Ergodicity of Combocontinuous Adaptive MCMC Algorithms

Jeffrey S. Rosenthal* and Jinyoung Yang*

July, 2015; last revised May, 2017

Abstract

This paper proves convergence to stationarity of certain adaptive MCMC algorithms, under certain assumptions including easily-verifiable upper and lower bounds on the transition densities and a continuous target density. In particular, the transition and proposal densities are not required to be continuous, thus improving on the previous ergodicity results of Craiu et al. [7].

1 Introduction

Markov Chain Monte Carlo (MCMC) algorithms are very widely used to analyze complex probability distributions (see e.g. [6]). Adaptive MCMC algorithms adjust the Markov chain transition probabilities on the fly, in an attempt to improve efficiency, based on the past and/or current information from the chain. Adaptive MCMC algorithms can be quite effective in practice (see e.g. [13, 12, 19, 11, 24, 23]), but the chain usually loses the Markovian property so that convergence to the target (stationary) distribution is no longer guaranteed (see e.g. [20]). Many papers present conditions which assure this convergence (e.g. [13, 12, 11, 24, 4, 2, 1, 18, 8]), but these conditions are usually difficult to verify in practice. By contrast, the results of [7] provide more easily checkable conditions that guarantee convergence of adaptive MCMC algorithms, however they require continuity of all of the transition densities which makes their application somewhat limited in practice.

It was shown in [18] that the convergence of an adaptive MCMC algorithm is implied by the two conditions of Diminishing Adaption and Containment

*Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu and jinyoung.yang@mail.utoronto.ca

(explained herein). In practice, Diminishing Adaption is easily satisfied simply by constructing the adaptive mechanism appropriately. Unfortunately, the Containment condition is a lot harder to establish (see e.g. [5]). [7] introduced several simple assumptions about upper and lower bounds on transition densities which, assuming continuity, guarantee the Containment condition and thus make ergodicity much easier to verify.

In this paper, we relax one of the assumptions in [7] which is required to guarantee the Containment of an adaptive MCMC algorithm. The results of [7] require the transition kernel densities (or proposal kernel densities for the Metropolis-Hastings algorithm) of an adaptive MCMC algorithm to be jointly continuous in x, y and γ. We here show that the joint continuity assumption on the kernel densities can be relaxed to a weaker assumption which we call "combocontinuity", for x and γ jointly, which includes the usual piecewise-continuity assumption as a special case, and which allows for e.g. truncated densities. For simplicity, we still assume that the target density π is continuous and positive throughout the state space. We prove our result by generalising Dini's Theorem (about uniform convergence of compactly-supported functions) to the combocontinuous case, and then applying that theorem to the case of combocontinuous transition densities.

Below, Sections 2 and 3 present background about Adaptive MCMC and about combocontinuity. Sections 4 and 5 present our general result, whose proof uses Section 7 which generalises Dini's Theorem, and Section 8 which proves a lemma about combocontinuity. Section 6 presents an accessible special case of our new algorithm, called Bounded Adaption Metropolis or "BAM", whose validity is proved in Section 9. Finally, Section 10 illustrates our results with two numerical examples of adaptive Metropolis-Hastings algorithms with combocontinuous proposal kernel densities.

2 Background about Adaptive MCMC

Consider a general state space $\mathcal{X}$ with σ-algebra $\mathcal{F}$, on which is defined a target probability distribution π. (In our applications below, we will also require a metric η on $\mathcal{X}$, and $\mathcal{F}$ will then be the corresponding Borel σ-algebra.) Suppose that for each γ in some index set $\mathcal{Y}$, P_γ is a valid MCMC algorithm, i.e. a time-homogeneous Markov chain kernel which leaves π stationary and is Harris ergodic so that $\lim_{n\to\infty} \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| = 0$ for each fixed $x \in \mathcal{X}$. (Here $\|P_\gamma^n(x,\cdot) - \pi(\cdot)\| := \sup_{A\in\mathcal{F}} |P_\gamma^n(x,A) - \pi(A)|$ is the total variation distance to the target distribution π after n iterations of P_γ.)

An *adaptive* MCMC algorithm $\{X_n\}$ uses some specified rule to select an index value Γ_n at each iteration, based on current and/or past information from the chain and/or auxiliary randomness. It then updates X_n according to the Markov kernel P_{Γ_n}, so that for each $x \in \mathcal{X}$ and $A \in \mathcal{F}$,

$$\mathbf{P}[X_{n+1} \in A \mid X_n = x, \Gamma_n = \gamma, X_0, \ldots, X_{n-1}, \Gamma_0, \ldots, \Gamma_{n-1}] \; = \; P_\gamma(x, A)\,.$$

If the adaption rule is chosen wisely, to attempt to achieve some sort of optimality, then adaptive MCMC algorithms sometimes provide very dramatic speed-ups in efficiency and convergence to stationarity (e.g. [13, 2, 1, 3, 11, 24, 19]). However, allowing Γ_n to depend on previous values of the $\{X_n\}$ can introduce biases so that the limiting distribution of X_n, if it exists at all, might be quite different than π (cf. [20], and Example 4 of [18]). This raises the question of what conditions assure convergence in distribution of $\{X_n\}$ to π, i.e. ensure that

$$\lim_{n\to\infty} \sup_{A\in\mathcal{F}} |\mathbf{P}(X_n \in A) - \pi(A)| = 0. \tag{2.1}$$

There have been many recent results about convergence of adaptive MCMC, as mentioned in the Introduction. Here we focus on the theorem of [18] which states that the convergence of an adaptive MCMC algorithm is ensured by two conditions: Diminishing Adaption and Containment. Diminishing Adaption requires the algorithm to adapt less and less as the chain moves along, or more formally that

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \; = \; 0 \text{ in probability}\,. \tag{2.2}$$

Containment requires the convergence times of the algorithm to remain bounded in probability, or more formally that for all $\epsilon > 0$,

$$\{M_\epsilon(X_n, \Gamma_n)\}_{n=1}^\infty \text{ is bounded in probability}\,, \tag{2.3}$$

where $M_\epsilon(x, \gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \leq \epsilon\}$ is the time required to get to within ϵ of the stationary distribution π when beginning at the state x and proceeding according to the fixed Markov chain kernal P_γ.

Now, since the adaptive rule can be specified by the user, Diminishing Adaption can usually be ensured by suitable adaption design or modification. On the other hand, the Containment condition is often difficult to verify in practice, requiring substantial specialised effort (e.g. [5]). The paper [7] provided some more easily verifiable conditions to ensure Containment, however they require awkward strong continuity assumptions as we discuss below. The purpose of this paper is to relax those continuity assumptions, so that Containment can be ensured more easily.

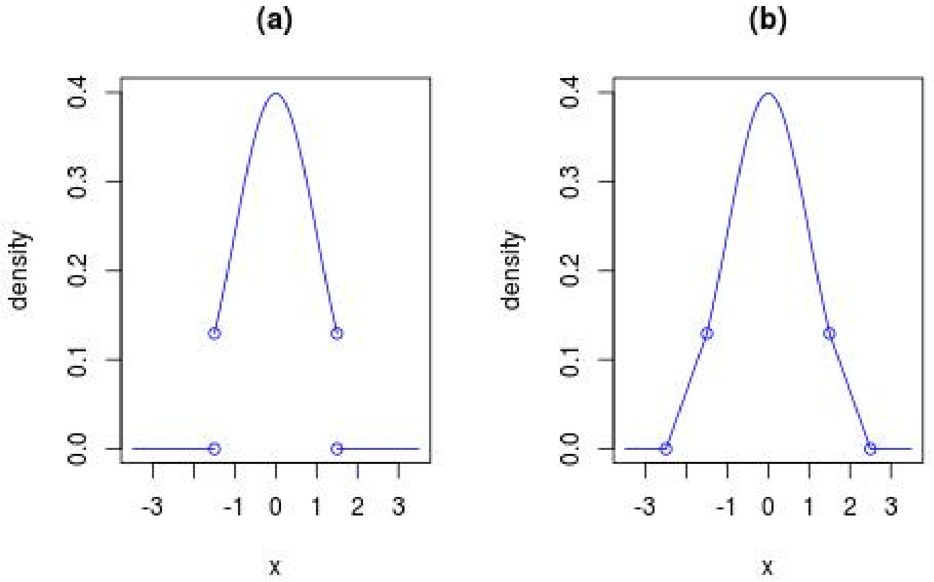


Figure 1: Two ways of truncating a normal density: with (a) a "firm" truncation (left), or (b) a "linear" truncation (right).

3 Combocontinuous Functions

The paper [7] introduced several simple assumptions about upper and lower bounds on transition densities which, assuming strong continuity conditions, guarantee the Containment condition and thus make ergodicity much easier to verify. However, the required continuity conditions are inconvenient. For one simple example, if using a truncated normal distribution (as is often used in this context), it is not permitted to use a "firm" truncation (Figure 1(a)), but rather it is necessary to linearly interpolate the truncation (Figure 1(b)), which is not difficult but which requires additional programming to implement. The present paper avoids this challenge by allowing for a more general notation of "combocontinuous" functions, a generalisation of piecewise-continuous functions.

We define a function f on a space S to be *combocontinuous* if it can be written as a finite combination of continuous functions, i.e. if $f(x) = g_{I(x)}(x)$ for some $m \in \mathbb{N}$, some (measurable) index function $I : S \to \{1, 2, \ldots, m\}$, and some finite collection $g_1, g_2, \ldots, g_m$ of continuous functions on S. Equivalently, this can be written as $f(x) = \sum_{i=1}^{m} g_i(x)\, \mathbf{1}(x \in T_i)$, where the g_i are continuous and the $T_i = \{x \in S : I(x) = i\}$ form a partition of S.

If I is constant on intervals, e.g. $I(x) = 1$ for $a \leq x \leq b$ and $I(x) = 2$ for

$b < x \le c$, then combocontinuity reduces to the usual notion of piecewise-continuity. In particular, firm truncations (as in Figure 1(a)) are always combocontinuous. On the other hand, if desired, a combocontinuous function could be much more complicated than a piecewise continuous function; for example, if $I(x) = 1$ for rational x, and $I(x) = 2$ for irrational x, then different proposals will be used from rational and from irrational states. In this paper, we mostly focus on the case of truncated densities.

Note that combocontinuous functions share many properties of continuous functions. For example, if the space S is compact, then a combocontinuous function f must be bounded above and below (since each g_i is), and if each g_i is positive then also $\inf_{x \in S} f(x) > 0$.

We will require certain extensions of combocontinuity to functions of several variables. We shall say that a function $f_\gamma(x)$ is *jointly combocontinuous* if $f_\gamma(x) = \alpha_{\gamma, I(x)}(x)$, or equivalently $f_\gamma(x) = \sum_{i=1}^m \alpha_{\gamma,i}(x)\,\mathbf{1}(x \in T_i)$ with $T_i = \{x \in \mathcal{X} : I(x) = i\}$, where again $I : \mathcal{X} \to \{1, 2, \ldots, m\}$ is a (measurable) indicator function, and for each fixed i the mapping $(x, \gamma) \mapsto \alpha_{\gamma,i}(x)$ is a jointly continuous function of x and γ.

We will also encounter functions of the form $f_\gamma(x, y)$ for $\gamma \in \mathcal{Y}$ and $x, y \in \mathcal{X}$, which *vanish* whenever $\eta(x, y) > D$ for some fixed metric η and positive constant D. For such functions, we generalise the notion of combocontinuous slightly, by saying that such a function is *truncated combocontinuous* if it can be written in the form $f_\gamma(x, y) = \beta_{\gamma, I(x)}(x, y)\,\mathbf{1}(\eta(x, y) \le D)$ where $I : \mathcal{X} \to \{1, 2, \ldots, m\}$ is a (measurable) index function, and for each fixed i the mapping $(x, y, \gamma) \mapsto \beta_{\gamma,i}(x, y)$ from $\mathcal{X} \times \mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ is jointly continuous. Equivalently, $f_\gamma(x, y) = \sum_{i=1}^m \beta_{\gamma,i}(x, y)\,\mathbf{1}(x \in T_i)\,\mathbf{1}(\eta(x, y) \le D)$ where again the $T_i = \{x \in \mathcal{X} : I(x) = i\}$ form a partition. (Note that this is more general than assuming simply that $f_\gamma(x, y) = \beta_{\gamma, I(x)}(x, y)$ where $\beta_{\gamma,i}(x, y) = 0$ whenever $\eta(x, y) > D$, since the indicator functions $\mathbf{1}(\eta(x, y) \le D)$ allow for "firm" truncations when y is at a distance D from x.)

4 Set-Up and Assumptions

Let $(\mathcal{X}, \eta)$ be a metric space, and let $\mathcal{F}$ be the corresponding Borel σ-algebra. Assume there is some "origin" point $0 \in \mathcal{X}$. Let P be a fixed transition kernel for a time-homogeneous Markov chain on $\mathcal{X}$, which is Harris ergodic to a stationary probability distribution π. Consider a stochastic process $\{X_n\}$ on $\mathcal{X}$ with the following properties:

(a) The process $\{X_n\}$ never moves more than some fixed finite distance $D >$

0 in any one step, i.e. the kernel P satisfies that

$$P\Big(x, \{y \in \mathcal{X} : \eta(x,y) \le D\}\Big) = 1, \quad x \in \mathcal{X}.$$

(b) The process $\{X_n\}$ moves by the fixed transition kernel P whenever the current state $X_n = x$ is outside of a fixed compact subset $K \subset \mathcal{X}$, i.e. $\mathbf{P}[X_{n+1} \in A \,|\, X_n = x, X_{n-1}, \ldots, X_0] = P(x, A)$ for $x \notin K$. Inside of K, the process can move arbitrarily in a non-anticipating way, subject only to measurability, to anywhere within K_D, where K_r is defined to be the set of $\forall x \in \mathcal{X}$ within a distance $r > 0$ of K.

(c) The fixed kernel P is bounded above by $P(x, dy) \le M\mu_*(dy)$ for some finite constant $M > 0$, for all $x \in K_D \backslash K$ and all $y \in K_{2D} \backslash K_D$, where μ_* is any probability measure concentrated on $K_{2D} \backslash K_D$.

(d) The fixed kernel P is bounded below by $P^{n_0}(x, A) \ge \epsilon \nu_*(A)$ (P^{n_0} is a n_0-step transition probability.) for some probability measure ν_* on $\mathcal{X}$, some $n_0 \in \mathbb{N}$, and some constant $\epsilon > 0$, for all $x \in K_{2D} \backslash K_D$ and all $A \in \mathcal{F}$. ν_* must be either (1) $\nu_* = \mu_*$ or (2) ν_* can be any probability measure on $\mathcal{X}$ if P is reversible with respect to π and $\mu_* = \pi|_{K_{2D} \backslash K_D}$. ($\mu_*$ here is the μ_* in (c) above.)

(Note that assumption (d) implies that $K_{2D} \backslash K_D$ is n_0-small for P. Also, in the case of a Metropolis-Hastings algorithm, if the corresponding proposal kernels Q_γ and Q satisfy the assumptions (a), (b) and (c), then P_γ and P automatically satisfy them too.)

By Theorem 5 of [7], if a stochastic process $\{X_n\}$ satisfies the above conditions, then it is bounded in probability, i.e.

$$\lim_{L \to \infty} \sup_{n \in \mathbb{N}} \mathbf{P}(\eta(X_n, 0) > L \,|\, X_0 = x_0) \;=\; 0.$$

Furthermore, by Proposition 6 of [7], if $\mathcal{X}$ is an open subset of $\mathbb{R}^d$, then condition (d) above, with $\nu_* = \text{Uniform}(K_{2D} \backslash K_D)$, is implied by the following condition (d'):

(d') The fixed kernel P is bounded below by $P(x, dy) \ge \epsilon Leb(dy)$ (Leb is the Lebesgue measure) whenever $x, y \in J$ with $|y - x| < \delta$ for some $\epsilon > 0$ and $\delta > 0$, where J is any bounded rectangle with $J \supset K_{2D} \backslash K_D$.

To prove Containment for an adaptive MCMC algorithm, an additional condition is needed besides $\{X_n\}$ being bounded in probability. One way to proceed is in terms of density functions. We shall assume that with respect to some reference measure $\lambda(\cdot)$ on $\mathcal{X}$, π has a density g so that $\pi(dy) = g(y)\lambda(dy)$, and furthermore *either* each kernel P_γ has a density p_γ with respect to $\lambda(\cdot)$ so $P_\gamma(x, dy) = p_\gamma(x, y)\lambda(dy)$, or each P_γ is a Metropolis-Hastings algorithm whose proposal kernel Q_γ has a density q_γ with respect to $\lambda(\cdot)$ so that $Q_\gamma(x, dy) = q_\gamma(x, y)\lambda(dy)$. We assume the reference measure $\lambda(\cdot)$ gives finite measure to every bounded set and for any $x \in \mathcal{X}$ and any $D > 0$, $\lambda(\{y \in \mathcal{X} | \eta(x, y) = D\}) = 0$.

In terms of these assumed densities and reference measure, we introduce an additional assumption (e) as follows.

(e) g is a continuous positive density function for π, and furthermore, either:

(e1) The mapping $(x, \gamma) \mapsto p_\gamma(x, y)$ is truncated combocontinuous, i.e. $p_\gamma(x, y) = \alpha_{\gamma,I(x)}(x, y)\mathbf{1}(\eta(x, y) \leq D)$ for some index function $I : \mathcal{X} \to \{1, 2, \ldots, m\}$ for some $m \in \mathbb{N}$, where each $\alpha_{\gamma,i}(x, y)$ is jointly continuous in x and γ for each fixed y. Furthermore, the $\alpha_{\gamma,i}(x, y)$ are uniformly bounded, so that $p_\gamma(x, y)$ is also uniformly bounded.

(e2) Or, in the case of a Metropolis-Hastings algorithm: The proposal density mapping $(x, \gamma) \mapsto q_\gamma(x, y)$ is truncated combocontinuous, i.e. $q_\gamma(x, y) = \beta_{\gamma,I(x)}(x, y)\mathbf{1}(\eta(x, y) \leq D)$ for some index function $I : \mathcal{X} \to \{1, 2, \ldots, m\}$ for some $m \in \mathbb{N}$, where $\beta_{\gamma,i}(x, y)$ is jointly continuous in x and γ for each fixed y, and $\int_{\mathcal{X}} \beta_{\gamma,i}(x, y)\,\lambda(dy) = 1$ for $i = 1, \ldots, m$. Furthermore, the $\beta_{\gamma,i}(x, y)$ are uniformly bounded, so that $a_\gamma(x, y)$ is also uniformly bounded.

Remark. If $\alpha_{\gamma,i}(x, y)$ is jointly combocontinuous in (x, y, γ), then $\alpha_{\gamma,i}(x, y)$ is uniformly bounded in a compact space and so is $p_\gamma(x, y)$ (or $q_\gamma(x, y)$).

In [7], a version of assumption (e) was used in which λ was assumed to be Lebesgue measure, and full continuity was assumed in place of combocontinuity, thus leading to inconvenient application as discussed in Section 3 above.

We will give a special attention to a Metropolis-Hastings algorithm ([15, 14]). The algorithm works as follows. At each iteration n, conditional on the current state X_n, the Markov chain proposes Y_{n+1} from some proposal subkernel $Q_\gamma(X_n, \cdot)$, whose subdensity is $q_\gamma(x, \cdot)$, with $q_\gamma(x, y) \geq 0$ and $\int_{y \in \mathcal{X}} q_\gamma(x, y)\,\lambda(dy) \leq 1$. The new proposal Y_{n+1} is accepted with probability

$$a(X_n, Y_{n+1}) = \min\big[1, \frac{\pi(Y_{n+1})q_\gamma(Y_{n+1}, X_n)}{\pi(X_n)q_\gamma(X_n, Y_{n+1})}\big],$$

otherwise Y_{n+1} is rejected with probability $1 - a(X_n, Y_{n+1})$. If Y_{n+1} is accepted, $X_{n+1} = Y_{n+1}$, if not, $X_{n+1} = X_n$. Hence, the Metropolis-Hastings algorithm has transition kernel

$$P_\gamma(x, dy) = a(x, y)Q_\gamma(x, dy) + r(x)\delta_x(dy),$$

where $r(x) = 1 - \int_{\mathcal{X}} a(x, y)Q_\gamma(x, dy)$, and $\delta_x(\cdot)$ is a point-mass at x. (Note also that if q_γ is *symmetric*, then the acceptance formula reduces to simply $a(X_n, Y_{n+1}) = \min\left[1, \frac{\pi(Y_{n+1})}{\pi(X_n)}\right]$.)

It is easily checked and well known (e.g. [15, 14, 22, 17]) that the above acceptance probability $a(X_n, Y_{n+1})$ ensures that the Markov chain is reversible with respect to π, i.e. that $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$, so the Metropolis-Hastings algorithm leaves π stationary, and assuming irreducibility it is Harris recurrent. Furthermore, if π has everywhere-positive density, then these facts are easily seen to remain true even if (as we shall do below) we reject all jumps of distance more than D.

5 Main Result

In terms of the above conditions, we have the following theorem which guarantees Containment, and hence also convergence provided Diminishing Adaption is satisfied.

Theorem 1. *Consider an adaptive MCMC algorithm as above. If the algorithm satisfies the assumptions (a), (b), (c), (d), and (e), and if the space $\mathcal{Y}$ of Markov kernel indices is compact, then the algorithm satisfies the Containment condition* (2.3)*. Hence, if it also satisfies the Diminishing Adaption condition* (2.2)*, then it converges to stationarity as in* (2.1)*.*

Proof. First, by Theorem 5 of [7], we know that the process $\{X_n\}$ is bounded in probability since it satisfies conditions (a), (b), (c), and (d).

Next, it follows from Lemma 5 in Section 8 below that for each $n \in \mathbb{N}$, the mapping

$$(x, \gamma) \mapsto \Delta(x, \gamma, n) := \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \tag{5.1}$$

is a jointly combocontinuous mapping in (x, γ), and each 'piece' is a non-increasing function of n which converges to 0. Then, by applying Theorem 4 (Generalised Dini's Theorem) in Section 7 below to the function $f_n(x, \gamma) = \Delta(x, \gamma, n + 1)$, we obtain that

$$\lim_{n\to\infty} \sup_{x\in\mathcal{C}} \sup_{\gamma\in\mathcal{Y}} \Delta(x, \gamma, n) = 0$$

for any compact set $\mathcal{C} \subset \mathcal{X}$.

The rest of the proof to show the Containment condition holds is the same as the last part of the proof of Proposition 23 in [7]. To repeat here, if $\{X_n\}$ is bounded in probability, then for any $\delta > 0$, there is a compact subset $\mathcal{C}$ such that $P(X_n \notin \mathcal{C}) \leq \delta$ for all n. With any $\epsilon > 0$, we can find n such that $\sup_{x \in \mathcal{C}} \sup_{\gamma \in \mathcal{Y}} \Delta(x, \gamma, n) < \epsilon$. Thus, $\sup_{x \in \mathcal{C}} \sup_{\gamma \in \mathcal{Y}} M_\epsilon(x, \gamma) < \infty$ for any $\epsilon > 0$.Let $L := \sup_{x \in \mathcal{C}} \sup_{\gamma \in \mathcal{Y}} M_\epsilon(x, \gamma)$. Then $P(M_\epsilon(X_n, \Gamma_n) > L) \leq \delta$ for all n. Therefore, the Containment condition holds, i.e. (2.3) is satisfied.

Finally, if the adaptive MCMC algorithm also satisfies the Diminishing Adaption condition, then by [18], the algorithm converges to π in total variation distance, i.e. (2.1) holds. □

It remains to prove the Generalised Dini's Theorem and Lemma 5 used in the proof of Theorem 1 above, which we do in Sections 7 and 8 below.

6 The Bounded Adaption Metropolis (BAM) Algorithm

To illustrate our results in a simple but useful case, consider the following Bounded Adaption Metropolis (BAM) Algorithm, which is an easier-to-implement version of the similarly-named algorithm presented in [7].

Let $\mathcal{X} = \mathbb{R}^d$, let $K \subseteq \mathcal{X}$ be a (large) bounded region, let π be a continuous positive density on $\mathcal{X}$, and let $D > 0$ be a (large) constant. Let $\mathcal{Y}$ be any compact collection of d-dimensional positive-definite matrices, and fix some specific matrix $\Sigma_* \in \mathcal{Y}$.

Define a process $\{X_n\}$ as follows: $X_0 = x_0$ for some fixed $x_0 \in K$. Then for $n = 0, 1, 2, \ldots$, given X_n, we generate a proposal Y_{n+1} by: (a) if $X_n \notin K$, then $Y_{n+1} \sim N(X_n,\ \Sigma_*)$; or (b) if $X_n \in K$, then $Y_{n+1} \sim N(X_n,\ \Sigma_{n+1})$, where the matrix $\Sigma_{n+1} \in \mathcal{Y}$ is selected in some fashion, perhaps depending on X_n and on the chain's entire history. Once Y_{n+1} is chosen, then if $|Y_{n+1} - X_n| > D$, the proposal is rejected so $X_{n+1} = X_n$. Otherwise, if $|Y_{n+1} - X_n| \leq D$, then with probability

$$a(X_n, Y_{n+1}) = \begin{cases} \min[1,\ \frac{\pi(Y_{n+1})}{\pi(X_n)}] & \text{if } X_n \in K \,\&\, Y_{n+1} \in K \text{ or } X_n \notin K \,\&\, Y_{n+1} \notin K \\ \min[1,\ \frac{\pi(Y_{n+1}) q_{\Sigma_*}(Y_{n+1}, X_n)}{\pi(X_n) q_{\Sigma_{n+1}}(X_n, Y_{n+1})}] & \text{if } X_n \in K \,\&\, Y_{n+1} \notin K \\ \min[1,\ \frac{\pi(Y_{n+1}) q_{\Sigma_{n+1}}(Y_{n+1}, X_n)}{\pi(X_n) q_{\Sigma_*}(X_n, Y_{n+1})}] & \text{if } X_n \notin K \,\&\, Y_{n+1} \in K \end{cases}$$

the proposal is accepted so $X_{n+1} = Y_{n+1}$, or with the remaining probability the proposal is rejected so $X_{n+1} = X_n$. (Here $q_\Sigma(x, y)$ is the density of $N(x, \Sigma)$ evaluated at y.) That is, at iteration n, BAM is a version of the

Metropolis-Hastings algorithm as defined earlier, with proposal subdistribution $N(x, \Sigma_{n+1})(dy)\, \mathbf{1}(\eta(x,y) \le D)$ for $X_n \in K$, or $N(x, \Sigma_*)(dy)\, \mathbf{1}(\eta(x,y) \le D)$ for $X_n \notin K$, and the above formula for $a(X_n, Y_{n+1})$ then corresponds to the usual Metropolis-Hastings acceptance probability for this proposal.

Special case. For example, Σ_{n+1} could be chosen to be $(2.38)^2 V_n/d$ where V_n is the empirical covariance matrix of $X_0, \ldots, X_n$ from the process's previous history (except restricted to some compact set $\mathcal{Y}$), since that choice approximates the optimal proposal covariance, cf. [13, 19]. One way to define V_n is let $V_n = \mathrm{Cov}(\langle X_0 \rangle, \langle X_1 \rangle, \ldots, \langle X_n \rangle) + \epsilon I_d$ for some arbitrarily small constant $\epsilon > 0$. $\langle X_i \rangle$ is a shrunken version of X_i, i.e. $\langle X_i \rangle_j \; = \max(-L, \min(L, X_{i,j}))$ for some (large) constant $L > 0$, with j indexing for j^{th} coordinate. This idea of defining V_n is from Section 12.3 of [7].

Proposition 2. *Consider the 'special case' described above. Let $\mathcal{Y} = \{\gamma \mid \gamma$ is a $d \times d$ positive definite matrix, $\epsilon\, I_d \le \gamma \le (8L^2 + \epsilon) d\, I_d\}$. Then $\mathcal{Y}$ is compact and every V_n is in $\mathcal{Y}$. Also, the 'special case' of BAM algorithm satisfies the Diminishing Adaption condition, i.e.* (2.2)

Proof. See proof of Proposition 6 in [25]. □

For the BAM algorithm, our results herein prove that the Containment condition holds, and hence convergence to stationarity also holds assuming Diminishing Adaption:

Theorem 3. *The above BAM algorithm satisfies Containment* (2.3)*. Hence, if the selection of the Σ_n satisfies Diminishing Adaption* (2.2)*, then convergence to stationarity* (2.1) *holds.*

This stands in contrast to other situations in which it is very difficult or impossible to establish convergence of adaptive MCMC algorithms. Theorem 3 is proved in Section 9 below, as a special case of our more general results.

7 Generalised Dini's Theorem

Dini's Theorem may be stated as follows (see e.g. Theorem 7.13 in [21]). Let $\{f_n\}$ be a sequence of continuous real-valued functions defined on a compact set C, which is non-decreasing (i.e. $f_n(x) \le f_{n+1}(x)$ for each fixed n and $x \in C$), and which converges pointwise to a continuous function f (i.e. $\lim_{n\to\infty} f_n(x) = f(x)$ for each fixed $x \in \mathcal{X}$). Then the convergence is uniform, i.e. $\lim_{n\to\infty} \sup_{x \in C} |f_n(x) - f(x)| = 0$.

In this section, we generalise Dini's Theorem to the combocontinuous case, so that the theorem can be applied to prove Theorem 1.

Theorem 4. (Generalised Dini's Theorem)

Suppose a set C is compact, and $\{f_n\}$ is a sequence of real-valued functions on C, and f is a continuous real-valued function on C, and:

1. *For each $n \in \mathbb{N}$, f_n can be expressed as $f_n(z) = f_{n,I(z)}(z)$ for some index function $I(z) \in \mathcal{J} = \{1, 2, \ldots, m\}$, and some $m \in \mathbb{N}$, and some collection $f_{n,i}$ of functions.*

2. *Each of these $f_{n,i}$ is a continuous real-valued function on $\overline{C}_i$, the closure of the subset $C_i = \{z \in C \mid I(z) = i\}$.*

3. *For each $i \in \mathcal{J}$, $\{f_{n,i}\}$ converges pointwise to f on $\overline{C}_i$.*

4. *For each $i \in \mathcal{J}$, $f_{n,i}(z) \geq f_{n+1,i}(z)$ for all $z \in \overline{C}_i$, $n = 1, 2, 3, \ldots.$*

Then $f_n \to f$ uniformly on C, i.e. $\lim_{n\to\infty} \sup_{x \in C} |f_n(x) - f(x)| = 0$.

Proof. This follows by applying the original Dini's Theorem separately on each subset $\overline{C}_i$. For a complete proof, let $g_{n,i} = f_{n,i} - f$ for each $i \in \mathcal{J}$. Since $\{f_{n,i}\}$ converges pointwise to f on $\overline{C}_i$, $\{g_{n,i}\}$ converges pointwise to 0 on $\overline{C}_i$. Also, for each $i \in \mathcal{J}$, since $f_{n,i}(z) \geq f_{n+1,i}(z)$ for all $z \in \overline{C}_i$, $g_{n,i} \geq g_{n+1,i}$ for all $z \in \overline{C}_i$.

Let $\epsilon > 0$ and $C_{n,i} = \{z \in \overline{C}_i | g_{n,i}(z) \geq \epsilon\}$, $i \in \mathcal{J}$. Then $C_{n,i}$ is closed, since $g_{n,i} = f_{n,i} - f$ is continuous on $\overline{C}_i$, and the continuous inverse image of any closed set is closed (e.g. Rudin, 1976, Theorem 4.8 Corollary). Hence, $C_{n,i}$ is compact, since closed subsets of compact sets are compact (e.g. Rudin, 1976, Theorem 2.35).

Next, note that $C_{n,i} \supset C_{n+1,i}$, since $g_{n,i} \geq g_{n+1,i}$ on $\overline{C}_i$. Pick $z \in \overline{C}_i$. Since $g_{n,i} \to 0$ on $\overline{C}_i$, $z \notin C_{n,i}$ if n is sufficiently large. Thus, for every $z \in \overline{C}_i$, we have that $z \notin \cap_{n=1}^{\infty} C_{n,i}$. It follows that $\cap_{n=1}^{\infty} C_{n,i}$ is the empty set. Hence, by the finite intersection property (e.g. Rudin, 1976, Theorem 2.36, Corollary), there must be some $N_i \in \mathbb{N}$ such that $C_{N_i,i}$ is empty.

Therefore, $0 \leq g_{n,i}(z) < \epsilon$ for all $z \in C_i$ and for all $n \geq N_i$. Hence, $0 \leq g_n(z) < \epsilon$ for all $z \in C$ and for all $n \geq \max(N_1, \ldots, N_m)$. Since ϵ is arbitrary, $f_n \to f$ uniformly on C. □

Remark. In Theorem 4, it does not suffice to assume only that $f_{n,i}$ converges pointwise to f on C_i, i.e. the closure $\overline{C_i}$ really is required. For example, let $C = [0, 2]$, and $m = 2$, with $I(x) = 1$ for $x \in [0, 1)$ and $I(x) = 2$ for $x \in [1, 2]$. Then let $f_{n,1}(x) = x^n$, and $f_{n,2}(x) = f(x) = 0$. Then $f_{n,1} \to 0$ pointwise on $C_1 := [0, 1)$, but $\sup_{C_1} f_{n,1} = 1$ for each n, so the convergence of f_n to f is not uniform.

8 Lemma About Combocontinuity

We here show that, under the assumptions of Theorem 1, the total variation distance mapping (5.1) is combocontinuous, and each 'piece' converges to 0. Then we can apply Theorem 4 (Generalised Dini's Theorem) to the mapping (5.1).

Lemma 5. *Consider an adaptive MCMC algorithm as in Section 2, with assumed densities as in Section 4. Assume condition (e). Then:*

1. *for each $n \geq 1$, the function $f_{n,\gamma}(x) := \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|$ is jointly combocontinuous in (x, γ) in the sense that: $f_{n,\gamma}(x) = f_{n,\gamma,I(x)}(x)$ for some index function $I : \mathcal{X} \to \{1, 2, \ldots, m\}$ for some $m \in \mathbb{N}$ where $f_{n,\gamma,i}(x)$ is jointly continuous in x and γ; and*

2. *for each fixed (x, γ, i), $f_{n,\gamma,i}(x)$ converges pointwise to 0 on $\mathcal{X}$ as $n \to \infty$ and is a non-increasing function in n.*

Proof. For simplicity we assume (e2); the proof for (e1) is similar but easier (e.g. replace $a_\gamma(x)$ by 1). Thus, the proposal densities $q_\gamma(x, y)$ of a Metropolis Hastings algorithm are truncated combocontinuous, with

$$q_\gamma(x,y) = \sum_{i=1}^{m} \beta_{\gamma,i}(x,y)\mathbf{1}(x \in T_i)\mathbf{1}(\eta(x,y) \leq D) \tag{8.1}$$

where $D > 0$ is some (large) constant, $\{T_1, \ldots, T_m\}$ is a partition of $\mathcal{X}$, and $\beta_{\gamma,i}(x, y)$ is jointly continuous in x and y and γ, and is uniformly bounded.

Let $a_\gamma(x)$ be the acceptance probability of a proposal from $x \in \mathcal{X}$ in the Metropolis-Hastings algorithm. Write $a_\gamma(x)$ as

$$a_\gamma(x) = \int_{\{y \in \mathcal{X} | \eta(x,y) \leq D\}} \min\left[1, \frac{g(y)q_\gamma(y,x)}{g(x)q_\gamma(x,y)}\right] q_\gamma(x,y)\lambda(dy)\,.$$

Define $w_\gamma(x, y)$ as

$$w_\gamma(x,y) = \min\left[1, \frac{g(y)q_\gamma(y,x)}{g(x)q_\gamma(x,y)}\right] q_\gamma(x,y)\,.$$

Notice that we can write

$$w_\gamma(x,y) = \sum_{i=1}^{m} w_{\gamma,i}(x,y)\mathbf{1}(x \in T_i)$$

where

$$w_{\gamma,i}(x,y) = \min\Big[1, \frac{g(y)q_\gamma(y,x)}{g(x)\beta_{\gamma,i}(x,y)}\Big]\beta_{\gamma,i}(x,y)\,.$$

Then the transition kernel $P_\gamma(x,dy)$ of this algorithm can be written as

$$P_\gamma(x,dy) = [1 - a_\gamma(x)]\delta_x(dy) + w_\gamma(x,y)\lambda(dy).$$

where $\delta_x(\cdot)$ is a point-mass at x.

Next, define a set $A^n_{(x,y,D)}$ as

$$A^n_{(x,y,D)} = \{(z_1, \ldots, z_n) \in \mathcal{X}^n \mid \eta(x,z_1) \leq D,\ \eta(z_1,z_2) \leq D,\ \eta(z_2,z_3) \leq D, \ldots,\ \eta(z_{n-1},z_n) \leq D, \eta(z_n,y) \leq D\}.$$

That is, $A^n_{(x,y,D)}$ is the set of paths moving from x to y in n steps while never moving more than a distance D on any one step. Then the n-step transition kernel $P^n_\gamma(x,dy)$ can be written as

$$P^n_\gamma(x,dy) = [1 - a_\gamma(x)]^n\delta_x(dy) + w^n_\gamma(x,y)\lambda(dy)$$

where we define

$$w^n_\gamma(x,y) = \sum_{S\neq\emptyset} w^{n,S}_\gamma(x,y).$$

Here the sum is over all non-empty subsets $S \subseteq \{1, 2, ..., n\}$, and $w^{n,S}_\gamma(x,y)$ is the sub-density corresponding to getting from x to y in n steps while accepting moves only at the times in S (while rejecting moves at all times not in S). For example, if $n = 5$ and $S = \{2,4,5\}$, then $w^{5,\{2,4,5\}}_\gamma(x,y)$ corresponds to transitioning from x to y in 5 steps, while the first and third proposals are rejected and the others are accepted. The transition density $w^{5,\{2,4,5\}}_\gamma(x,y)$ can be thus written as

$$w^{5,\{2,4,5\}}_\gamma(x,y) = \iint\limits_{A^2_{(x,y,D)}} [1 - a_\gamma(x)]w_\gamma(x,y_1)[1 - a_\gamma(y_1)]w_\gamma(y_1,y_2)w_\gamma(y_2,y)\,\lambda(dy_1)\lambda(dy_2)\,.$$

First, we prove that $[1 - a_\gamma(x)]^n$ is jointly combocontinuous in x and γ. We write

$$a_\gamma(x) = \sum_{i=1}^m a_{\gamma,i}(x)\mathbf{1}(x \in T_i)$$

Then, $a_{\gamma,i}(x)$ can be written as

$$a_{\gamma,i}(x) = \int_{\{y \in \mathcal{X} | \eta(x,y) \leq D\}} w_{\gamma,i}(x,y)\lambda(dy) \tag{8.2}$$

for $i = 1, \ldots, m$. Fix (x_0, γ_0). Let a sequence $\{(x_k, \gamma_k)\}_{k=1}^{\infty} \to (x_0, \gamma_0)$. Group possible y values into

$$\begin{aligned} U_1 &= \{y \in \mathcal{X} | \eta(x_k, y) \leq D \; \& \; \eta(x_0, y) \leq D\} \\ U_2 &= \{y \in \mathcal{X} | \eta(x_k, y) \leq D \; \& \; \eta(x_0, y) > D\} \\ U_3 &= \{y \in \mathcal{X} | \eta(x_k, y) > D \; \& \; \eta(x_0, y) \leq D\} \end{aligned}$$

When $k \to \infty$, $\lambda(U_2), \lambda(U_3) \to 0$; $w_{\gamma,i}(x,y)$ is jointly continuous in x and γ for each fixed y and uniformly bounded; $\lambda(U_1)$ is finite. Thus, $a_{\gamma_k,i}(x_k) \to a_{\gamma_0,i}(x_0)$ as $k \to \infty$ by the Bounded Convergence Theorem; i.e. $a_{\gamma,i}(x)$ is jointly continuous in x and γ and so are $[1 - a_{\gamma,i}(x)]$ and $[1 - a_{\gamma,i}(x)]^n$. We conclude that $[1 - a_\gamma(x)]^n$ is jointly combocontinuous in x and γ.

Second, we prove $w_\gamma^{n,S}(x,y)$ is jointly combocontinuous in x and γ, and is uniformly bounded when restricted to the subset $\{(x,y) : \eta(x,y) \leq \ell D\}$ where $|S| = \ell$. Fix $x_0 \in T_i$ and $\gamma_0 \in \mathcal{Y}$. Let a sequence $\{(x_k, \gamma_k)\}_{k=1}^{\infty} \to (x_0, \gamma_0)$, with each $x_k \in \overline{T}_i$. Let y_1 be the very first proposal accepted by either of two chains started at x_k and x_0. We group possible y_1 values into the same subsets U_1 and U_2 and U_3 as above. As $k \to \infty$, $\lambda(U_2), \lambda(U_3) \to 0$. Also, $[1 - a_{\gamma,i}(x)]$ and $w_{\gamma,i}(x,y)$ are jointly continuous in x and γ. Now, $\lambda(A^{l-1}_{(x_k,y,D)})$ and $\lambda(A^{l-1}_{(x_0,y,D)})$ are finite so that we can apply the Bounded Convergence Theorem for the set $A^{l-1}_{(x_k,y,D)} \cap A^{l-1}_{(x_0,y,D)}$. Hence, $w_{\gamma_k}^{n,S}(x_k, y) \to w_{\gamma_0}^{n,S}(x_0, y)$ as $k \to \infty$. It follows that $w_\gamma^{n,S}(x,y)$ is jointly continuous in x and γ when restricted to $\overline{T}_i$. This means we can write $w_\gamma^{n,S}(x,y)$ as

$$w_\gamma^{n,S}(x,y) = \sum_{i=1}^{m} w_{\gamma,i}^{n,S}(x,y)\mathbf{1}(x \in T_i) \tag{8.3}$$

where $w_{\gamma,i}^{n,S}(x,y) = w_\gamma^{n,S}(x,y)$ for $x \in \overline{T}_i$, and $w_{\gamma,i}^{n,S}(x,y)$ is extended to other $x \in \mathcal{X}$ arbitrarily subject to preserving its continuity. This shows that $w_\gamma^{n,S}(x,y)$ is jointly combocontinuous in x and γ. Furthermore, the function g is continuous, and $\beta_{\gamma,i}$ is (by assumption) uniformly bounded. So, a_γ and w_γ are uniformly bounded when restricted to a compact set. Therefore, $w_\gamma^{n,S}(x,y)$ is uniformly bounded when restricted to the subset $\{(x,y) : \eta(x,y) \leq \ell D\}$.

Third, we prove $\|P_\gamma^n(x,\cdot)-\pi(\cdot)\|$ is jointly combocontinuous in x and γ. We write $\|P_\gamma^n(x,\cdot)-\pi(\cdot)\|$ as

$$\|P_\gamma^n(x,\cdot)-\pi(\cdot)\| = [1-a_\gamma(x)]^n + 0.5 * \int_{\mathcal{X}} |w_\gamma^n(x,y)-g(y)|\lambda(dy). \tag{8.4}$$

We rewrite (8.4) as

$$\begin{aligned}\|P_\gamma^n(x,\cdot)-\pi(\cdot)\| &= \sum_{i=1}^m \Big\{[1-a_{\gamma,i}(x)]^n + 0.5 * \int_{\mathcal{X}} |w_{\gamma,i}^n(x,y)-g(y)|\lambda(dy)\Big\}\mathbf{1}(x\in T_i)\\ &= \sum_{i=1}^m f_{n,\gamma,i}(x)\mathbf{1}(x\in T_i)\,, \qquad (8.5)\end{aligned}$$

where we set $f_{n,\gamma,i}(x) = [1-a_{\gamma,i}(x)]^n + 0.5 * \int_{\mathcal{X}} |w_{\gamma,i}^n(x,y)-g(y)|\lambda(dy)$. We proved above that $[1-a_\gamma(x)]^n$ is jointly combocontinuous in x and γ. Now we prove the joint combocontinuity for the latter part of (8.4). Again, we group possible $y's$. This time for some n, we group $y's$ based on the distance ℓD for every $\ell \le n$, creating $3n$ groups of y. i.e.

$$\begin{aligned}
U_{1,1} &= \{y\in\mathcal{X} | \eta(x_k,y)\le D \;\&\; \eta(x_0,y)\le D\}\\
U_{1,2} &= \{y\in\mathcal{X} | \eta(x_k,y)\le D \;\&\; D<\eta(x_0,y)\le 2D\}\\
U_{1,3} &= \{y\in\mathcal{X} | D<\eta(x_k,y)\le 2D \;\&\; \eta(x_0,y)\le D\}\\
U_{2,1} &= \{y\in\mathcal{X} | D<\eta(x_k,y)\le 2D \;\&\; D<\eta(x_0,y)\le 2D\}\\
U_{2,2} &= \{y\in\mathcal{X} | D<\eta(x_k,y)\le 2D \;\&\; 2D<\eta(x_0,y)\le 3D\}\\
U_{2,3} &= \{y\in\mathcal{X} | 2D<\eta(x_k,y)\le 3D \;\&\; D<\eta(x_0,y)\le 2D\}\\
&\vdots\\
U_{n,1} &= \{y\in\mathcal{X} | (n-1)D<\eta(x_k,y)\le nD \;\&\; (n-1)D<\eta(x_0,y)\le nD\}\\
U_{n,2} &= \{y\in\mathcal{X} | (n-1)D<\eta(x_k,y)\le nD \;\&\; \eta(x_0,y)> nD\}\\
U_{n,3} &= \{y\in\mathcal{X} | \eta(x_k,y)> nD \;\&\; (n-1)D<\eta(x_0,y)\le nD\}
\end{aligned}$$

$\lambda(U_{1,2}),\lambda(U_{1,3}),\lambda(U_{2,2}),\lambda(U_{2,3}),\ldots,\lambda(U_{n,2}),\lambda(U_{n,3}) \to 0$ as $k\to 0$. With jointly continuity and uniform boundedness of the function $w_{\gamma,i}^{n,S}(x,y)$ proved above, by Bounded Convergence Theorem the latter part of (8.4) is combocontinuous in x and γ as $\lambda(U_{1,1}\cup U_{2,1}\cup\ldots\cup U_{n,1})$ is finite. Thus, $\|P_\gamma^n(x,\cdot)-\pi(\cdot)\|$ is jointly combocontinuous in x and γ.

Lastly, we want to prove each $f_{n,\gamma,i}$ from (8.5) converges to 0 on $\mathcal{X}$ as $n\to\infty$, and is a non-increasing function of n. Now, each $f_{n,\gamma,i}$ is the total variation distance of a Metropolis-Hastings algorithm, which moves according to a fixed transition kernel after the first iteration. No matter which state the chain is at after the first iteration (which must be within distance D of the initial state), from that point onwards it converges to its stationary distribution π. Thus, for each fixed (x, γ, i), the function $f_{n,\gamma,i}(x)$ converges pointwise to 0 on $\mathcal{X}$, and is a non-increasing function in n for $n \geq 2$ by e.g. Proposition 3(c) of [16]. The result follows. □

Remark. In the last paragraph of the above proof, note that we need each $f_{n,\gamma,i}$ from (8.5) to converge to 0 on $\mathcal{X}$, not just on T_i. This is because when we apply Theorem 4 (Generalised Dini's Theorem), we need the convergence on $\overline{T}_i$, not just on T_i. Hence, it is not sufficient to just state that each P_γ is Harris ergodic to π.

9 Proof of Theorem 3

Denote the density of π as g with respect to Lebesgue measure. Denote the proposal kernels for $x \in K$ as $\{Q^*_\gamma(x, \cdot)\}_{\gamma \in \mathcal{Y}}$ and the proposal kernel for $x \notin K$ as $Q(x, \cdot)$. We then define Q_γ as

$$Q_\gamma(x, \cdot) = \begin{cases} Q^*_\gamma(x, \cdot) & x \in K \\ Q(x, \cdot) & x \notin K \end{cases}$$

Since we reject a proposal y from x if $|x-y| > D$, $Q_\gamma(x, dy) = 0$ if $|x-y| > D$. Let $P_\gamma(x, \cdot)$ be a corresponding transition kernel for $Q_\gamma(x, \cdot)$.

The BAM algorithm with a fixed proposal kernel Q_γ is reversible with respect to π. Thus, π is a stationary distribution for the algorithm. As noted in Section 4, a full-dimensional Metropolis-Hastings algorithm with a centered truncated normal proposal kernel is Harris recurrent. Thus, the BAM algorithm with a fixed proposal kernel Q_γ is Harris recurrent. Since it is also aperiodic, it follows that the BAM algorithm with each fixed Q_γ is Harris ergodic to π.

As constructed, the BAM algorithm determines γ (or Σ_{n+1}) of $\{Q_\gamma(x, \cdot)\}_{\gamma \in \mathcal{Y}}$ at each iteration n based on the past and present states from the Markov chain. And $\mathcal{Y}$ is compact. It follows that the BAM algorithm follows the set-up of Section 2.

We also need to check if the algorithm satisfies the assumptions (a), (b), (c), (d) and (e) from Section 4. As noted in Section 4, we only need the corresponding Q_γ and Q to satisfy the assumptions (a), (b) and (c), to ensure that P_γ and P also satisfy them. The proposal kernels of the above BAM algorithm have bounded jumps since $Q_\gamma(x, dy) = 0$ when $|x - y| > D$. Thus, it satisfies assumption (a), with metric $\eta(x, y) = |x - y|$. The chain moves by a fixed transition kernel, Q, outside of a compact subset K, satisfying the assumption (b). The fixed proposal kernel Q is bounded above as described in (c) since it is a normal distribution.

The state space $\mathcal{X}$ of the algorithm is an open subset of $\mathbb{R}^d$, so we can use the assumption (d') to imply the assumption (d). Since Q is a normal distribution and π is continuous on $\mathcal{X}$, the assumption (d') is satisfied. We can easily see that the proposal kernel densities $q_\gamma(x, y)$ of $Q_\gamma(x, dy)$ is continuous in γ and jointly combocontinuous in x and γ as in (e2).

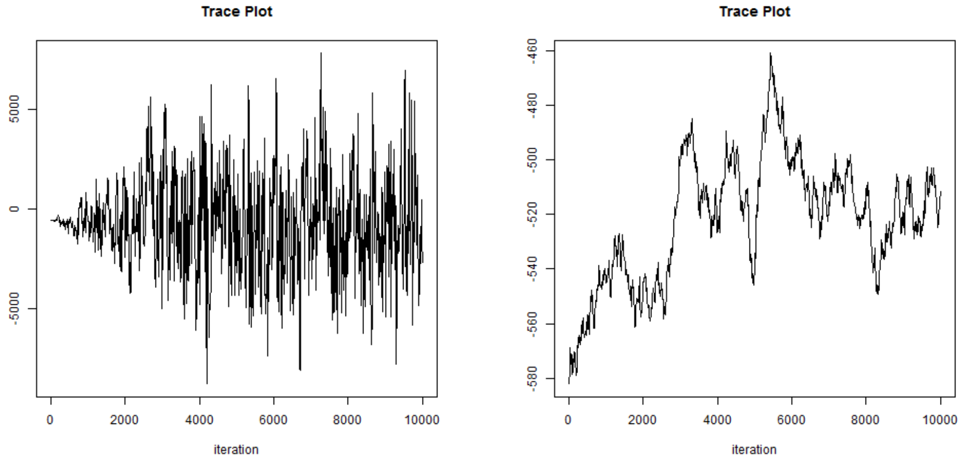
Therefore, by Theorem 1, the BAM algorithm satisfies Containment (2.3), thus proving Theorem 3. □

10 Numerical examples

In this section, we run the 'special case' of BAM algorithm (with $\epsilon = 0.001$) described in Section 6 above, on two specific statistical examples, and compare its performances with those of non-adaptive Metropolis algorithms. The first example is in dimension $d = 9$, and the second one is in dimension $d = 12$. In both cases, our compact set K is defined as the d-dimensional cube $[-100000, 100000]^d$, and our step size bound is $D = 100000$. The fixed proposal kernel for the BAM algorithm when $X_n \notin K$ is $Q \sim N(X_n, I_d)$.

10.1 Application: 9-dimensional Multivariate Normal Distribution

We first consider the target distribution $N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ with $d = 9$ and $\Sigma \in \mathbb{R}^{d \times d}$ are fixed and arbitrary (subject to Σ being positive-definite; in fact we set $\Sigma = M\,M^t$ where the matrix M was generated with iid normal entries). The starting value for the MCMC algorithm is μ itself. The trace plots (of coordinate 7) for a BAM algorithm and a standard Metropolis algorithm (with proposal kernal $N(X_n, I_d)$) are shown in Figure 2.



Bounded Adaption Metropolis

Standard Metropolis, $N(X_n, I_d)$

Figure 2: Trace plots (of coordinate 7) for a Bounded Adaption Metropolis (left) versus a Standard Metropolis algorithm with proposal kernel $N(X_n, I_d)$ (right), on a 9-dimensional multivariate normal target distribution, showing the superiority of the BAM algorithm.

We can see the mixing of the BAM algorithm is a lot better than the standard Metropolis. In the trace plots of the BAM algorithm, we see an increase in the average jumping distance from one state to the next state in the first a few thousands iterations, which implies the adaption was indeed effective.

10.2 Application: Pump Failure Model

We next consider a BAM algorithm for a true Bayesian statistical model, applied to the number of failures of pumps at a nuclear power plant. This model was first introduced by [9]; we use the slightly different set-up from [10]. The resulting posterior density is

$$f(\lambda_1, ..., \lambda_n, \alpha, \beta | y_1, ..., y_n) \propto e^{-\alpha}\beta^{0.1-1}e^{-\beta}\prod_{i=1}^{n}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}(\lambda_i t_i)^{y_i}e^{-\lambda_i t_i}.$$

This follows from the assumption that the number of failures follow the distribution

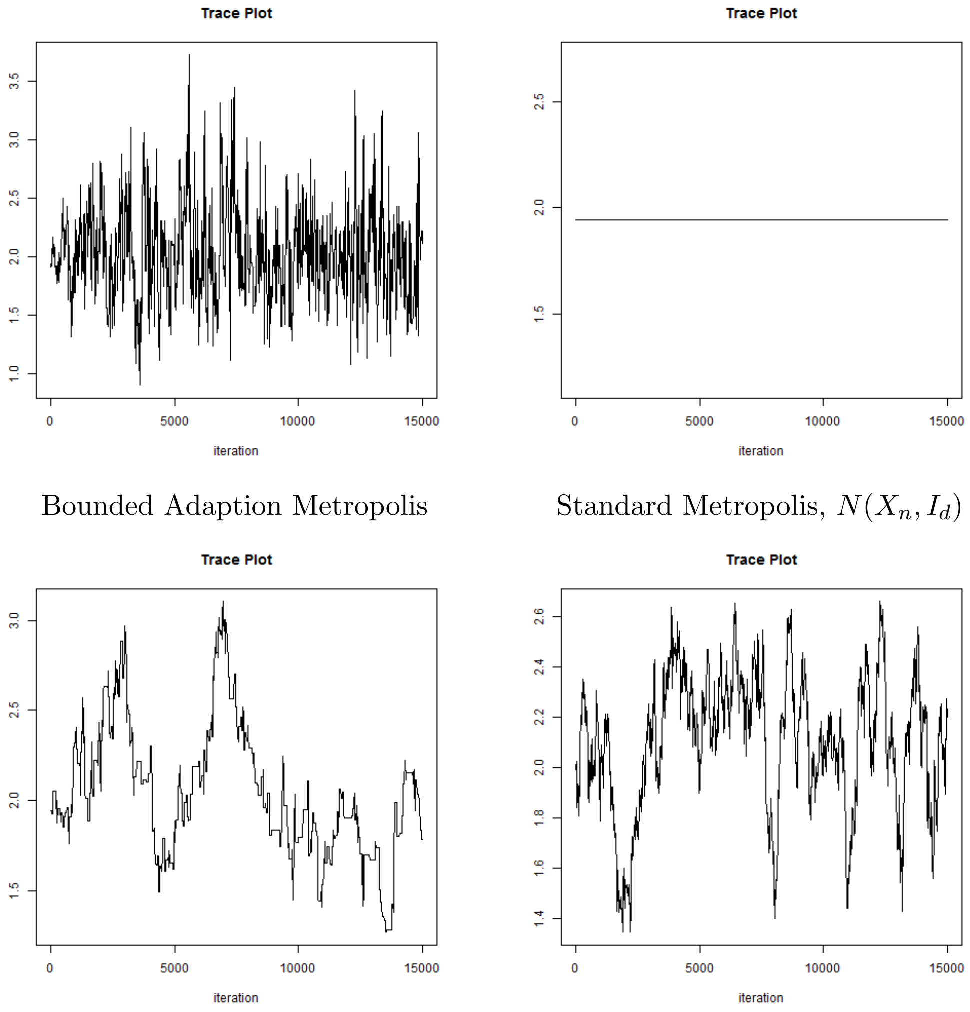
$$f(y_1, ..., y_n | \lambda_1, ..., \lambda_n) = \prod_{i=1}^{n}\text{Poisson}(\lambda_i t_i),$$

Table 1: Pump Failure Data

Obs. no.	1	2	3	4	5	6	7	8	9	10
y_i	5	1	5	14	3	19	1	1	4	22
t_i	94.320	15.720	62.880	125.760	5.240	31.440	1.048	1.048	2.096	10.480

where λ_i is the failure rate of pump i, t_i is the length of operation time (in thousands of hours), and n is the number of pumps, and furthermore $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, $\alpha \sim \text{Exp}(1)$, and $\beta \sim \text{Gamma}(0.1, 1)$. Here $n = 10$, so since $\lambda_1, ..., \lambda_n, \alpha, \beta > 0$, the state space is $(0, \infty)^d$ with dimension $d = n + 2 = 12$. The data for the model are the values of the y_i and t_i, which are reproduced in Table 1 herein.

We run both Bounded Adaption Metropolis and standard Metropolis algorithm to compare them. For ease of comparison, each run uses initial values given by the estimates of each parameter obtained from a previous standard MCMC run.



Bounded Adaption Metropolis

Standard Metropolis, $N(X_n, I_d)$

Standard Metropolis, $N(X_n, 0.01 I_d)$

Standard Metropolis, $N(X_n, 0.001 I_d)$

Figure 3: Trace plots (of coordinate 10) for the Pump Failure Model example for a Bounded Adaption Metropolis algorithm (top left), compared to Standard Metropolis algorithms with proposal distributions whose Gaussian covariance matrices are the d-dimensional identity (top right), 0.01 times this identity (bottom left), and 0.001 times this identity (bottom right).

Figure 3 shows the trace plots for the pump failure model with the BAM algorithm, and with the standard Metropolis algorithm with proposal kernels $N(X_n, I_d)$, $N(X_n, 0.01\, I_d)$, and $N(X_n, 0.001\, I_d)$.

For the non-adaptive algorithm with a fixed proposal kernel $N(X_n, I_d)$, not a single proposal was accepted for the whole 15000 iterations. This is a combination of a couple of factors. First, it is a rare event for 12 univariate normal proposal kernels to suggest all positive numbers when the variance for the each proposal kernel is 1 and the starting value, X_0, for each coordinate ranges from 0.6 to 2. Also, X_0 is the estimates from a previous MCMC run, so the evaluation of X_0 under the target density would be greater than most of the other points in the state space. Thus, accepting a new proposal, a point in the state space, over X_0 does not exactly have a high probability of happening unless the move is really small. The BAM algorithm overcomes this problem as it adjusts the proposal variance to suit for the target distribution of interest.

If we reduce down the scale of our proposal kernel for non-adaptive algorithms to $N(X_n, 0.01I_d)$ or $N(X_n, 0.001I_d)$, then the proposals are accepted more often. However, the mixing of the chains for these non-adaptive algorithms are still clearly not as good as for BAM. This indicates that our new Bounded Adaption Metropolis (BAM) adaptive MCMC algorithm performs better than standard Metropolis algorithms, even if their proposal scalings are adjusted manually to allow for reasonable acceptance rates.

It is our hope that the easily-verifiable ergodicity conditions presented herein will allow MCMC practitioners to make more widespread use of such adaptive MCMC algorithms, and thus benefit from their computational speedups without suffering from burdensome or uncheckable technical conditions.

Acknowledgements. We thank the anonymous reviewer for very helpful comments, which led to many improvements.

References

[1] C. Andrieu and Y. F. Atchadé. On the efficiency of adaptive MCMC algorithms. *Electronic Communications in Probability*, 12(33):336–349, 2007.

[2] C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.

[3] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18:343–373, 2008.

[4] Y. F. Atchadé and J. S. Rosenthal. On adaptive Markov Chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

[5] Y. Bai, G. O. Roberts, and J. S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances and Applications in Statistics*, 21(1):1–54, 2011.

[6] S. Brooks, A. Gelman, G. L. Jones, and X. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Taylor & Francis, 2011.

[7] R. V. Craiu, L. Gray, K. Latuszynski, N. Madras, G. O. Roberts, and J. S. Rosenthal. Stability of adversarial markov chains, with an application to adaptive mcmc algorithms. *Annals of Applied Probability*, 25(6):3592–3623, 2015.

[8] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289, 2011.

[9] D. P. Gaver and I. G. O'Muircheartaigh. Robust empirical Bayes analyses of event rates. *Technometrics*, 29(1):1–15, 1987.

[10] E. I. George, U. E. Makov, and A. F. M. Smith. Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, 20(2):147–156, 1993.

[11] P. Giordani and R. Kohn. Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics*, 19(2):243–259, 2010.

[12] H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

[13] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

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

[15] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[16] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[17] G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16(4):2123–2139, 2006.

[18] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475, 2007.

[19] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[20] J. S. Rosenthal. Adaptive MCMC Java applet. 2004. URL: `http://probability.ca/jeff/java/adapt.html`.

[21] W. Rudin. *Principles of mathematical analysis.* McGraw-Hill New York, 3rd edition, 1976.

[22] L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[23] E. Turro, N. Bochkina, A. M. K. Hein, and S. Richardson. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC bioinformatics*, 8(1):439–448, 2007.

[24] M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008, 2012.

[25] J. Yang. *Convergence and efficiency of adaptive MCMC.* PhD thesis, Department of Statistical Sciences, University of Toronto, 2016. Unpublished thesis.