Complexity Bounds for MCMC via Diffusion Limits

by

Gareth O. Roberts[1] and Jeffrey S. Rosenthal[2]

(August 2014; revised March 2015 and May 2015.)

Abstract. We connect known results about diffusion limits of Markov chain Monte Carlo (MCMC) algorithms to the computer science notion of algorithm complexity. Our main result states that any weak limit of a Markov process implies a corresponding complexity bound (in an appropriate metric). We then combine this result with previously-known MCMC diffusion limit results to prove that under appropriate assumptions, the Random-Walk Metropolis (RWM) algorithm in d dimensions takes $O(d)$ iterations to converge to stationarity, while the Metropolis-Adjusted Langevin Algorithm (MALA) takes $O(d^{1/3})$ iterations to converge to stationarity.

1. Introduction.

In the computer science literature, algorithms are often analysed in terms of "complexity" bounds. In the Markov chain Monte Carlo (MCMC) literature, algorithms are sometimes understood in terms of diffusion limits. The purpose of this note is to connect these two approaches, and in particular to show that diffusion limits (and other weak limits) can imply algorithm complexity bounds.

Complexity results in computer science go back at least to Cobham (1964), and took on greater focus with the pioneering *NP-complete* work of Cook (1971). In the Markov chain context, computer scientists have been bounding convergence times of Markov chain algorithms since at least Jerrum and Sinclair (1989), focusing largely on spectral gap bounds for Markov chains on finite state spaces. More recently, attention has turned to bounding spectral gaps of modern Markov chain algorithms on general (e.g. uncountable) state spaces, again primarily via spectral gaps (e.g. Woodard et al., 2009a, 2009b). These bounds often focus on the order of the convergence time in terms of some particular parameter, such as the dimension d of the corresponding state space.

Meanwhile, in statistics, MCMC algorithms are extremely widely used and studied (see e.g. Brooks et al., 2011, and the many references therein), and their running times are an extremely important practical issue. They have been studied from a variety of perspectives,

[1]Department of Statistics, University of Warwick, CV4 7AL, Coventry, U.K. Email: `g.o.roberts@lancaster.ac.uk`. Supported in part by EPSRC grants EP/20620/01 and EP/S61577/01.

[2]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.

including directly bounding the convergence in total variation distance (see e.g. Rosenthal, 1995b, 1996, 2002; Jones and Hobert, 2001, 2004; and references therein), convergence "diagnostics" via statistical analysis of the Markov chain output (e.g. Gelman and Rubin, 1992), and most notably by proving weak convergence limits of sped-up versions of the algorithms to diffusion limits (e.g. Roberts et al., 1997; Roberts and Rosenthal, 1998).

Direct total variation bounds for MCMC are sometimes presented in terms of the convergence order (e.g. see Rosenthal, 1995a, for order bounds for a Gibbs sampler for a variance components model). And, the MCMC diffusion limits often involve speeding up the original algorithm by a certain order, and then proving weak convergence to a fixed process which converges in $O(1)$ iterations, thus giving them the flavour of complexity order bounds too. However, these MCMC results are typically not stated precisely in terms of convergence time complexity results, and (perhaps because of this) they are often overlooked by the computer science complexity community.

In this paper, we attempt to connect these two streams of Markov chain convergence time bounds. In particular, we establish (Theorem 1) that results about weak limits do directly imply corresponding complexity bounds (using an appropriate convergence metric as described below). We then apply our theorem to previous results about diffusion limits of MCMC algorithms (Section 3), to establish running time complexity order bounds for such MCMC algorithms as the Random-Walk Metropolis algorithm (Theorem 2) and the Metropolis-adjusted Langevin algorithm (Theorem 3).

2. Assumptions and Main Result.

In this section, we state our general result about obtaining convergence complexity bounds from weak limits. To set it up, let $(\mathcal{X}, \mathcal{F}, \rho)$ be a general measurable metric space, i.e. a non-empty (and possibly uncountable) set $\mathcal{X}$ endowed with a metric ρ which induces a Borel σ-algebra $\mathcal{F}$ of measurable subsets. We wish to bound the convergence of a stochastic process $\{X_t\}$ on $(\mathcal{X}, \mathcal{F})$ to its stationary probability distribution π. To measure the distance to stationarity, on finite state spaces one often (see e.g. Aldous and Fill, 2002, Section 2.4.1) uses the total variation distance defined by

$$\|\mathcal{L}_x(X_t) - \pi\|_{TV} \ := \ \sup_{|f|\leq 1} \Big|\mathbf{E}_x[f(X_t)] - \pi(f)\Big|$$

where the supremum is taken over all measurable functions $f : \mathcal{X} \to \mathbf{R}$ with $|f(x)| \leq 1$ for all $x \in \mathcal{X}$. Here $\mathcal{L}_x(X_t)$ is the law of X_t conditional on starting at $X_0 = x$, and $\mathbf{E}_x[f(X_t)]$ is the expected value of f with respect to this law, and $\pi(f) = \int f(x)\, \pi(dx)$ is the expected value of f with respect to π.

This total variation distance can also be used on general state spaces in many instances (see e.g. Rosenthal, 1995). However, it is not appropriate for bounding the weak convergence which arises in the diffusion context, since it may not go to zero for processes which converge only weakly to stationarity, so we do not use it here. Instead, we let

$$\mathrm{Lip}_1^1 \;=\; \Big\{f : \mathcal{X} \to \mathbf{R},\ |f(x) - f(y)| \le \rho(x, y)\ \forall x, y \in \mathcal{X},\ |f| \le 1\Big\}$$

be the set of all functions from $\mathcal{X}$ to $\mathbf{R}$ with Lipschitz constant ≤ 1 *and* with $|f(x)| \le 1$ for all $x \in \mathcal{X}$, and use the distance function

$$\|\mathcal{L}_x(X_t) - \pi\|_{KR} \;:=\; \sup_{f \in \mathrm{Lip}_1^1} \Big|\mathbf{E}_x[f(X_t)] - \pi(f)\Big|\,.$$

(Here "KR" stands for "Kantorovich-Rubinstein"; see the proof of Proposition 6 below.) The distance $\|\cdots\|_{KR}$ is similar to, but more restrictive than, the total variation distance, and as discussed below (see Proposition 6), it metrises weak convergence and so is appropriate for our purposes.

We also note that many approaches to stationary instead directly bound the spectral gap of the corresponding Markov operator (e.g. Woodard et al., 2009b). However, on general state spaces, the spectral gap is zero for Markov chains which are not "geometrically ergodic" (see e.g. Theorem 2 of Roberts and Rosenthal, 1997), even if they do converge to stationarity. Furthermore, many MCMC algorithms are not geometrically ergodic (e.g. the Random-Walk Metropolis algorithm on target distributions with heavier-than-exponential tails, see Theorem 3.3 of Mengersen and Tweedie, 1996). They also are often not reversible, which makes spectral gaps harder to study or interpret. For these reasons, we do not wish to restrict attention to spectral gaps, which is another reason that we use the metric $\|\cdots\|_{KR}$.

A related issue is what initial states X_0 should be considered. On finite state spaces, one often (e.g. Jerrum and Sinclair, 1989, Section 2) considers the worst case, by taking a supremum over all initial states x, i.e. using something like $\sup_{x\in\mathcal{X}} \|\mathcal{L}_x(X_t) - \pi\|_{TV}$. But this supremum is also frequently inappropriate on general state spaces. For instance, if $\mathcal{X}$ is unbounded, then as t increases one can start from worse and worse states X_0 so that the supremum might never go to 0. Instead, we need to specify more precisely which initial state(s) X_0 to consider. As a concrete choice, we will take the π-average of the distances to stationarity from all initial states X_0 in $\mathcal{X}$. That is, for any Markov chain $\{X_t\}$ on $(\mathcal{X}, \mathcal{F})$ with stationary distribution π, we measure the distance to stationarity at time t by the distance function

$$\mathbf{E}_{X_0\sim\pi}\|\mathcal{L}_{X_0}(X_t) - \pi\|_{KR} \;:=\; \int_{x\in\mathcal{X}} \pi(dx)\ \|\mathcal{L}_x(X_t) - \pi\|_{KR}\,.$$

Using this distance function, we can state our main result, about bounding convergence to stationarity using weak convergence of a sequence of processes to another fixed process. To avoid technicalities, we assume that this limiting process is *càdlàg*, i.e. has sample paths which are continuous on the right with limits on the left. (In our MCMC examples, the limiting process will in fact be a diffusion with *continuous* sample paths, so this is not a problem.)

Theorem 1. *Let $X^{(d)} = \{X_t^{(d)}\}_{t\geq 0}$ be a stochastic process on $(\mathcal{X}, \mathcal{F}, \rho)$, for each $d \in \mathbf{N}$, which converges weakly in the Skorokhod topology as $d \to \infty$ to a càdlàg process $X^{(\infty)} = \{X_t^{(\infty)}\}_{t\geq 0}$, i.e. $X^{(d)} \Rightarrow X^{(\infty)}$. Assume these processes all have the same stationary probability distribution π, and that $X^{(\infty)}$ converges (either weakly or in total variation distance) to π. Then for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that*

$$\mathbf{E}_{X_0^{(d)}\sim\pi} \|\mathcal{L}_{X_0^{(d)}}(X_t^{(d)}) - \pi\|_{KR} \; < \; \epsilon, \quad t \geq T, \quad d \geq D\,.$$

Theorem 1 may be summarised as saying that if a sequence $\{X^{(d)}\}$ of Markov processes converges weakly to a limiting ergodic process, then we can bound the convergence of the sequence of processes uniformly over all sufficiently large d, i.e. the processes converge in $O(1)$ iterations with respect to d. We will next apply this result to previously known diffusion limits of common MCMC algorithms.

3. Application to MCMC.

Our primarily interest is in the use of Theorem 1 to bound the complexity of MCMC algorithms. We begin with the most popular MCMC algorithm, the Random-Walk Metropolis (RWM) algorithm. This algorithm proceeds, given a positive target probability density π_d on the state space $\mathbf{R}^d$, by running a Markov chain $\{\mathbf{Z}_n^d\}_{n=0}^\infty$ as follows. Given the value $\mathbf{Z}_n^d$, a proposed new state $\mathbf{Y}_{n+1}^d \sim MVN(\mathbf{Z}_n^d,\, \sigma_d^2)$ is chosen from a multivariate normal distribution centered at $\mathbf{Z}_n^d$, and then with probability $\min[1,\; \pi_d(\mathbf{Y}_{n+1}^d)/\pi_d(\mathbf{Z}_n^d)]$ the proposal is accepted and $\mathbf{Z}_{n+1}^d = \mathbf{Y}_{n+1}^d$, otherwise with the remaining probability the proposal is rejected and $\mathbf{Z}_{n+1}^d = \mathbf{Z}_n^d$. This algorithm is easily seen to be irreducible and aperiodic and to leave π_d stationary, so it will converge asymptotically to π_d. The question then becomes how quickly it will converge, and what choice of proposal variance σ_d^2 is optimal.

In this context, Roberts et al. (1997) proved the result that $U^d \Rightarrow U$ as $d \to \infty$, where $U_t^d = \mathbf{Z}_{\lfloor dt \rfloor,\,1}^d$ is the first coordinate of the RWM algorithm sped up by a factor of d, and U is a limiting ergodic Langevin diffusion, and $\Rightarrow$ indicates weak convergence in the usual Skorokhod topology. They proved this result under certain strong technical assumptions, namely that π_d takes on the special product form $\pi_d(\mathbf{x}) = \prod_{i=1}^{d} h(x_i)$ for some fixed function $h : \mathbf{R} \to (0, \infty)$ with h'/h Lipschitz continuous, and (A1) $\int [h'(x)/h(x)]^8 h(x)dx < \infty$, and (A2) $\int [h''(x)/h(x)]^4 h(x)dx < \infty$. They also assumed the processes $\mathbf{Z}^d$ are in stationarity, and that $\sigma_d^2 = \ell^2/(d-1)$ for some fixed $\ell > 0$.

This theorem of Roberts et al. (1997) allowed them to study the limiting diffusion U as a function of the proposal variance parameter ℓ, and optimise it to prove that the algorithm converges fastest when its asymptotic acceptance rate is equal to 0.234... (see also Roberts and Rosenthal, 2001). Furthermore, since their process U^d involved speeding up the original algorithm by a factor of d, their results seemed to imply that RWM required $O(d)$ iterations to converge. However, a precise statement of such a complexity bound was not provided.

In light of Theorem 1 above, we are now able to use the diffusion limit of Roberts et al. (1997) to give an actual complexity bound on the RWM algorithm. We need one slight technical extension, namely to replace (A1) and (A2) above by the slightly stronger conditions (A1*) $\int [h'(x)/h(x)]^{12} h(x)dx < \infty$ and (A2*) $\int [h''(x)/h(x)]^6 h(x)dx < \infty$. We then have the following result, proved in Section 5 below.

Theorem 2. *Let $Z^{(d)}$ be a RWM algorithm satisfying the above technical assumptions of Roberts et al. (1997), except with (A1) and (A2) replaced by (A1*) and (A2*). Then for any $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that*

$$\mathbf{E}_{Z_{0,1}^{(d)} \sim h} \| \mathcal{L}_{Z_{0,1}^{(d)}}(Z_{\lfloor dt \rfloor,1}^{(d)}) - h \|_{KR} \; < \; \epsilon \,, \quad t \geq T, \quad d \geq D \,.$$

(Here $\mathcal{L}_{Z_{0,1}^{(d)}}(Z_{\lfloor dt \rfloor,1}^{(d)})$ represents the probability distribution of the first coordinate of $Z^{(d)}$ at iteration number equal to the greatest integer not exceeding dt, conditional on the process starting with its first coordinate equal to the specified state $Z_{0,1}^{(d)}$, and with all other coordinates of $Z_0^{(d)}$ chosen independently according to the density h.) Hence, the RWM algorithm takes $O(d)$ iterations to converge to within ϵ of stationarity in its first (or any one) coordinate.

We believe this to be the first precise general result about the complexity order of the RWM algorithm. It does require strong technical assumptions, but it still applies to a fairly general collection of densities on $\mathbf{R}^d$. Furthermore, empirical studies (see e.g. Roberts and Rosenthal, 2001) indicate that even when RWM algorithms do not satisfy the technical assumptions, they still exhibit similar limiting behaviour.

Another MCMC diffusion limit concerns the Metropolis-Adjusted Langevin Algorithm (MALA). This algorithm is similar to the above Random-Walk Metropolis algorithm, except that now the proposal state $\mathbf{Y}_{n+1}^d \sim MVN(\mathbf{Z}_n^d + \frac{1}{2}\sigma_d^2 \nabla \log \pi_d(Z_n^d),\, \sigma_d^2)$ is chosen from a multivariate normal distribution centered at $\mathbf{Z}_n^d + \frac{1}{2}\sigma_d^2 \nabla \log \pi_d(Z_n^d)$ (to better approximate π_d), and the above acceptance probability is modified by the ratio of the corresponding proposal normal distributions. In this context, Roberts and Rosenthal (1999) proved that $U^d \Rightarrow U$, where $U_t^d = \mathbf{Z}^d_{\lfloor d^{1/3}t \rfloor, 1}$ is the first coordinate of the MALA algorithm sped up by a factor of $d^{1/3}$, and U is again a limiting ergodic Langevin diffusion. This result again required strong technical assumptions, this time that $\pi_d(\mathbf{x}) = \prod_{i=1}^d h(x_i)$ for some fixed function $h : \mathbf{R} \to (0, \infty)$ with polynomially-bounded log-derivatives of all orders, and finite moments of all orders, with h'/h Lipschitz continuous. They also assumed the processes $\mathbf{Z}^d$ are in stationarity, and that $\sigma_d^2 = \ell^2 \, d^{-1/3}$ for some fixed $\ell > 0$.

This theorem of Roberts and Rosenthal (1999) allowed them to optimise the limiting diffusion U as a function of ℓ, and to prove that the algorithm converges fastest when its asymptotic acceptance rate is equal to $0.574\ldots$. Also, since their process U^d involved speeding up the original algorithm by a factor of $d^{1/3}$, their results seemed to imply that MALA required $O(d^{1/3})$ iterations to converge. Once again, we can use Theorem 1 above to obtain the following more formal complexity bound (proved in Section 5 below).

Theorem 3. *Let $Z^{(d)}$ be a MALA algorithm on a product density in d dimensions satisfying the above technical assumptions of Roberts and Rosenthal (1999). Then for any $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that (with notation as in Theorem 2)*

$$\mathbf{E}_{Z_{0,1}^{(d)} \sim h} \|\mathcal{L}_{Z_{0,1}^{(d)}}(Z^{(d)}_{\lfloor d^{1/3}t \rfloor, 1}) - h\|_{KR} \;<\; \epsilon\,, \quad t \geq T, \quad d \geq D\,.$$

Hence, the MALA algorithm takes $O(d^{1/3})$ iterations to converge to within ϵ of stationarity in its first (or any one) coordinate.

Finally, we note that a number of other diffusion limits have been proven for MCMC algorithms in other contexts. For example, Bédard (2007, 2008) and Sherlock and Roberts (2009) have extended the original RWM diffusion limit to more general target distributions; Roberts (1998) and Neal and Roberts (2006, 2008, 2011) and Jourdain et al. (2013a, 2013b) have extended it to other related cases; and Neal et al. (2012) have established diffusion limits for RWM algorithms on discontinuous target densities. Each of these diffusion limit results could also be combined with Theorem 1 above to yield complexity order bounds in new contexts.

4. Proof of Theorem 1.

In this section, we prove Theorem 1. Along the way, we establish that $\| \cdots \|_{KR}$ metrises weak convergence (Proposition 6), and that $\mathbf{E}_{X_0 \sim \pi} \| \mathcal{L}_{X_0}(X_t^{(d)}) - \pi \|_{KR}$ is a non-increasing function of t (Lemma 11). We first establish that $\| \cdots \|_{KR}$ is a norm:

Lemma 4. *Let S be any non-empty collection of functionals $\mathcal{X} \to \mathbf{R}$ which is symmetric (i.e. if $f \in S$ then $-f \in S$). Let $\|\mu\| = \sup_{f \in S} \mu(f)$. Then $\| \ldots \|$ is a (possibly infinite) norm function on the set of all signed measures on $(\mathcal{X}, \mathcal{F})$. In particular, $\| \cdots \|_{KR}$ is a norm.*

Proof. It is immediate that $\|0\| = 0$, and that $\|a\,\mu\| = a\,\|\mu\|$ for $a > 0$. The symmetry of S implies that $\| -\mu\| = \|\mu\|$. Finally, for the triangle inequality, we check that

$$\|\mu + \nu\| \;=\; \sup_{f \in S} \Big(\mu(f) + \nu(f) \Big) \;\leq\; \Big(\sup_{f \in S} \mu(f) \Big) + \Big(\sup_{f \in S} \nu(f) \Big) \;=\; \|\mu\| + \|\nu\| \,.$$

Hence, $\| \ldots \|$ is a norm. The claim about $\| \cdots \|_{KR}$ then follows by taking $S = \mathrm{Lip}_1^1$. ∎

We next show that truncating the metric ρ does not change Lip_1^1:

Lemma 5. *Let $\rho^* = \min(2, \rho)$. Then*

$$\mathrm{Lip}_1^1 \;=\; \{ f : \mathcal{X} \to \mathbf{R},\; |f(x) - f(y)| \leq \rho^*(x, y)\; \forall x, y \in \mathcal{X},\; |f| \leq 1 \} \,.$$

Proof. This is immediate since we always have $|f(x) - f(y)| \leq 2$ for $f \in \mathrm{Lip}_1^1$. ∎

Proposition 6. *The metric $\Delta(\mu, \nu) := \|\mu - \nu\|_{KR}$ metrises weak convergence of probability measures on $(\mathcal{X}, \mathcal{F}, \rho)$. That is, if $\{\mu_t\}$ and μ are probability measures on $(\mathcal{X}, \mathcal{F}, \rho)$, then $\{\mu_t\} \Rightarrow \mu$ if and only if $\lim_{t \to \infty} \Delta(\mu_t, \mu) = 0$.*

Proof. Let ρ^* be as in Lemma 5. We first note that since ρ and ρ^* agree for distances ≤ 2, they give rise to precisely the same open subsets. Therefore, $(\mathcal{X}, \rho^*)$ induces the same Borel σ-algebra $\mathcal{F}$ that $(\mathcal{X}, \rho)$ does, and thus gives rise to the same Skorokhod topology. Hence, weak convergence on $(\mathcal{X}, \mathcal{F}, \rho)$ is precisely equivalent to weak convergence on $(\mathcal{X}, \mathcal{F}, \rho^*)$. Furthermore, by Lemma 5, the metric $\| \cdots \|_{KR}$ is the same on $(\mathcal{X}, \mathcal{F}, \rho^*)$ as on $(\mathcal{X}, \mathcal{F}, \rho)$. Hence, it suffices to prove the result on the truncated space $(\mathcal{X}, \mathcal{F}, \rho^*)$.

Now, since $(\mathcal{X}, \mathcal{F}, \rho^*)$ is a bounded metric space, it is known (see e.g. Givens and Shortt, 1984, Proposition 4) that weak convergence on $(\mathcal{X}, \mathcal{F}, \rho^*)$ is metrised by the Wasserstein metric W_1 on $(\mathcal{X}, \rho^*)$, defined by

$$W_1(\mu, \nu) \ := \ \inf \mathbf{E}[\rho^*(X, Y)]$$

where the infimum is taken over all pairs (X, Y) of random variables on $(\mathcal{X}, \mathcal{F})$ such that $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. On the other hand, again since $(\mathcal{X}, \mathcal{F}, \rho^*)$ is a bounded metric space, it is known (Kantorovich and Rubinstein, 1958; see e.g. Givens and Shortt, 1984, p. 233) that for probability measures μ and ν on $(\mathcal{X}, \mathcal{F}, \rho^*)$, the Wasserstein metric $W_1(\mu, \nu)$ is precisely equal to $\|\mu - \nu\|_{KR}$. Combining these two facts, the result follows for $(\mathcal{X}, \mathcal{F}, \rho^*)$, and hence also for $(\mathcal{X}, \mathcal{F}, \rho)$. ■

Lemma 7. *If $X^{(\infty)}$ converges to π, either weakly or in total variation distance, then for all $x \in \mathcal{X}$ and $\epsilon > 0$ there is $T < \infty$ such that $\|\mathcal{L}_x(X_T^{(\infty)}) - \pi\|_{KR} \leq \epsilon/2$ for all $t \geq T$.*

Proof. If the convergence is weak, then this follows from Proposition 6. If the convergence is in total variation distance, then this still follows since $\| \ldots \|_{KR} \leq \| \ldots \|_{TV}$. ■

We are now in a position to prove convergence of $X_T^{(d)}$ for certain *fixed* times T:

Proposition 8. *Under the assumptions of Theorem 1, for any $x \in \mathcal{X}$ and $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that*

$$\|\mathcal{L}_x(X_T^{(d)}) - \pi\|_{KR} \ < \ \epsilon\,, \quad d \geq D\,.$$

Proof. Using Lemma 4, we have by the triangle inequality that

$$\|\mathcal{L}_x(X_t^{(d)}) - \pi\|_{KR} \ \leq \ \|\mathcal{L}_x(X_t^{(d)}) - \mathcal{L}_x(X_t^{(\infty)})\|_{KR} \ + \ \|\mathcal{L}_x(X_t^{(\infty)}) - \pi\|_{KR}\,. \tag{1}$$

To continue, we recall that since $X^{(d)}$ converges weakly to $X^{(\infty)}$, it follows that $X_t^{(d)}$ converges weakly to $X_t^{(\infty)}$ for all fixed times $t > 0$ such that $X^{(\infty)}$ has probability 0 of *jumping* at time t, i.e. for all but at most a countable number of times t (since $X^{(\infty)}$ is càdlàg). By Lemma 7, there is $T < \infty$ such that $\|\mathcal{L}_x(X_T^{(\infty)}) - \pi\|_{KR} \leq \epsilon/2$, and by increasing T as necessary we can assume that $X^{(\infty)}$ has probability 0 of jumping at time T. Then $X_T^{(d)}$ converges weakly to $X_T^{(\infty)}$, so by Proposition 6 there is $D < \infty$ such that for all $d \geq D$, $\|\mathcal{L}_x(X_T^{(d)}) - \mathcal{L}_x(X_T^{(\infty)})\|_{KR} < \epsilon/2$. The result then follows from (1). ■

Remark 9. If the weak convergence of $X^{(d)}$ to $X^{(\infty)}$ is assumed to be uniform over bounded time intervals, then we can strengthen Proposition 8 to say that for any $x \in \mathcal{X}$ and $\epsilon > 0$ and $S < \infty$, there are $D < \infty$ and $T < \infty$ such that $\|\mathcal{L}_x(X_t^{(d)}) - \pi\|_{KR} < \epsilon$ for all $t \in [T, T+S]$.

Corollary 10. *Under the assumptions of Theorem 1, for any $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that*

$$\mathbf{E}_{X_0 \sim \pi} \|\mathcal{L}_{X_0}(X_T^{(d)}) - \pi\|_{KR} \ < \ \epsilon \,, \quad d \geq D \,.$$

Proof. We first let

$$A_m \ = \ \{x \in \mathcal{X} : \|\mathcal{L}_x(X_t^{(\infty)}) - \pi\|_{KR} \ < \ \epsilon/4 \ \ \forall t \geq m\} \,.$$

Then $A_{m+1} \supseteq A_m$ by inspection, and $\bigcup_m A_m = \mathcal{X}$ by Lemma 7. Hence, by continuity of probabilities (see e.g. Proposition 3.3.1 of Rosenthal, 2000), $\lim_{m \to \infty} \pi(A_m) = 1$. We can therefore find $T < \infty$ such that $\pi(A_T) \geq 1 - (\epsilon/8)$. As in the proof of Proposition 8, by increasing T as necessary we can assume that $X^{(\infty)}$ has probability 0 of jumping at time T.

Next, for this fixed T, let

$$B_m \ = \ \{x \in \mathcal{X} : \|\mathcal{L}_x(X_T^{(d)}) - \mathcal{L}_x(X_T^{(\infty)})\|_{KR} \ < \ \epsilon/4 \ \ \forall d \geq m\} \,.$$

Then $B_{m+1} \supseteq B_m$ by inspection, and $\bigcup_m B_m = \mathcal{X}$ since $X_T^{(d)} \Rightarrow X_T^{(\infty)}$, so again by continuity of probabilities we can find $D \in \mathbf{N}$ such that $\pi(B_D) \geq 1 - (\epsilon/8)$.

We then compute that for this fixed T and D, and for any $d \geq D$,

$$\begin{aligned} &\mathbf{E}_{X_0 \sim \pi} \|\mathcal{L}_{X_0}(X_T^{(d)}) - \pi\|_{KR} \\ &= \ \mathbf{E}_{X_0 \sim \pi} \Big(\mathbf{1}_{X_0 \in A_T \cap B_D} \, \|\mathcal{L}_{X_0}(X_T^{(d)}) - \pi\|_{KR} \Big) \ + \ \mathbf{E}_{X_0 \sim \pi} \Big(\mathbf{1}_{X_0 \notin A_T \cap B_D} \, \|\mathcal{L}_{X_0}(X_T^{(d)}) - \pi\|_{KR} \Big) \\ &\qquad \leq \ [(\epsilon/4) + (\epsilon/4)] + [(\epsilon/8) + (\epsilon/8)] \times 2 \ = \ \epsilon \,, \end{aligned}$$

where for the first term we have used the triangle inequality, and for the second term we have used the fact that by definition we always have $\|\mathcal{L}_x(X_T^{(d)}) - \pi\|_{KR} \leq 2$ for any x and d. This gives the result. ∎

Corollary 10 is nearly what we need to prove Theorem 1. However, for Theorem 1 we want the convergence to be within ϵ for *all* $t \geq T$, not just for one fixed T (nor just for all t in some bounded time interval, as in Remark 9). Unfortunately, $\|\mathcal{L}_x(X_t^{(d)}) - \pi\|_{KR}$ might not be a non-increasing function of t (though $\|\mathcal{L}_x(X_t^{(d)}) - \pi\|_{TV}$ always is, see e.g. Proposition 3(c) of Roberts and Rosenthal, 2004). On the other hand, fortunately the quantity $\mathbf{E}_{X_0 \sim \pi} \|\mathcal{L}_{X_0}(X_t^{(d)}) - \pi\|_{KR}$ is indeed non-increasing:

Lemma 11. *Let $\|\ldots\|$ be any norm function on signed measures on $(\mathcal{X}, \mathcal{F})$. Let $P^t(x, \cdot)$ be the transition probabilities for a Markov chain on $(\mathcal{X}, \mathcal{F})$ with stationary probability distribution π. Let $\operatorname{dist}(t) = \mathbf{E}_{X_0 \sim \pi} \|P^t(X_0, \cdot) - \pi\|$. Then $\operatorname{dist}(t)$ is a non-increasing function of t. In particular, in the context of Theorem 1, $\mathbf{E}_{X_0 \sim \pi} \|\mathcal{L}_{X_0}(X_t^{(d)}) - \pi\|_{KR}$ is a non-increasing function of t.*

Proof. We compute by stationarity that for $s, t > 0$,

$$
\begin{aligned}
\operatorname{dist}(s+t) &= \mathbf{E}_{X_0 \sim \pi} \|P^{s+t}(X_0, \cdot) - \pi\| \\
&= \mathbf{E}_{X_0 \sim \pi} \left\| \int_{y \in \mathcal{X}} P^s(X_0, dy) \; P^t(y, \cdot) - \pi \right\| \\
&\leq \mathbf{E}_{X_0 \sim \pi} \int_{y \in \mathcal{X}} P^s(X_0, dy) \left\| P^t(y, \cdot) - \pi \right\| \\
&= \mathbf{E}_{Y_0 \sim \pi} \|P^t(Y_0, \cdot) - \pi\| \;=\; \operatorname{dist}(t)\,,
\end{aligned}
$$

thus proving the first claim. The claim about $\mathbf{E}_{x \sim \pi} \|\mathcal{L}_x(X_t^{(d)}) - \pi\|_{KR}$ then follows by Lemma 4 upon setting $P^t(x, A) = \mathbf{P}[X_t^{(d)} \in A \,|\, X_0^{(d)} = x]$. ∎

Theorem 1 then follows by combining Corollary 10 and Lemma 11.

5. Proofs of Theorems 2 and 3.

Theorems 2 and 3 *nearly* follow immediately by applying Theorem 1 to the diffusion limit results of Roberts et al. (1997) and of Roberts and Rosenthal (1999), respectively. However, there is one technical issue. The previous diffusion limit results assume that the process begins in the stationary distribution. By contrast, Theorem 1 involves $\mathcal{L}_x(X_t)$, i.e. *conditioning* on the stochastic processes' first coordinate beginning at a specific state $x \in \mathcal{X}$. So, to prove Theorems 2 and 3, we need to establish that the diffusion limit results remain valid even upon conditioning on the starting value of the processes.

For the RWM algorithm, this does indeed follow, at least upon strengthening (A1) and (A2) to (A1*) and (A2*) as above:

Proposition 12. *Let $Z^{(d)}$ be a RWM algorithm satisfying the above technical assumptions of Roberts et al. (1997), except with (A1) and (A2) replaced by (A1*) and (A2*). Then for π-a.e. $x \in \mathcal{X}$, ${}_xU^d \Rightarrow {}_xU$ as $d \to \infty$, where ${}_xU_t^d = (\mathbf{Z}^d_{\lfloor dt \rfloor, 1} \,|\, \mathbf{Z}^d_{0,1} = x)$ is the first coordinate of the RWM algorithm sped up by a factor of d, conditional on starting at the state x, and ${}_xU$ is the limiting ergodic Langevin diffusion U also conditional on starting at x.*

Proof. The proof is very similar to the proof of the unconditioned diffusion limit theorem of Roberts et al. (1997), using generators and cores (cf. Ethier and Kurtz, 1986). The only point of departure between our proof and theirs concerns their Lemma 2.1, which states that for each fixed $t > 0$,

$$\lim_{d\to\infty} \mathbf{P}[Z_s^{(d)} \in F_d \text{ for all } 0 \le s \le t] \;=\; 1\,,$$

where F_d is the event that both $|A_d| < d^{-1/8}$ and $|B_d| < d^{-1/8}$, where

$$A_d := \frac{1}{d-1}\sum_{i=2}^{d}\Big(((\log h(x_i))')^2 - \mathbf{E}_{x_i\sim h}[((\log h(x_i))')^2]\Big)$$

and

$$B_d := \frac{1}{d-1}\sum_{i=2}^{d}\Big((\log h(x_i))'' - \mathbf{E}_{x_i\sim h}[(\log h(x_i))'']\Big)\,.$$

To complete our proof, we need to show that this statement remains valid, even when conditioning on starting at a specific state $x \in \mathcal{X}$. (In fact, it would suffice to replace $1/8$ by any other power $\alpha \in (0, 1/2)$, but we do not need to do that.)

To this end, fix $t > 0$, and let

$$p(d,x) \;=\; \mathbf{P}[Z_s^{(d)} \notin F_d \text{ for some } 0 \le s \le t \mid Z_{0,1}^{(d)} = x]\,,$$

and let $r(d) = \mathbf{E}_{x\sim h} p(d,x)$ be its expected value when averaged over x in stationarity. Also, let

$$v_j \;=\; \mathbf{E}_{x_i\sim h}\Big[\Big(((\log h(x_i))')^2 - \mathbf{E}_{x_i\sim h}[((\log h(x_i))')^2]\Big)^j\Big]$$

and recall that $|v_j| < \infty$ for $1 \le j \le 6$ by (A1*). Then, by expanding out the power $(A_d)^6$, while omitting terms involving v_1 since clearly $v_1 = 0$, we conclude that

$$\mathbf{E}_\pi[(A_d)^6] \;=\; (d-1)^{-6}\left[\binom{d-1}{1}v_6 + (d-1)(d-2)\binom{6}{2}v_2v_4 + \binom{d-1}{2}\binom{6}{3}v_3^2 + \binom{d-1}{3}\binom{6}{2,2,2}v_2^3\right]$$

$$=\; (d-1)^{-6}\Big[(d-1)v_6 + 15(d-1)(d-2)v_2v_4 + 10(d-1)(d-2)v_3^2 + 15(d-1)(d-2)(d-3)v_2^3\Big]\,.$$

In particular, $\mathbf{E}_\pi[(A_d)^6] = O(d^{-3})$ as $d \to \infty$. Hence, by Markov's inequality,

$$\mathbf{P}_\pi(|A_d| > d^{-1/8}) \;\le\; \mathbf{E}[(A_d)^6]/(d^{-1/8})^6 \;=\; O(d^{-3+(6/8)}) \;=\; O(d^{-9/4})\,.$$

Similarly, $\mathbf{P}_\pi(|B_d| > d^{-1/8}) \le O(d^{-9/4})$, so that also $\mathbf{P}[Z_s^{(d)} \notin F_d] \le O(d^{-9/4})$.

Next, we note that there are $O(dt)$ different RWM iterations corresponding to times s with $0 \le s \le t$. Hence, by subadditivity,

$$r(d) \;:=\; \mathbf{P}_\pi[Z_s^{(d)} \notin F_d \text{ for some } 0 \le s \le t] \;\le\; O(dt\, d^{-9/4}) \;=\; O(t\, d^{-5/4})\,.$$

In particular, $\sum_{d=2}^{\infty} r(d) < \infty$, which is the key.

Finally, we wish to show that $\lim_{d\to\infty} p(d,x) = 0$ with probability 1. To that end, let $\epsilon > 0$, and set $S_d = \{x \in \mathcal{X} : p(d,x) \geq \epsilon\}$. Then writing $\mathbf{P}_h(A)$ for $\int_A h(x)\,dx$, it follows by Markov's inequality that $\mathbf{P}_h(S_d) \leq \mathbf{E}_h[p(d,x)]/\epsilon = r(d)/\epsilon$. Hence, $\sum_{d=2}^{\infty} \mathbf{P}_h(S_d) \leq \sum_{d=2}^{\infty} r(d)/\epsilon < \infty$. Hence, by the Borel-Cantelli Lemma, $\mathbf{P}_h(S_d \text{ infinitely often}) = 0$, i.e. the set of x with an infinite sequence of d with $p(d,x) \geq \epsilon$ has probability zero. This means that with probability 1, $\limsup_{d\to\infty} p(d,x) < \epsilon$. Since this is true for all $\epsilon > 0$, it follows that with probability 1, $\lim p(d,x) = 0$, as desired, thus completing the proof. ∎

Theorem 2 then follows by combining Theorem 1 and Proposition 12.

Finally, we prove a similar result for the MALA algorithm. In this case, no strengthening of the assumptions is required:

Proposition 13. *Let $Z^{(d)}$ be a MALA algorithm on a product density in d dimensions satisfying the above technical assumptions of Roberts and Rosenthal (1999). Then for π-a.e. $x \in \mathcal{X}$, ${}_xU^d \Rightarrow {}_xU$ as $d \to \infty$, where ${}_xU_t^d = (\mathbf{Z}^d_{\lfloor d^{1/3}t\rfloor,1} \,|\, \mathbf{Z}^d_{0,1} = x)$ is the first coordinate of the RWM algorithm sped up by a factor of $d^{1/3}$, conditional on starting at x, and ${}_xU$ is the limiting ergodic Langevin diffusion U also conditional on starting at x.*

Proof. The proof involves modifying the weak convergence proof of Roberts and Rosenthal (1999), along lines very similar to that of Proposition 12, so we omit the details. Furthermore, since Roberts and Rosenthal (1999) assume finite moments of *all* polynomial orders, there is no need to strengthen any of their assumptions as was necessary for Proposition 12. ∎

Theorem 3 then follows by combining Theorem 1 and Proposition 13.

Acknowledgements. We thank Dawn Woodard and Alexandre Thiéry for helpful discussions of these matters, and thank the anonymous referee for several very valuable comments.

References

D. Aldous and J.A. Fill (2002), Reversible Markov Chains and Random Walks on Graphs. Unfinished monograph, available at: http://www.stat.berkeley.edu/~aldous/RWG/book.html

M. Bédard (2007), Weak Convergence of Metropolis Algorithms for Non-iid Target Distributions. Ann. Appl. Prob. **17**, 1222–1244.

M. Bédard (2008), Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234. Stoch. Proc. Appl. **118**, 2198–2222.

S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall / CRC Press.

A. Cobham (1964), The intrinsic computational difficulty of functions. In Proceedings of the 1964 International Congress for Logic, Methodology, and Philosophy of Science, Y. Bar-Hille, ed., Elsevier/North-Holland, Amsterdam, 24-30.

S. Cook (1971), The complexity of theorem-proving procedures. Third Annual ACM Symposium on Theory of Computing, 151-158.

S.N. Ethier and T.G. Kurtz (1986), Markov processes, characterization and convergence. Wiley, New York.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci. **7(4)**, 457–472.

C.R. Givens and R.M. Shortt (1984), A class of Wasserstein metrics for probability distributions. Michigan Math. J. **31(2)**, 231–240.

G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science **16**, 312–334.

G.L. Jones and J.P. Hobert (2004), Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. Ann. Stat. **32**, 784–817.

B. Jourdain, T. Lelièvre, and B. Miasojedow (2013a), Optimal scaling for the transient phase of Metropolis Hastings algorithms: the longtime behavior. Bernoulli, to appear.

B. Jourdain, T. Lelièvre, and B. Miasojedow (2013b), Optimal scaling for the transient phase of Metropolis Hastings algorithms: the mean-field limit. Preprint.

L. Kantorovich and G. Rubinstein (1958), On a space of completely additive functions. (In Russian.) Vestnik Leningrad. Univ. **13**, 52–59.

K.L. Mengersen and R.L. Tweedie (1996), Rates of convergence of the Hastings and Metropolis algorithms. Ann. Stat. **24**, 101–121.

P. Neal and G.O. Roberts (2006), Optimal Scaling for partially updating MCMC algorithms. Ann. Appl. Prob. **16(2)**, 474–515.

P. Neal and G.O. Roberts (2008), Optimal Scaling for Random Walk Metropolis on spherically constrained target densities. Meth. Comput. Appl. Prob. **10**, 277-297.

P. Neal and G.O. Roberts (2011), Optimal scaling of random walk Metropolis algorithms with non-Gaussian proposals. Meth. Comput. Appl. Prob. **13(3)**, 583–601.

P. Neal, G.O. Roberts, and W.K. Yuen (2012), Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. Ann. Appl. Prob. **22(5)**, 1880–1927.

G.O. Roberts (1998), Optimal Metropolis algorithms on the hypercube. Stochastics **62**, 275–284.

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 (1997), Geometric ergodicity and hybrid Markov chains. Elec. Comm. Prob. **2(2)**, 13–25.

G.O. Roberts and J.S. Rosenthal (1998), Optimal scaling of discrete approximations to Langevin diffusions. J. Roy. Stat. Soc. B **60**, 255–268.

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 (2004), General state space Markov chains and MCMC algorithms. Prob. Surv. **1**, 20–71.

J.S. Rosenthal (1995a), Rates of convergence for Gibbs sampler for variance components models. Ann. Stat. **23**, 740–761.

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

J.S. Rosenthal (1996), Convergence of Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.

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

J.S. Rosenthal (2002), Quantitative convergence rates of Markov chains: A simple account. Elec. Comm. Prob. **7**, No. 13, 123–128.

C. Sherlock and G.O. Roberts (2009), Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. Bernoulli **15(3)**, 774–798.

D.B. Woodard, S.C. Schmidler, M.L. Huber (2009a), Sufficient conditions for torpid mixing of parallel and simulated tempering. Elec. J. Prob. **14**, 780–804.

D.B. Woodard, S.C. Schmidler, M.L. Huber (2009b), Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. Ann. Appl. Prob. **19**, 617–640.