1

Markov Chain Monte Carlo Algorithms: Theory and Practice

Jeffrey S. Rosenthal

Department of Statistics
University of Toronto
Toronto, Ontario, Canada
http://probability.ca/jeff/

Summary. We describe the importance and widespread use of Markov chain Monte Carlo (MCMC) algorithms, with an emphasis on the ways in which theoretical analysis can help with their practical implementation. In particular, we discuss how to achieve rigorous quantitative bounds on convergence to stationarity using the coupling method together with drift and minorisation conditions. We also discuss recent advances in the field of adaptive MCMC, where the computer iteratively selects from among many different MCMC algorithms. Such adaptive MCMC algorithms may fail to converge if implemented naively, but they will converge correctly if certain conditions such as Diminishing Adaptation are satisfied.

1.1 Introduction

Markov chain Monte Carlo (MCMC) algorithms were first introduced in statistical physics [17], and gradually found their way into image processing [12] and statistical inference [15, 32, 11, 33]. Their main use is to sample from a complicated probability distribution $\pi(\cdot)$ on a state space $\mathcal{X}$ (which is usually high-dimensional, and often continuous, e.g. an open subset of $\mathbf{R}^d$). In particular, MCMC has revolutionized the field of Bayesian statistical inference, where $\pi(\cdot)$ would usually be a posterior distribution which is otherwise intractable but which can (hopefully) be easily sampled using MCMC.

In brief, MCMC proceeds as follows. We define a Markov chain $P(x, \cdot)$ on $\mathcal{X}$ that leaves $\pi(\cdot)$ stationary. We first sample X_0 from some (simple) *initial distribution* on $\mathcal{X}$. We then iteratively sample X_n from $P(X_{n-1}, \cdot)$, for $n = 1, 2, 3, \ldots$. The hope is that for "large enough" n, the distribution of X_n will be approximately equal to $\pi(\cdot)$, i.e. $P(X_n \in A) \approx \pi(A)$ for all measurable $A \subseteq \mathcal{X}$. If so, then X_n is approximately a *sample* from $\pi(\cdot)$. And, once we can generate samples from $\pi(\cdot)$, then we can easily use those samples to approximately compute any quantities of interesting involving probabilities or expectations with respect to $\pi(\cdot)$.

Such algorithms have become extremely popular in Bayesian statistics and other areas. At last count, the *MCMC Preprint Service* lists about seven thousand research papers, and the phrase "Markov chain Monte Carlo" elicits over three hundred thousand hits in *Google*. As a result of this popularity, many people are using MCMC algorithms without possessing much knowledge of the theory of Markov chains or probability, and there has been some divorce between theoreticians and practitioners of MCMC.

Despite this, there are a number of ways in which theory has had, and continues to have, important implications for the practical use of MCMC. In this paper, we concentrate on two areas: theoretical bounds on time to stationarity (Section 1.3), and validity of adaptive MCMC algorithms (Section 1.4); for additional background see e.g. [24] and the references therein.

1.2 Asymptotic Convergence

The first and most basic question about MCMC is whether it converges asymptotically, i.e. whether it is true that for sufficiently large n, the distribution of X_n is close to $\pi(\cdot)$. This is a bare minimal requirement for an MCMC algorithm to be "valid".

On a finite state space $\mathcal{X}$, it is well known that if a time-homogeneous Markov chain is irreducible and aperiodic, then it has a unique stationarity distribution $\pi(\cdot)$, to which it will converge in distribution as $n \to \infty$.

In this context, "irreducible" means that for all $x, y \in \mathcal{X}$, y is *accessible* from x, i.e. there is $n \in \mathbf{N}$ such that $P^n(x, \{y\}) \equiv \mathbf{P}(X_n \in \{y\} \,|\, X_0 = x) > 0$. This is clearly impossible on a continuous (uncountable) state space $\mathcal{X}$, since the subset $\{y \in \mathcal{X} : \exists\, n \in \mathbf{N},\ P^n(x, \{y\}) > 0\}$ is always countable. However, it is possible to weaken the condition "irreducible" to that of *ϕ-irreducible*, meaning there exists a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all measurable $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and all $x \in \mathcal{X}$, there exists $n \in \mathbf{N}$ such that $P^n(x, A) > 0$. It is then well known (see e.g. [18, 33, 24]) that if a Markov chain (on a general countably-generated state space $\mathcal{X}$) is ϕ-irreducible and aperiodic, and possesses an stationarity probability distribution $\pi(\cdot)$ (which is no longer guaranteed), then asymptotic convergence still holds, and in fact

$$\lim_{n\to\infty} \sup_{A\subseteq\mathcal{X}} |P^n(x, A) - \pi(A)| \;=\; 0\,, \qquad \pi\text{-}a.e.\ x \in \mathcal{X}\,. \tag{1.1}$$

For example, if the Markov chain transition probabilities all have positive densities with respect to Lebesgue measure on $\mathbf{R}^d$, then we can simply let $\phi(\cdot)$ be Lebesgue measure, to see that ϕ-irreducibility is satisfied (and aperiodicity follows immediately as well). More generally, ϕ-irreducibility follows if the n-step transitions $P^n(x, \cdot)$ have positive densities on subsets which expand to $\mathcal{X}$ as $n \to \infty$.

Such considerations are usually sufficient to easily guarantee asymptotic convergence of MCMC algorithms which arise in practice. However, results such as (1.1) only apply when $n \to \infty$. This leads to numerous questions, such as: How large must n be before $P^n(x, A) \approx \pi(A)$? And, how can the Markov chain be modified to make this convergence faster? Each of these questions can be approached experimentally, through repeated simulation and analysis of output for specific examples. However, they can also be considered theoretically, as we now discuss.

1.3 Quantitative Convergence Bounds

In this section, we consider the question of how to obtain rigorous, quantitative bounds on the total variation distance to stationarity of a Markov chain $\{X_n\}$ to its stationary distribution $\pi(\cdot)$, i.e. how to bound

$$\|\mathcal{L}(X_n) - \pi\| \; := \; \sup_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A) - \pi(A)| \, .$$

Of course, if the Markov chain is complicated and high dimensional (as we assume here), then $\mathcal{L}(X_n)$ is complicated too, so our task is non-trivial.

While there are many approaches to this problem, the one we shall consider here is based on the *coupling inequality.* Specifically, let $\{X_n\}$ and $\{X'_n\}$ be two different copies of the Markov chain, each marginally following the transition probabilities $P(x, \cdot)$. Assume that $\{X'_n\}$ was started in stationarity, so that $\mathbf{P}(X'_n \in A) = \pi(A)$ for all n and A. Then by writing $\mathbf{P}(X_n \in A) = \mathbf{P}(X_n \in A, \; X_n = X'_n) + \mathbf{P}(X_n \in A, \; X_n \neq X'_n)$, and similarly for X'_n, it follows that

$$\begin{aligned} \|\mathcal{L}(X_n) - \pi\| &= \; \sup\nolimits_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A) - \pi(A)| \\ &= \; \sup\nolimits_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A) - \mathbf{P}(X'_n \in A)| \\ &= \; \sup\nolimits_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A, \; X_n \neq X'_n) - \mathbf{P}(X'_n \in A, \; X_n \neq X'_n)| \\ &\leq \; \mathbf{P}(X_n \neq X'_n) \, . \end{aligned}$$

In other words, to bound $\|\mathcal{L}(X_n) - \pi\|$, it suffices to "force" $X_n = X'_n$ with high probability. However, this presents its own challenges. In particular, if $\mathcal{X}$ is continuous, then if $\{X_n\}$ and $\{X'_n\}$ proceed independently, then we will usually have $\mathbf{P}(X_n = X'_n) = 0$, which is of no help. On the other hand, if we can define $\{X_n\}$ and $\{X'_n\}$ *jointly* in a way that increases $\mathbf{P}(X_n = X'_n)$, then this can help to bound convergence. One way to accomplish this is with *small sets*, as we discuss next.

1.3.1 Minorisation conditions (small sets)

Suppose we know that $P(x, \cdot) \geq \epsilon \, \nu(\cdot)$, for all $x \in C \subseteq \mathcal{X}$, for some "overlap" probability measure $\nu(\cdot)$. That is,

$$P(x, A) \; \geq \; \epsilon \, \nu(A) \, , \qquad x \in C, \; A \subseteq \mathcal{X} \, . \tag{1.2}$$

Such inequalities are called *minorisation conditions*, and the subset C is called a *small set*. For background, see e.g. [18, 24].

For example, if $P(x, dy)$ has a density $f(x, y)$ with respect to Lebesgue measure $\lambda(\cdot)$, and $f(x, y) \geq \delta$ for $x \in C$ and $y \in B$, then (1.2) is satisfied with $\epsilon = \delta\, \lambda(B)$ and $\nu(A) = \lambda(A \cap B) \,/\, \lambda(B)$. In particular, it is often easy enough to verify (1.2) even if the details of the transitions $P(x, \cdot)$ are quite complicated.

If (1.2) holds, then whenever $(X_{n-1}, X'_{n-1}) \in C \times C$, we can use *Nummelin splitting* [20, 18, 24] to jointly update X_n and X'_n in such a way that $X_n = X'_n$ with probability at least ϵ. Thus, we have managed to "force" $X_n = X'_n$ with non-zero probability, as desired.

Putting this together, it follows that for any $j \in \mathbf{N}$,

$$\|\mathcal{L}(X_n) - \pi\| \;\leq\; (1-\epsilon)^j + \mathbf{P}(N_{n-1} < j)\,, \tag{1.3}$$

where $N_{n-1} = \#\{m : 0 \leq m \leq n-1,\ (X_m, X'_m) \in C \times C\}$ is the number of "opportunities" that the two chains have had to couple by time n.

If $C = \mathcal{X}$, then $N_{n-1} = n$, and (1.3) reduces (with $j = n$) simply to $\|\mathcal{L}(X_n) - \pi\| \leq (1-\epsilon)^n$. This is a very precise and useful inequality, which gives an exponentially-decreasing upper bound on the distance to stationarity, depending only on the value of ϵ from (1.2).

However, in typical MCMC applications it will not be possible to take $C = \mathcal{X}$ due to the inherently "unbounded" nature of the Markov chain. In this case, we need other methods to control N_n. One idea is through a *drift condition*, as we now discuss.

Remark. Of course, strictly speaking, MCMC algorithms are always run on real computers which are finite-state machines, so in some sense the state space $\mathcal{X}$ is always finite. But it is much more useful to model the state spaces as being truly infinite, rather than try to obtain bounds based on some machine-imposed truncation.

1.3.2 Drift conditions

Suppose there is some function $V : \mathcal{X} \to [0, \infty)$, and $\lambda < 1$ and $\Lambda < \infty$, such that

$$\mathbf{E}\Big(V(X_n) \,|\, X_{n-1} = x\Big) \;\leq\; \lambda\, V(x) + \Lambda\,, \qquad x \in \mathcal{X}\,. \tag{1.4}$$

Such inequalities are called *drift conditions*. Intuitively, (1.4) means that when the chain is at large values of V, it will tend to "drift" towards smaller V values.

For this to be useful, we need to be able to couple the chains when they are at small values of V. So, suppose further that (1.2) is satisfied with $C = \{x \in \mathcal{X} : V(x) \leq D\}$ for some $D > 0$, i.e. that

$$P(x, \cdot) \;\geq\; \epsilon\, \nu(\cdot)\,, \qquad \forall x \text{ with } V(x) \leq D\,. \tag{1.5}$$

Condition (1.4) then implies that the pair $\{(X_n, X'_n)\}$ will tend to “drift” towards $C \times C$, so hopefully $\mathbf{P}(N_{n-1} < j)$ will be small, thus making the bound (1.3) useful.

1.3.3 An explicit convergence bound

Putting this all together proves the following bound [28, 30]. (For related results and discussion see [19, 27, 10, 8, 5, 24].)

Theorem 1. *If the drift condition (1.4) and minorisation condition (1.5) hold, with $D > \frac{2\Lambda}{1-\lambda}$, then for any integer $0 \le j \le n$,*

$$\|\mathcal{L}(X_n) - \pi\| \;\le\; (1-\epsilon)^j \;+\; \alpha^{-n+j-1}\Delta^j \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}\big(V(X_0)\big)\right), \quad (1.6)$$

where $\alpha = \frac{1+D}{1+2\Lambda+\lambda D} > 1$ and $\Delta = 1 + 2(\lambda D + \Lambda)$.

If we set $j = \lfloor cn \rfloor$ in (1.6) for appropriate small $c > 0$, then this provides a quantitative, exponentially-decreasing upper bound on $\|\mathcal{L}(X_n) - \pi\|$, easily computed in terms of only the quantities ϵ from (1.5) and λ and Λ from (1.4).

The question remains whether the bound (1.6) is useful in genuinely complicated MCMC algorithms. We now consider an example.

1.3.4 A 20-dimensional example

We now consider a specific 20-dimensional MCMC algorithm. It corresponds to a model for a James-Stein shrinkage estimator, and is a version of a Gibbs sampler related to “variance components models” and “random-effects models”, as applied to data from baseball hitting percentages; for details see [29] and the references therein.

For present purposes, we need know only that the Markov chain’s state space is given by

$$\mathcal{X} \;=\; [0,\infty) \times \mathbf{R} \times \mathbf{R}^{18} \;\subseteq\; \mathbf{R}^{20}\,,$$

and that if we write the chain’s state at time n as $X_n = (A^{(n)}, \mu^{(n)}, \theta_1^{(n)}, \ldots \theta_{18}^{(n)})$, then given X_{n-1}, the chain generates X_n by:

$$A^{(n)} \;\sim\; IG\left(\frac{15}{2},\; 2 + \frac{1}{2}\sum(\theta_i^{(n-1)} - \overline{\theta}^{(n-1)})^2\right)\;;$$

$$\mu^{(n)} \;\sim\; N\left(\overline{\theta}^{(n-1)},\; A^{(n)}/18\right)\;;$$

$$\theta_i^{(n)} \;\sim\; N\left(\frac{\mu^{(n)}\beta + Y_i A^{(n)}}{\beta + A^{(n)}},\; \frac{A^{(n)}\beta}{\beta + A^{(n)}}\right)\,, \quad 1 \le i \le 18\,;$$

where β is a known positive constant, $\{Y_i\}$ are the known actual data values, and $\overline{\theta}^{(n)} = \frac{1}{18}\sum_{i=1}^{18}\theta_i^{(n)}$. Here $N(m,v)$ is a normal distribution with mean m

and variance v, while $IG(a,b)$ is an inverse-gamma distribution with density proportional to $e^{-b/x}x^{-(a+1)}$. This chain is specifically designed so that it will have a stationary probability distribution $\pi(\cdot)$ equal to the posterior distribution for the particular Bayesian statistical model of interest.

This chain represents a typical statistical application of MCMC. In particular, the state space is high-dimensional, and the transition densities are known but messy functions of data values without any particularly nice structure or symmetry. We know the chain has a stationarity distribution $\pi(\cdot)$, but know little else.

On the positive side, it is easily seen that the transition densities for this chain are positive throughout $\mathcal{X}$. So, the asymptotic convergence (1.1) must hold. However, quantitative bounds on the time to stationarity are more challenging, and might at first glance appear intractable. However, using Theorem 1, we are able to achieve this.

Our first challenge is to verify the drift condition (1.4). To do this, we choose the drift function

$$V(A,\mu,\theta_1,\ldots,\theta_{18}) = \sum_{i=1}^{18}(\theta_i - \overline{Y})^2 \,,$$

where $\overline{Y} = \frac{1}{18}\sum_{i=1}^{18} Y_i$. (Intuitively, V measures how far our current vector of values are from the "center" of the given data.) It is then messy but reasonably straightforward to compute [29] that (1.4) is satisfied with $\lambda = 0.000289$ and $\Lambda = 0.161$.

Our next challenge is to verify the minorisation condition (1.5). To do this, we take $D = 1$, and compute [29] that (1.5) is satisfied with $\epsilon = 0.0656$.

We then apply Theorem 1 to conclude that, starting with $\theta_i^{(0)} = \overline{Y}$ for all i (say), and setting $j = n/2$ (for n even, say), we have

$$\|\mathcal{L}(X_n) - \pi\| \;\leq\; (0.967)^n + (0.935)^n(1.17)\,.$$

This is the precise quantitative bound that we sought. In particular, with $n = 140$, we have that

$$\|\mathcal{L}(X_{140}) - \pi(\cdot)\| \;\leq\; 0.009 \;<\; 0.01\,.$$

In other words, we have proved that the chain will "converge" (to within 1% of stationarity) after at most 140 iterations.

Although this is just an *upper* bound on convergence (and, indeed, convergence is probably actually achieved after 10 or fewer iterations), it is the only known rigorous bound. And, since it is very quick and easy to run the Markov chain for 140 iterations on a computer, this bound is of clear practical benefit. Similar bounds have been obtained for other practical examples of MCMC, see e.g. [28, 16].

Remark. We refer to this example as 20-dimensional since the state space is an open subset of $\mathbf{R}^{20}$. However, since the θ_i are conditionally independent given A and μ, one could also say that this Gibbs sampler has just three components, A, μ, and θ, where θ happens to live in $\mathbf{R}^{18}$ instead of $\mathbf{R}$.

1.4 Adaptive MCMC

For a given state space $\mathcal{X}$ and target probability distribution $\pi(\cdot)$, there are many possible MCMC algorithms which will converge asymptotically. An important practical question is, which MCMC choice is "best", or at least good enough to converge after a feasible number of iterations?

Even within a given class of MCMC algorithms, choices of related tuning parameters can be crucial in the algorithms success. A number of recent papers [14, 1, 3, 25, 26, 34, 4, 2] have considered the possibility of having the computer modify the Markov chain transitions while the chain runs, in an effort to seek better convergence. This raises a number of theoretical and practical issues, which we now discuss.

1.4.1 A toy example

Suppose $\pi(\cdot)$ is a simple distribution on the trivial state space $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, with $\pi(x) > 0$ for all $x \in \mathcal{X}$. (For definiteness, take $\pi(x) = 0$ for $x \notin \mathcal{X}$.) Fix $\gamma \in \mathbf{N}$, e.g. $\gamma = 2$. Consider a "random-walk Metropolis" (RWM) algorithm, defined as follows:

- Given X_n, first propose a state $Y_{n+1} \in \mathbf{Z}$, with

$$Y_{n+1} \sim \text{Uniform}\{X_n - \gamma, \ldots, X_n - 1, X_n + 1, \ldots, X_n + \gamma\}\,.$$

- Then, with probability $\min[1,\ \pi(Y_{n+1})/\pi(X_n)]$, *accept* this proposal by setting $X_{n+1} = Y_{n+1}$.
- Otherwise, with probability $1 - \min[1,\ \pi(Y_{n+1})/\pi(X_n)]$, *reject* this proposal by setting $X_{n+1} = X_n$.

It is easily seen that these transition probabilities have $\pi(\cdot)$ as a stationary distribution, and are irreducible and aperiodic, so we have asymptotic convergence as in (1.1), for any choice of $\gamma \in \mathbf{N}$. (This example is discussed in [3, 25]; for an interactive display see [31].)

However, this still leaves the question of choice of γ. If $\gamma = 1$, the chain will move at most one unit at each iteration, leading to slow convergence. On the other hand, if say $\gamma = 50$, then the chain will usually propose values outside of $\mathcal{X}$ which will all be rejected, again leading to slow convergence. Best is a "moderate" value of γ, e.g. $\gamma = 4$.

In a more complicated example, the best choice of a tuning parameter (like γ) will be far less obvious. So, we consider the possibility of automating the choice of γ. As an example, we might adapt γ as follows:

- Start with γ set to $\Gamma_0 = 2$ (say).

- Each time a proposal is accepted, set $\Gamma_{n+1} = \Gamma_n + 1$ (so γ increases, and the acceptance rate decreases).
- Each time a proposal is rejected, set $\Gamma_{n+1} = \max(\Gamma_n - 1, 1)$ (so γ decreases, and the acceptance rate increases).

This appears to be a logical way for the computer to seek out good choices of γ, and in simulations [31] it appears to work well for a while. However, if (say) $\pi\{2\}$ is very small, then the chain will eventually get "stuck" with $X_n = \Gamma_n = 1$ for long stretches of time. This is due to a certain *asymmetry*: for the adaptive chain, *entering* the region $\{X_n = \Gamma_n = 1\}$ is much easier than *leaving* it. In particular, this adaptive chain does not converge to $\pi(\cdot)$ at all, but rather may converge to a different distribution giving far too much weight to the state 1. That is, the adaption – which attempted to *improve* the convergence – actually ruined the convergence entirely.

1.4.2 An adaptive MCMC convergence theorem

In light of counter-examples like the above, we seek conditions which guarantee that adaptive MCMC schemes will in fact converge. One such result is the following, from [25]; for related results see e.g. [1, 3, 34, 4, 2]. To state it, define the "ϵ convergence time function" $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to \mathbf{N}$ by

$$M_\epsilon(x,\gamma) \;=\; \inf\left\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \leq \epsilon\right\}.$$

Theorem 2. *An adaptive scheme $\{(X_n, \Gamma_n)\}$, using transition kernels $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ will converge, i.e. $\lim_{n\to\infty} \|\mathcal{L}(X_n) - \pi(\cdot)\| = 0$, assuming (i) $\pi(\cdot)$ is stationary for each individual P_γ, and (ii) the "Diminishing Adaptation" property that $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| = 0$ in probability, and (iii) the "Containment" property that for all $\epsilon > 0$, the values $\{M_\epsilon(X_n, \Gamma_n)\}$ remains bounded in probability as $n \to \infty$.*

In this theorem, condition (i) is basic to any adaptive MCMC algorithm, and (ii) can be ensured by careful design of the adaption, while (iii) is an unfortunate technical condition though it is nearly always satisfied in practical examples [4]. Furthermore, these same conditions also guarantee central limit theorems (CLTs) for adaptive MCMC with bounded functionals, though not necessarily with unbounded functionals [34].

In light of this theorem, we see that the toy example of 1.4.1 satisfies conditions (i) and (iii), but not (ii). However, (ii) will be satisfied, and the chain will converge to $\pi(\cdot)$, if we modify the adaption so that at time n, it only adapts with probability $p(n)$ for some probabilities $p(n) \to 0$, otherwise the value of γ is left unchanged. In particular, we could choose, say, $p(n) = 1/n$, in which case we would still have $\sum_n p(n) = \infty$ and thus still have an infinite amount of adaptation, and yet still guarantee convergence.

1.4.3 A 100-dimensional example

For complicated examples in high dimensions, adaption is not as trivial as for the example of Section 1.4.1, but it is still quite feasible.

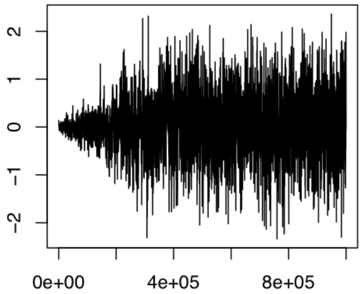
For example, it is known (see [23] and the references therein) that if target distribution $\pi(\cdot)$ is (approximately) a high-dimensional normal distribution with covariance Σ, then the optimal Gaussian proposal distribution for a RWM algorithm is equal to $N(x,\, (2.38)^2 d^{-1}\, \Sigma)$.

Now, the target covariance Σ is generally unknown, but it can be approximated by Σ_n, the empirical covariance of the first n iterations of the Markov chain. This suggests [14, 26] an adaptive MCMC algorithm with proposal distribution at the n^{th} iteration given by the mixture distribution

$$Q_n(x,\cdot) \;=\; 0.95\; N\Big(x,\, (2.38)^2 d^{-1}\, \Sigma_n\Big) \;+\; 0.05\; N\Big(x,\; (0.1)^2 d^{-1}\, I_d\Big)$$

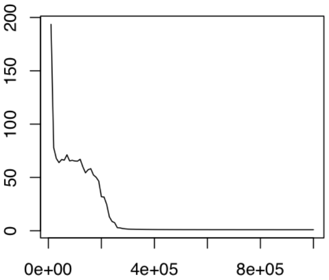
[if Σ_n is non-singular, otherwise say $Q_n(x,\cdot) \;=\; N\big(x,\; (0.1)^2 d^{-1}\, I_d\big)$]. Such algorithms will generally satisfy condition (ii) of Theorem 2, and furthermore will satisfy condition (iii) provided the tails of $\pi(\cdot)$ are not too heavy [4].

If we run [26] this algorithm on an example in dimension $d = 100$, then a trace plot of the first coordinate (plotted against iteration number) looks as follows:



A close inspection of this plot shows that the first coordinate is initially “stuck” at values very close to 0. Then, after about 300,000 iterations, the empirical Σ_n gets close to the true Σ, so the adaptive algorithm “finds” good proposal distributions and starts mixing well. At this point, the first coordinate mixes nicely and efficiently over values concentrated between about -1 and 1, corresponding to accurate samples of the first coordinate from the true target distribution $\pi(\cdot)$.

This interpretation can be confirmed by looking at a plot of the sub-optimality factor $b_n \;\equiv\; d\left(\sum_{i=1}^{d} \lambda_{in}^{-2}\right) / \left(\sum_{i=1}^{d} \lambda_{in}^{-1}\right)^2$, where $\{\lambda_{in}\}$ are the eigenvalues of the matrix $\Sigma_n^{1/2}\Sigma^{-1/2}$. This quantity b_n is known [23] to measure the convergence slow-down factor of a chain using the covariance estimate Σ_n obtained after n iterations, compared to a chain using the true covariance Σ. The plot clearly shows that the values of b_n are initially very large, and then get close to 1 after about 300,000 iterations:



This further confirms that after about 300,000 iterations, the adaptive scheme "finds" good values of Σ_n which accurately approximate Σ, leading to fast and accurate convergence. And, since the 100×100 covariance matrix Σ involves 5,050 unknown values, it seems clear that this optimisation could not have been done manually, and that the adaptive MCMC scheme really was essential to achieving fast convergence to $\pi(\cdot)$. Similar success has been found in other high-dimensional examples (see e.g. [14, 26]), and we expect that adaptive MCMC will be used more often in the years ahead.

1.5 Connection with QMC?

In the context of a conference on "MCQMC", it is reasonable to ask about the placement of MCMC algorithms in the Monte Carlo (MC) / Quasi-Monte Carlo (QMC) divide.

For the most part, MCMC algorithms are squarely on the MC side, using pseudorandom number generators to power the iterations according to (approximately) the laws of probability. Furthermore, much of the theoretical analysis, including that discussed herein, uses probability theory and assumes the algorithms follow probabilistic laws. However, it has been observed [21, 6, 22] that it is also possible to power MCMC algorithms using quasi-random sequences.

In principle, QMC is "smarter" than just using (pseudo)random numbers, so should be better. Furthermore, it is known [13, 7, 6] that using e.g. *antithetic* or other not-entirely-random variates can sometimes speed up MCMC convergence. So, it seems that future MCMC work – both applied and theoretical – might make more use of quasi-randomness and thus make more of a leap towards the QMC world.

However, many of the ideas considered herein – ideas like "irreducible", "coupling", "minorisation", "drift", "Diminishing Adaptation", "Containment", etc. – all use *probabilistic intuition* and it is not clear how to translate them into QMC ideas. Furthermore, in many cases we may not know enough about the (complicated, messy, high-dimensional) target distribution to design QMC effectively, and it might be easier to verify "weak" conditions like minorisation and drift.

Thus, in this paper, we have treated the algorithms as being “truly random”, i.e. within the context of traditional Monte Carlo. However, we look forward to more QMC ideas finding their way into MCMC in the future.

1.6 Summary

The main points of this article may be summarised as follows:

- MCMC algorithms are extremely widely used, in Bayesian statistics and elsewhere.
- *Quantitative convergence bounds* are a very important topic for MCMC, with both practical and theoretical implications.
- An approach using the *coupling inequality*, together with *minorisation* and *drift conditions*, can provide specific, useful bounds (like “140”) on the convergence times even of rather complicated Markov chains on continuous, high-dimensional state spaces.
- For a given problem, many different MCMC algorithms are available, and it can be difficult (though very important) to choose among them.
- *Adaptive MCMC* is a promising recent method of getting the computer to help find better MCMC algorithms during the course of a run.
- Naive application of adaptive MCMC may fail to converge to $\pi(\cdot)$.
- However, theorems are available which prove the validity of adaptive MCMC under certain conditions which can often be verified for specific adaptive schemes.
- Adaptive MCMC works well in some high-dimensional statistics-related examples, including an adaptive random-walk Metropolis (RWM) algorithm in dimension 100.
- While MCMC is traditionally on the “MC” side of the MC / QMC divide, we anticipate greater connections between MCMC algorithms and quasi-Monte Carlo ideas in the future.

And more generally:

- Theory informs the applied use of MCMC in many ways, thus providing an excellent arena in which mathematical results can have a genuine and widespread impact on applications of algorithms.

It is to be hoped that many experts in MC and QMC will get more interested in MCMC algorithms, and make further theoretical contributions to this interesting and widely applicable area.

1.7 Acknowledgements

The author’s research is supported in part by NSERC of Canada. All of the author’s papers, applets, and software are freely available at his web page `probability.ca/jeff`.

References

1. C. Andrieu and E. Moulines (2006), On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. Ann. Appl. Prob. **16**, 1462–1505.
2. Y. Atchadé and G. Fort (2008), Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. Preprint.
3. Y.F. Atchadé and J.S. Rosenthal (2005), On Adaptive Markov Chain Monte Carlo Algorithms. Bernoulli **11**, 815–828.
4. Y. Bai, G.O. Roberts, and J.S. Rosenthal (2008), On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms. Preprint.
5. P.H. Baxendale (2005), Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Ann. Appl. Prob. **15**, 700–738.
6. R.V. Craiu and C. Lemieux (2007), Acceleration of the Multiple-try Metropolis Algorithm using Antithetic and Stratified sampling. Stat. and Comput. **17(2)**, 109–120.
7. R.V. Craiu and X.-L. Meng (2005), Multi-process parallel antithetic coupling for forward and backward Markov chain Monte Carlo. Ann. Stat. **33(2)**, 661–697.
8. R. Douc, E. Moulines, and J.S. Rosenthal (2002), Quantitative bounds on convergence of time-inhomogeneous Markov Chains. *Annals of Applied Probability* **14**, (2004), 1643–1665.
9. R. Douc, E. Moulines, and P. Soulier (2007), Computable convergence rates for sub-geometric ergodic Markov chains. Bernoulli **13**, 831–848.
10. G. Fort and E. Moulines (2000), Computable Bounds For Subgeometrical And Geometrical Ergodicity. Unpublished manuscript. Available at: http://citeseer.ist.psu.edu/fort00computable.html
11. A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.
12. S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.
13. P.J. Green and X.-L. Han (1992), Metropolis methods, Gaussian proposals, and antithetic variables. In Stochastic Models, Statistical Methods and Algorithms in Image Analysis (P. Barone et al., Eds.). Springer, Berlin.
14. H. Haario, E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm. Bernoulli **7**, 223–242.
15. W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.
16. G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science **16**, 312–334.
17. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.
18. S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at: http://probability.ca/MT/
19. S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981–1011.
20. E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.
21. A.B. Owen and S.D. Tribble (2005), A quasi-Monte Carlo Metropolis algorithm. PNAS **102(25)**, 8844–8849.

22. A.B. Owen and S.D. Tribble (2008), Constructions of weakly CUD sequences for MCMC. Elec. J. Stat. **2**, 634–660.
23. G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**, 351–367.
24. G.O. Roberts and J.S. Rosenthal (2004), General state space Markov chains and MCMC algorithms. Prob. Surv. **1**, 20–71.
25. G.O. Roberts and J.S. Rosenthal (2007), Coupling and Ergodicity of Adaptive MCMC. J. Appl. Prob. **44**, 458–475.
26. G.O. Roberts and J.S. Rosenthal (2006), Examples of Adaptive MCMC. J. Comp. Graph. Stat., to appear.
27. G.O. Roberts and R.L. Tweedie (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229. Corrigendum, Stoch. Proc. Appl. **91** (2001), 337–338.
28. J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.
29. J.S. Rosenthal (1996), Convergence of Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.
30. J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. Elec. Comm. Prob. **7**, No. 13, 123–128.
31. J.S. Rosenthal (2004), Adaptive MCMC Java Applet. Available at: http://probability.ca/jeff/java/adapt.html
32. M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528–550.
33. L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.
34. C. Yang (2008), On The Weak Law of Large Numbers for Unbounded Functionals for Adaptive MCMC. Preprint.