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 in the relative supremum norm of the Metropolis–Hastings simulation algorithm with a general generating function are established. An explicit expression for the convergence rate is given.

1. Introduction. This paper discusses the convergence rate for the Metropolis–Hastings simulation algorithm proposed in Hastings (1970). The Metropolis–Hastings simulation algorithm is used for sampling from a distribution $f(x)$. There is currently a lot of interest in MCMC both theoretically and in a large number of applications, see Geyer (1992). The challenge in Metropolis-Hastings is to find a good generating function. The explicit formula for the convergence rate given in this paper may be used to compare different generating functions.

Meyn & Tweedie (1993) prove that the Doeblin condition is equivalent to uniform ergodic, i.e. uniform convergence in total

Date: April 29, 1998.

1991 *Mathematics Subject Classification.* 60J27.

Key words and phrases. Metropolis–Hastings algorithm, convergence.

Revised version will appear in Statistics and Probability Letters. Research supported by Research Council of Norway. Pres. adr.: Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway, E-mail:Lars.Holden@nr.no.

variation norm. The total variation norm is always bounded by the relative supremum norm. Hence, the requirement in the theorem in this paper implies that the Doeblin condition is obtained in finite number of steps. The relative supremum norms is used in this paper since it gives a simple expression for the convergence rate. Convergence in other norms may be derived from the convergence in the relative supremum norm.

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 $i = 1, \dots, n$:
    (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, $q(x \mid y) > 0$ 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:

$$\begin{aligned}
\Omega(y) &= \{x \in \Omega;\ q(x \mid y) > 0\},\\
h(x,y) &= \min\bigl\{f(x)\,q(y \mid x), f(y)\,q(x \mid y)\bigr\},\\
Q(x,y) &= \frac{h(x,y)}{f(y)} = \min\left\{q(x \mid y), q(y \mid x)\frac{f(x)}{f(y)}\right\}\\
R^j(x) &= \frac{p^j(x)}{f(x)} - 1 \quad \text{and} \quad R^j_M = \sup_{x\in\Omega}\left\{\left|\frac{p^j(x)}{f(x)} - 1\right|\right\}
\end{aligned}$$

where $p^j(x)$ is the density after j iterations. 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. 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

$$R^{i+1}(y) = R^i(y)\left(1 - \int_{\Omega(y)} Q(x,y)\,dx\right) + \int_{\Omega(y)} R^i(x)Q(x,y)\,dx,$$

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

Proof The definition of the Metropolis–Hastings algorithm gives

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

where it is used that $\alpha(x,y) = h(x,y)/(f(y)q(x \mid y))$ and that $h(x,y)$ is symmetric. The rest of the lemma follows trivially from the above calculation. □

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.

4. Convergence for positive generating function. If the generating 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. Mengersen & Tweedie (1994) prove a similar proposition with the stronger assumption $q(x \mid y) = q(x)$.

Proposition Assume that $q(x \mid y) \geq af(x)$ is satisfied for all $x, y \in \Omega$ where the constant $a \in [0,1]$. Then the relative error of the Metropolis–Hastings simulation algorithm satisfies $R_M^{i+1} \leq (1-a)R_M^i$.

Proof The assumption in the proposition implies $Q(x, y) \geq af(x)$. The lemma gives

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

The above calculation is also valid for $\tilde{R}^i(x) = -R^i(x)$. Then $|R^{i+1}(y)| \leq R_M^i(1-a)$. Since both $f(\cdot)$ and $q(\cdot \mid y)$ are densities, $a \in [0,1]$. □

5. Vanishing generating function. When the generating 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 using the 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 $j = 0, \dots, s-1$. Let S_j be the set which consists of element j in all the sequences in $S_y^{x^0}$.

Theorem Let the state space Ω be an open subset of $\mathbb{R}^n$ and assume that $\inf_{z\in\Omega} \int_{\Omega(z)} f(x)\, dx > 0$ and that $\sup_{z\in\Omega} \int_{\Omega(z)} f(x)\, dx$ is finite. 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

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

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

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

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 $R_M^{i+s} \leq (1-c)R_M^i$ where $c \in (0,1]$. If such a set S_y^x does not exist, then there exsists $\epsilon > 0$ such that $R_M^j \geq \epsilon$ for all j.

Proof First the following lower bounds on $Q(\cdot,\cdot)$ are needed. Equation (1) implies that $Q(x^j, x^{j+1}) \geq a_j \, f(x^j)$ for $x^j \in D_j(x^{j+1})$. Then the integral is bounded:

$$1 \geq \int_{D_j(x^{j+1})} Q(x^j, x^{j+1}) \, dx^j \geq a_j \int_{D_j(x^{j+1})} f(x^j) \, dx^j \geq a_j b_j \tag{2}$$

for $j = 1, \ldots, s-1$. This also shows that $c = a_0 \prod_{j=1}^{s-1}(a_j b_j) \in (0,1]$. The proposition implies that $R^j(x) \leq R_M$. The lemma gives

$$R^{j+1}(x^{j+1}) \leq R_M - \int_{\Omega(x^{j+1})} \Big(R_M - R^j(x^j)\Big) \, Q(x^j, x^{j+1}) \, dx^j.$$

Notice that the integration is with respect to $\Omega(x^{j+1})$ which may have a lower dimension than Ω. This gives

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

In the calculation we have used the lower bound on $Q(\cdot,\cdot)$, changed the order of integration using the fact that S spans Ω. Before the order is shifted it is integrated over all possible sequences $\{x^j\}_{j=0}^{j=s}$ fixing only $x^s = y$, afterwards it is only integrated over the sets $D_j(x^{j+1})$ with both $x^s = y$ and x^0 fixed. Then (2) is used. Notice that the integration domain $D_j(x^{j+1})$ depends on both $x^s = y$ and x^0. Similarly $\tilde{R}^{i+s}(x) = -R^{i+s}(y)$ is bounded, which proves the first part of the theorem.

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 $q(x^j \mid x^{j+1}) \geq af(x^j)$, $A_y^{a,s} = \operatorname{span}(S_y^{a,s}) \subset \Omega$. and $A_y = \sup_{a,s}\bigl\{\operatorname{span}(S_y^{a,s})\bigr\}$. If $A_y^{a,s}$ does not have positive measure in $\mathbb{R}^n$, y is replaced by another state in Ω.

Assume first $A_y \neq \Omega$. Then there will be no jumps between A_y and $\Omega \setminus A_y$, which implies $R_M^i \geq \epsilon > 0$.

Assume then $A_y = \Omega$. Then 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. Let

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

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 $j \leq s$ which implies $R_M^i \geq \epsilon > 0$. □

Example 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

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

then $R_M^i = \epsilon$ for all i. The algorithm converges in L_1, L_∞ and $T.V.$ norm. In this example the assumption $\inf_y \int_{\Omega(y)} f(x)\, dx > 0$ is violated for $|y|$ sufficiently large.

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

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

Then $R_M^i = \epsilon$ for all i. The algorithm converges in L_1, and $T.V.$ norm but not in L_∞. In this example (1) is violated.

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

$$\Omega = \big\{(x_1, x_2) \in \mathbb{R}^2; \quad x_1 \geq 1 \text{ and } 0 < x_2 \leq x_1^{-2}\big\}.$$

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

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

and

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

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 $R_M^j = \epsilon$ for all $j \geq 0$. In this example $\sup_y \int_{\Omega(y)} f(x)\,dx$ is not finite.

Example 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 $0 < \beta < 1/2$. Setting $D_j\big(x^{j+1}\big) = \big(j\frac{y-x}{s} + x - \gamma, j\frac{y-x}{s} + x + \gamma\big)$ where $\gamma = (\beta - (1/s))/2$, $s > 1/\beta$, gives $R_M^{i+s} \leq \left(1 - \frac{1}{2\beta}\left(\frac{1}{2} - \frac{1}{2\beta s}\right)^{s-1}\right) R_M^i$.

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

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

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

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

For μ large this example is similar to a Strauss process with strong attraction. The movement of a chain between domains with high density is only possible by passing through domains with low density.

Let x, y be in opposite corners and $s = n$. The theorem gives the following slow convergence for n large. $R_M^{i+s} \leq \left(1 - \frac{n!}{\nu n^n}\right) R_M^i$. This may be the exact convergence of the first n steps. When the number of iterations increases, R_M^i decreases faster. This is illustrated by assuming that $\mu = \nu$ 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 $p^{i+1}(x) = \left(1 + \frac{j-k}{n}\epsilon\left(1 - \frac{1}{n}\right)\right) f(x)$which gives

$R_M^{i+s} \leq \left(1 - \frac{1}{n}\right) R_M^i$. 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.

Acknowledgment. The author thanks Fred Espen Benth and Øivind Skare for valuable discussions.

References

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

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

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.

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