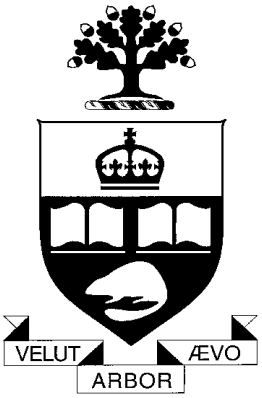

Coupling and Ergodicity of Adaptive MCMC

by

Gareth O. Roberts
Department of Mathematics and Statistics
Lancaster University

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 0501 March 1, 2005

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Coupling and Ergodicity of Adaptive MCMC

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(March 2005; last revised March 2007.)

Abstract. We consider basic ergodicity properties of adaptive MCMC algorithms under minimal assumptions, using coupling constructions. We prove convergence in distribution and a weak law of large numbers. We also give counter-examples to demonstrate that the assumptions we make are not redundant.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms are a widely used method of approximately sampling from complicated probability distributions. However, a wide variety of different MCMC algorithms are available, and it is often necessary to *tune* the scaling and other parameters before the algorithm will converge efficiently.

It is tempting to automate and improve this tuning through the use of adaptive MCMC algorithms, which attempt to "learn" the best parameter values while they run. In this paper, we consider the extent to which ergodicity and stationarity of the specified target distribution are preserved under adaptation.

Adaptive MCMC methods using regeneration times and other complicated constructions have been proposed by Gilks et al. (1998), Brockwell and Kadane (2005), and elsewhere. On the other hand, related adaptive schemes can often fail to preserve stationarity of the target distribution (see e.g. Proposition 4 below). This leads to the question of what conditions on natural (non-regenerative) adaptive MCMC algorithms guarantee that the stationarity of $\pi(\cdot)$ will be preserved.

* Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England. Email: g.o.roberts@lancaster.ac.uk.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Web: http://probability.ca/jeff/ Supported in part by NSERC of Canada.

A significant step in this direction was made by Haario et al. (2001). They proposed an Adaptive Metropolis algorithm which attempts to optimise the proposal distribution of a Metropolis algorithm to be approximately $(2.38)^2\Sigma\,/\,d$, where d is the dimension, and Σ is the $d \times d$ covariance matrix of the d coordinates under stationarity. (Such a proposal is optimal in certain settings according to the results of Roberts et al., 1997; see also Roberts and Rosenthal, 2001, and Bédard, 2006.) They do so by estimating Σ from the empirical distribution of the Markov chain output so far, thus adapting the estimate of Σ while the algorithm runs (but less and less as time goes on). They prove that a particular version of this algorithm (which involves adding a small multiple of the identity matrix to each proposal covariance) correctly converges in distribution to the target $\pi(\cdot)$.

It was observed by Andrieu and Robert (2002) that the algorithm of Haario et al. (2001) can be viewed as a version of the Robbins-Monro stochastic control algorithm (Robbins and Monro, 1951). The results of Haario et al. were then generalised by Atchadé and Rosenthal (2005) and Andrieu and Moulines (2003), proving convergence of more general adaptive MCMC algorithms. (Andrieu and Moulines, 2003, also prove a central limit theorem result.) Those two papers removed many restrictions and limitations of the Haario et al. result, but at the expense of requiring other technical hypotheses which may be difficult to verify in practice.

In this paper, we present somewhat simpler conditions, which still ensure ergodicity and stationarity of the specified target distribution. After introducing our notation and terminology (Section 2), and considering some special cases (Section 3), we present a running example (Section 4) which illustrates adaptive MCMC's potential pitfalls. We then use a bivariate coupling construction to prove the validity of adaptive MCMC in uniform (Section 5) and non-uniform (Section 6) settings. We make connections to drift conditions (Section 7) and recurrence properties (Section 8), and prove a weak law of large numbers (Section 9), before presenting some general discussion of adaptive MCMC (Section 10).

2. Preliminaries.

We let $\pi(\cdot)$ be a fixed "target" probability distribution, on a state space $\mathcal{X}$ with σ-algebra $\mathcal{F}$. The goal of MCMC is to approximately sample from $\pi(\cdot)$ through the use of Markov chains, particularly when $\pi(\cdot)$ is too complicated and high-dimensional to facilitate more direct sampling.

We let $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels on $\mathcal{X}$, each of which has $\pi(\cdot)$ as a stationary distribution: $(\pi\, P_\gamma)(\cdot) = \pi(\cdot)$.

Assuming P_γ is ϕ-irreducible and aperiodic (which it usually will be), this implies (see e.g. Meyn and Tweedie, 1993) that P_γ be *ergodic* for $\pi(\cdot)$, i.e. that for all x, $\lim_{n\to\infty} \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| = 0$, where $\|\mu(\cdot) - \nu(\cdot)\| = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|$ is the usual total variation distance. That is, P_γ represents a "valid" MCMC algorithm, i.e. defines a Markov chain which will converge in distribution to the target $\pi(\cdot)$. So, if we keep γ fixed, then the Markov chain algorithm described by P_γ will eventually converge to $\pi(\cdot)$.

However, some choices of γ may lead to far less efficient algorithms than others, and it may be difficult to know in advance which choices of γ are preferable. To deal with this, *adaptive* MCMC proposes that at each time n we let the choice of γ be given by a $\mathcal{Y}$-valued random variable Γ_n, updated according to specified rules.

Formally, for $n = 0, 1, 2, \ldots$, we have a $\mathcal{X}$-valued random variable X_n representing the state of the algorithm at time n, and a $\mathcal{Y}$-valued random variable Γ_n representing the choice of kernel to be used when updating from X_n to X_{n+1}. We let

$$\mathcal{G}_n = \sigma(X_0, \ldots, X_n, \Gamma_0, \ldots, \Gamma_n)$$

be the filtration generated by $\{(X_n, \Gamma_n)\}$. Thus,

$$\mathbf{P}[X_{n+1} \in B \,|\, X_n = x, \Gamma_n = \gamma, \mathcal{G}_{n-1}] \;=\; P_\gamma(x, B)\,, \qquad x \in \mathcal{X},\ \gamma \in \mathcal{Y},\ B \in \mathcal{F}\,, \tag{1}$$

while the conditional distribution of Γ_{n+1} given $\mathcal{G}_n$ is to be specified by the particular adaptive algorithm being used. We let

$$A^{(n)}((x,\gamma), B) = \mathbf{P}[X_n \in B \,|\, X_0 = x, \Gamma_0 = \gamma]\,, \qquad B \in \mathcal{F}$$

record the conditional probabilities for X_n for the adaptive algorithm, given the initial conditions $X_0 = x$ and $\Gamma_0 = \gamma$. Note that $A^{(n)} \neq \prod_{i=0}^{n-1} P_{\Gamma_i}$, since $A^{(n)}$ represents the *unconditional* distribution of the algorithm, equivalent to *integrating* over the distributions of $\Gamma_1, \ldots, \Gamma_{n-1}$.

Finally, we let

$$T(x, \gamma, n) \;=\; \|A^{(n)}((x,\gamma),\cdot) - \pi(\cdot)\| \;\equiv\; \sup_{B \in \mathcal{F}} |A^{(n)}((x,\gamma), B) - \pi(B)|$$

denote the total variation distance between the distribution of our adaptive algorithm at time n, and the target distribution $\pi(\cdot)$. Call the adaptive algorithm *ergodic* if $\lim_{n\to\infty} T(x, \gamma, n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$. We can then ask, will the adaptive chain necessarily be ergodic? Since each P_γ converges to $\pi(\cdot)$, one might expect that our adaptive algorithm does too. However, we shall see below (Section 4) that this is not always the case.

3. Some Special Cases.

Adaptive MCMC, in the sense that we have defined it above, includes as special cases a number of previously considered schemes, including most obviously:

- Traditional MCMC: $\Gamma_n \equiv 1$ for all n.
- Systematic-scan hybrid algorithm: $(\Gamma_n) = (1, 2, \ldots, d, 1, 2, \ldots, d, 1, 2, \ldots)$, where e.g. P_i moves only the i^{th} coordinate.
- Random-scan hybrid algorithm: $\{\Gamma_n\}$ are i.i.d. $\sim$ Uniform$\{1, 2, \ldots, d\}$.

In this section, we make some observations about these and other special cases, to provide context for the more general results to come.

To begin, call an adaptive MCMC algorithm an *independent adaptation* if for all n, Γ_n is independent of X_n. (This includes the traditional and hybrid cases described above.) For independent adaptations, stationarity of $\pi(\cdot)$ is guaranteed:

Proposition 1. *Consider an independent adaptation algorithm $A^{(n)}((x,\gamma),\cdot)$, where $\pi(\cdot)$ is stationary for each $P_\gamma(x,\cdot)$. Then $\pi(\cdot)$ is also stationary for $A^{(n)}((x,\gamma),\cdot)$, i.e.*

$$\int_{x\in\mathcal{X}} \mathbf{P}[X_{n+1} \in B \,|\, X_n = x, \mathcal{G}_{n-1}]\; \pi(dx) \;=\; \pi(B)\,, \qquad B \in \mathcal{F}\,.$$

Proof. Using (1), and the independence of Γ_n and X_n, and the stationarity of $\pi(\cdot)$ for P_γ, we have:

$$\int_{x\in\mathcal{X}} \mathbf{P}[X_{n+1}\in B \,|\, X_n = x, \mathcal{G}_{n-1}]\,\pi(dx)$$

$$= \int_{x\in\mathcal{X}}\int_{\gamma\in\mathcal{Y}} \mathbf{P}[X_{n+1}\in B \,|\, X_n = x, \Gamma_n=\gamma, \mathcal{G}_{n-1}]\,\mathbf{P}[\Gamma_n\in d\gamma \,|\, X_n = x, \mathcal{G}_{n-1}]\,\pi(dx)$$

$$= \int_{x\in\mathcal{X}}\int_{\gamma\in\mathcal{Y}} P_\gamma(x,B)\,\mathbf{P}[\Gamma_n\in d\gamma \,|\, \mathcal{G}_{n-1}]\,\pi(dx)$$

$$= \int_{\gamma\in\mathcal{Y}} \mathbf{P}[\Gamma_n\in d\gamma \,|\, \mathcal{G}_{n-1}] \int_{x\in\mathcal{X}} P_\gamma(x,B)\,\pi(dx)$$

$$= 1\cdot\pi(B) \;=\; \pi(B)\,.$$

■

On the other hand, it is well known that even for independent adaptations, irreducibility may be destroyed:

Example 2. Let $\mathcal{X} = \{1,2,3,4\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 2/7$, and $\pi\{4\} = 1/7$. Let $P_1(1,\{2\}) = P_1(3,\{1\}) = P_1(4,\{3\}) = 1$, and $P_1(2,\{3\}) = P_1(2,\{4\}) = 1/2$. Similarly, let $P_2(2,\{1\}) = P_2(3,\{2\}) = P_2(4,\{3\}) = 1$, and $P_2(1,\{3\}) = P_2(1,\{4\}) = 1/2$. Then it is easily checked that each of P_1 and P_2 are irreducible and aperiodic, with stationary distribution $\pi(\cdot)$. On the other hand, $(P_1P_2)(1,\{1\}) = 1$, so when beginning in state 1, the systematic-scan adaptive chain P_1P_2 alternates between states 1 and 2 but never reaches the state 3. Hence, this adaptive algorithm fails to be irreducible, and also $T(x,\gamma,n) \not\to 0$ as $n\to\infty$, even though each individual P_i is ergodic. ■

Another special case of adaptive MCMC is to introduce some stopping time τ with $\mathbf{P}(\tau<\infty) = 1$, such that no adaptations are done after time τ, i.e. such that $\Gamma_n = \Gamma_\tau$ whenever $n\geq\tau$. This scheme, which we refer to as *finite adaptation*, has been proposed by e.g. Pasarica and Gelman (2003) and E. Moulines (personal communication). It is analogous to the common MCMC practice of using a number initial "trial" Markov chain runs with different tunings, to determine good parameter values, and then using a final MCMC run with fixed parameters to accomplish the sampling. Finite sampling schemes always preserve asymptotic convergence:

Proposition 3. *Consider a finite adaptation MCMC algorithm, in which each individual P_γ is ergodic, i.e., $\lim_{n\to\infty} \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| = 0$ for all $\gamma \in \mathcal{Y}$ and $x \in \mathcal{X}$. Then the finite adaptation MCMC algorithm is also ergodic.*

Proof. Let $d(x,\gamma,n) = \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|$. It follows from the assumptions that $\lim_{n\to\infty} d(x,\gamma,n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$. Hence, conditional on X_τ and Γ_τ, $\lim_{n\to\infty} d(X_\tau,\Gamma_\tau,n) = 0$. The result follows from integrating over the distributions of X_τ and Γ_τ, and using the Bounded Convergence Theorem. ∎

Both finite and independent adaptive chains represent "safe" methods of implementing adaptation, in the sense that they provide some adaptation without destroying the stationarity of $\pi(\cdot)$. However, of greater interest are *dependent*, *infinite* adaptations, i.e. adaptations which continue to modify the Γ_n, by continuing to learn based on the values of X_n. In such cases, typically the pair sequence $\{(X_n,\Gamma_n)\}_{n=0}^\infty$ is Markovian, in which case we call the algorithm a *Markovian adaptation*, but no assumptions about independent or finite adaptations can be made. This leads to the question of when such adaptations preserve the stationarity of $\pi(\cdot)$, and the asymptotic distributional convergence of the algorithm.

4. Running Example.

To illustrate the limitations of adaptive MCMC, and the application of our theorems, we present the following running example. This example was discussed in Atchadé and Rosenthal (2005); an animated Java applet version is also available (Rosenthal, 2004).

Let $K \geq 4$ be an integer, and let $\mathcal{X} = \{1,2,\ldots,K\}$. Let $\pi\{2\} = b > 0$ be very small, and $\pi\{1\} = a > 0$, and $\pi\{3\} = \pi\{4\} = \ldots = \pi\{K\} = (1-a-b)/(K-2) > 0$. Let $\mathcal{Y} = \mathbf{N}$. For $\gamma \in \mathcal{Y}$, let P_γ be the kernel corresponding to a random-walk Metropolis algorithm for $\pi(\cdot)$, with proposal distribution

$$Q_\gamma(x,\cdot) = \text{Uniform}\{x-\gamma, x-\gamma+1, \ldots, x-1, x+1, x+2, \ldots, x+\gamma\},$$

i.e. uniform on all the integers within γ of x, aside from x itself. The kernel P_γ then proceeds, given X_n and Γ_n, by first choosing a proposal state $Y_{n+1} \sim Q_{\Gamma_n}(X_n,\cdot)$. With

probability $\min[1,\ \pi(Y_{n+1}) \,/\, \pi(X_n)]$ it then accepts this proposal by setting $X_{n+1} = Y_{n+1}$. Otherwise, with probability $1 - \min[1,\ \pi(Y_{n+1}) \,/\, \pi(X_n)]$, it rejects this proposal by setting $X_{n+1} = X_n$. (If $Y_{n+1} \notin \mathcal{X}$, then the proposal is always rejected; this corresponds to setting $\pi(y) = 0$ for $y \notin \mathcal{X}$.)

We define the adaptive scheme as follows. Begin with $\Gamma_0 = 1$ (say). Let $M \in \mathbf{N} \cup \{\infty\}$ and let $p : \mathbf{N} \to [0, 1]$. For $n = 0, 1, 2, \ldots$, given X_n and Γ_n, if the next proposal is accepted (i.e., if $X_{n+1} \neq X_n$) and $\Gamma_n < M$, then with probability $p(n)$ let $\Gamma_{n+1} = \Gamma_n + 1$, otherwise let $\Gamma_{n+1} = \Gamma_n$. Otherwise, if the next proposal is rejected (i.e., if $X_{n+1} = X_n$) and $\Gamma_n > 1$, then with probability $p(n)$ let $\Gamma_n = \Gamma_{n-1} - 1$, otherwise let $\Gamma_{n+1} = \Gamma_n$. In words, with probability $p(n)$, we increase γ (to a maximum of M) each time a proposal is accepted, and decrease γ (to a minimum of 1) each time a proposal is rejected.

We record a few specific versions of this scheme:

- The "original running example" has $M = \infty$ and $p(n) \equiv 1$, i.e. it modifies Γ_n in every iteration except when $\Gamma_n = 1$ and the next proposal is rejected.
- The "singly-modified running example" has $M = \infty$ but arbitrary $p(n)$.
- The "doubly-modified running example" has $M < \infty$ and arbitrary $p(n)$.
- The "One-Two" version has $M = 2$ and $p(n) \equiv 1$.

The intuition for these schemes is that accepted proposals indicate there may be room for γ to grow, while rejected proposals indicate γ may be too large. Indeed, this scheme is somewhat analogous to the Adaptive Metropolis algorithm of Haario et al. (2001), in that it attempts to search for an optimal proposal scaling to obtain a reasonable acceptance rate (not too close to either 0 or 1). However, and perhaps surprisingly, this simple adaptive scheme can completely destroy convergence to $\pi(\cdot)$:

Example 4. Let $\epsilon > 0$, and consider One-Two version with $K = 4$, $a = \epsilon$, and $b = \epsilon^3$. Then it is easily verified that there is $c > 0$ such that $\mathbf{P}[X_3 = \Gamma_3 = 1 \,|\, X_0 = x, \Gamma_0 = \gamma] \geq c\,\epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, i.e. the algorithm has $O(\epsilon)$ probability of reaching the configuration $\{x = \gamma = 1\}$. On the other hand, $\mathbf{P}[X_1 = \Gamma_1 = 1 \,|\, X_0 = \Gamma_0 = 1] = 1 - \epsilon^2/2$, i.e. the algorithm has just $O(\epsilon^2)$ probability of leaving the configuration $\{x = \gamma = 1\}$ once it is

there. This probabilistic asymmetry implies that $\lim_{\epsilon \searrow 0} \lim_{n\to\infty} \mathbf{P}[X_n = \Gamma_n = 1] = 1$. Hence,

$$\lim_{\epsilon \searrow 0} \lim_{n\to\infty} T(x, \gamma, n) \geq \lim_{\epsilon \searrow 0} (1 - \pi\{1\}) = \lim_{\epsilon \searrow 0} (1 - \epsilon) = 1 .$$

In particular, for any $\delta > 0$, there is $\epsilon > 0$ with $\lim_{n\to\infty} T(x, \gamma, n) \geq 1 - \delta$, so the algorithm does not converge at all. ∎

Hence, for this running example, ergodicity of the adaptive algorithm does *not* hold. On the other hand, below we shall prove some theorems giving sufficient conditions to ensure ergodicity. Along the way, we shall prove (Corollary 7) that the doubly-modified running example is ergodic, provided $p(n) \to 0$. We shall then prove (Corollary 16) that the singly-modified running example is also ergodic, again provided that $p(n) \to 0$.

5. Uniformly Converging Case.

Our next result requires that the convergence to $\pi(\cdot)$ of the various P_γ kernels all be uniformly bounded (though we shall relax this condition in Section 6). It also requires that the amount of adapting diminishes as $n \to \infty$, which can be achieved either by modifying the parameters by smaller and smaller amounts (as in the Adaptive Metropolis algorithm of Haario et al., 2001), or by doing the adaptations with smaller and smaller probability (as in our singly-modified running example, above, with adaptation probabilities $p(n) \to 0$). In either case, it is still permitted to have an infinite total amount of adaptation (e.g., to have $\sum_n p(n) = \infty$ in our example, or to have $\sum_n D_n = \infty$ in the theorem below). In particular, there is no requirement that the Γ_n converge.

Theorem 5. *Consider an adaptive MCMC algorithm, on a state space $\mathcal{X}$, with adaptation index $\mathcal{Y}$, so $\pi(\cdot)$ is stationary for each kernel P_γ for $\gamma \in \mathcal{Y}$. Assume that:*

(a) [Simultaneous Uniform Ergodicity] For all $\epsilon > 0$, there is $N = N(\epsilon) \in \mathbf{N}$ such that $\|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$; and

(b) [Diminishing Adaptation] $\lim_{n\to\infty} D_n = 0$ in probability, where $D_n = \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ is a $\mathcal{G}_{n+1}$-measurable random variable (depending on the random values Γ_n and Γ_{n+1}).

Then the adaptive algorithm is ergodic.

Proof. Let $\epsilon > 0$. Choose $N = N(\epsilon)$ as in condition (a). Then let $H_n = \{D_n \geq \epsilon/N^2\}$, and use condition (b) to choose $n^* = n^*(\epsilon) \in \mathbf{N}$ large enough so that

$$\mathbf{P}(H_n) \leq \epsilon/N\,, \qquad n \geq n^*\,. \tag{2}$$

To continue, fix a "target time" $K \geq n^* + N$. We shall construct a coupling which depends on the target time K (cf. Roberts and Rosenthal, 2002), to prove that $\mathcal{L}(X_K) \approx \pi(\cdot)$.

Define the event $E = \bigcap_{i=n+1}^{n+N} H_i^c$. It follows from (2) that for $n \geq n^*$, we have $\mathbf{P}(E) \geq 1-\epsilon$. By the triangle inequality and induction, on the event E we have $\sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+k}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| < \epsilon/N$ for all $k \leq N$, and in particular

$$\|P_{\Gamma_{K-N}}(x,\cdot) - P_{\Gamma_m}(x,\cdot)\| < \epsilon/N \quad \text{on } E\,, \qquad x \in \mathcal{X}, \quad K-N \leq m \leq K\,. \tag{3}$$

To construct the coupling, first construct the original adaptive chain $\{X_n\}$ together with its adaptation sequence $\{\Gamma_n\}$, starting with $X_0 = x$ and $\Gamma_0 = \gamma$. We claim that on E, we can construct a second chain $\{X'_n\}_{n=K-N}^{K}$ such that $X'_{K-N} = X_{K-N}$, and $X'_n \sim P_{\Gamma_{K-N}}(X'_{n-1},\cdot)$ for $K-N+1 \leq n \leq K$, and $\mathbf{P}[X'_i = X_i \text{ for } K-N \leq i \leq m] \geq 1 - [m-(K-N)]\,\epsilon/N$ for $K-N \leq m \leq K$.

Indeed, the claim is trivially true for $m = K-N$. Suppose it is true for some value m. Then conditional on $\mathcal{G}_m$ and the event that $X'_i = X_i$ for $K-N \leq i \leq m$, we have $X_{m+1} \sim P_{\Gamma_m}(X_m,\cdot)$ and $X'_{m+1} \sim P_{\Gamma_{K-N}}(X_m,\cdot)$. It follows from (3) that the conditional distributions of X_{m+1} and X'_{m+1} are within ϵ/N of each other. Hence, by e.g. Roberts and Rosenthal (2004, Proposition 3(g)), we can ensure that $X'_{m+1} = X_{m+1}$ with probability $\geq 1-\epsilon/N$. It follows that $\mathbf{P}[X'_i = X_i \text{ for } K-N \leq i \leq m+1] \geq \mathbf{P}[X'_i = X_i \text{ for } K-N \leq i \leq m]\,(1-\epsilon/N) \geq \big(1-[m-(K-N)]\,\epsilon/N\big)(1-\epsilon/N) \geq 1-[m+1-(K-N)]\,\epsilon/N$. The claim thus follows by induction.

In particular, this shows that on E, $\mathbf{P}[X'_K = X_K] \geq 1-(K-(K-N))\epsilon/N = 1-\epsilon$. That is, $\mathbf{P}[X'_K \neq X_K,\ E] < \epsilon$.

On the other hand, conditioning on X_{K-N} and using condition (a), we have $\|P^N_{\Gamma_{K-N}}(X_{K-N},\cdot) - \pi(\cdot)\| < \epsilon$. Integrating over the distribution of X_{K-N} gives that $\|\mathcal{L}(X'_K) - \pi(\cdot)\| < \epsilon$. It follows (again from e.g. Roberts and Rosenthal, 2004, Proposition 3(g)) that we can construct $Z \sim \pi(\cdot)$ such that $\mathbf{P}[X'_K \neq Z] < \epsilon$. Furthermore, we can construct all of $\{X_n\}$,

$\{X'_n\}$, and Z jointly on a common probability space, by first constructing $\{X_n\}$ and $\{X'_n\}$ as above, and then constructing Z conditional on $\{X_n\}$ and $\{X'_n\}$ from any conditional distribution satisfying that $Z \sim \pi(\cdot)$ and $\mathbf{P}[X'_K \neq Z] < \epsilon$. (This joint construction can always be achieved, though it may require enlarging the underlying probability space; see e.g. Fristedt and Gray, 1997, p. 430.)

We then have

$$\mathbf{P}[X_K \neq Z] \;\leq\; \mathbf{P}[X_K \neq X'_K,\; E] + \mathbf{P}[X'_K \neq Z,\; E] + \mathbf{P}[E^c] \;<\; \epsilon + \epsilon + \epsilon \;=\; 3\,\epsilon\,.$$

Hence, $\|\mathcal{L}(X_K) - \pi(\cdot)\| < 3\,\epsilon$, i.e. $T(x, \gamma, K) < 3\,\epsilon$. Since $K \geq n^* + N$ was arbitrary, this means that $T(x, \gamma, K) \leq 3\epsilon$ for all sufficiently large K. Hence, $\lim_{K\to\infty} T(x, \gamma, K) = 0$. ∎

Even with the uniformity assumption (a), Theorem 5 still applies in many situations, as the following corollaries show. We begin with the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite:

Corollary 6. *Suppose an adaptive MCMC algorithm satisfies Diminishing Adaptation, and also that each P_γ is ergodic for $\pi(\cdot)$ (i.e., $\lim_{n\to\infty} \|P^n_\gamma(x, \cdot) - \pi(\cdot)\| = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$). Suppose further that $\mathcal{X}$ and $\mathcal{Y}$ are finite. Then the adaptive algorithm is ergodic.*

Proof. Let $\epsilon > 0$. By assumption, for each $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there is $N(x, \gamma, \epsilon)$ such that $\|P_\gamma^{N(x,\gamma,\epsilon)}(x, \cdot) - \pi(\cdot)\| \leq \epsilon$. Letting $N(\epsilon) = \max_{x\in\mathcal{X},\, \gamma\in\mathcal{Y}} N(x, \gamma, \epsilon)$, we see that condition (a) of Theorem 5 is satisfied. The result follows. ∎

We can apply the above corollary to one version of our running example:

Corollary 7. *The doubly-modified running example (presented in Section 4 above) is ergodic provided that the adaptation probabilities $p(n)$ satisfy $\lim_{n\to\infty} p(n) = 0$.*

Proof. In that example, each P_γ is π-irreducible and aperiodic, and hence ergodic for $\pi(\cdot)$. Furthermore, both $\mathcal{X}$ and $\mathcal{Y}$ are finite. Also, Diminishing Adaptation holds since $\|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| \leq p(n) \to 0$. Hence, the result follows from Corollary 6. ∎

Often, $\mathcal{X}$ and $\mathcal{Y}$ will not be finite. However, under compactness and continuity assumptions, similar reasoning applies:

Corollary 8. *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that $\mathcal{X} \times \mathcal{Y}$ is compact in some topology, with respect to which the mapping $(x, \gamma) \mapsto T(x, \gamma, n)$ is continuous for each fixed $n \in \mathbf{N}$. Then the adaptive algorithm is ergodic.*

Proof. Fix $\epsilon > 0$. For $n \in \mathbf{N}$, let $\mathcal{W}_n \subseteq \mathcal{X} \times \mathcal{Y}$ be the set of all pairs (x, γ) such that $\|P_\gamma^n(x, \cdot) - \pi(\cdot)\| < \epsilon$. Since each P_γ is ergodic, this means that every pair (x, γ) is in $\mathcal{W}_n$ for all sufficiently large n. In particular, $\bigcup_n \mathcal{W}_n = \mathcal{X} \times \mathcal{Y}$.

On the other hand, by continuity, each $\mathcal{W}_n$ is an open set. Thus, by compactness, there is a finite set $\{n_1, \ldots, n_r\}$ such that $\mathcal{W}_{n_1} \cup \ldots \cup \mathcal{W}_{n_r} = \mathcal{X} \times \mathcal{Y}$. Letting $N = N(\epsilon) = \max[n_1, \ldots, n_r]$, we see that condition (a) of Theorem 5 is satisfied. The result follows. ∎

In applying Corollary 8, the following lemma is sometimes useful:

Lemma 9. *Suppose the mapping $(x, \gamma) \mapsto P_\gamma(x, \cdot)$ is continuous with respect to a product metric space topology, meaning that for each $x \in \mathcal{X}$, $\gamma \in \mathcal{Y}$, and $\epsilon > 0$, there is $\delta = \delta(x, \gamma, \epsilon) > 0$ such that $\|P_{\gamma'}(x', \cdot) - P_\gamma(x, \cdot)\| < \epsilon$ for all $x' \in \mathcal{X}$ and $\gamma' \in \mathcal{Y}$ satisfying $\mathrm{dist}(x', x) + \mathrm{dist}(\gamma', \gamma) < \delta$ (for some distance metrics on $\mathcal{X}$ and $\mathcal{Y}$). Then for each $n \in \mathbf{N}$, the mapping $(x, \gamma) \mapsto T(x, \gamma, n)$ is continuous.*

Proof. Given $x \in \mathcal{X}$, $\gamma \in \mathcal{Y}$, $n \in \mathbf{N}$, and $\epsilon > 0$, find $\delta > 0$ with $\|P_{\gamma'}(x', \cdot) - P_\gamma(x, \cdot)\| < \epsilon/n$ whenever $\mathrm{dist}(x', x) + \mathrm{dist}(\gamma', \gamma) < \delta$. Then given x' and γ' with $\mathrm{dist}(x', x) + \mathrm{dist}(\gamma', \gamma) < \delta$, as in the proof of Theorem 5 we can construct X_n' and X_n with $X_n' \sim P_{\gamma'}^n(x', \cdot)$ and $X_n \sim P_\gamma^n(x, \cdot)$, such that $\mathbf{P}[X_n' = X_n] \geq 1 - \epsilon$. Hence, $\|\mathcal{L}(X_n') - \mathcal{L}(X_n)\| < \epsilon$. The triangle inequality then implies that $\|\mathcal{L}(X_n') - \pi(\cdot)\|$ and $\|\mathcal{L}(X_n) - \pi(\cdot)\|$ are within ϵ of each other, thus giving the result. ∎

The required continuity conditions follow if the transition kernels have bounded densities with continuous dependencies:

Corollary 10. *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, $P_\gamma(x, dz) = f_\gamma(x, z)\,\lambda(dz)$ has a density $f_\gamma(x, \cdot)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$. Finally, suppose the $f_\gamma(x, z)$ are uniformly bounded, and that for each fixed $z \in \mathcal{X}$, the mapping $(x, \gamma) \mapsto f_\gamma(x, z)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then the adaptive algorithm is ergodic.*

Proof. We have (e.g. Roberts and Rosenthal, 2004, Proposition 3(f)) that

$$\|P_{\gamma'}(x', \cdot) - P_\gamma(x, \cdot)\| \;=\; \frac{1}{2} \int_{\mathcal{X}} [M(y) - m(y)]\,\lambda(dy)\,, \tag{4}$$

where $M(y) = \max[f_\gamma(x, y),\, f_{\gamma'}(x', y)]$ and $m(y) = \min[f_\gamma(x, y),\, f_{\gamma'}(x', y)]$. By continuity of the mapping $(x, \gamma) \mapsto f_\gamma(x, y)$, and the finiteness of $\lambda(\cdot)$, it follows from the Bounded Convergence Theorem that the mapping $(x, \gamma) \mapsto P_\gamma(x, \cdot)$ is continuous. The result then follows by applying Lemma 9 to Corollary 8. ∎

Metropolis-Hastings algorithms do not have densities (since they have positive probability of rejecting the proposal and not moving). In particular, the expression in (4) does not diminish to 0 as x' approaches x. However, if the proposal kernels have densities, then a similar result still holds:

Corollary 11. *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, P_γ represents a Metropolis-Hastings algorithm with proposal kernel $Q_\gamma(x, dy) = f_\gamma(x, y)\,\lambda(dy)$ having a density $f_\gamma(x, \cdot)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$, with corresponding density g for $\pi(\cdot)$ so that $\pi(dy) = g(y)\,\lambda(dy)$. Finally, suppose that the $f_\gamma(x, y)$ are uniformly bounded, and for each fixed $y \in \mathcal{X}$, the mapping $(x, \gamma) \mapsto f_\gamma(x, y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then the adaptive algorithm is ergodic.*

Proof. In this case, the probability of accepting a proposal from x is given by:

$$a_\gamma(x) \;=\; \int_{\mathcal{X}} \min\left[1,\; \frac{g(y)\, f_\gamma(y,x)}{g(x)\, f_\gamma(x,y)}\right]\; f_\gamma(x,y)\, \lambda(dy)\,,$$

which is a jointly continuous function of $(x,\gamma) \in \mathcal{X} \times \mathcal{Y}$ by the Bounded Convergence Theorem. We decompose $P_\gamma(x,\cdot)$ as:

$$P_\gamma(x,dz) \;=\; [1 - a_\gamma(x)]\, \delta_x(dz) + p_\gamma(x,z)\, \lambda(dz)$$

where $p_\gamma(x,z)$ is jointly continuous in x and γ. Iterating this, we can write the n-step transition law as:

$$P_\gamma^n(x,dz) \;=\; [1 - a_\gamma(x)]^n\, \delta_x(dz) + p_\gamma^n(x,z)\, \lambda(dz)$$

for appropriate jointly continuous $p_\gamma^n(x,z)$.

We can assume without loss of generality that $a_\gamma(x) = 1$ whenever $\lambda\{x\} > 0$, i.e. that $\delta_x(\cdot)$ and $\pi(\cdot)$ are orthogonal measures. (Indeed, if $\lambda\{x\} > 0$, then we can modify the proposal densities so as to include $[1 - a_\gamma(x)]\, \delta_x(dz)$ as part of $p_\gamma(x,z)\, \lambda(dz)$.) It then follows that:

$$\|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \;=\; [1 - a_\gamma(x)]^n + \frac{1}{2} \int_{\mathcal{X}} |p^n(x,z) - g(z)|\, \lambda(dz)\,.$$

This quantity is jointly continuous in x and γ, again by the Bounded Convergence Theorem. Moreover, by ergodicity, it converges to zero as $n \to \infty$ for each fixed x and γ. Hence, by compactness, the convergence is uniform in x and γ, i.e. condition (a) of Theorem 5 is satisfied. The result follows. ∎

Remark. The strong conditions imposed in Corollary 10 and Corollary 11 can of course be relaxed using more specialised arguments in specific examples.

We now consider the Adaptive Metropolis algorithm of Haario et al. (2001). In that algorithm, it is assumed that $\mathcal{X} \subseteq \mathbf{R}^d$ is compact, with finite reference measure $\lambda(\cdot)$ given by Lebesgue measure restricted to $\mathcal{X}$. Also, the proposal kernels are multivariate normal,

of the form $Q_\gamma(x, \cdot) = MVN(x, \gamma)$ where γ is a non-negative-definite $d \times d$ matrix. This ensures that each P_γ is ergodic for $\pi(\cdot)$, and that the density mappings $(x, \gamma) \mapsto f_\gamma(x, y)$ are continuous and bounded. Furthermore, the specific details of their algorithm (including that $\mathcal{X}$ is bounded, and that $\epsilon\, I_d$ is added to each empirical covariance matrix at each iteration of the algorithm) ensure (their eqn. (14)) that there are $c_1, c_2 > 0$ such that $c_1\, I_d \leq \gamma \leq c_2\, I_d$ (i.e., both $\gamma - c_1\, I_d$ and $c_2\, I_d - \gamma$ are non-negative-definite) for all γ, which implies that we can take $\mathcal{Y}$ (and hence also $\mathcal{X} \times \mathcal{Y}$) to be compact. Corollary 11 therefore implies:

Corollary 12. *The Adaptive Metropolis algorithm of Haario et al. (2001) is ergodic.*

This provides an alternative analysis to the mixingale approach of Haario et al. (2001). Haario et al. actually prove a law of large numbers for their algorithm (for bounded functionals), which we consider in Section 9 below.

Remark. For the Adaptive Metropolis algorithm, Haario et al. (2001) in fact show (their Corollary 3) that the covariance matrices stabilise, i.e. there is $\gamma_* \in \mathcal{Y}$ such that $\Gamma_n \to \gamma_*$ with probability 1. On the other hand, Theorem 5 and its corollaries (aside from Corollary 12) apply even in cases where $\{\Gamma_n\}$ has infinite oscillation.

6. Non-Uniformly Converging Case.

In this section, we relax the uniform convergence rate condition (a) of Theorem 5. Indeed, an examination of the proof of Theorem 5 shows that condition (a) was used only to ensure that $P^N_{\Gamma_{K-N}}(X_{K-N}, \cdot)$ was close to $\pi(\cdot)$. This suggests that we can generalise to the case where $P^N_{\Gamma_{K-N}}(X_{K-N}, \cdot)$ is "usually" close to $\pi(\cdot)$. To proceed, for $\epsilon > 0$, define the "ϵ convergence time function" $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to \mathbf{N}$ by

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

If each individual P_γ is ergodic, then $M_\epsilon(x, \gamma) < \infty$.

Theorem 13. *Consider an adaptive MCMC algorithm with Diminishing Adaptation (i.e.,* $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| = 0$ *in probability). Let* $x_* \in \mathcal{X}$ *and* $\gamma_* \in \mathcal{Y}$. *Then* $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$ *provided that for all* $\epsilon > 0$, *the sequence* $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^\infty$ *is bounded in probability given* $X_0 = x_*$ *and* $\Gamma_0 = \gamma_*$, *i.e. for all* $\delta > 0$, *there is* $N \in \mathbf{N}$ *such that* $\mathbf{P}[M_\epsilon(X_n, \Gamma_n) \le N \,|\, X_0 = x_*, \Gamma_0 = \gamma_*] \ge 1 - \delta$ *for all* $n \in \mathbf{N}$.

Proof. From the proof of Theorem 5, we conclude that for all $\epsilon > 0$ there is $n_* \in \mathbf{N}$ such that for all $N \in \mathbf{N}$ and all $K \ge n_* + N$, we can simultaneously construct the original chain $\{X_n\}$, and $Z \sim \pi(\cdot)$, such that (writing $\mathcal{G}_0$ for $\{X_0 = x_*,\, \Gamma_0 = \gamma_*\}$)

$$T(x_*, \gamma_*, n) \;<\; 3\,\epsilon + \mathbf{P}\big[M_\epsilon(X_n, \Gamma_n) > N \,|\, \mathcal{G}_0\big]\,.$$

Find $m \in \mathbf{N}$ such that $\mathbf{P}[M_\epsilon(X_n, \Gamma_n) > m \,|\, \mathcal{G}_0] \le \epsilon$ for all $n \in \mathbf{N}$. Then setting $N = m$, we conclude that

$$T(x_*, \gamma_*, K) \;\le\; 3\,\epsilon + \epsilon \;=\; 4\,\epsilon\,, \qquad K \ge n_* + m\,.$$

The result follows. ∎

We shall use the following two easily-verified lemmas. The first follows by induction, the second by Markov's inequality.

Lemma 14. *Let* $\{e_n\}_{n=0}^\infty$ *be a sequence of real numbers. Suppose* $e_{n+1} \le \lambda\, e_n + b$ *for some* $0 \le \lambda < 1$ *and* $0 \le b < \infty$, *for all* $n = 0, 1, 2, 3, \ldots$. *Then* $\sup_n e_n \le \max[e_0,\ b/(1-\lambda)]$.

Lemma 15. *Let* $\{W_n\}_{n=0}^\infty$ *be a sequence of non-negative random variables. If* $\sup_n \mathbf{E}(W_n) < \infty$, *then* $\{W_n\}$ *is bounded in probability.*

We can then prove:

Corollary 16. *The singly-modified running example (presented in Section 4) is ergodic provided that the adaptation probabilities* $p(n)$ *satisfy* $\lim_{n\to\infty} p(n) = 0$.

Proof. Let $V(\gamma) = \exp(\gamma)$. Then it is easily verified (since the probability of accepting from proposal $Q_\gamma(x, \cdot)$ is always $\leq K/2\gamma$) that $\mathbf{E}[V(\Gamma_{n+1}) \,|\, \Gamma_n = \gamma] \leq \lambda\, V(\gamma) + b\mathbf{1}_C(\gamma)$ where $\lambda = 2/e$, $C = \{\gamma \in \mathcal{Y} : \gamma \leq \gamma_*\}$, $\gamma_* = K(e^2 - 1)/2$, and $b = (1 - \lambda)\gamma_* + 1$. Hence, $\mathbf{E}[V(\Gamma_{n+1})] \leq \lambda\, \mathbf{E}[V(\Gamma_n)] + b$. It follows from Lemma 14 that $\sup_n \mathbf{E}[V(\Gamma_n)] \leq b/(1-\lambda) < \infty$. Since $\gamma \leq V(\gamma)$, $\sup_n E[\Gamma_n] < \infty$, so $\{\Gamma_n\}$ is bounded in probability by Lemma 15. But since each set $\{(x, \gamma) : x \in \mathcal{X}, \gamma \leq G\}$ is finite, and since each individual $M_\epsilon(x, \gamma)$ is finite, it follows that $\{M_\epsilon(X_n, \Gamma_n)\}$ is also bounded in probability, for each $\epsilon > 0$. The result then follows from Theorem 13. ∎

7. Connections to Drift and Minorisation Conditions.

The quantity $M_\epsilon(x, \gamma)$ is rather abstract. It can be made somewhat more concrete using the theory of quantitative convergence rate bounds (e.g. Meyn and Tweedie, 1994; Rosenthal, 1995, 2002; Roberts and Tweedie, 1999; Baxendale, 2005). For example, Theorem 2.3 of Meyn and Tweedie (1994) implies the following:

Proposition 17. *Consider a Markov chain kernel P on a state space $(\mathcal{X}, \mathcal{F})$ with stationary probability distribution $\pi(\cdot)$. Suppose there is $C \in \mathcal{F}$, $V : \mathcal{X} \to [1, \infty)$, $\delta > 0$, $\lambda < 1$, and $b < \infty$, such that $\sup_C V = v < \infty$, and*
(i) [strongly aperiodic minorisation condition] there exists a probability measure $\nu(\cdot)$ on C with $P(x, \cdot) \geq \delta\, \nu(\cdot)$ for all $x \in C$; and
(ii) [geometric drift condition] $PV \leq \lambda\, V + b\, \mathbf{1}_C$, i.e. $(PV)(x) \leq \lambda\, V(x) + b\, \mathbf{1}_C(x)$ for all $x \in \mathcal{X}$ (where $(PV)(x) = \mathbf{E}[V(X_1) \,|\, X_0 = x]$).
Then there are $K < \infty$ and $\rho < 1$, depending <u>only</u> on the constants δ, λ, b, and v, such that $\|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq K\, V(x)\, \rho^n$ for all $\gamma \in \mathcal{Y}$.

To make use of this Proposition, we consider a notion related to the *simultaneous geometrically ergodicity* studied by Roberts, Rosenthal, and Schwartz (1998). Say a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of Markov chain kernels is *simultaneously strongly aperiodically geometrically ergodic* if there is $C \in \mathcal{F}$, $V : \mathcal{X} \to [1, \infty)$, $\delta > 0$, $\lambda < 1$, and $b < \infty$, such that $\sup_C V = v < \infty$, and
(i) for each $\gamma \in \mathcal{Y}$, there exists a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma(x, \cdot) \geq \delta\, \nu_\gamma(\cdot)$ for

all $x \in C$; and
(ii) $(P_\gamma)V \leq \lambda\, V + b\, \mathbf{1}_C$.

We then have:

Theorem 18. *Consider an adaptive MCMC algorithm with Diminishing Adaptation, such that the family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is simultaneously strongly aperiodically geometrically ergodic with $\mathbf{E}[V(X_0)] < \infty$. Then the adaptive algorithm is ergodic.*

Proof. By Theorem 13 and Lemma 15 and Proposition 17, it suffices to show that $\sup_n \mathbf{E}[V(X_n)] < \infty$. Now, we have by assumption that $(P_\gamma)V \leq \lambda\, V + b\, \mathbf{1}_C$ for all $\gamma \in \mathcal{Y}$, so $\mathbf{E}[V(X_{n+1}) \,|\, X_n = x, \Gamma_n = \gamma] \leq \lambda\, V(x) + b$. Integrating over the distribution of Γ_n, we conclude that $\mathbf{E}[V(X_{n+1}) \,|\, X_n = x] \leq \lambda\, V(x) + b$. Hence, from the double-expectation formula, $\mathbf{E}[V(X_{n+1})] \leq \lambda\, \mathbf{E}[V(X_n)] + b$. Then, from Lemma 14,
$\sup_n \mathbf{E}[V(X_n)] \leq \max\big[\mathbf{E}[V(X_0)],\ b/(1-\lambda)\big] < \infty$. ∎

Remark. In Theorem 18, the strong aperiodicity condition $\nu_\gamma(C) = 1$ can be dropped if $\inf_{x\notin C} V(x) > 2b/(1-\lambda)$ (Rosenthal, 1995).

The results of Meyn and Tweedie (1994) and Rosenthal (1995) give *geometric* quantitative bounds on convergence, which is quite a strong property. For present purposes, all that is required is that $M(x,\gamma) \leq V(x)\, a(n)$ where $a(n) \to 0$ (at any rate), uniformly in γ. So, the hypotheses of Theorem 18 are overly strong in this sense.

To weaken these hypotheses, we consider *polynomial ergodicity*. While some results about polynomial ergodicity (e.g. Jarner and Roberts, 2002; Fort and Moulines, 2003) do not provide explicit quantitative convergence bounds, Theorems 3 and 4 of the paper of Fort and Moulines (2000) do. To make use of their result, call a family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ of Markov chain kernels *simultaneously polynomially ergodic* if each P_γ is π-irreducible with stationary distribution $\pi(\cdot)$, and there is $C \in \mathcal{F}$ and $m \in \mathbf{N}$ and $\delta > 0$ and probability measures $\nu_\gamma(\cdot)$ on $\mathcal{X}$ such that $\pi(C) > 0$, and $P_\gamma^m(x,\cdot) \geq \delta\, \nu_\gamma(\cdot)$ for all $x \in C$ and $\gamma \in \mathcal{Y}$, and there is $q \in \mathbf{N}$ and measurable functions $V_0, V_1, \ldots, V_k : \mathcal{X} \to (0,\infty)$, such that for $k = 0, 1, \ldots, q-1$, there are $0 < \alpha_k < 1$, $b_k < \infty$, and $c_k > 0$ such that:

$(P_\gamma V_{k+1})(x) \leq V_{k+1}(x) - V_k(x) + b_k \mathbf{1}_C(x)$ for $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$; $V_k(x) \geq c_k$ for $x \in \mathcal{X}$; $V_k(x) - b_k \geq \alpha_k V_k(x)$ for $x \in \mathcal{X} \setminus C$; and $\sup_C V_q < \infty$ and $\pi(V_q^\beta) < \infty$ for some $0 < \beta \leq 1$. These conditions are rather technical, however they are weaker than assuming geometric ergodicity. Analogous to Theorem 18, we then have the following:

Theorem 19. *Consider an adaptive MCMC algorithm with Diminishing Adaptation, such that the family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ is simultaneously polynomially ergodic. Then the adaptive algorithm is ergodic.*

Continuing in this direction, we note that Theorem 13.0.1 of Meyn and Tweedie (1993) indicates that to merely prove convergence (as opposed to geometric convergence), it suffices to have an even weaker drift condition of the form $PV \leq V - 1 + b \mathbf{1}_C$. So, perhaps it suffices for the validity of adaptive MCMC algorithms that such drift conditions hold uniformly for all P_γ. Unfortunately, the available results (e.g. Meyn and Tweedie, 1993) appear not to provide any explicit *quantitative* bounds on convergence. Furthermore, conditional on failing to couple, the sequence $\{\mathbf{E}_\gamma[V(X_n)]\}$ may not remain bounded in probability even for a fixed chain P_γ (see e.g. Pemantle and Rosenthal, 1998), which necessitates the additional assumption that the sequence $\{V(X_n)\}$ remains bounded in probability for the adaptive chain. We have not yet been able to draw any firm conclusions based only on these weakest drift conditions, so we state this as an open problem:

Open Problem 20. *Consider an adaptive MCMC algorithm with Diminishing Adaptation, with $C \in \mathcal{F}$, $V : \mathcal{X} \to [1, \infty)$, $\delta > 0$, and $b < \infty$, with $\sup_C V = v < \infty$, and:*
(i) for each $\gamma \in \mathcal{Y}$, there exists a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma(x, \cdot) \geq \delta \nu_\gamma(\cdot)$ for all $x \in C$; and
(ii) $P_\gamma V \leq V - 1 + b \mathbf{1}_C$ for all $\gamma \in \mathcal{Y}$.
Suppose further that the sequence $\{V(X_n)\}_{n=0}^\infty$ is bounded in probability, given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$. Does this imply that $\lim_{n \to \infty} T(x_*, \gamma_*, n) = 0$?

8. Relation to Recurrence.

The above results indicate that an adaptive MCMC algorithm with Diminishing Adaptation is ergodic provided that it is "recurrent in probability" in some sense. This leads to the following recurrence-related open problem:

Open Problem 21. *Consider an adaptive MCMC algorithm with Diminishing Adaptation. Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$. Suppose that for all $\epsilon > 0$, there is $m \in \mathbf{N}$ such that $\mathbf{P}[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.} \,|\, X_0 = x_*, \Gamma_0 = \gamma_*] = 1$ (where "i.o." means "infinitely often", i.e. for an infinite number of $n \in \mathbf{N}$). Does this imply that $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$?*

It may be possible to approach Open Problem 21 along similar lines to the proof of Theorem 13. The difficulty is that, even if $M_\epsilon(X_n, \Gamma_n) < m$ infinitely often, this does not control the probability that $M_\epsilon(X_n, \Gamma_n) < m$ for a specific time like $n = K - N$. Thus, while we can approximately couple X_n to $\pi(\cdot)$ for infinitely many times n, it is not clear that we can accomplish this at a particular time $n = K$. (This is related to the notion of *faithfulness* of couplings; see Rosenthal, 1997 and Häggström, 2001.)

Related to recurrence, we also have the following.

Example 22. ("Stairway to Heaven") Let $\mathcal{X} = \{(i, j) \in \mathbf{N} \times \mathbf{N} : i = j \text{ or } i = j + 1\}$ be an infinite staircase, with target distribution given by $\pi(i, j) \propto j^{-2}$. Given a state x, write h for the (left or right) horizontal neighbour of x, v for the (up or down) vertical neighbour of x, hv for the vertical neighbour of the horizontal neighbour of x, and vh for the horizontal neighbour of the vertical neighbour of x. [Special case: if $x = (1, 1)$, then take $v = (1, 1)$.]

The adaptive space is $\mathcal{Y} = \{0, 1\}$, consisting of the following two "exclusive Metropolis within Gibbs" algorithms specifying both where to move and how to adapt, from the current state $x = X_n$ and current adaptation parameter $\gamma = \Gamma_n$:

$\gamma = 0$: Let $\alpha = \min[1,\ \pi(hv) \,/\, \pi(h)]$. With probability α move to hv (and leave $\Gamma_{n+1} = 0$), otherwise move to h (and set $\Gamma_{n+1} = 1$). [Intuitive description: First move to horizontal neighbour h. Then, propose moving from h to hv; accept proposal with probability α.]

$\gamma = 1$: Let $\alpha = \min[1,\ \pi(v)\,/\,\pi(x)]$. With probability α move to vh (and leave $\Gamma_{n+1} = 1$), otherwise move to h (and set $\Gamma_{n+1} = 0$). [Intuitive description: First propose moving to vertical neighbour v; accept proposal with probability α. Either way, then move to current horizontal neighbour.]

It is easily checked that both P_0 and P_1 preserve stationarity of $\pi(\cdot)$, and are irreducible and aperiodic.

On the other hand, if the chain is at $X_n = (i, i)$, and $\Gamma_n = 0$, then

$$\mathbf{P}[X_{n+1} \neq (i+1, i+1)] \;=\; 1 - \pi((i+1, i+1))\,/\,\pi((i+1, i))$$

$$=\; 1 - i^2/(i+1)^2 \;=\; 1 - i^2/(i^2+2i+1) \;=\; (2i+1)/(i^2+2i+1) \;\asymp\; 2/i\,.$$

Furthermore, even if the chain rejects, so $X_{n+1} = (i+1, i)$, then also $\Gamma_{n+1} = 1$, and the chain will then attempt to move up on the next step, thus continuing its voyage up the staircase. In other words, the only way the voyage up the staircase can be reversed is if the chain rejects on *two consecutive steps*, which has probability $\asymp 2/i^2$. Since $\sum_i 2/i^2 < \infty$, it follows from the Borel-Cantelli Lemma (e.g. Rosenthal, 2000, Theorem 3.4.2) that $\mathbf{P}[X_n^{(1)}$ is non-decreasing$] > 0$. Hence, $\mathbf{P}[\lim_{n\to\infty} X_n^{(1)} = \infty] > 0$, i.e. there is positive probability that the chain will climb the infinite stairway (in search for "all that glitters is gold") without ever rejecting. Furthermore, even if the chain does reject twice in succession, then it will decrease to $(1, 1)$ and then begin its attempted climb again. We conclude that $\mathbf{P}[\lim_{m\to\infty} X_m^{(1)} = \infty] = 1$, i.e. the chain is transient. ■

Remark. In the above example, one possible drift function for the $\gamma = 0$ algorithm is given by $V(i, i) = 4i$ and $V(i+1, i) = i$. For $\gamma = 1$, one possible drift function is $V(i, i) = i$ and $V(i+1, i) = 4i$. However, simultaneous drift conditions cannot be found.

9. Laws of Large Numbers.

When MCMC is used in practice, often entire sequences $X_1, X_2, \ldots, X_n$ of Markov chain output are combined together to form averages of the form $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$ to estimate the mean $\pi(g) = \int g(x)\,\pi(dx)$ of a function $g : \mathcal{X} \to \mathbf{R}$. To justify such approximations, we require laws of large numbers for ergodic averages of the form:

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

either in probability or almost surely, for suitably regular functions g. For traditional MCMC algorithms this topic is well-studied, see e.g. Tierney (1994) and Meyn and Tweedie (1993). For adaptive MCMC, this topic is dealt with in some detail in other papers in the literature, under somewhat stronger assumptions than those used here (see in particular Andrieu and Moulines, 2003).

In this section, we consider the extent to which laws of large numbers hold for adaptive MCMC under weaker assumptions. We shall concentrate on the simultaneous uniform ergodicity case, as in Theorem 5 above.

Theorem 23. *[Weak Law of Large Numbers] Consider an adaptive MCMC algorithm. Suppose that conditions (a) and (b) of Theorem 5 hold. Let $g : \mathcal{X} \to \mathbf{R}$ be a bounded measurable function. Then for any starting values $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$ we have*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$.

Proof. Assume without loss of generality that $\pi(g) = 0$. Let $a = \sup_{x\in\mathcal{X}} |g(x)| < \infty$. Write $\mathbf{E}_{\gamma,x}$ for expectations with respect to the Markov chain kernel P_γ when started from $X_0 = x$, and write $\mathbf{P}_{\gamma,x}$ for the corresponding probabilities. Write $\mathbf{E}$ and $\mathbf{P}$ (without subscripts) for expectations and probabilities with respect to the adaptive chain.

The usual law of large numbers for Markov chains (see e.g. Tierney, 1994) implies that for each fixed $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, $\lim_{n\to\infty} \mathbf{E}_{\gamma,x} \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right| \to \pi(g) = 0$. Condition (a) implies that this convergence can be bounded uniformly over choices of x and γ, i.e. given

$\epsilon > 0$ we can find an integer N such that

$$\mathbf{E}_{\gamma,x}\left(\left|\frac{\sum_{i=1}^{N} g(X_i)}{N}\right|\right) < \epsilon\,, \qquad x \in \mathcal{X},\ \gamma \in \mathcal{Y}\,.$$

In terms of this N, we use condition (b) to find $n^* \in \mathbf{N}$ satisfying (2). The coupling argument in the proof of Theorem 5 then implies that on the event E defined there (which has probability $\geq 1-\epsilon$), for all $n \geq n_*$, the adaptive chain sequence $X_{n+1}, \ldots, X_{n+N}$ can be coupled with probability $\geq 1-\epsilon$ with a corresponding sequence arising from the fixed Markov chain P_{Γ_n}. In other words, since $|g| \leq a$,

$$\mathbf{E}\left(\frac{1}{N}\left|\sum_{i=n+1}^{n+N} g(X_i)\right| \;\middle|\; \mathcal{G}_n\right) \;\leq\; \mathbf{E}_{\Gamma_n, X_n}\left(\left|\frac{\sum_{i=1}^{N} g(X_i)}{N}\right|\right) + a\,\epsilon + a\,\mathbf{P}(E^c) \;\leq\; (1+2a)\,\epsilon\,. \quad (5)$$

Now consider any integer T sufficiently large that

$$\max\left[\frac{an^*}{T}, \frac{aN}{T}\right] \leq \epsilon\,. \quad (6)$$

Then (writing $\lfloor r \rfloor$ for the greatest integer not exceeding r) we have:

$$\left|\frac{1}{T}\sum_{i=1}^{T} g(X_i)\right| \;\leq\; \left|\frac{1}{T}\sum_{i=1}^{n^*} g(X_i)\right| + \left|\frac{1}{\lfloor \frac{T-n^*}{N} \rfloor} \sum_{j=1}^{\lfloor \frac{T-n^*}{N} \rfloor} \frac{1}{N}\sum_{k=1}^{N} g(X_{n^*+(j-1)N+k})\right|$$

$$+ \left|\frac{1}{T} \sum_{n^*+\lfloor \frac{T-n^*}{N} \rfloor N+1}^{T} g(X_i)\right|\,. \quad (7)$$

By (6), the first and last terms on the right-hand side of (7) are each $\leq \epsilon$. By (5), the middle term is an average of terms each of which has absolute expectation $\leq (1+2a)\epsilon$. Hence, taking expectations and using the triangle inequality, we have that

$$\mathbf{E}\left(\left|T^{-1}\sum_{i=1}^{T} g(X_i)\right|\right) \;\leq\; \epsilon + (1+2a)\epsilon + \epsilon \;=\; \epsilon(3+2a)\,.$$

Markov's inequality then gives that

$$\mathbf{P}\left(\left|T^{-1}\sum_{i=1}^{T} g(X_i)\right| \geq \epsilon^{1/2}\right) \;\leq\; \epsilon^{1/2}(3+2a)\,.$$

Since this holds for all sufficiently large T, and $\epsilon > 0$ was arbitrary, the result follows. ∎

On the other hand, a strong law does not hold under the conditions of Theorem 5:

Example 24. We begin with the "One-Two" version of the Running Example used in Example 4, with the parameters a and b chosen so that $\lim_{n\to\infty} \mathbf{P}[X_n = 1] = p$ for some $p > \pi\{1\}$. Let $\{I_k\}_{k=0}^{\infty}$ be a sequence of independent binary variables, with $\mathbf{P}[I_k = 1] = 1/k$ and $\mathbf{P}[I_k = 0] = 1 - 1/k$.

Consider the following new adaptation scheme. Set $\Gamma_0 = \Gamma_1 = \Gamma_2 = 1$ (say). Then for each iteration $n \geq 3$, find $k \in \mathbf{N}$ with $2^{k^2} + 1 \leq n \leq 2^{(k+1)^2}$. If $I_k = 1$ then at time n we proceed according to the One-Two adaptation scheme (i.e., set $\Gamma_n = 2$ if the previous proposal was accepted, otherwise set $\Gamma_n = 1$). If $I_k = 0$ then we simply set $\Gamma_n = 1$ (regardless of whether the previous proposal was accepted or rejected).

This scheme ensures that the probability of adaptation at any particular iteration n goes to 0 as $n \to \infty$, so that condition (b) of Theorem 5 is satisfied. Also, since $\mathcal{X}$ and $\mathcal{Y}$ are finite, and each P_γ is irreducible and aperiodic, condition (a) of Theorem 5 is satisfied. On the other hand, with probability 1, there will still be an infinite number of k with $I_k = 1$, say $I_{k_1} = I_{k_2} = \ldots = 1$, so the fully adaptive strategy will still be adopted infinitely often.

Now, as $k \to \infty$, we have that

$$\frac{1}{2^{k^2}} \sum_{i=1}^{2^{k^2}} \mathbf{1}(X_i = 1) \;\approx\; \frac{1}{2^{k^2} - 2^{(k-1)^2}} \sum_{i=2^{(k-1)^2}+1}^{2^{k^2}} \mathbf{1}(X_i = 1)\,.$$

It follows that along the subsequence $\{2^{(k_i)^2}\}$,

$$\lim_{i\to\infty} \frac{1}{2^{(k_i)^2}} \sum_{i=1}^{2^{(k_i)^2}} \mathbf{1}(X_i = 1) \;=\; p \;>\; \pi\{1\}\,.$$

Hence, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = 1) \neq \pi\{1\}$, i.e. a strong law of large numbers fails for the function $g(x) = \mathbf{1}(x = 1)$. ∎

10. Discussion.

This paper has investigated the validity of adaptive MCMC algorithms. We emphasised (Example 4) that natural-seeming adaptive schemes may destroy convergence. On the other hand, we showed (Theorems 5 and 13) that under the assumption of Diminishing Adaptation, together with some sort of near-uniform control of the convergence rates, adaptive MCMC can be shown to be ergodic. We also provided (Theorem 23) a weak law of large numbers for bounded functions in this general setting.

This leads to the question of what adaptations should be used in practice. We believe the most natural and useful adaptive MCMC scheme proposed to date is the Adaptive Metropolis algorithm of Haario et al. (2001) discussed earlier. We have performed simulations on variations of this algorithm (by means of a Cholesky Decomposition), and have found its performance to be quite promising and worthy of further investigation.

Now, algorithms which adapt acceptance probabilities are clearly limited by the crudeness of such a global criterion. More ambitious schemes might involve adapting acceptance probabilities in different ways in different parts of the state space. For example, a proposal distribution might be of the form $Y_{n+1} \sim N(x, \sigma_x^2)$, where σ_x^2 is a function of the current state x involving unknown parameters, e.g. $\sigma_x^2 = e^a(1+|x|)^b$. The parameters (e.g. a and b) can then be modified adaptively based on the chain's previous output, provided only that Diminishing Adaption and "convergence time bounded in probability" properties hold. We have done some simulations with this scheme for simple one-dimensional target distributions, and found it to be very promising; we are currently considering higher-dimensional analogues.

Another simple adaptation scheme is to simultaneously run two chains $\{X_n\}$ and $\{X'_n\}$, and have the chain $\{X_n\}$ adapt its values of $\{\Gamma_n\}$ based on information learned not from $\{X_n\}$ itself, but rather from $\{X'_n\}$. If the updates of $\{X'_n\}$ are made independently of the values of $\{X_n\}$, then the $\{\Gamma_n\}$ will also be chosen independently of the $\{X_n\}$, so that $\{X_n\}$ will preserve stationary by Proposition 1. This represents a sort of generalisation of the traditional scheme of first doing a "trial run" to tune the parameters, and then basing inferences on a non-adaptive main run after the parameters are tuned. In this case, the "trial run" $\{X'_n\}$ continues simultaneously with the "main run" $\{X_n\}$, and the main run

continues to tune itself – independently of its own chain values – as it proceeds.

Ongoing work is currently investigating these and related ideas. We look forward to continuing these investigations, and to seeing many significant advances in adaptive MCMC methodology in the years ahead.

Acknowledgements. We thank Christophe Andrieu, Heikki Haario, Antonietta Mira, and Christian Robert for organising the excellent January 2005 "Adap'ski" workshop in Bormio, Italy, which provided inspiration related to this paper. We thank Christophe Andrieu, Eric Moulines, and Eero Saksman for helpful comments. We thank the anonymous referee for a careful reading that led to many improvements.

REFERENCES

C. Andrieu and E. Moulines (2003), On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. Preprint.

C. Andrieu and C.P. Robert (2002), Controlled MCMC for optimal sampling. Preprint.

Y.F. Atchadé and J.S. Rosenthal (2005), On Adaptive Markov Chain Monte Carlo Algorithms. Bernoulli **11(5)**, 815–828.

P.H. Baxendale (2005), Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Ann. Appl. Prob. **15**, 700–738.

M. Bédard (2006), On the robustness of optimal scaling for Metropolis-Hastings algorithms. Ph.D. dissertation, University of Toronto. In preparation.

A.E. Brockwell and J.B. Kadane (2005), Identification of regeneration times in MCMC simulation, with application to adaptive schemes. J. Comp. Graph. Stat. **14**, 436–458.

G. Fort and E. Moulines (2000), Computable Bounds For Subgeometrical And Geometrical Ergodicity. Available at: http://citeseer.ist.psu.edu/fort00computable.html

G. Fort and E. Moulines (2003), Polynomial ergodicity of Markov transition kernels. Stoch. Proc. Appl. **103**, 57–99.

B. Fristedt and L. Gray (1997), A modern approach to probability theory. Birkhauser, Boston.

W.R. Gilks, G.O. Roberts, and S.K. Sahu (1998), Adaptive Markov Chain Monte Carlo. J. Amer. Stat. Assoc. **93**, 1045–1054.

H. Haario, E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm. Bernoulli **7**, 223–242.

O. Häggström (2001), A note on disagreement percolation. Rand. Struct. Alg. **18**, 267–278.

S.F. Jarner and G.O. Roberts (2002), Polynomial convergence rates of Markov chains. Ann. Appl. Prob., 224–247, 2002.

S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at `probability.ca/MT`.

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981–1011.

C. Pasarica and A. Gelman (2003), Adaptively scaling the Metropolis algorithm using the average squared jumped distance. Preprint.

H. Robbins and S. Monro (1951), A stochastic approximation method. Ann. Math. Stat. **22**, 400–407.

G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. **7**, 110–120.

G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**, 351–367.

G.O. Roberts and J.S. Rosenthal (2002), One-shot coupling for certain stochastic recursive sequences. Stoch. Proc. Appl. **99**, 195–208.

G.O. Roberts and J.S. Rosenthal (2004), General state space Markov chains and MCMC algorithms. Prob. Surveys **1**, 20–71.

G.O. Roberts, J.S. Rosenthal, and P.O. Schwartz (1998), Convergence properties of perturbed Markov chains. J. Appl. Prob. **35**, 1–11.

G.O. Roberts and R.L. Tweedie (1996), Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika **83**, 95–110.

G.O. Roberts and R.L. Tweedie (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229. Correction: **91** (2001), 337–338.

J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (1997), Faithful couplings of Markov chains: now equals forever. Adv. Appl. Math. **18**, 372–381.

J.S. Rosenthal (2000), A first look at rigorous probability theory. World Scientific Publishing Company, Singapore.

J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. Electr. Comm. Prob. **7**, 123–128.

J.S. Rosenthal (2004), Adaptive MCMC Java Applet. Available at:

http://probability.ca/jeff/java/adapt.html

L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.