movement of a chain between two of the opposite corners in this example is similar to the movement of a cluster in a Strauss process. In a Strauss process the cluster may also move slowly. This problem is known to converge very slowly.

Let x, y be in opposite corners. Choose the set of sequences S with $s = n$ as follows. Let $D_j(x^{j+1})$ be the $j+1$ states sequentially determined by j coordinates equal the coordinates of y and $n-j$ coordinates equal the coordinates of x. The $j+1$ coordinates in $D_{j+1}(x^{j+2})$, which are equal to the coordinates of y, are the same as the coordinates as $D_j(x^{j+1})$ have equal to the coordinates of y and one of the the other coordinates. This gives

$$a_j b_j = a_j \int_{D_j(x^{j+1})} f(x^j)\, dx^j = \frac{j+1}{n}.$$

It is easily seen that $a_0 = 1/n\nu$. Then the theorem states

$$\left|\frac{p^{i+n}(y)}{f(y)} - 1\right| \leq \left(1 - \frac{n!}{\nu n^n}\right) \sup_{x\in\Omega}\left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\}, \tag{7}$$

which indicates very slow convergence for n large.

If we set $\nu = \beta^n/2$ and

$$p^i(x) = \begin{cases} (1-\epsilon)\, f(x) & \text{for } x_i \geq 1-\beta \text{ for all } i, \\ (1+\epsilon)\, f(x) & \text{else,} \end{cases}$$

then the relative supremum norm does not decrease the first $n-1$ iterations and the theorem gives a good description of the convergence rate the first n iterations. However, when the number of iterations increases $p^i(x) - f(x)$ will change more gradually, moving from one corner to the opposite one. The relative reduction in the relative supremum norm will then increase. This is illustrated by assuming that $\mu = \nu = 1$ and $p^i(x) = (1 + \epsilon(j-k)/n)\ f(x)$, where j is the number of $x_i < \beta$ and k is the number of $x_i \geq 1-\beta$. The lemma implies that

$$p^{i+1}(x) = \left(1 + \frac{j-k}{n}\epsilon\Big(1 - \frac{1}{n}\Big)\right) f(x). \tag{8}$$

This gives

$$\left|\frac{p^{i+1}(y)}{f(y)} - 1\right| \leq \left(1 - \frac{1}{n}\right) \sup_{x\in\Omega} \left\{\left|\frac{p^i(x)}{f(x)} - 1\right|\right\},$$

which is considerably better than (7). The critical difference between these two cases is that in the first case $p^i(x) - f(x)$ is only positive for x in a small corner while in the second there is a gradual change.

Equation (8) is proved as follows: Let

$$R^i_{j,k} = \frac{p^i(x)}{f(x)} - 1 = \epsilon\frac{j-k}{n},$$

where j is the number of $x_i < \beta$ and k is the number of $x_i \geq 1-\beta$ in x. By invoking the lemma, the calculation goes as follows:

$$\begin{aligned} R^{i+1}_{j,k} &= R^i_{j,k} + \frac{j(1-2\beta)}{n}\Big(R^i_{j-1,k} - R^i_{j,k}\Big) \\ &\quad + \frac{(n-j-k)\beta}{n}\Big(R^i_{j+1,k} - R^i_{j,k}\Big) + \frac{k(1-2\beta)}{n}\Big(R^i_{j,k-1} - R^i_{j,k}\Big) \\ &\quad + \frac{(n-j-k)\beta}{n}\Big(R^i_{j,k+1} - R^i_{j,k}\Big) + \frac{j\beta}{n}\Big(R^i_{j-1,k+1} - R^i_{j,k}\Big) \\ &\quad + \frac{k\beta}{n}\Big(R^i_{j+1,k-1} - R^i_{j,k}\Big) \\ &= \epsilon\,\frac{j-k}{n}\left(1 - \frac{1}{n}\right). \end{aligned}$$

5. Closing remarks. In this paper geometric convergence of the Metropolis–Hastings simulation algorithm is proved under weak assumptions. If the assumptions for geometric convergence are not satisfied, then in general the algorithm does not converge in the relative supremum norm. The first assumption is that it is possible to jump between any two states in the state space in a finite number of jumps where the number of jumps is independent of the position. The second

main assumption is that the ratio between the generation function and the limiting function is bounded by a positive constant in a domain with positive measure. The convergence rate depends heavily on the size of this constant and the corresponding domain in addition to the number of jumps necessary to move between the two different states in the state space.

The results in the paper also give a good qualitative understanding of the convergence. The lemma shows that the error in the relative supremum norm in a point y in iteration $i+1$ is the average of the error in iteration i weighted by $Q(x,y)$. Assume that $\Omega = \mathbb{R}$ and that the change in each iteration is limited. Then the lemma states that the high frequency error in $p^0(x)/f(x)$ is reduced quickly and the low frequency error is reduced more slowly.

Let $A = \{x \in \Omega; p^0(x) > f(x)\}$ and $B = \{x \in \Omega; p^0(x) < f(x)\}$. If the probability of jumping from $x^i \in A$ to $x^{i+1} \in B$ is high, then the convergence is fast. On the other hand, if several jumps are necessary in order to move from A to B, the convergence is slow, particularly to begin with. This is illustrated in Example 8 where there is no reduction in the relative supremum norm the first n iterations for a particular $p^0(x)$. After several iterations there will be a gradual change in the error such that $p^i(x)/f(x)$ will be quite similar at points where the probability density for jumping between the points in one or a few iterations is large. Then the relative supremum norm decreases in each iteration and the convergence is faster than the bound described in the theorem. If s is small, the theorem describes this slow convergence since the bound on the convergence rate in the theorem is independent of $p^i(x)$. A better bound on the convergence rate is found by increasing s and choosing the set D_j as large as possible. This gives a good estimate for the convergence rate for the worst possible $p^0(x)$. Alternatively, the lemma may be used as illustrated in the last part of Example 8. This gives a good estimate for the convergence rate for a particular $p^0(x)$

Acknowledgment. The author thanks Fred Espen Benth and Øivind Skare for valuable discussions and the Research Council of Norway for financial support.

References

Billingsley, P. (1986), *Probability and Measure*, 2nd edn, John Wiley & Sons, New York.

Diaconis, P. & Saloff-Coste, L. (1995), What do we know about the metropolis algorith?, *in* 'Proc. 27 Symp. Theory Comp.', pp. 112–129.

Frigessi, A., Hwang, C., Stefano, P. & Sheu, S. (1993), 'Convergence rates of the gibbs sampler, the metropolis algorithm and other single-site updating dynamics', *J. Roy. Statist. Soc. Ser. B* **55**(1), 205–219.

Geyer, C. J. (1992), 'Practical markov chain monte carlo', *Statistical Science* **7**, 473–483.

Geyer, C. J. & Thompson, E. T. (1995), 'Annealing markov chain monte carlo with applications to ancestral inference', *J. Amer. Statist. Assoc.* **90**, 909–920.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika*.

Hektoen, A. & Holden, L. (1996), Bayesian modelling of sequence stratigraphic bounding surfaces, *in* 'Proc. '5th Inter. Geostat. Congr.', Wollongong, Australia', Kluwer Academic Publishers.

Mengersen, K. L. & Tweedie, R. L. (1994), Rates of convergence of the hastings and metropolis algorithms, Preprint, Quensland University of Technology and Colorado State University.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. & Teller, E. (1953), 'Equations of state calculations by fast computing machines', *J. Chem. Phys.* **21**, 1087–1092.

Meyn, S. P. & Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer Verlag, London.

Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons, New York.

Roberts, G. O. & Tweedie, R. L. (1996), 'Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms', *Biometrika* **83**(1), 95–110.

Skare, Ø., Skorstad, A., Hauge, R. & Holden, L. (1996), Conditioning a fluvial model on seismic data, *in* 'Proc. '5th Inter. Geostat. Congr.', Wollongong, Australia', Kluwer Academic Publishers.

Norwegian Computing Center
P.O. Box 114 Blindern
N-0314 Oslo
Norway
E-mail:Lars.Holden@nr.no