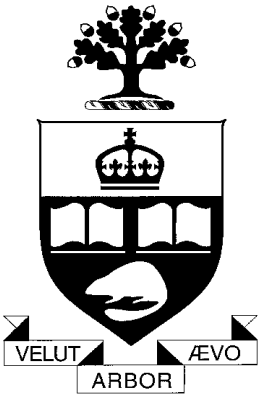

GENERAL STATE SPACE MARKOV CHAINS AND MCMC ALGORITHMS

by

Gareth O. Roberts
Department of Mathematics and Statistics
Lancaster University

and

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

Technical Report No. 0402 March 8, 2004

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

GENERAL STATE SPACE MARKOV CHAINS AND MCMC ALGORITHMS

by

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

(March 2004; revised August 2004)

Abstract. This paper surveys various results about Markov chains on general (non-countable) state spaces. It begins with an introduction to Markov chain Monte Carlo (MCMC) algorithms, which provide the motivation and context for the theory which follows. Then, sufficient conditions for geometric and uniform ergodicity are presented, along with quantitative bounds on the rate of convergence to stationarity. Many of these results are proved using direct coupling constructions based on minorisation and drift conditions. Necessary and sufficient conditions for Central Limit Theorems (CLTs) are also presented, in some cases proved via the Poisson Equation or direct regeneration constructions. Finally, optimal scaling and weak convergence results for Metropolis-Hastings algorithms are discussed. None of the results presented is new, though many of the proofs are. We also describe some Open Problems.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms – such as the Metropolis-Hastings algorithm ([52], [36]) and the Gibbs sampler (e.g. Geman and Geman [32]; Gelfand and Smith [30]) – have become extremely popular in statistics, as a way of approximately sampling from complicated probability distributions in high dimensions (see for example the reviews [92], [88], [33], [70]). Most dramatically, the existence of MCMC algorithms has transformed Bayesian inference, by allowing practitioners to sample from posterior distributions of complicated statistical models.

* 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.

In addition to their importance to applications in statistics and other subjects, these algorithms also raise numerous questions related to probability theory and the mathematics of Markov chains. In particular, MCMC algorithms involve Markov chains $\{X_n\}$ having a (complicated) stationary distribution $\pi(\cdot)$, for which it is important to understand as precisely as possible the nature and speed of the convergence of the law of X_n to $\pi(\cdot)$ as n increases.

This paper attempts to explain and summarise MCMC algorithms and the probability theory questions that they generate. After introducing the algorithms (Section 2), we discuss various important theoretical questions related to them. In Section 3 we present various convergence rate results commonly used in MCMC. Most of these are proved in Section 4, using direct coupling arguments and thereby avoiding many of the analytic technicalities of previous proofs. We consider MCMC central limit theorems in Section 5, and optimal scaling and weak convergence results in Section 6. Numerous references to the MCMC literature are given throughout. We also describe some Open Problems.

1.1. The problem.

The problem addressed by MCMC algorithms is the following. We're given a density function π_u, on some state space $\mathcal{X}$, which is possibly unnormalised but at least satisfies $0 < \int_{\mathcal{X}} \pi_u < \infty$. (Typically $\mathcal{X}$ is an open subset of $\mathbf{R}^d$, and the densities are taken with respect to Lebesgue measure, though other settings – including discrete state spaces – are also possible.) This density gives rise to a probability measure $\pi(\cdot)$ on $\mathcal{X}$, by

$$\pi(A) \;=\; \frac{\int_A \pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}\,. \tag{1}$$

We want to (say) estimate expectations of functions $f : \mathcal{X} \to \mathbf{R}$ with respect to $\pi(\cdot)$, i.e. we want to estimate

$$\pi(f) \;=\; \mathbf{E}_\pi[f(X)] \;=\; \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}\,. \tag{2}$$

If $\mathcal{X}$ is high-dimensional, and π_u is a complicated function, then direct integration (either analytic or numerical) of the integrals in (2) is infeasible.

The classical Monte Carlo solution to this problem is to simulate i.i.d. random variables $Z_1, Z_2, \ldots, Z_N \sim \pi(\cdot)$, and then estimate $\pi(f)$ by

$$\hat{\pi}(f) \;=\; (1/N) \sum_{i=1}^{N} f(Z_i)\,. \tag{3}$$

This gives an unbiased estimate, having standard deviation of order $O(1/\sqrt{N})$. Furthermore, if $\pi(f^2) < \infty$, then by the classical Central Limit Theorem, the error $\hat{\pi}(f) - \pi(f)$ will have a limiting normal distribution, which is also useful. The problem, however, is that if π_u is complicated, then it is very difficult to directly simulate i.i.d. random variables from $\pi(\cdot)$.

The Markov chain Monte Carlo (MCMC) solution is to instead construct a *Markov chain* on $\mathcal{X}$ which is easily run on a computer, and which has $\pi(\cdot)$ as a stationary distribution. That is, we want to define easily-simulated Markov chain transition probabilities $P(x, dy)$ for $x, y \in \mathcal{X}$, such that

$$\int_{x \in \mathcal{X}} \pi(dx)\, P(x, dy) \;=\; \pi(dy)\,. \tag{4}$$

Then hopefully (see Subsection 3.2), if we run the Markov chain for a long time (started from anywhere), then for large n the distribution of X_n will be approximately stationary: $\mathcal{L}(X_n) \approx \pi(\cdot)$. We can then (say) set $Z_1 = X_n$, and then restart and rerun the Markov chain to obtain Z_2, Z_3, etc., and then do estimates as in (3).

It may seem at first to be even more difficult to find such a Markov chain, then to estimate $\pi(f)$ directly. However, we shall see in the next section that constructing (and running) such Markov chains is often surprisingly straightforward.

Remark. In the practical use of MCMC, rather than start a fresh Markov chain for each new sample, often an entire tail of the Markov chain run $\{X_n\}$ is used to create an estimate such as $(N-B)^{-1} \sum_{i=B+1}^{N} f(X_i)$, where the *burn-in* value B is hopefully chosen large enough that $\mathcal{L}(X_B) \approx \pi(\cdot)$. In that case the different $f(X_i)$ are not independent, but the estimate can be computed more efficiently. Since many of the mathematical issues which arise are similar in either implementation, we largely ignore this modification herein.

Remark. MCMC is, of course, not the only way to sample or estimate from complicated probability distributions. Other possible sampling algorithms include "rejection sampling" and "importance sampling", not reviewed here; but these alternative algorithms only work well in certain particular cases and are not as widely applicable as MCMC algorithms.

1.2. Motivation: Bayesian Statistics Computations.

While MCMC algorithms are used in many fields (statistical physics, computer science), their most widespread application is in Bayesian statistical inference.

Let $L(\mathbf{y}|\theta)$ be the likelihood function (i.e., density of data $\mathbf{y}$ given unknown parameters θ) of a statistical model, for $\theta \in \mathcal{X}$. (Usually $\mathcal{X} \subseteq \mathbf{R}^d$.) Let the "prior" density of θ be $p(\theta)$. Then the "posterior" distribution of θ given $\mathbf{y}$ is the density which is proportional to

$$\pi_u(\theta) \;\equiv\; L(\mathbf{y} \,|\, \theta)\, p(\theta)\,.$$

(Of course, the normalisation constant is simply the density for the data $\mathbf{y}$, though that constant may be impossible to compute.) The "posterior mean" of any functional f is then given by:

$$\pi(f) \;=\; \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}\,.$$

For this reason, Bayesians are anxious (even desperate!) to estimate such $\pi(f)$. Good estimates allow Bayesian inference can be used to estimate a wide variety of parameters, probabilities, means, etc. MCMC has proven to be extremely helpful for such Bayesian estimates, and MCMC is now extremely widely used in the Bayesian statistical community.

2. Constructing MCMC Algorithms.

We see from the above that an MCMC algorithm requires, given a probability distribution $\pi(\cdot)$ on a state space $\mathcal{X}$, a Markov chain on $\mathcal{X}$ which is easily run on a computer, and which has $\pi(\cdot)$ as its stationary distribution as in (4). This section explains how such Markov chains are constructed. It thus provides motivation and context for the theory which follows; however, for the reader interested purely in the mathematical results, this section can be omitted with little loss of continuity.

A key notion is *reversibility*, as follows.

Definition. A Markov chain on a state space $\mathcal{X}$ is *reversible* with respect to a probability distribution $\pi(\cdot)$ on $\mathcal{X}$, if

$$\pi(dx)\, P(x, dy) \;=\; \pi(dy)\, P(y, dx)\,, \qquad x, y \in \mathcal{X}\,.$$

A very important property of reversibility is the following.

Proposition 1. *If a Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary for the chain.*

Proof. We compute that

$$\int_{x\in\mathcal{X}} \pi(dx)\, P(x, dy) \;=\; \int_{x\in\mathcal{X}} \pi(dy)\, P(y, dx) \;=\; \pi(dy) \int_{x\in\mathcal{X}} P(y, dx) \;=\; \pi(dy)\,. \qquad \blacksquare$$

We see from this lemma that, when constructing an MCMC algorithm, it suffices to create a Markov chain which is easily run, and which is *reversible* with respect to $\pi(\cdot)$. The simplest way to do so is to use the Metropolis-Hastings algorithm, as we now discuss.

2.1. The Metropolis-Hastings Algorithm.

Suppose again that $\pi(\cdot)$ has a (possibly unnormalised) density π_u, as in (1). Let $Q(x, \cdot)$ be essentially any other Markov chain, whose transitions also have a (possibly unnormalised) density, i.e. $Q(x, dy) \propto q(x, y)\, dy$.

The Metropolis-Hastings algorithm proceeds as follows. First choose some X_0. Then, given X_n, generate a *proposal* Y_{n+1} from $Q(X_n, \cdot)$. Also flip an independent coin, whose probability of heads equals $\alpha(X_n, Y_{n+1})$, where

$$\alpha(x, y) = \min\left[1,\; \frac{\pi_u(y)\, q(y, x)}{\pi_u(x)\, q(x, y)}\right].$$

(To avoid ambiguity, we set $\alpha(x, y) = 1$ whenever $\pi(x)\, q(x, y) = 0$.) Then, if the coin is heads, "accept" the proposal by setting $X_{n+1} = Y_{n+1}$; if the coin is tails then "reject" the proposal by setting $X_{n+1} = X_n$. Replace n by $n + 1$ and repeat.

The reason for the unusual formula for $\alpha(x, y)$ is the following:

Proposition 2. *The Metropolis-Hastings algorithm (as described above) produces a Markov chain $\{X_n\}$ which is reversible with respect to $\pi(\cdot)$.*

Proof. We need to show

$$\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)\,.$$

It suffices to assume $x \neq y$ (since if $x = y$ then the equation is trivial). But for $x \neq y$, setting $c = \int_{\mathcal{X}} \pi_u(x)\, dx$,

$$\begin{aligned}
\pi(dx)\, P(x, dy) &= [c^{-1}\pi_u(x)\, dx]\, [q(x, y)\, \alpha(x, y)\, dy] \\
&= c^{-1}\pi_u(x)\, q(x, y)\, \min\left[1,\ \frac{\pi_u(y)q(y, x)}{\pi_u(x)q(x, y)}\right] dx\, dy \\
&= c^{-1}\min[\pi_u(x)\, q(x, y),\ \pi_u(y)q(y, x)]\, dx\, dy\,,
\end{aligned}$$

which is symmetric in x and y. ∎

To run the Metropolis-Hastings algorithm on a computer, we just need to be able to run the proposal chain $Q(x, \cdot)$ (which is easy, for appropriate choices of Q), and then do the accept/reject step (which is easy, provided we can easily compute the densities at individual points). Thus, running the algorithm is quite feasible. Furthermore we need to compute only *ratios* of densities [e.g. $\pi_u(y) \,/\, \pi_u(x)$], so we don't require the *normalising constants* $c = \int_{\mathcal{X}} \pi_u(x) dx$.

However, this algorithm in turn suggests further questions. Most obviously, how should we choose the proposal distributions $Q(x, \cdot)$? In addition, once $Q(x, \cdot)$ is chosen, then will we really have $\mathcal{L}(X_n) \approx \pi(\cdot)$ for large enough n? How large is large enough? We will return to these questions below.

Regarding the first question, there are many different classes of ways of choosing the proposal density, such as:

- **Symmetric Metropolis Algorithm.** Here $q(x, y) = q(y, x)$, and the acceptance probability simplifies to

$$\alpha(x, y) = \min\left[1,\ \frac{\pi_u(y)}{\pi_u(x)}\right]$$

- **Random walk Metropolis-Hastings.** Here $q(x, y) = q(y - x)$. For example, perhaps $Q(x, \cdot) = N(x, \sigma^2)$, or $Q(x, \cdot) = \text{Uniform}(x - 1,\, x + 1)$.

- **Independence sampler.** Here $q(x, y) = q(y)$, i.e. $Q(x, \cdot)$ does not depend on x.

- **Langevin algorithm.** Here the proposal is generated by

$$Y_{n+1} \sim N(X_n + (\delta/2)\, \nabla \log \pi(X_n),\ \delta)\,,$$

  for some (small) $\delta > 0$. (This is motivated by a discrete approximation to a Langevin diffusion processes.)

More about optimal choices of proposal distributions will be discussed in a later section, as will the second question about time to stationarity (i.e. how large does n need to be).

2.2. Combining Chains.

If P_1 and P_2 are two different chains, each having stationary distribution $\pi(\cdot)$, then the new chain $P_1 P_2$ also has stationary distribution $\pi(\cdot)$.

Thus, it is perfectly acceptable, and quite common (see e.g. Tierney [92] and [68]), to make new MCMC algorithms out of old ones, by specifying that the new algorithm applies first the chain P_1, then the chain P_2, then the chain P_1 again, etc. (And, more generally, it is possible to combine many different chains in this manner.)

Note that, even if each of P_1 and P_2 are reversible, the combined chain $P_1 P_2$ will in general *not* be reversible. It is for this reason that it is important, when studying MCMC, to allow for non-reversible chains as well.

2.3. The Gibbs Sampler.

The Gibbs sampler is also known as the "heat bath" algorithm, or as "Glauber dynamics". Suppose again that $\pi_u(\cdot)$ is d-dimensional density, with $\mathcal{X}$ an open subset of $\mathbf{R}^d$, and write $\mathbf{x} = (x_1, \ldots, x_d)$.

The i^{th} **component Gibbs sampler** is defined such that P_i leaves all components besides i unchanged, and replaces the i^{th} component by a draw from the full conditional distribution of $\pi(\cdot)$ conditional on <u>all</u> the other components.

More formally, let

$$S_{x,i,a,b} \;=\; \{y \in \mathcal{X};\; y_j = x_j \text{ for } j \neq i, \text{ and } a \leq y_i \leq b\}\,.$$

Then

$$P_i(x,\, S_{x,i,a,b}) \;=\; \frac{\int_a^b \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)\, dt}{\int_{-\infty}^{\infty} \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)\, dt}\,, \quad a \leq b\,.$$

It follows immediately (from direct computation, or from the definition of conditional density), that P_i, is reversible with respect to $\pi(\cdot)$. (In fact, P_i may be regarded as a special case of a Metropolis-Hastings algorithm, with $\alpha(x, y) \equiv 1$.) Hence, P_i has $\pi(\cdot)$ as a stationary distribution.

We then construct the full Gibbs sampler out of the various P_i, by combining them (as in the previous subsection) in one of two ways:

- **The deterministic-scan Gibbs sampler** is

$$P \;=\; P_1 P_2 \ldots P_d\,.$$

  That is, it performs the d different Gibbs sampler components, in sequential order.

- **The random-scan Gibbs sampler** is

$$P \;=\; \frac{1}{d} \sum_{i=1}^{d} P_i\,.$$

  That is, it does one of the d different Gibbs sampler components, chosen uniformly at random.

Either version produces an MCMC algorithm having $\pi(\cdot)$ as its stationary distribution. The output of a Gibbs sampler is thus a "zig-zag pattern", where the components get updated one at a time. (Also, the random-scan Gibbs sampler is reversible, while the deterministic-scan Gibbs sampler usually is not.)

2.4. Detailed Bayesian Example: Variance Components Model.

We close this section by presenting a typical example of a target density π_u that arises in Bayesian statistics, in an effort to illustrate the problems and issues which arise.

The model involves fixed constant μ_0 and positive constants a_1, b_1, a_2, b_2, and σ_0^2. It involves three hyperparameters, σ_θ^2, σ_e^2, and μ, each having priors based upon these constants as follows: $\sigma_\theta^2 \sim IG(a_1, b_1)$; $\sigma_e^2 \sim IG(a_2, b_2)$; and $\mu \sim N(\mu_0, \sigma_0^2)$. It involves K further parameters $\theta_1, \theta_2, \ldots, \theta_K$, conditionally independent given the above hyperparameters, with $\theta_i \sim N(\mu, \sigma_\theta^2)$. In terms of these parameters, the data $\{Y_{ij}\}$ ($1 \le i \le K$, $1 \le j \le J$) are assumed to be distributed as $Y_{ij} \sim N(\theta_i, \sigma_e^2)$, conditionally independently given the parameters. A graphical representation of the model is as follows:

$$
\begin{array}{ccccc}
 & & \mu & & \\
 & \swarrow & \downarrow & \searrow & \\
\theta_1 & \ldots\ldots & & \theta_K & \qquad \theta_i \sim N(\mu, \sigma_\theta^2) \\
\downarrow & & & \downarrow & \\
Y_{11}, \ldots, Y_{1J} & & & Y_{K1}, \ldots, Y_{KJ} & \qquad Y_{ij} \sim N(\theta_i, \sigma_e^2)
\end{array}
$$

The Bayesian paradigm then involves conditioning on the values of the data $\{Y_{ij}\}$, and considering the joint distribution of all $K+3$ parameters given this data. That is, we are interested in the distribution

$$\pi(\cdot) = \mathcal{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K \,|\, \{Y_{ij}\}),$$

defined on the state space $\mathcal{X} = (0, \infty)^2 \times \mathbf{R}^{K+1}$. We would like to sample from this distribution $\pi(\cdot)$. We compute that this distribution's unnormalised density is given by

$$
\begin{aligned}
&\pi_u(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \propto \\
&\qquad e^{-b_1/\sigma_\theta^2} {\sigma_\theta^2}^{-a_1-1} e^{-b_2/\sigma_e^2} {\sigma_e^2}^{-a_2-1} e^{-(\mu-\mu_0)^2/2\sigma_0^2} \\
&\times \prod_{i=1}^{K} [e^{-(\theta_i-\mu)^2/2\sigma_\theta^2}/\sigma_\theta] \times \prod_{i=1}^{K}\prod_{j=1}^{J} [e^{-(Y_{ij}-\theta_i)^2/2\sigma_e^2}/\sigma_e] .
\end{aligned}
$$

This is a very typical target density for MCMC in statistics, in that it is high-dimensional $(K+3)$, its formula is messy and irregular, it is positive throughout $\mathcal{X}$, and it is larger in "center" of $\mathcal{X}$ and smaller in "tails" of $\mathcal{X}$.

We now consider constructing MCMC algorithms to sample from the target density π_u. We begin with the Gibbs sampler. To run a Gibbs sampler, we require the full conditionals distributions, computed (without difficulty since they are all one-dimensional) to be as follows:

$$\mathcal{L}(\sigma_\theta^2 \mid \mu, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; IG\left(a_1 + \frac{1}{2}K,\; b_1 + \frac{1}{2}\sum_i (\theta_i - \mu)^2\right);$$

$$\mathcal{L}(\sigma_e^2 \mid \mu, \sigma_\theta^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; IG\left(a_2 + \frac{1}{2}KJ,\; b_2 + \frac{1}{2}\sum_{i,j} (Y_{ij} - \theta_i)^2\right);$$

$$\mathcal{L}(\mu \mid \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}) \;=\; N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2\sum_i \theta_i}{\sigma_\theta^2 + K\sigma_0^2},\; \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right);$$

$$\mathcal{L}(\theta_i \mid \mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K, Y_{ij}) \;=\; N\left(\frac{J\sigma_\theta^2\overline{Y}_i + \sigma_e^2\mu}{J\sigma_\theta^2 + \sigma_e^2},\; \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right),$$

where $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^J Y_{ij}$, and the last equation holds for $1 \le i \le K$. The Gibbs sampler then proceeds by updating the $K+3$ variables, in turn (either deterministic or random scan), according to the above conditional distributions. This is feasible since the conditional distributions are all easily simulated (IG and N). In fact, it appears to work well, both in practice and according to various theoretical results; this model was one of the early statistical applications of the Gibbs sampler by Gelfand and Smith [30], and versions of it have been used and studied often (see e.g. [78], [56], [81], [20], [43], [44]).

Alternatively, we can run a Metropolis-Hastings algorithm for this model. For example, we might choose a symmetric random-walk Metropolis algorithm with proposals of the form $N(X_n, \sigma^2 I_{K+3})$ for some $\sigma^2 > 0$ (say). Then, given X_n, the algorithm would proceed as follows:

1. Choose $Y_{n+1} \sim N(X_n, \sigma^2 I_{K+3})$;
2. Choose $U_{n+1} \sim \text{Uniform}[0,1]$;

3. If $U_{n+1} < \pi_u(Y_{n+1}) \,/\, \pi_u(X_n)$, then set $X_{n+1} = Y_{n+1}$ (accept). Otherwise set $X_{n+1} = X_n$ (reject).

This MCMC algorithm also appears to work well for this model, at least if the value of σ^2 is chosen appropriately (as discussed in Section 6). We conclude that, for such "typical" target distributions $\pi(\cdot)$, both the Gibbs sampler and appropriate Metropolis-Hastings algorithms perform well in practice, and allow us to sample from $\pi(\cdot)$.

3. Bounds on Markov Chain Convergence Times.

Once we know how to construct (and run) lots of different MCMC algorithms, other questions arise. Most obviously, do they converge to the distribution $\pi(\cdot)$? And, how quickly does this convergence take place?

To proceed, write $P^n(x, A)$ for the n-step transition law of the Markov chain:

$$P^n(x, A) = \mathbf{P}[X_n \in A \,|\, X_0 = x] \,.$$

The main MCMC convergence questions are, is $P^n(x, A)$ "close" to $\pi(A)$ for large enough n? And, how large is large enough?

3.1. Total Variation Distance.

We shall measure the distance to stationary in terms of total variation distance, defined as follows:

Definition. The *total variation distance* between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_A |\nu_1(A) - \nu_2(A)| \,.$$

We can then ask, is $\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0$? And, given $\epsilon > 0$, how large must n be so that $\|P^n(x,\cdot) - \pi(\cdot)\| < \epsilon$? We consider such questions herein.

We first pause to note some simple properties of total variation distance.

Proposition 3. *(a)* $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_{f:\mathcal{X}\to[0,1]} |\int f d\nu_1 - \int f d\nu_2|$.
(b) $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[a,b]} |\int f d\nu_1 - \int f d\nu_2|$ *for any* $a < b$, *and in particular* $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} |\int f d\nu_1 - \int f d\nu_2|$.

(c) If $\pi(\cdot)$ is stationary for a Markov chain kernel P, then $\|P^n(x,\cdot)-\pi(\cdot)\|$ is non-increasing in n, i.e. $\|P^n(x,\cdot)-\pi(\cdot)\| \le \|P^{n-1}(x,\cdot)-\pi(\cdot)\|$ for $n \in \mathbf{N}$.

(d) More generally, letting $(\nu_i P)(A) = \int \nu_i(dx)\, P(x,A)$, we always have $\|(\nu_1 P)(\cdot) - (\nu_2 P)(\cdot)\| \le \|\nu_1(\cdot) - \nu_2(\cdot)\|$.

(e) Let $t(n) = 2\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\|$, where $\pi(\cdot)$ is stationary. Then t is sub-multiplicative, i.e. $t(m+n) \le t(m)\, t(n)$ for $m, n \in \mathbf{N}$.

(f) If $\mu(\cdot)$ and $\nu(\cdot)$ have densities g and h, respectively, with respect to some σ-finite measure $\rho(\cdot)$, and $M = \max(g,h)$ and $m = \min(g,h)$, then

$$\|\mu(\cdot) - \nu(\cdot)\| \;=\; \frac{1}{2}\int_{\mathcal{X}} (M-m)\, d\rho \;=\; 1 - \int_{\mathcal{X}} m\, d\rho\,.$$

(g) Given probability measures $\mu(\cdot)$ and $\nu(\cdot)$, there are jointly defined random variables X and Y such that $X \sim \mu(\cdot)$, $Y \sim \nu(\cdot)$, and $\mathbf{P}[X = Y] = 1 - \|\mu(\cdot) - \nu(\cdot)\|$.

Proof. For (a), let $\rho(\cdot)$ be any σ-finite measure such that $\nu_1 \ll \rho$ and $\nu_2 \ll \rho$ (e.g. $\rho = \nu_1 + \nu_2$), and set $g = d\nu_1/d\rho$ and $h = d\nu_2/d\rho$. Then $|\int f d\nu_1 - \int f d\nu_2| = |\int f(g-h)\, d\rho|$. This is maximised (over all $0 \le f \le 1$) when $f = 1$ on $\{g > h\}$ and $f = 0$ on $\{h > g\}$ (or vice-versa), in which case it equals $|\nu_1(A) - \nu_2(A)|$ for $A = \{g > h\}$ (or $\{g < h\}$), thus proving the equivalence.

Part (b) follows very similarly to (a), except now $f = b$ on $\{g > h\}$ and $f = a$ on $\{g < h\}$ (or vice-versa), leading to $|\int f d\nu_1 - \int f d\nu_2| = (b-a)\,|\nu_1(A) - \nu_2(A)|$.

For part (c), we compute that

$$|P^{n+1}(x,A) - \pi(A)| \;=\; \left|\int_{y\in\mathcal{X}} P^n(x,dy)P(y,A) - \int_{y\in\mathcal{X}} \pi(dy)P(y,A)\right|$$

$$=\; \left|\int_{y\in\mathcal{X}} P^n(x,dy)f(y) - \int_{y\in\mathcal{X}} \pi(dy)f(y)\right| \;\le\; \|P^n(x,\cdot) - \pi(\cdot)\|\,,$$

where $f(y) = P(y,A)$, and where the inequality comes from part (a).

Part (d) follows very similarly to part (c).

Part (e) follows since $t(n)$ is an L^∞ operator norm of P^n (cf. Meyn and Tweedie [53], Lemma 16.1.1). More specifically, let $\hat{P}(x,\cdot) = P^n(x,\cdot) - \pi(\cdot)$ and $\hat{Q}(x,\cdot) = P^m(x,\cdot) - \pi(\cdot)$,

so that

$$(\hat{P}\,\hat{Q}f)(x) \;\equiv\; \int_{y\in\mathcal{X}} f(y) \int_{z\in\mathcal{X}} [P^n(x,dz) - \pi(dz)]\,[\mathbf{P}^m(z,dy) - \pi(dy)]$$

$$= \int_{y\in\mathcal{X}} f(y)\,[P^{n+m}(x,dy) - \pi(dy) - \pi(dy) + \pi(dy)]$$

$$= \int_{y\in\mathcal{X}} f(y)\,[P^{n+m}(x,dy) - \pi(dy)]\,.$$

Then let $f : \mathcal{X} \to [0,1]$, let $g(x) = (\hat{Q}f)(x) \equiv \int_{y\in\mathcal{X}} \hat{Q}(x,dy)f(y)$, and let $g^* = \sup_{x\in\mathcal{X}} |g(x)|$. Then $g^* \le \frac{1}{2}\,t(m)$ by part (a). Now, if $g^* = 0$, then clearly $\hat{P}\hat{Q}f = 0$. Otherwise, we compute that

$$2 \sup_{x\in\mathcal{X}} |(\hat{P}\,\hat{Q}f)(x)| = 2\,g^* \sup_{x\in\mathcal{X}} |(\hat{P}[g/g^*])(x)| \le t(m) \sup_{x\in\mathcal{X}} (\hat{P}[g/g^*])(x)|\,. \tag{5}$$

Since $-1 \le g/g^* \le 1$, we have $(\hat{P}[g/g^*])(x) \le 2\,\|P^n(x,\cdot) - \pi(\cdot)\|$ by part (b), so that $\sup_{x\in\mathcal{X}}(\hat{P}[g/g^*])(x) \le t(n)$. The result then follows from part (a) together with (5).

The first equality of part (f) follows since, as in the proof of part (b) with $a = -1$ and $b = 1$, we have

$$\|\mu(\cdot) - \nu(\cdot)\| \;=\; \frac{1}{2}\Big(\int_{g>h} (g-h)\,d\rho + \int_{h>g} (h-g)\,d\rho\Big) \;=\; \frac{1}{2}\int_{\mathcal{X}} (M-m)\,d\rho\,.$$

The second equality of part (f) then follows since $M+m = g+h$, so that $\int_{\mathcal{X}}(M+m)\,d\rho = 2$, and hence

$$\frac{1}{2}\int_{\mathcal{X}} (M-m)\,d\rho \;=\; 1 - \frac{1}{2}\Big(2 - \int_{\mathcal{X}} (M-m)\,d\rho\Big)$$

$$= \;1 - \frac{1}{2}\int_{\mathcal{X}} \big((M+m) - (M-m)\big)\,d\rho \;=\; 1 - \int_{\mathcal{X}} m\,d\rho\,.$$

For part (g), we let $a = \int_{\mathcal{X}} m\,d\rho$, $b = \int_{\mathcal{X}}(g-m)\,d\rho$, and $c = \int_{\mathcal{X}}(h-m)\,d\rho$. The statement is trivial if any of a, b, c equal zero, so assume they are all positive. We then jointly construct random variables Z, U, V, I such that Z has density m/a, U has density $(g-m)/b$, V has density $(h-m)/c$, and I is independent of Z, U, V with $\mathbf{P}[I=1] = a$ and $\mathbf{P}[I=0] = 1-a$. We then let $X = Y = Z$ if $I = 1$, and $X = U$ and $Y = V$ if $I = 0$.

Then it is easily checked that $X \sim \mu(\cdot)$ and $Y \sim \nu(\cdot)$. Furthermore U and V have disjoint support, so $\mathbf{P}[U = V] = 0$. Then using part (f),

$$\mathbf{P}[X = Y] \;=\; \mathbf{P}[I = 1] \;=\; a \;=\; 1 \,-\, \|\mu(\cdot) - \nu(\cdot)\|\,,$$

as claimed. ∎

Remark. Proposition 3(e) is false without the factor of 2. For example, suppose $\mathcal{X} = \{1, 2\}$, with $P(1, \{1\}) = 0.3$, $P(1, \{2\}) = 0.7$, $P(2, \{1\}) = 0.4$, $P(2, \{2\}) = 0.6$, $\pi(1) = \frac{4}{11}$, and $\pi(2) = \frac{7}{11}$. Then $\pi(\cdot)$ is stationary, and $\sup_{x \in \mathcal{X}} \|P(x, \cdot) - \pi(\cdot)\| = 0.0636$, and $\sup_{x \in \mathcal{X}} \|P^2(x, \cdot) - \pi(\cdot)\| = 0.00636$, but $0.00636 > (0.0636)^2$. On the other hand, some authors instead define total variation distance as *twice* the value used here, in which case the factor of 2 in Proposition 3(e) is not written explicitly.

3.2. Asymptotic Convergence.

Even if a Markov chain has stationary distribution $\pi(\cdot)$, it may still fail to converge to stationarity:

Example 1. Suppose $\mathcal{X} = \{1, 2, 3\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{1\}) = P(1, \{2\}) = P(2, \{1\}) = P(2, \{2\}) = 1/2$, and $P(3, \{3\}) = 1$. Then $\pi(\cdot)$ is stationary. However, if $X_0 = 1$, then $X_n \in \{1, 2\}$ for all n, so $P(X_n = 3) = 0$ for all n, so $P(X_n = 3) \not\to \pi\{3\}$, and the distribution of X_n does not converge to $\pi(\cdot)$. (In fact, here the stationary distribution is not *unique*, and the distribution of X_n converges to a different stationary distribution defined by $\pi\{1\} = \pi\{2\} = 1/2$.)

The above example is "reducible", in that the chain can never get from state 1 to state 3, in any number of steps. Now, the classical notion of "irreducibility" is that the chain has positive probability of eventually reaching any state from any other state, but if $\mathcal{X}$ is uncountable then that condition is impossible. Instead, we demand the weaker condition of ϕ-irreducibility:

Definition. A chain is *ϕ-irreducible* if there exists a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.

For example, if $\phi(A) = \delta_{x_*}(A)$, then this requires that x_* has positive probability of eventually being reached from any state x. Thus, if a chain has any one state which is reachable from anywhere (which on a finite state space is equivalent to being *indecomposible*), then it is ϕ-irreducible. However, if $\mathcal{X}$ is uncountable then often $P(x, \{y\}) = 0$ for all x and y. In that case, $\phi(\cdot)$ might instead be e.g. Lebesgue measure on $\mathbf{R}^d$, so that $\phi(\{x\}) = 0$ for all singleton sets, but such that all subsets A of positive Lebesgue measure are eventually reachable with positive probability from any $x \in \mathcal{X}$.

Running Example. Here we introduce a running example, to which we shall return several times. Suppose that $\pi(\cdot)$ is a probability measure having unnormalised density function π_u with respect to d-dimensional Lebesgue measure. Consider the Metropolis-Hastings algorithm for π_u with proposal density $q(\mathbf{x}, \cdot)$ with respect to d-dimensional Lebesgue measure. Then if $q(\cdot, \cdot)$ is positive and continuous on $\mathbf{R}^d \times \mathbf{R}^d$, and π_u is positive everywhere, then the algorithm is π-irreducible. Indeed, let $\pi(A) > 0$. Then there exists $R > 0$ such that $\pi(A_R) > 0$, where $A_R = A \cap B_R(\mathbf{0})$, and $B_R(\mathbf{0})$ represents the ball of radius R centred at $\mathbf{0}$. Then by continuity, for any $\mathbf{x} \in \mathbf{R}^d$, $\inf_{\mathbf{y} \in A_R} \min\{q(\mathbf{x}, \mathbf{y}), q(\mathbf{y}, \mathbf{x})\} \geq \epsilon$ for some $\epsilon > 0$, and thus we have that

$$P(\mathbf{x}, A) \;\geq\; P(\mathbf{x}, A_R) \;=\; \int_{A_R} q(\mathbf{x}, \mathbf{y}) \, \min\left[1, \frac{\pi_u(\mathbf{y})\, q(\mathbf{y}, \mathbf{x})}{\pi_u(\mathbf{x})\, q(\mathbf{x}, \mathbf{y})}\right] d\mathbf{y}$$

$$\geq\; \epsilon \, \mathrm{Leb}\big(\{\mathbf{y} \in A_R : \; \pi_u(\mathbf{y}) \geq \pi_u(\mathbf{x})\}\big) \;+\; \frac{\epsilon\, K}{\pi_u(\mathbf{x})} \, \pi\big(\{\mathbf{y} \in A_R : \; \pi_u(\mathbf{y}) < \pi_u(\mathbf{x})\}\big)\,,$$

where $K = \int_{\mathcal{X}} \pi_u(\mathbf{x})\, d\mathbf{x} > 0$. Since $\pi(\cdot)$ is mutually absolutely continuous with Lebesgue measure, and since $Leb(A_R) > 0$, it follows that the terms in this final sum cannot both be 0, so that we must have $P(x, A) > 0$. Hence, the chain is π-irreducible.

Even ϕ-irreducible chains might not converge in distribution, due to periodicity problems, as in the following simple example.

Example 2. Suppose again $\mathcal{X} = \{1, 2, 3\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$. Then $\pi(\cdot)$ is stationary, and the chain is ϕ-irreducible [e.g. with $\phi(\cdot) = \delta_1(\cdot)$]. However, if $X_0 = 1$ (say), then $X_n = 1$ whenever n is a multiple of 3, so $P(X_n = 1)$ oscillates between 0 and 1, so again $P(X_n = 1) \not\to \pi\{3\}$, and there is again no convergence to $\pi(\cdot)$.

To avoid this problem, we require *aperiodicity*, and we adopt the following definition (which suffices for the ϕ-irreducible chains with stationary distributions that we shall study; for more general relationships see e.g. Meyn and Tweedie [53], Theorem 5.4.4):

Definition. A Markov chain with stationary distribution $\pi(\cdot)$ is *aperiodic* if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ with $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$, such that $\pi(\mathcal{X}_1) > 0$ (and hence $\pi(\mathcal{X}_i) > 0$ for all i). (Otherwise, the chain is *periodic*, with *period* equal to the largest such value of d, and corresponding *periodic decomposition* $\mathcal{X}_1, \ldots, \mathcal{X}_d$.)

Running Example, Continued. Here we return to the Running Example introduced above, and demonstrate that no additional assumptions are necessary to ensure aperiodicity. To see this, suppose that $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint subsets of $\mathcal{X}$ both of positive π measure, with $P(\mathbf{x}, \mathcal{X}_2) = 1$ for all $\mathbf{x} \in \mathcal{X}_1$. But just take any $\mathbf{x} \in \mathcal{X}_1$, then since $\mathcal{X}_1$ must have positive Lebesgue measure,

$$P(\mathbf{x}, \mathcal{X}_1) \;\geq\; \int_{\mathbf{y} \in \mathbf{X}_1} q(\mathbf{x}, \mathbf{y})\, \alpha(\mathbf{x}, \mathbf{y})\, d\mathbf{y} \;>\; 0$$

for a contradiction. Therefore aperiodicity must hold. (It is possible to demonstrate similar results for other MCMC algorithms, such as the Gibbs sampler, see e.g. Tierney [92]. Indeed, it is rather rare for MCMC algorithms to be periodic.)

Now we can state the main asymptotic convergence theorem, whose proof is described in Section 4. (This theorem assumes that the state space's σ-algebra is *countably generated*, but this is a very weak assumption which is true for e.g. any countable state space, or any subset of $\mathbf{R}^d$ with the usual Borel σ-algebra, since that σ-algebra is generated by the balls with rational centers and rational radii.)

Theorem 4. *If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$,*

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| \;=\; 0\,.$$

In particular, $\lim_{n\to\infty} P^n(x,A) = \pi(A)$ for all measurable $A \subseteq \mathcal{X}$.

Fact 5. In fact, under the conditions of Theorem 4, if $h : \mathcal{X} \to \mathbf{R}$ with $\pi(|h|) < \infty$, then a "strong law of large numbers" also holds (see e.g. Meyn and Tweedie [53], Theorem 17.0.1), as follows:

$$\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} h(X_i) \;=\; \pi(h) \quad w.p.\ 1\,. \tag{6}$$

Theorem 4 requires that the chain be ϕ-irreducible and aperiodic, and have stationary distribution $\pi(\cdot)$. Now, MCMC algorithms are created precisely so that $\pi(\cdot)$ is stationary, so this requirement is not a problem. Furthermore, it is usually straightforward to verify that chain is ϕ-irreducible, where e.g. ϕ is Lebesgue measure on an appropriate region. Also, aperiodicity almost always holds, e.g. for virtually any Metropolis algorithm or Gibbs sampler. Hence, Theorem 4 is widely applicable to MCMC algorithms.

It is worth asking why the convergence in Theorem 4 is just from π-a.e. $x \in \mathcal{X}$. The problem is that the chain may have unpredictable behaviour on a "null set" of π-measure 0, and fail to converge there. Here is a simple example due to C. Geyer (personal communication):

Example 3. Let $\mathcal{X} = \{1, 2, \ldots\}$. Let $P(1, \{1\}) = 1$, and for $x \geq 2$, $P(x, \{1\}) = 1/x^2$ and $P(x, \{x+1\}) = 1 - (1/x^2)$. Then chain has stationary distribution $\pi(\cdot) = \delta_1(\cdot)$, and it is π-irreducible and aperiodic. On the other hand, if $X_0 = x \geq 2$, then $\mathbf{P}[X_n = x + n \text{ for all } n] = \prod_{j=x}^{\infty}(1 - (1/j^2)) > 0$, so that $\|P^n(x,\cdot) - \pi(\cdot)\| \not\to 0$. Here Theorem 4 holds for $x = 1$ which is indeed π-a.e. $x \in \mathcal{X}$, but it does not hold for $x \geq 2$.

Remark. The transient behaviour of the chain on the null set in Example 3 is not accidental. If instead the chain converged on the null set to some other stationary distribution, but still had positive probability of escaping the null set (as it must to be ϕ-irreducible), then with probability 1 the chain would eventually exit the null set, and would thus converge to $\pi(\cdot)$ from the null set after all.

It is reasonable to ask under what circumstances the conclusions of Theorem 4 will hold for all $x \in \mathcal{X}$, not just π-a.e. Obviously, this will hold if the transition kernels $P(x, \cdot)$ are all absolutely continuous with respect to $\pi(\cdot)$ (i.e., $P(x, dy) = p(x, y)\, \pi(dy)$ for some function $p : \mathcal{X} \times \mathcal{X} \to [0, \infty)$), or for any Metropolis algorithm whose proposal distributions $Q(x, \cdot)$ are absolutely continuous with respect to $\pi(\cdot)$. It is also easy to see that this will hold for our Running Example described above. More generally, it suffices that the chain be *Harris recurrent*, meaning that for all $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, and all $x \in \mathcal{X}$, the chain will eventually reach B from x with probability 1, i.e. $\mathbf{P}[\exists\, n : \ X_n \in B \,|\, X_0 = x] = 1$. This condition is stronger than π-irreducibility (as evidenced by Example 3); for further discussions of this see e.g. Orey [60], Tierney [92], Chan and Geyer [15], and [74].

Finally, we note that periodic chains occasionally arise in MCMC (see e.g. Neal [57]), and much of the theory can be applied to this case. For example, we have the following.

Corollary 6. *If a Markov chain is ϕ-irreducible, with period $d \geq 2$, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$,*

$$\lim_{n \to \infty} \left\| (1/d) \sum_{i=n}^{n+d-1} P^i(x, \cdot) - \pi(\cdot) \right\| = 0\,, \tag{7}$$

and also the strong law of large numbers (6) continues to hold without change.

Proof. Let the chain have periodic decomposition $\mathcal{X}_1, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$, and let P' be the d-step chain P^d restricted to the state space $\mathcal{X}_1$. Then P' is ϕ-irreducible and aperiodic on $\mathcal{X}_1$, with stationary distribution $\pi'(\cdot)$ which satisfies that $\pi(\cdot) = (1/d) \sum_{j=0}^{d-1} (\pi' P^j)(\cdot)$. Now, from Proposition 3(c), it suffices to prove (7) when $n = md$ with $m \to \infty$, and for simplicity we assume without loss of generality that $x \in \mathcal{X}_1$. From Proposition 3(d), we

have $\|P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)\| \leq \|P^{md}(x,\cdot) - \pi'(\cdot)\|$ for $j \in \mathbf{N}$. Then, by the triangle inequality,

$$\left\|(1/d)\sum_{i=md}^{md+d-1} P^i(x,\cdot) - \pi(\cdot)\right\| \;=\; \left\|(1/d)\sum_{j=0}^{d-1} P^{md+j}(x,\cdot) - (1/d)\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)\right\|$$

$$\leq\; (1/d)\sum_{j=0}^{d-1}\|P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)\| \;\leq\; (1/d)\sum_{j=0}^{d-1}\|P^{md}(x,\cdot) - \pi'(\cdot)\|\,.$$

But applying Theorem 4 to P', we obtain that $\lim_{m\to\infty}\|P^{md}(x,\cdot) - \pi'(\cdot)\| = 0$ for π'-a.e. $x \in \mathcal{X}_1$, thus giving the first result.

To establish (6), let $\overline{P}$ be the transition kernel for the Markov chain on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ corresponding to the sequence $\{(X_{md}, X_{md+1}, \ldots, X_{md+d-1})\}_{m=0}^{\infty}$, and let $\overline{h}(x_0,\ldots,x_{d-1}) = (1/d)(h(x_0)+\ldots+h(x_{d-1}))$. Then just like P', we see that $\overline{P}$ is ϕ-irreducible and aperiodic, with stationary distribution given by

$$\overline{\pi} \;=\; \pi' \times (\pi' P) \times (\pi' P^2) \times \ldots \times (\pi' P^{d-1})\,.$$

Applying Fact 5 to $\overline{P}$ and $\overline{h}$ establishes that (6) holds without change. ∎

Remark. In particular, both (7) and (6) hold (without further assumptions re periodicity) for any irreducible (or indecomposible) Markov chain on a *finite* state space.

A related question for periodic chains, not considered here, is to consider quantitative bounds on the difference of average distributions, $\|(1/n)\sum_{i=1}^{n} P^i(x,\cdot) - \pi(\cdot)\|$, through the use of *shift-coupling*; see Aldous and Thorisson [3], and [67].

3.3. Uniform Ergodicity.

Theorem 4 implies asymptotic convergence to stationarity, but does not say anything about the *rate* of this convergence. One "qualitative" convergence rate property is uniform ergodicity:

Definition. A Markov chain having stationary distribution $\pi(\cdot)$ is *uniformly ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; M\,\rho^n\,, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$ and $M < \infty$.

One equivalence of uniform ergodicity is:

Proposition 7. *A Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| < 1/2$ for some $n \in \mathbf{N}$.*

Proof. If the chain is uniformly ergodic, then

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; \lim_{n\to\infty} M\,\rho^n \;=\; 0\,,$$

so $\sup_{x\in\mathcal{X}} \|P^n(x,\cdot)-\pi(\cdot)\| < 1/2$ for all sufficiently large n. Conversely, if $\sup_{x\in\mathcal{X}} \|P^n(x,\cdot)-\pi(\cdot)\| < 1/2$ for some $n \in \mathbf{N}$, then in the notation of Proposition 3(e), we have that $t(n) \equiv \beta < 1$, so that for all $j \in \mathbf{N}$, $t(jn) \leq \big(t(n)\big)^j = \beta^j$. Hence, from Proposition 3(c),

$$\|P^m(x,\cdot) - \pi(\cdot)\| \leq \|P^{\lfloor m/n\rfloor n}(x,\cdot) - \pi(\cdot)\| \leq \frac{1}{2}t\left(\lfloor m/n\rfloor n\right) \leq \beta^{\lfloor m/n\rfloor} \leq \beta^{-1}\,\left(\beta^{1/n}\right)^m\,,$$

so the chain is uniformly ergodic with $M = \beta^{-1}$ and $\rho = \beta^{1/n}$. ∎

Remark. The above Proposition of course continues to hold if we replace $1/2$ by δ for any $0 < \delta < 1/2$. However, it is false for $\delta \geq 1/2$. For example, if $\mathcal{X} = \{1, 2\}$, with $P(1, \{1\}) = P(2, \{2\}) = 1$, and $\pi(\cdot)$ is uniform on $\mathcal{X}$, then $\|P^n(x,\cdot) - \pi(\cdot)\| = 1/2$ for all $x \in \mathcal{X}$ and $n \in \mathbf{N}$.

To develop further conditions which ensure uniform ergodicity, we require a definition.

Definition. A subset $C \subseteq \mathcal{X}$ is *small* (or, (n_0, ϵ, ν)-small) if there exists a positive integer n_0, $\epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that the following *minorisation condition* holds:

$$P^{n_0}(x,\cdot) \;\geq\; \epsilon\,\nu(\cdot) \qquad x \in C\,, \tag{8}$$

i.e. $P^{n_0}(x, A) \geq \epsilon\, \nu(A)$ for all $x \in C$ and all measurable $A \subseteq \mathcal{X}$.

Remark. Some authors (e.g. Meyn and Tweedie [53]) also require that C have positive stationary measure, but for simplicity we don't explicitly require that here. In any case, $\pi(C) > 0$ follows under the additional assumption of the drift condition (10) considered in the next section.

Intuitively, this condition means that all of the n_0-step transitions from within C, all have an "ϵ-overlap", i.e. a component of size ϵ in common. (This concept goes back to Doeblin [22]; for further background, see e.g. [23], [8], [59], [4], and [53]; for applications to convergence rates see e.g. [54], [79], [81], [70], [76], [24], [84].) We note that if $\mathcal{X}$ is countable, and if

$$\epsilon_{n_0} \;\equiv\; \sum_{y \in \mathcal{X}} \inf_{x \in C} P^{n_0}(x, \{y\}) \;>\; 0\,, \tag{9}$$

then C is $(n_0, \epsilon_{n_0}, \nu)$-small where $\nu\{y\} = \epsilon_{n_0}^{-1} \inf_{x \in C} P^{n_0}(x, \{y\})$. (Furthermore, for an irreducible (or just indecomposible) and aperiodic chain on a *finite* state space, we always have $\epsilon_{n_0} > 0$ for sufficiently large n_0 (see e.g. [80]), so this method always applies in principle.) Similarly, if the transition probabilities have densities with respect to some measure $\eta(\cdot)$, i.e. if $P^{n_0}(x, dy) = p_{n_0}(x, y)\, \eta(dy)$, then we can take $\epsilon_{n_0} = \int_{y \in \mathcal{X}} \big(\inf_{x \in \mathcal{X}} p_{n_0}(x, y) \big)\, \eta(dy)$.

Remark. As observed in [71], small-set conditions of the form $P(x, \cdot) \geq \epsilon\, \nu(\cdot)$ for all $x \in C$, can be replaced by *pseudo-small* conditions of the form $P(x, \cdot) \geq \epsilon\, \nu_{xy}(\cdot)$ and $P(y, \cdot) \geq \epsilon\, \nu_{xy}(\cdot)$ for all $x, y \in C$, without affecting any bounds which use pairwise coupling (which includes all of the bounds considered here before Section 5. Thus, all of the results stated in this section remain true without change if "small set" is replaced by "pseudo-small set" in the hypotheses. For ease of exposition, we do not emphasise this point herein.

The main result guaranteeing uniform ergodicity, which goes back to Doeblin [22] and Doob [23] and in some sense even to Markov [49], is the following.

Theorem 8. *Consider a Markov chain with invariant probability distribution $\pi(\cdot)$. Suppose the minorisation condition (8) is satisfied for some $n_0 \in \mathbf{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \mathcal{X}$ (i.e., the entire state space is small). Then the*

chain is uniformly ergodic, and in fact $\|P^n(x,\cdot)-\pi(\cdot)\| \le (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ *for all* $x \in \mathcal{X}$*, where* $\lfloor r \rfloor$ *is the greatest integer not exceeding* r*.*

Theorem 8 is proved in Section 4. We note also that Theorem 8 provides a *quantitative* bound on the distance to stationarity $\|P^n(x,\cdot) - \pi(\cdot)\|$, namely that it must be $\le (1-\epsilon)^{\lfloor n/n_0 \rfloor}$. Thus, once n_0 and ϵ are known, we can find n_* such that, say, $\|P^{n_*}(x,\cdot)-\pi(\cdot)\| \le 0.01$, a fact which can be applied in certain MCMC contexts (see e.g. [77]). We can then say that n_* iterations "suffices for convergence" of the Markov chain. On a discrete state space, we have that $\|P^n(x,\cdot) - \pi(\cdot)\| \le (1-\epsilon_{n_0})^{\lfloor n/n_0 \rfloor}$ with ϵ_{n_0} as in (9).

Running Example, Continued. Recall our Running Example, introduced above. Since we have imposed strong continuity conditions on q, it is natural to conjecture that compact sets are small. However this is not true without extra regularity conditions. For instance, consider dimension $d = 1$, and suppose that $\pi_u(x) = \mathbf{1}_{0<|x|<1}|x|^{-1/2}$, and let $q(x,y) \propto \exp\{-(x-y)^2/2\}$, then it is easy to check that any neighbourhood of 0 is not small. However in the general setup of our Running Example, all compact sets on which π_u is bounded are small. To see this, suppose C is a compact set on which π_u is bounded by $k < \infty$. Let $\mathbf{x} \in C$, and let D be any compact set of positive Lebesgue and π measure, such that $\inf_{\mathbf{x},\mathbf{y}\in C\cup D} q(\mathbf{x},\mathbf{y}) = \epsilon > 0$ and $\sup_{\mathbf{x}\in C,\mathbf{y}\in D} q(\mathbf{x},\mathbf{y}) = M < \infty$. We then have,

$$P(\mathbf{x}, d\mathbf{y}) \;\ge\; q(\mathbf{x},\mathbf{y})\, d\mathbf{y}\, \min\left\{1, \frac{\pi_u(\mathbf{y})\, q(\mathbf{y},\mathbf{x})}{\pi_u(\mathbf{x})\, q(\mathbf{x},\mathbf{y})}\right\} \;\ge\; \epsilon\, d\mathbf{y} \min\left\{1, \frac{\epsilon\, \pi_u(\mathbf{y})}{Mk}\right\},$$

which is a positive measure independent of $\mathbf{x}$. Hence, C is small. (This example also shows that if π_u is continuous, the state space $\mathcal{X}$ is compact, and q is continuous and positive, then $\mathcal{X}$ is small, and so the chain must be uniformly ergodic.)

If a Markov chain is *not* uniformly ergodic (as few MCMC algorithms on unbounded state spaces are), then Theorem 8 cannot be applied. However, it is still of great importance, given a Markov chain kernel P and an initial state x, to be able to find n_* so that, say, $\|P^{n_*}(x,\cdot) - \pi(\cdot)\| \le 0.01$. This issue is discussed further below.

3.4. Geometric ergodicity.

A weaker condition than uniform ergodicity is geometric ergodicity, as follows (for background and history, see e.g. Nummelin [59], and Meyn and Tweedie [53]):

Definition. A Markov chain with stationary distribution $\pi(\cdot)$ is *geometrically ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; M(x)\,\rho^n\,, \qquad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, where $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.

The difference between geometric ergodicity and uniform ergodicity is that now the constant M may depend on the initial state x.

Of course, if the state space $\mathcal{X}$ is *finite*, then all irreducible and aperiodic Markov chains are geometrically (in fact, uniformly) ergodic. However, for infinite $\mathcal{X}$ this is not the case. For example, it is shown by Mengersen and Tweedie [51] (see also [75]) that a symmetric random-walk Metropolis algorithm is geometrically ergodic essentially if and only if $\pi(\cdot)$ has finite exponential moments. (For chains which are not geometrically ergodic, it is possible also to study *polynomial ergodicity*, not considered here; see Fort and Moulines [29], and Jarner and Roberts [41].) Hence, we now discuss conditions which ensure geometric ergodicity.

Definition. Given Markov chain transition probabilities P on a state space $\mathcal{X}$, and a measurable function $f : \mathcal{X} \to \mathbf{R}$, define the function $Pf : \mathcal{X} \to \mathbf{R}$ such that $(Pf)(x)$ is the conditional expected value of $f(X_{n+1})$, given that $X_n = x$. In symbols, $(Pf)(x) = \int_{y\in\mathcal{X}} f(y)\,P(x,dy)$.

Definition. A Markov chain satisfies a *drift condition* (or, univariate geometric drift condition) if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$, such that

$$PV \;\leq\; \lambda V + b\mathbf{1}_C\,, \tag{10}$$

i.e. such that $\int_{\mathcal{X}} P(x,dy)V(y) \leq \lambda V(x) + b\mathbf{1}_C(x)$ for all $x \in \mathcal{X}$.

The main result guaranteeing geometric ergodicity is the following.

Theorem 9. *Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose the minorisation condition (8) is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition (10) is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$ with $V(x) < \infty$ for at least one (and hence for π-a.e.) $x \in \mathcal{X}$. Then then chain is geometrically ergodic.*

Theorem 9 is usually proved by complicated analytic arguments (see e.g. [59], [53], [7]). In Section 4, we describe a proof of Theorem 9 which uses direct coupling constructions instead. Note also that Theorem 9 provides no *quantitative* bounds on $M(x)$ or ρ, though this is remedied in Theorem 12 below.

Fact 10. In fact, it follows from Theorems 15.0.1, 16.0.1, and 14.3.7 of Meyn and Tweedie [53], and Proposition 1 of [68], that the minorisation condition (8) and drift condition (10) of Theorem 9 are equivalent (assuming ϕ-irreducibility and aperiodicity) to the apparently stronger property of "V-uniform ergodicity", i.e. that there is $C < \infty$ and $\rho < 1$ such that

$$\sup_{|f| \le V} |P^n f(x) - \pi(f)| \;\le\; C\, V(x)\, \rho^n\,, \qquad x \in \mathcal{X}\,,$$

where $\pi(f) = \int_{x \in \mathcal{X}} f(x)\, \pi(dx)$. That is, we can take $\sup_{|f| \le V}$ instead of just $\sup_{0 < f < 1}$ (compare Proposition 3 parts (a) and (b)), and we can let $M(x) = C\, V(x)$ in the geometric ergodicity bound. Furthermore, we always have $\pi(V) < \infty$. (The term "V-uniform ergodicity", as used in [53], perhaps also implies that $V(x) < \infty$ for all $x \in \mathcal{X}$, rather than just for π-a.e. $x \in \mathcal{X}$, though we do not consider that distinction further here.)

Open Problem #1. *Can direct coupling methods, similar to those used below to prove Theorem 9, also be used to provide an alternative proof of Fact 10? (For some progress in this direction, see [24].)*

Example 4. Here we consider a simple example of geometric ergodicity of Metropolis algorithms on $\mathbf{R}$ (see Mengersen and Tweedie [51], and [75]). Suppose that $\mathcal{X} = \mathbf{R}^+$ and $\pi_u(x) = e^{-x}$. We will use a symmetric (about x) proposal distribution $q(x, y) = q(|y - x|)$

with support contained in $[x-a, x+a]$. In this simple situation, a natural drift function to take is $V(x) = e^{cx}$ for some $c > 0$. For $x \geq a$, we compute:

$$PV(x) = \int_{x-a}^{x} V(y)q(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)dy\frac{\pi_u(y)}{\pi_u(x)}$$

$$+V(x)\int_{x}^{x+a} q(x,y)dy(1-\pi_u(y)/\pi_u(x))\,.$$

By the symmetry of q, this can be written as

$$\int_{x}^{x+a} I(x,y)\,q(x,y)\,dy\,,$$

where

$$I(x,y) = \frac{V(y)\pi_u(y)}{\pi_u(x)} + V(2x-y) + V(x)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right)$$

$$= e^{cx}\left[e^{(c-1)u} + e^{-cu} + 1 - e^{-u}\right] = e^{cx}\left[2-(1+e^{(c-1)u})(1-e^{-cu})\right]\,,$$

and where $u = y - x$. For $c < 1$, this is equal to $2(1-\epsilon)V(x)$ for some positive constant ϵ. Thus in this case we have shown that for all $x > a$

$$PV(x) \leq \int_{x}^{x+a} 2V(x)(1-\epsilon)q(x,y)dy = (1-\epsilon)V(x)\,.$$

Furthermore, it is easy to show that $PV(x)$ is bounded on $[0,a]$ and that $[0,a]$ is in fact a small set. Thus, we have demonstrated that the drift condition (10) holds. Hence, the algorithm is geometrically ergodic by Theorem 9. (It turns out that for such Metropolis algorithms, a certain condition, which essentially requires an exponential bound on the tail probabilities of $\pi(\cdot)$, is in fact *necessary* for geometric ergodicity; see [75].)

Implications of geometric ergodicity for central limit theorems are discussed in Section 5. In general, it believed by practitioners of MCMC that geometric ergodicity is a useful property. But does geometric ergodicity really matter? Consider the following examples.

Example 5. ([70]) Consider an independence sampler, with $\pi(\cdot)$ an Exponential(1) distribution, and $Q(x, \cdot)$ an Exponential(λ) distribution. Then if $0 < \lambda \leq 1$, the sampler is geometrically ergodic, has central limit theorems (see Section 5), and generally behaves fairly well even for very small λ. On the other hand, for $\lambda > 1$ the sampler fails to be geometrically ergodic, and indeed for $\lambda \geq 2$ it fails to have central limit theorems, and generally behaves quite poorly. For example, the simulations in [70] indicate that with $\lambda = 5$, when started in stationarity and averaged over the first million iterations, the sampler will usually return an average value of about 0.8 instead of 1, and then occasionally return a very large value instead, leading to very unstable behaviour. Thus, this is an example where the property of geometric ergodicity does indeed correspond to stable, useful convergence behaviour.

However, geometric ergodicity does not always guarantee a useful Markov chain algorithm, as the following two examples show.

Example 6. ("Witch's Hat", e.g. Matthews [50]) Let $\mathcal{X} = [0, 1]$, let $\delta = 10^{-100}$ (say), let $0 < a < 1 - \delta$, and let $\pi_u(\mathbf{x}) = \delta + \mathbf{1}_{[a,a+\delta]}(\mathbf{x})$. Then $\pi([a, a + \delta]) \approx 1/2$. Now, consider running a typical Metropolis algorithm on π_u. Unless $X_0 \in [a, a + \delta]$, or the sampler gets "lucky" and achieves $X_n \in [a, a + \delta]$ for some moderate n, then the algorithm will likely miss the tiny interval $[a, a + \delta]$ entirely, over any feasible time period. The algorithm will thus "appear" (to the naked eye or to any statistical test) to converge to the Uniform($\mathcal{X}$) distribution, even though Uniform($\mathcal{X}$) is very different from $\pi(\cdot)$. Nevertheless, this algorithm is still geometrically ergodic (in fact uniformly ergodic). So in this example, geometric ergodicity does not guarantee a well-behaved sampler.

Example 7. Let $\mathcal{X} = \mathbf{R}$, and let $\pi_u(x) = 1/(1 + x^2)$ be the (unnormalised) density of the Cauchy distribution. Then a random-walk Metropolis algorithm for π_u (with, say, $X_0 = 0$ and $Q(x, \cdot) = \text{Uniform}[x - 1, x + 1]$) is ergodic but is *not* geometrically ergodic. And, indeed, this sampler has very slow, poor convergence properties. On the other hand, let $\pi'_u(x) = \pi_u(x)\, \mathbf{1}_{|x| \leq 10^{100}}$, i.e. π'_u corresponds to π_u truncated at $\pm$ one googol. Then the same random-walk Metropolis algorithm for π'_u *is* geometrically ergodic, in fact uniformly

ergodic. However, the two algorithms are *indistinguishable* when run for any remotely feasible number of iterations. Thus, this is an example where geometric ergodicity does not in any way indicate improved performance of the algorithm.

In addition to the above two examples, there are also numerous examples of important Markov chains on *finite* state spaces (such as the single-site Gibbs sampler for the Ising model at low temperature on a large but finite grid) which are irreducible and aperiodic, and hence uniformly (and thus also geometrically) ergodic, but which converge to stationarity extremely slowly.

The above examples illustrate a limitation of *qualitative* convergence properties such as geometric ergodicity. It is thus desirable where possible to instead obtain *quantitative* bounds on Markov chain convergence. We consider this issue next.

3.5. Quantitative Convergence Rates.

In light of the above, we ideally want *quantitative* bounds on convergence rates, i.e. bounds of the form $\|P^n(x,\cdot) - \pi(\cdot)\| \leq g(x,n)$ for some *explicit* function $g(x,n)$, which (hopefully) is small for large n. Such questions now have a substantial history in MCMC, see e.g. [54], [79], [81], [48], [20], [76], [43], [44], [24], [11], [84], [28], [9], [85], [86].

We here present a result from [84], which follows as a special case of [24]; it is based on the approach of [79] while also taking into account a small improvement from [76].

Our result requires a *bivariate drift condition* of the form

$$\overline{P}h(x,y) \;\leq\; h(x,y) \,/\, \alpha\,, \qquad (x,y) \notin C \times C \tag{11}$$

for some function $h : \mathcal{X} \times \mathcal{X} \to [1,\infty)$ and some $\alpha > 1$, where

$$\overline{P}h(x,y) \;\equiv\; \int_{\mathcal{X}} \int_{\mathcal{X}} h(z,w)\, P(x,dz)\, P(y,dw)\,.$$

(Thus, $\overline{P}$ represents running two independent copies of the chain.) Of course, (11) is closely related to (10); for example we have the following (see also [79], and Proposition 2 of [86]):

Proposition 11. *Suppose the univariate drift condition (10) is satisfied for some* $V : \mathcal{X} \to [1, \infty]$, $C \subseteq \mathcal{X}$, $\lambda < 1$, *and* $b < \infty$. *Let* $d = \inf_{x \in C^c} V(x)$. *Then if* $d > [b/(1-\lambda)] - 1$, *then the bivariate drift condition (11) is satisfied for the same* C, *with* $h(x,y) = \frac{1}{2}[V(x) + V(y)]$ *and* $\alpha^{-1} = \lambda + b/(d+1) < 1$.

Proof. If $(x,y) \notin C \times C$, then either $x \notin C$ or $y \notin C$ (or both), so $h(x,y) \geq (1+d)/2$, and $PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$. Then

$$\overline{P}h(x,y) = \frac{1}{2}[PV(x) + PV(y)] \leq \frac{1}{2}[\lambda V(x) + \lambda V(y) + b]$$

$$= \lambda\, h(x,y) + b/2 \leq \lambda\, h(x,y) + (b/2)[h(x,y)/((1+d)/2)] = [\lambda + b/(1+d)]\, h(x,y)\,.$$

Furthermore, $d > [b/(1-\lambda)] - 1$ implies that $\lambda + b/(1+d) < 1$. ∎

Finally, we let

$$B_{n_0} \;=\; \max\left[1,\; \alpha^{n_0}(1-\epsilon) \sup_{C \times C} \overline{R}h\right], \tag{12}$$

where for $(x,y) \in C \times C$,

$$\overline{R}h(x,y) \;=\; \int_{\mathcal{X}} \int_{\mathcal{X}} (1-\epsilon)^{-2} h(z,w)\, (P^{n_0}(x,dz) - \epsilon\nu(dz))\, (P^{n_0}(y,dw) - \epsilon\nu(dw))\,.$$

In terms of these assumptions, we state our result as follows.

Theorem 12. *Consider a Markov chain on a state space* $\mathcal{X}$, *having transition kernel* P. *Suppose there is* $C \subseteq \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, *a probability distribution* $\nu(\cdot)$ *on* $\mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbf{N}$, *and* $\epsilon > 0$, *such that (8) and (11) hold. Define* B_{n_0} *by (12). Then for any joint initial distribution* $\mathcal{L}(X_0, X_0')$, *and any integers* $1 \leq j \leq k$, *if* $\{X_n\}$ *and* $\{X_n'\}$ *are two copies of the Markov chain started in the joint initial distribution* $\mathcal{L}(X_0, X_0')$, *then*

$$\|\mathcal{L}(X_k) - \mathcal{L}(X_k')\|_{TV} \;\leq\; (1-\epsilon)^j + \alpha^{-k}\, (B_{n_0})^{j-1}\, \mathbf{E}[h(X_0, X_0')]\,.$$

In particular, by choosing $j = \lfloor rk \rfloor$ *for sufficiently small* $r > 0$, *we obtain an explicit, quantitative convergence bound which goes to 0 exponentially quickly as* $k \to \infty$.

Theorem 12 is proved in Section 4. Versions of this theorem have been applied to various realistic MCMC algorithms, including for versions of the variance components model described earlier, resulting in bounds like $\|P^n(x,\cdot) - \pi(\cdot)\| < 0.01$ for $n = 140$ or $n = 3415$; see e.g. [81], and Jones and Hobert [44]. Thus, while it is admittedly hard work to apply Theorem 12 to realistic MCMC algorithms, it is indeed possible and often can establish rigorously that perfectly feasible numbers of iterations are sufficient to ensure convergence.

Remark. For complicated Markov chains, it might be difficult to apply Theorem 12 successfully. In such cases, MCMC practitioners instead use "convergence diagnostics", i.e. do statistical analysis of the realised output $X_1, X_2, \ldots$, to see if the distributions of X_n appear to be "stable" for large enough n. Many such diagnostics involve running the Markov chain repeatedly from different initial states, and checking if the chains all converge to approximately the same distribution (see e.g. Gelman and Rubin [31], and Cowles and Carlin [18]). This technique often works well in practice. However, it provides no rigorous guarantees and can sometimes be fooled into prematurely claiming convergence (see e.g. [50]), as is likely to happen for the examples at the end of Section 3. Furthermore, convergence diagnostics can also introduce bias into the resulting estimates (see [19]). Overall, despite the extensive theory surveyed herein, the "convergence time problem" remains largely unresolved for practical application of MCMC. (This is also the motivation for "perfect MCMC" algorithms, originally developed by Propp and Wilson [62] and not discussed here; for further discussion see e.g. Kendall and Møller [45], Thönnes [91], and Fill et al. [27].)

4. Convergence Proofs using Coupling Constructions.

In this section, we prove some of the theorems stated earlier. There are of course many methods available for bounding convergence of Markov chains, appropriate to various settings (see e.g. [1], [21], [87], [2], [89], and Subsection 5.4 herein), including the setting of large but finite state spaces that often arises in computer science (see e.g. Sinclair [87] and Randall [63]) but is not our emphasis here. In this section, we focus on the method of *coupling*, which seems particularly well-suited to analysing MCMC algorithms on general

(uncountable) state spaces. It is also particularly well-suited to incorporating small sets (though small sets can also be combined with *regeneration theory*, see e.g. [8], [4], [56], [37]). Some of the proofs below are new, and avoid many of the long analytic arguments of some previous proofs (e.g. Nummelin [59], and Meyn and Tweedie [53]).

4.1. The Coupling Inequality.

The basic idea of coupling is the following. Suppose we have two *random variables* X and Y, defined jointly on some space $\mathcal{X}$. If we write $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ for their respective probability distributions, then we can write

$$\begin{aligned}
\|\mathcal{L}(X) - \mathcal{L}(Y)\| &= \sup_A |P(X \in A) - P(Y \in A)| \\
&= \sup_A |P(X \in A, X = Y) + P(X \in A, X \neq Y) \\
&\qquad - P(Y \in A, Y = X) - P(Y \in A, Y \neq X)| \\
&= \sup_A |P(X \in A, X \neq Y) - P(Y \in A, Y \neq X)| \\
&\leq P(X \neq Y)\,,
\end{aligned}$$

so that

$$\|\mathcal{L}(X) - \mathcal{L}(Y)\| \leq P(X \neq Y)\,. \tag{13}$$

That is, *the variation distance between the laws of two random variables is bounded by the probability that they are unequal.* For background, see e.g. Pitman [61], Lindvall [47], and Thorisson [90].

4.2. Small Sets and Coupling.

Suppose now that C is a small set. We shall use the following coupling construction, which is essentially the "splitting technique" of Nummelin [58] and Athreya and Ney [8]; see also Nummelin [59], and Meyn and Tweedie [53]. The idea is to run two copies $\{X_n\}$ and $\{X'_n\}$ of the Markov chain, each of which marginally follows the updating rules $P(x, \cdot)$, but whose joint construction (using C) gives them as high a probability as possible of becoming equal to each other.

THE COUPLING CONSTRUCTION:

Start with $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, and $n = 0$, and repeat the following loop forever.

Beginning of Loop. Given X_n and X'_n:

1. If $X_n = X'_n$, choose $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$, and replace n by $n+1$.
2. Else, if $(X_n, X'_n) \in C \times C$, then:
   (a) w.p. ϵ, choose $X_{n+n_0} = X'_{n+n_0} \sim \nu(\cdot)$;
   (b) else, w.p. $1-\epsilon$, conditionally independently choose

$$X_{n+n_0} \sim \frac{1}{1-\epsilon}[P^{n_0}(X_n, \cdot) - \epsilon\,\nu(\cdot)]\,,$$

$$X'_{n+n_0} \sim \frac{1}{1-\epsilon}[P^{n_0}(X'_n, \cdot) - \epsilon\,\nu(\cdot)]\,.$$

   In the case $n_0 > 1$, for completeness go back and construct $X_{n+1}, \ldots, X_{n+n_0-1}$ from their correct conditional distributions given X_n and X_{n+n_0}, and similarly (and conditionally independently) construct $X'_{n+1}, \ldots, X'_{n+n_0-1}$ from their correct conditional distributions given X'_n and X'_{n+n_0}. In any case, replace n by $n+n_0$.
3. Else, conditionally independently choose $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$, and replace n by $n+1$.

Then return to Beginning of Loop.

Under this construction, it is easily checked that X_n and X'_n are each marginally updated according to the correct transition kernel P. It follows that $\mathbf{P}[X_n \in A] = P^n(x, \cdot)$ and $\mathbf{P}[X'_n \in A] = \pi(A)$ for all n. Moreover the two chains are run independently until they both enter C at which time the minorisation *splitting* construction (step 2) is utilised. Without such a construction, on uncountable state spaces, we would not be able to ensure successful coupling of the two processes.

The *coupling inequality* then says that $\|P^n(x, \cdot) - \pi(\cdot)\| \le \mathbf{P}[X_n \neq X'_n]$. The question is, can we use this to obtain useful bounds on $\|P^n(x, \cdot) - \pi(\cdot)\|$? In fact, we shall now provide proofs (nearly self-contained) of all of the theorems stated earlier, in terms of this coupling construction. This allows for intuitive understanding of the theorems, while also avoiding various analytic technicalities of the previous proofs of some of these theorems.

4.3. Proof of Theorem 8.

In this case, $C = \mathcal{X}$, so every n_0 iterations we have probability at least ϵ of making X_n and X'_n equal. It follows that if $n = n_0 m$, then $\mathbf{P}[X_n \neq X'_n] \leq (1-\epsilon)^m$. Hence, from the coupling inequality, $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^m = (1-\epsilon)^{n/n_0}$ in this case. It then follows from Proposition 3(c) that $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for any n. ∎

4.4. Proof of Theorem 12.

We follow the general outline of [84]. We again begin by assuming that $n_0 = 1$ in the minorisation condition for the small set C (and thus write B_{n_0} as B), and indicate at the end what changes are required if $n_0 > 1$.

Let

$$N_k \;=\; \#\{m \;:\; 0 \leq m \leq k,\; (X_m, X'_m) \in C \times C\}\,,$$

and let $\tau_1, \tau_2, \ldots$ be the times of the successive visits of $\{(X_n, X'_n)\}$ to $C \times C$. Then for any integer j with $1 \leq j \leq k$,

$$\mathbf{P}[X_k \neq X'_k] \;=\; \mathbf{P}[X_k \neq X'_k,\; N_{k-1} \geq j] \;+\; \mathbf{P}[X_k \neq X'_k,\; N_{k-1} < j]\,. \tag{14}$$

Now, the event $\{X_k \neq X'_k,\; N_{k-1} \geq j\}$ is contained in the event that the first j coin flips all came up tails. Hence, $\mathbf{P}[X_k \neq X'_k,\; N_{k-1} \geq j] \leq (1-\epsilon)^j$. which bounds the first term in (14).

To bound the second term in (14), let

$$M_k \;=\; \alpha^k B^{-N_{k-1}} h(X_k, X'_k)\, \mathbf{1}(X_k \neq X'_k)\,, \qquad k = 0, 1, 2, \ldots$$

(where $N_{-1} = 0$).

Lemma 13. *We have*

$$\mathbf{E}[M_{k+1} \,|\, X_0, \ldots, X_k, X'_0, \ldots, X'_k] \;\leq\; M_k\,,$$

i.e. $\{M_k\}$ *is a supermartingale.*

Proof. If $(X_k, X'_k) \notin C \times C$, then $N_k = N_{k-1}$, so

$$\begin{aligned}
&\mathbf{E}[M_{k+1} \,|\, X_0, ..., X_k, X'_0, ..., X'_k] \\
&\quad = \alpha^{k+1} B^{-N_{k-1}} \mathbf{E}[h(X_{k+1}, X'_{k+1})\, \mathbf{1}(X_{k+1} \neq X'_{k+1}) \,|\, X_k, X'_k] \\
&\qquad\qquad \text{(since our coupling construction is Markovian)} \\
&\quad \leq \alpha^{k+1} B^{-N_{k-1}} \mathbf{E}[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k]\, \mathbf{1}(X_k \neq X'_k) \\
&\quad = M_k\, \alpha\, \mathbf{E}[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k] \,/\, h(X_k, X'_k) \\
&\quad \leq M_k\,,
\end{aligned}$$

by (11). Similarly, if $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$, so assuming $X_k \neq X'_k$ (since if $X_k = X'_k$, then the result is trivial), we have

$$\begin{aligned}
&\mathbf{E}[M_{k+1} \,|\, X_0, ..., X_k, X'_0, ..., X'_k] \\
&\quad = \alpha^{k+1} B^{-N_{k-1}-1} \mathbf{E}[h(X_{k+1}, X'_{k+1})\, \mathbf{1}(X_{k+1} \neq X'_{k+1}) \,|\, X_k.X'_k] \\
&\quad = \alpha^{k+1} B^{-N_{k-1}-1} (1-\epsilon)(\bar{R}h)(X_k, X'_k) \\
&\quad = M_k \alpha B^{-1} (1-\epsilon)(\bar{R}h)(X_k, X'_k)/h(X_k, X'_k) \\
&\quad \leq M_k\,,
\end{aligned}$$

by (12). Hence, $\{M_k\}$ is a supermartingale. ∎

To proceed, we note that since $B \geq 1$,

$$\mathbf{P}[X_k \neq X'_k,\, N_{k-1} < j] \;=\; \mathbf{P}[X_k \neq X'_k,\, N_{k-1} \leq j-1] \;\leq\; \mathbf{P}[X_k \neq X'_k,\, B^{-N_{k-1}} \geq B^{-(j-1)}]$$

$$\begin{aligned}
&= \mathbf{P}[\mathbf{1}(X_k \neq X'_k)\, B^{-N_{k-1}} \geq B^{-(j-1)}] \\
&\leq B^{j-1} \mathbf{E}[\mathbf{1}(X_k \neq X'_k)\, B^{-N_{k-1}}] \qquad \text{(by Markov's inequality)} \\
&\leq B^{j-1} \mathbf{E}[\mathbf{1}(X_k \neq X'_k)\, B^{-N_{k-1}}\, h(X_k, X'_k)] \qquad \text{(since } h \geq 1\text{)} \\
&= \alpha^{-k} B^{j-1} \mathbf{E}[M_k] \qquad \text{(by defn of } M_k\text{)} \\
&\leq \alpha^{-k} B^{j-1} \mathbf{E}[M_0] \qquad \text{(since } \{M_k\} \text{ is supermartingale)}
\end{aligned}$$

$$= \alpha^{-k} B^{j-1} \mathbf{E}[h(X_0, X_0')] \qquad \text{(by defn of } M_0\text{)}\,.$$

Theorem 12 now follows (in the case $n_0 = 1$), by combining these two bounds with (14) and (13).

Finally, we consider the changes required if $n_0 > 1$. In this case, the main change is that we do not wish to count visits to $C \times C$ during which the joint chain could not try to couple, i.e. visits which correspond to the "filling in" times for going back and constructing $X_{n+1}, \ldots, X_{n+n_0}$ [and similarly for X'] in step 2 of the coupling construction. Thus, we instead let N_k count the number of visits to $C \times C$, and $\{\tau_i\}$ the actual visit times, *avoiding* all such "filling in" times. Also, we replace N_{k-1} by N_{k-n_0} in (14) and in the definition of M_k. Finally, what is a supermartingale is not $\{M_k\}$ but rather $\{M_{t(k)}\}$, where $t(k)$ is the latest time $\leq k$ which does not correspond to a "filling in" time. (Thus, $t(k)$ will take the value k, unless the joint chain visited $C \times C$ at some time between $k - n_0$ and $k - 1$.) With these changes, the proof goes through just as before. ∎

4.5. Proof of Theorem 9.

Here we give a direct coupling proof of Theorem 9, thereby somewhat avoiding the technicalities of e.g. Meyn and Tweedie [53] (though admittedly with a slightly weaker conclusion; see Fact 10). Our approach shall be to make use of Theorem 12. To begin, set $h(x, y) = \frac{1}{2}[V(x) + V(y)]$. Our proof will use the following technical result.

Lemma 14. *We may assume without loss of generality that*

$$\sup_{x \in C} V(x) \; < \; \infty\,. \tag{15}$$

Specifically, given a small set C and drift function V satisfying (8) and (10), we can find a small set $C_0 \subseteq C$ such that (8) and (10) still hold (with the same n_0 and ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that (15) also holds.

Proof. Let λ and b be as in (10). Choose δ with $0 < \delta < 1 - \lambda$, let $\lambda_0 = 1 - \delta$, let $K = b/(1 - \lambda - \delta)$, and set

$$C_0 \;=\; C \cap \{x \in \mathcal{X} : V(x) \leq K\}\,.$$

Then clearly (8) continues to hold on C_0, since $C_0 \subseteq C$. It remains to verify that (10) holds with C replaced by C_0, and λ replaced by λ_0. Now, (10) clearly holds for $x \in C_0$ and $x \notin C$, by inspection. Finally, for $x \in C \setminus C_0$, we have $V(x) \geq K$, and so using the original drift condition (10), we have

$$(PV)(x) \;\leq\; \lambda\, V(x) + b\, \mathbf{1}_C(x) \;=\; (1-\delta)V(x) - (1-\lambda-\delta)V(x) + b$$

$$\leq\; (1-\delta)V(x) - (1-\lambda-\delta)K + b \;=\; (1-\delta)V(x) \;=\; \lambda_0\, V(x)\,,$$

showing that (10) still holds, with C replaced by C_0 and λ replaced by λ_0. ∎

As an aside, we note that in Lemma 14, it may not be possible to satisfy (15) by instead modifying V and leaving C unchanged:

Proposition 15. *There exists a geometrically ergodic Markov chain, with small set C and drift function V satisfying (8) and (10), such that there does not exist a drift function $V_0 : \mathcal{X} \to [0, \infty]$ with the property that upon replacing V by V_0, (8) and (10) continue to hold, and (15) also holds.*

Proof. Consider the Markov chain on $\mathcal{X} = (0, \infty)$, defined as follows. For $x \geq 2$, $P(x, \cdot) = \delta_{x-1}(\cdot)$, a point-mass at $x - 1$. For $1 < x < 2$, $P(x, \cdot)$ is uniform on $[1/2,\ 1]$. For $0 < x \leq 1$, $P(x, \cdot) = \frac{1}{2}\,\lambda(\cdot) + \frac{1}{2}\,\delta_{h(x)}(\cdot)$, where λ is Lebesgue measure on $(0, 1)$, and $h(x) = 1 + \sqrt{\log(1/x)}$.

For this chain, the interval $C = (0, 1)$ is clearly $(1,\ 1/2,\ \lambda)$-small. Furthermore, since $\int_0^1 \sqrt{\log(1/x)}\, dx = \sqrt{\pi}/2 < \infty$, the return times to C have finite mean, so the chain has a stationary distribution by standard renewal theory arguments (see e.g. Asmussen [4]). In addition, with drift function $V(x) = \max(e^x, x^{-1/2})$, we can compute $(PV)(x)$ explicitly,

and verify directly that (say) $PV(x) \leq 0.8\, V(x) + 4\, \mathbf{1}_C(x)$ for all $x \in \mathcal{X}$, thus verifying (10) with $\lambda = 0.8$ and $b = 4$. Hence, by Theorem 9, the chain is geometrically ergodic.

On the other hand, suppose we had a some drift function V_0 satisfying (10), such that $\sup_{x \in C} V_0(x) < \infty$. Then since $PV_0(x) = \frac{1}{2} V_0(h(x)) + \frac{1}{2} \int_0^1 V_0(y)\, dy$, this would imply that $\sup_{x \in C} V_0((h(x)) < \infty$, i.e. that $V_0(h(x))$ is bounded for all $0 < x \leq 1$, which would in turn imply that V_0 were bounded everywhere on $\mathcal{X}$. But then Fact 10 would imply that the chain is uniformly ergodic, which it clearly is not. This gives a contradiction. ■

Thus, for the remainder of this proof, we can (and do) assume that (15) holds. This, together with (10), implies that

$$\sup_{(x,y) \in C \times C} \overline{R} h(x,y) \;<\; \infty\,, \tag{16}$$

which in turn ensures that the quantity B_{n_0} of (12) is finite.

To continue, let $d = \inf_{C^c} V$. Then we see from Proposition 11 that the bivariate drift condition (11) will hold, *provided* that $d > b/(1-\lambda) - 1$. In that case, Theorem 9 follows immediately (in fact, in a *quantitative* version) by combining Proposition 11 with Theorem 12.

However, if $d \leq b/(1-\lambda) - 1$, then this argument does not go through. This is not merely a technicality; the condition $d > b/(1-\lambda) - 1$ ensures that the chain is *aperiodic*, and without this condition we must somehow use the assumption of aperiodicity more directly in the proof.

Our plan shall be to *enlarge* C so that the new value of d satisfies $d > b/(1-\lambda) - 1$, and to use aperiodicity to show that C remains a small set (i.e., that (8) still holds though perhaps for uncontrollably larger n_0 and smaller $\epsilon > 0$). Theorem 9 will then follow from Proposition 11 and Theorem 12 as above. (Note that we will have no direct control over the new values of n_0 and ϵ, which is why this approach does not provide a *quantitative* convergence rate bound.)

To proceed, choose any $d' > b/(1-\lambda) - 1$, let $S = \{x \in \mathcal{X};\ V(x) \leq d'\}$, and set $C' = C \cup S$. This ensures that $\inf_{x \in C'^c} V(x) \geq d' > b/(1-\lambda) - 1$. Furthermore, since V is bounded on S by construction, we see that (15) will still hold with C replaced by C'. It

then follows from (16) and (10) that we will still have $B_{n_0} < \infty$ even upon replacing C by C'. Thus, Theorem 9 will follow from Proposition 11 and Theorem 12 if we can prove:

Lemma 16. *C' is a small set.*

To prove Lemma 16, we use the notion of "petite set", following [53].

Definition. A subset $C \subseteq \mathcal{X}$ is *petite* (or, (n_0, ϵ, ν)-petite), relative to a Markov chain P, if there exists a positive integer n_0, $\epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that

$$\sum_{i=1}^{n_0} P^i(x, \cdot) \;\geq\; \epsilon\, \nu(\cdot) \qquad x \in C\,. \tag{17}$$

Intuitively, the definition of petite set is like that of small set, except that it allows the different states in C to cover the minorisation measure $\epsilon\,\nu(\cdot)$ at different times i. Obviously, any small set is petite. The converse is false in general, as the petite set condition does not itself rule out *periodic* behaviour of the chain (for example, perhaps some of the states $x \in C$ cover $\epsilon\,\nu(\cdot)$ only at odd times, and others only at even times). However, for an *aperiodic*, ϕ-irreducible Markov chain, we have the following result, whose proof is presented in the Appendix.

Lemma 17. *(Meyn and Tweedie [53], Theorem 5.5.7) For an aperiodic, ϕ-irreducible Markov chain, all petite sets are small sets.*

To make use of Lemma 17, we use the following.

Lemma 18. *Let $C' = C \cup S$ where $S = \{x \in \mathcal{X};\; V(x) \leq d'\}$ for some $d' < \infty$, as above. Then C' is petite.*

Proof. To begin, choose N large enough that $r \equiv 1 - \lambda^N d' > 0$. Let $\tau_C = \inf\{n \geq 1;\; X_n \in C\}$ be the first return time to C. Let $Z_n = \lambda^{-n} V(X_n)$, and let $W_n = Z_{\min(n,\tau_C)}$. Then the drift condition (10) implies that W_n is a supermartingale. Indeed, if $\tau_C \leq n$, then

$$\mathbf{E}[W_{n+1} \,|\, X_0, X_1, \ldots, X_n] \;=\; \mathbf{E}[Z_{\tau_c} \,|\, X_0, X_1, \ldots, X_n] \;=\; Z_{\tau_c} \;=\; W_n\,,$$

while if $\tau_C > n$, then $X_n \notin C$, so using (10),

$$\begin{aligned}\mathbf{E}[W_{n+1} \,|\, X_0, X_1, ...X_n] &= \lambda^{-(n+1)}(PV)(X_n) \\ &\leq \lambda^{-(n+1)}\lambda V(X_n) \\ &= \lambda^{-n} V(X_n) \\ &= W_n \,.\end{aligned}$$

Hence, for $x \in S$, using Markov's inequality and the fact that $V \geq 1$,

$$\mathbf{P}[\tau_C \geq N \,|\, X_0 = x] = \mathbf{P}[\lambda^{-\tau_C} \geq \lambda^{-N} \,|\, X_0 = x]$$

$$\leq \lambda^N \,\mathbf{E}[\lambda^{-\tau_C} \,|\, X_0 = x] \leq \lambda^N \,\mathbf{E}[W_{\tau_C} \,|\, X_0 = x]$$

$$\leq \lambda^N \,\mathbf{E}[W_0 \,|\, X_0 = x] = \lambda^N \, V(x) \leq \lambda^N \, d' \,,$$

so that $\mathbf{P}[\tau_C < N \,|\, X_0 = x] \geq r$.

On the other hand, recall that C is $(n_0, \epsilon, \nu(\cdot))$-small, so that $P^{n_0}(x, \cdot) \geq \epsilon \, \nu(\cdot)$ for $x \in C$. It follows that for $x \in S$, $\sum_{i=1+n_0}^{N+n_0} P^i(x, \cdot) \geq r \, \epsilon \, \nu(\cdot)$. Hence, for $x \in S \cup C$, $\sum_{i=n_0}^{N+n_0} P^i(x, \cdot) \geq r \, \epsilon \, \nu(\cdot)$. This shows that $S \cup C$ is petite. ∎

Combining Lemmas 18 and 17, we see that C' must be small, proving Lemma 16, and hence proving Theorem 9. ∎

4.6. Proof of Theorem 4.

Theorem 4 does not assume the existence of any small set C, so it is not clear how to make use of our coupling construction in this case. However, help is at hand in the form of a remarkable result about the existence of small sets, due to Jain and Jameson [40] (see also Orey [60]). We shall not prove it here; for modern proofs see e.g. [59], p. 16, or [53], Theorem 5.2.2. The key idea (see e.g. Meyn and Tweedie [53], Theorem 5.2.1) is to extract the part of $P^{n_0}(x, \cdot)$ which is absolutely continuous with respect to the measure ϕ, and then to find a C with $\phi(C) > 0$ such that this density part is at least $\delta > 0$ throughout C.

Theorem 19. *(Jain and Jameson [40]) Every ϕ-irreducible Markov chain, on a state space with countably generated σ-algebra, contains a small set $C \subseteq \mathcal{X}$ with $\phi(C) > 0$. (In fact, each $B \subseteq \mathcal{X}$ with $\phi(B) > 0$ in turn contains a small set $C \subseteq B$ with $\phi(C) > 0$.) Furthermore, the minorisation measure $\nu(\cdot)$ may be taken to satisfy $\nu(C) > 0$.*

In terms of our coupling construction, if we can show that the pair (X_n, X'_n) will hit $C \times C$ infinitely often, then they will have infinitely many opportunities to couple, with probability $\geq \epsilon > 0$ of coupling each time. Hence, they will eventually couple with probability 1, thus proving Theorem 4.

We prove this following the outline of [83]. We begin with a lemma about return probabilities:

Lemma 20. *Consider a Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$. Suppose that for some $A \subseteq \mathcal{X}$, we have $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$. Then for π-almost-every $x \in \mathcal{X}$, $\mathbf{P}_x(\tau_A < \infty) = 1$.*

Proof. Suppose to the contrary that the conclusion does not hold, i.e. that

$$\pi\big\{x \in \mathcal{X} : \ \mathbf{P}_x(\tau_A = \infty) > 0\big\} > 0\,. \tag{18}$$

Then we make the following claims (proved below):

Claim 1. Condition (18) implies that there are constants $\ell, \ell_0 \in \mathbf{N}$, $\delta > 0$, and $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, such that

$$\mathbf{P}_x\left(\tau_A = \infty,\ \sup\{k \geq 1;\ X_{k\ell_0} \in B\} < \ell\right) \geq \delta\,, \qquad x \in B\,.$$

Claim 2. Let B, ℓ, ℓ_0, and δ be as in Claim 1. Let $L = \ell\ell_0$, and let $S = \sup\{k \geq 1;\ X_{kL} \in B\}$, using the convention that $S = -\infty$ if the set $\{k \geq 1;\ X_{kL} \in B\}$ is empty. Then for all integers $1 \leq r \leq j$,

$$\int_{x \in \mathcal{X}} \pi(dx)\, \mathbf{P}_x[S = r,\ X_{jL} \notin A] \ \geq\ \pi(B)\, \delta\,.$$

Assuming the claims, we complete the proof as follows. We have by stationarity that for any $j \in \mathbf{N}$,

$$\pi(A^C) \ =\ \int_{x \in \mathcal{X}} \pi(dx)\, P^{jL}(x, A^C) \ =\ \int_{x \in \mathcal{X}} \pi(dx)\, \mathbf{P}_x[X_{jL} \notin A]$$

$$\geq \sum_{r=1}^{j} \int_{x\in\mathcal{X}} \pi(dx)\, \mathbf{P}_x[S = r,\ X_{jL} \notin A] \ \geq \ \sum_{r=1}^{j} \pi(B)\, \delta \ = \ j\, \pi(B)\, \delta\,.$$

For $j > 1\,/\,\pi(B)\,\delta$, this gives $\pi(A^C) > 1$, which is impossible. This gives a contradiction, and hence completes the proof of Lemma 20, subject to the proofs of Claims 1 and 2 below. ∎

Proof of Claim 1. By (18), we can find δ_1 and a subset $B_1 \subseteq \mathcal{X}$ with $\pi(B_1) > 0$, such that $\mathbf{P}_x(\tau_A < \infty) \leq 1 - \delta_1$ for all $x \in B_1$. On the other hand, since $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$, we can find $\ell_0 \in \mathbf{N}$ and $\delta_2 > 0$ and $B_2 \subseteq B_1$ with $\pi(B_2) > 0$ and with $P^{\ell_0}(x, A) \geq \delta_2$ for all $x \in B_2$.

Set $\eta = \#\{k \geq 1;\ X_{k\ell_0} \in B_2\}$. Then for any $r \in \mathbf{N}$ and $x \in \mathcal{X}$, we have $\mathbf{P}_x(\tau_A = \infty,\ \eta = r) \leq (1 - \delta_2)^r$. In particular, $\mathbf{P}_x(\tau_A = \infty,\ \eta = \infty) = 0$. Hence, for $x \in B_2$, we have

$$\mathbf{P}_x(\tau_A = \infty,\ \eta < \infty) \ = \ 1 - \mathbf{P}_x(\tau_A = \infty,\ \eta = \infty) - \mathbf{P}_x(\tau_A < \infty) \geq 1 - 0 - (1 - \delta_1) = \delta_1\,.$$

Hence, there is $\ell \in \mathbf{N}$, $\delta > 0$, and $B \subseteq B_2$ with $\pi(B) > 0$, such that

$$\mathbf{P}_x\left(\tau_A = \infty,\ \sup\{k \geq 1;\ X_{k\ell_0} \in B_2\} < \ell\right) \geq \delta\,, \qquad x \in B\,.$$

Finally, since $B \subseteq B_2$, we have $\sup\{k \geq 1;\ X_{k\ell_0} \in B_2\} \geq \sup\{k \geq 1;\ X_{k\ell_0} \in B\}$, thus establishing the claim. ∎

Proof of Claim 2. We compute using stationarity and then Claim 1 that

$$\int_{x\in\mathcal{X}} \pi(dx)\, \mathbf{P}_x[S = r,\ X_{jL} \notin A]$$

$$= \int_{x\in\mathcal{X}} \pi(dx) \int_{y\in B} P^{rL}(x, dy)\, \mathbf{P}_y[S = -\infty,\ X_{(j-r)L} \notin A]$$

$$= \int_{y\in B}\int_{x\in\mathcal{X}} \pi(dx)P^{rL}(x,dy)\,\mathbf{P}_y[S=-\infty,\ X_{(j-r)L}\notin A]$$
$$= \int_{y\in B} \pi(dy)\,\mathbf{P}_y[S=-\infty,\ X_{(j-r)L}\notin A]$$
$$\geq \int_{y\in B}\pi(dy)\,\delta = \pi(B)\,\delta\,. \qquad \blacksquare$$

To proceed, we let C be a small set as in Theorem 19. Consider again the coupling construction $\{(X_n,Y_n)\}$. Let $G\subseteq\mathcal{X}\times\mathcal{X}$ be the set of (x,y) for which $\mathbf{P}_{(x,y)}\big(\exists\, n\geq 1;\ X_n=Y_n\big)=1$. From the coupling construction, we see that if $(X_0,X_0')\equiv(x,X_0')\in G$, then $\lim_{n\to\infty}\mathbf{P}[X_n=X_n']=1$, so that $\lim_{n\to\infty}\|P^n(x,\cdot)-\pi(\cdot)\|=0$, proving Theorem 4. Hence, it suffices to show that for π-a.e. $x\in\mathcal{X}$, we have $\mathbf{P}[(x,X_0')\in G]=1$.

Let G be as above, let $G_x=\{y\in\mathcal{X};\ (x,y)\in G\}$ for $x\in\mathcal{X}$, and let $\overline{G}=\{x\in\mathcal{X};\ \pi(G_x)=1\}$. Then Theorem 4 follows from:

Lemma 21. $\pi(\overline{G})=1$.

Proof. We first prove that $(\pi\times\pi)(G)=1$. Indeed, since $\nu(C)>0$ by Theorem 19, it follows from Lemma 34 that, from any $(x,y)\in\mathcal{X}\times\mathcal{X}$, the joint chain has positive probability of eventually hitting $C\times C$. It then follows by applying Lemma 20 to the joint chain, that the joint chain will return to $C\times C$ with probability 1 from $(\pi\times\pi)$-a.e. $(x,y)\notin C\times C$. Once the joint chain reaches $C\times C$, then conditional on not coupling, the joint chain will update from $\overline{R}$ which must be absolutely continuous with respect to $\pi\times\pi$, and hence (again by Lemma 20) will return again to $C\times C$ with probability 1. Hence, the joint chain will repeatedly return to $C\times C$ with probability 1, until such time as $X_n=X_n'$. And by the coupling construction, each time the joint chain is in $C\times C$, it has probability $\geq\epsilon$ of then forcing $X_n=X_n'$. Hence, eventually we will have $X_n=X_n'$, thus proving that $(\pi\times\pi)(G)=1$.

Now, if we had $\pi(\overline{G})<1$, then we would have

$$(\pi\times\pi)(G^C) \;=\; \int_{\mathcal{X}}\pi(dx)\pi(G_x^C) \;=\; \int_{\overline{G}^C}\pi(dx)[1-\pi(G_x)] \;>\; 0\,,$$

contradicting the fact that $(\pi\times\pi)(G)=1$. $\blacksquare$

5. Central Limit Theorems for Markov Chains.

Suppose $\{X_n\}$ is a Markov chain on a state space $\mathcal{X}$ which is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$. Assume the chain begins in stationarity, i.e. that $X_0 \sim \pi(\cdot)$. Let $h : \mathcal{X} \to \mathbf{R}$ be some functional with finite stationary mean $\pi(h) \equiv \int_{x \in \mathcal{X}} h(x)\, \pi(dx)$.

We say that h satisfies a Central Limit Theorem (CLT) (or, $\sqrt{n}$-CLT) if there is some $\sigma^2 < \infty$ such that the normalised sum $n^{-1/2} \sum_{i=1}^{n} [h(X_i) - \pi(h)]$ converges weakly to a $N(0,\, \sigma^2)$ distribution. (We allow for the special case $\sigma^2 = 0$, corresponding to convergence to the constant 0.) It then follows under reversibility or uniform integrability assumptions (see e.g. [46], [15]) that

$$\sigma^2 = \lim_{n \to \infty} \frac{1}{n} \mathbf{E}\Big[\Big(\sum_{i=1}^{n} [h(X_i) - \pi(h)] \Big)^2 \Big] , \tag{19}$$

and also $\sigma^2 = \tau\, \mathbf{Var}_\pi(h)$, where $\tau = \sum_{k \in \mathbf{Z}} \mathrm{Corr}(h(X_0), h(X_k))$ is the *integrated autocorrelation time*. (In the reversible case this is also related to spectral measures; see e.g. [46], [34], [68].) Clearly $\sigma^2 < \infty$ requires that $\mathbf{Var}_\pi(h) < \infty$, i.e. that $\pi(h^2) < \infty$.

Such CLTs are helpful for understanding the *errors* which arise from Monte Carlo estimation, and are thus the subject of considerable discussion in the MCMC literature (e.g. [34], [92], [15], [53], [75], [37], [6], [42]).

5.1. A Negative Result.

One might expect that CLTs always hold when $\pi(h^2)$ is finite, but this is false. For example, it is shown in [65] that Metropolis-Hastings algorithms whose acceptance probabilities are too low may get so "stuck" that $\tau = \infty$ and they will not have a $\sqrt{n}$-CLT. More specifically, the following is proved:

Theorem 22. *Consider a reversible Markov chain, beginning in its stationary distribution $\pi(\cdot)$, and let $r(x) = \mathbf{P}[X_{n+1} = X_n \,|\, X_n = x]$. Then if*

$$\lim_{n \to \infty} \; n\, \pi\big([h - \pi(h)]^2\, r^n \big) \;=\; \infty \,, \tag{20}$$

then a $\sqrt{n}$-CLT does <u>not</u> hold for h.

Proof. We compute directly from (19) that

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\Big[\Big(\sum_{i=1}^{n}[h(X_i) - \pi(h)\Big)^2\Big]$$

$$\geq \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\Big[\Big(\sum_{i=1}^{n}[h(X_i) - \pi(h)\Big)^2 \mathbf{1}(X_0 = X_1 = \ldots = X_n)\Big]$$

$$= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\Big[\Big(n[h(X_0) - \pi(h)]\Big)^2 r(X_0)^n\Big]$$

$$= \lim_{n\to\infty} n\,\pi\Big([h - \pi(h)]^2\, r^n\Big) = \infty\,,$$

by (20). Hence, a $\sqrt{n}$-CLT cannot exist. ∎

In particular, Theorem 22 is used in [65] to prove that for the independence sampler with target $\mathbf{Exp}(1)$ and i.i.d. proposals $\mathbf{Exp}(\lambda)$, the identity function has no $\sqrt{n}$-CLT for any $\lambda \geq 2$.

The question then arises of what conditions on the Markov chain transitions, and on the functional h, guarantee a $\sqrt{n}$-CLT for h.

5.2. Conditions Guaranteeing CLTs.

Here we present various positive results about the existence of CLTs. Some, though not all, of these results are then proved in the following two sections.

For i.i.d. samples, classical theory guarantees a CLT provided the second moments are finite (e.g. [13], Theorem 27.1; [82], p. 110). For *uniformly ergodic* chains, an identical result exists; it is shown in Corollary 4.2(ii) of Cogburn [17] (cf. Theorem 5 of Tierney [92]) that:

Theorem 23. *If a Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.*

If a chain is just *geometrically ergodic* but not uniformly ergodic, then a similar result holds under the slightly stronger assumption of a finite $2 + \delta$ moments. That is, it is shown in Theorem 18.5.3 of Ibragimov and Linnik [39] (see also Theorem 2 of Chan and Geyer [15], and Theorem 2 of Hobert et al. [37]) that:

Theorem 24. *If a Markov chain with stationary distribution $\pi(\cdot)$ is geometrically ergodic, then a $\sqrt{n}$-CLT holds for h whenever $\pi(|h|^{2+\delta}) < \infty$ for some $\delta > 0$.*

It follows, for example, that the independence sampler example mentioned above (which fails to have a $\sqrt{n}$-CLT, but which has finite moments of all orders) is not geometrically ergodic.

It is shown in Corollary 3 of [68] that Theorem 24 can be strengthened if the chain is *reversible*:

Theorem 25. *If the Markov chain is geometrically ergodic and reversible, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.*

Comparing Theorems 25 and 24 leads to the following yes-or-no open question (see [6]):

Open Problem #2. *Consider a Markov chain which is geometrically ergodic, but not necessarily reversible. Let $h : \mathcal{X} \to \mathbf{R}$ with $\pi(h^2) < \infty$. Does a $\sqrt{n}$-CLT necessarily exist for h? [However, see the Note at the end of this paper.]*

To prove positive results about this open question, a good first step would be to consider chains of the form $P = P_1 P_2$, where each of P_1 and P_2 is reversible with respect to $\pi(\cdot)$, but P is not reversible. Showing that $\sqrt{n}$-CLT's must exist whenever $\pi(h^2) < \infty$ would be quite interesting even in this first case. To prove a negative result requires a counter-example, involving a Markov chain which is geometrically ergodic but not reversible, and a functional $h : \mathcal{X} \to \mathbf{R}$ such that $\pi(h^2) < \infty$ but $\pi(|h|^{2+\delta}) = \infty$ for all $\delta > 0$, which does not have a $\sqrt{n}$-CLT. We have not succeeded in either direction – not even for countable state space chains.

If P is *reversible*, then it was proved by Kipnis and Varadhan [46] that finiteness of σ^2 is all that is required:

Theorem 26. *For a ϕ-irreducible and aperiodic Markov chain which is reversible, a $\sqrt{n}$-CLT holds for h whenever $\sigma^2 < \infty$, where σ^2 is given by (19).*

In a different direction, we have the following:

Theorem 27. *Suppose a Markov chain is geometrically ergodic, satisfying (10) for some $V : \mathcal{X} \to [1, \infty]$ which is finite π-a.e. Let $h : \mathcal{X} \to \mathbf{R}$ with $h^2 \leq K\,V$ for some $K < \infty$. Then a $\sqrt{n}$-CLT holds for h.*

Before proving some of these results, we consider two extensions which are straightforward mathematically, but which may be of practical importance.

Proposition 28. *The above CLT results (i.e., Theorems 23, 24, 25, 26, and 27) all remain true if instead of beginning with $X_0 \sim \pi(\cdot)$, as above, we begin with $X_0 = x$, for π-a.e. $x \in \mathcal{X}$.*

Proof. The hypotheses of the various CLT results all imply that the chain is ϕ-irreducible and aperiodic, with stationary distribution $\pi(\cdot)$. Hence, by Theorem 4, there is convergence to $\pi(\cdot)$ from π-a.e. $x \in \mathcal{X}$. For such x, let $\epsilon > 0$, and find $m \in \mathbf{N}$ such that $\|P^m(x, \cdot) - \pi(\cdot)\| \leq \epsilon$. It then follows from Proposition 3(g) that we can jointly construct copies $\{X_n\}$ and $\{X'_n\}$ of the Markov chain, with $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, such that

$$\mathbf{P}[X_n = X'_n \text{ for all } n \geq m] \geq 1 - \epsilon\,.$$

But this means that for any $A \subseteq \mathcal{X}$,

$$\limsup_{n\to\infty} \left| \mathbf{P}\Big(n^{-1/2} \sum_{i=1}^{n} [h(X_i) - \pi(h)] \in A\Big) \;-\; \mathbf{P}\Big(n^{-1/2} \sum_{i=1}^{n} [h(X'_i) - \pi(h)] \in A\Big) \right| \;\leq\; \epsilon\,.$$

Since $\epsilon > 0$ is arbitrary, and since we know that $n^{-1/2} \sum_{i=1}^{n} [h(X'_i) - \pi(h)]$ converges in distribution to $N(0, \sigma^2)$, hence so does $n^{-1/2} \sum_{i=1}^{n} [h(X_i) - \pi(h)]$. ∎

Proposition 29. *The CLT Theorems 23 and 24 remain true if the chain is periodic of period $d \geq 2$, provided that the d-step chain $P' = P^d\big|_{\mathcal{X}_1}$ (as in the proof of Corollary 6) has all the other properties required of P in the original result (i.e. ϕ-irreducibility, and uniform or geometric ergodicity), and that the function h still satisfies the same moment condition.*

Proof. As in the proof of Corollary 6, let $\overline{P}$ be the d-step chain defined on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, and $\overline{h}(x_0, \ldots, x_{d-1}) = h(x_0) + \ldots + h(x_{d-1})$. Then $\overline{P}$ inherits the irreducibility and ergodicity properties of P' (formally, since P' is *de-initialising* for $\overline{P}$; see [72]). Then, Theorem 23 or 24 establishes a CLT for $\overline{P}$ and $\overline{h}$. However, this is easily seen to be equivalent to the corresponding CLT for the original P and h, thus giving the result. ∎

Remark. In particular, combining Theorem 23 with Proposition 29, we see that a $\sqrt{n}$-CLT holds for any function h for any irreducible (or indecomposible) Markov chain on a *finite* state space, without any assumption of aperiodicity. (See also the Remark following Corollary 6 above.)

Remark. We note that for periodic chains as in Proposition 29, the formula (19) for the asymptotic variance σ^2 continues to hold without change. The relation $\sigma^2 = \tau \, \mathbf{Var}_\pi(h)$ also continues to hold, except that now the formula for the integrated autocorrelation time τ requires that the sum be taken over ranges whose lengths are multiples of d, i.e. the flexibly-ordered infinite sum $\tau = \sum_{k \in \mathbf{Z}} \mathrm{Corr}(X_0, X_k)$ must be replaced by the more precisely limited sum $\tau = \lim_{m,\ell \to \infty} \sum_{k=-\ell d}^{md} \mathrm{Corr}(X_0, X_k)$ (otherwise the sum will not converge, since now the individual terms do not go to 0).

5.3. CLT Proofs using the Poisson Equation.

Here we provide proofs of *some* of the results stated in the previous subsection.

We begin by stating a version of the *martingale central limit theorem*, which was proved independently by Billingsley [12] and Ibragimov [38]; see e.g. p. 375 of Durrett [25].

Theorem 30. *(Billingsley [12] and Ibragimov [38]) Let $\{Z_n\}$ be a stationary ergodic sequence, with $\mathbf{E}[Z_n \mid Z_1, \ldots, Z_{n-1}] = 0$ and $\mathbf{E}[(Z_n)^2] < \infty$. Then $n^{-1/2} \sum_{i=1}^n Z_i$ converges weakly to a $N(0, \sigma^2)$ distribution for some $\sigma^2 < \infty$.*

To make use of Theorem 30, consider the *Poisson equation*: $h - \pi(h) = g - Pg$. A useful result is the following (see e.g. Theorem 17.4.4 of Meyn and Tweedie [53]):

Theorem 31. *Let P be a transition kernel for an aperiodic, ϕ-irreducible Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$, with $X_0 \sim \pi(\cdot)$. Let $h : \mathcal{X} \to \mathbf{R}$ with $\pi(h^2) < \infty$, and suppose there exists $g : \mathcal{X} \to \mathbf{R}$ with $\pi(g^2) < \infty$ which solves the Poisson equation, i.e. such that $h - \pi(h) = g - Pg$. Then h satisfies a $\sqrt{n}$-CLT.*

Proof. Let $Z_n = g(X_n) - Pg(X_{n-1})$. Then $\{Z_n\}$ is stationary since $X_0 \sim \pi(\cdot)$. Also $\{Z_n\}$ is ergodic since the Markov chain converges asymptotically (by Theorem 4). Furthermore, $\mathbf{E}[Z_n^2] \leq 4\,\pi(g^2) < \infty$. Also,

$$\begin{aligned}\mathbf{E}[g(X_n) - Pg(X_{n-1}) \,|\, X_0, \ldots, X_{n-1}] \;&=\; \mathbf{E}[g(X_n) \,|\, X_{n-1}] - Pg(X_{n-1}) \\ &=\; Pg(X_{n-1}) - Pg(X_{n-1}) \;=\; 0\,.\end{aligned}$$

Since $Z_1, \ldots, Z_{n-1} \in \sigma(X_0, \ldots, X_{n-1})$, it follows that $\mathbf{E}_\pi[Z_n \,|\, Z_1, \ldots, Z_{n-1}] = 0$. Hence, by Theorem 30, $n^{-1/2} \sum_{i=1}^n Z_i$ converges weakly to $N(0, \sigma^2)$. But

$$\begin{aligned}n^{-1/2} \sum_{i=1}^n [h(X_i) - \pi(h)] \;&=\; n^{-1/2} \sum_{i=1}^n [g(X_i) - Pg(X_i)] \\ =\; n^{-1/2} \sum_{i=1}^n [g(X_i) - Pg(X_{i-1})] &+ n^{-1/2} Pg(X_0) - n^{-1/2} Pg(X_n) \\ =\; n^{-1/2} \sum_{i=1}^n Z_i &+ n^{-1/2} Pg(X_0) - n^{-1/2} Pg(X_n)\,.\end{aligned}$$

The result follows since $n^{-1/2} g(X_0)$ and $n^{-1/2} Pg(X_n)$ both converge to zero in probability as $n \to \infty$. ∎

Corollary 32. *If $\sum_{k=0}^\infty \sqrt{\pi\big((P^k[h - \pi(h)])^2\big)} < \infty$, then h satisfies a $\sqrt{n}$-CLT.*

Proof. Let

$$g_k(x) \;=\; P^k h(x) - \pi(h) \;=\; P^k[h - \pi(h)](x)\,,$$

where by convention $P^0 h(x) = h(x)$, and let $g(x) = \sum_{k=0}^\infty g_k(x)$. Then we compute directly that

$$(g - Pg)(x) \;=\; \sum_{k=0}^\infty g_k(x) - \sum_{k=0}^\infty Pg_k(x) \;=\; \sum_{k=0}^\infty g_k(x) - \sum_{k=1}^\infty g_k(x)$$

$$= \; g_0(x) \; = \; P^0 h(x) - \pi(h) \; = \; h(x) - \pi(h)\,.$$

Hence, the result follows from Theorem 31, *provided* that $\pi(g^2) < \infty$. On the other hand, it is known (in fact, since $\mathbf{Cov}(X,Y) \leq \sqrt{\mathbf{Var}(X)\,\mathbf{Var}(Y)}$) that the $L^2(\pi)$ norm satisfies the triangle inequality, so that

$$\sqrt{\pi(g^2)} \; \leq \; \sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)}\,,$$

so that $\pi(g^2) < \infty$ provided $\sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)} < \infty$. ∎

Proof of Theorem 25. Let

$$\|P\|_{L^2(\pi)} = \sup_{\substack{\pi(f)=0 \\ \pi(f^2)=1}} \pi\big((Pf)^2\big) = \sup_{\substack{\pi(f)=0 \\ \pi(f^2)=1}} \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{X}} f(y)\, P(x,dy) \right)^2 \pi(dx)$$

be the usual $L^2(\pi)$ operator norm for P, when restricted to those functionals f with $\pi(f) = 0$ and $\pi(f^2) < \infty$. Then it is shown in Theorem 2 of [68] that reversible chains are geometrically ergodic if and only if they satisfy $\|P\|_{L^2(\pi)} < 1$, i.e. there is $\beta < 1$ with $\pi((Pf)^2) \leq \beta^2 \pi(f^2)$ whenever $\pi(f) = 0$ and $\pi(f^2) < \infty$. Furthermore, reversibility implies self-adjointness of P in $L^2(\pi)$, so that $\|P^k\|_{L^2(\pi)} = \|P\|^k_{L^2(\pi)}$, and hence $\pi((P^k f)^2) \leq \beta^{2k}\pi(f^2)$.

Let $g_k = P^k h - \pi(h)$ as in the proof of Corollary 32. Then this implies that $\pi((g_k)^2) \leq \beta^{2k}\,\pi((h - \pi(h)^2)$, so that

$$\sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)} \leq \sqrt{\pi((h-\pi(h))^2)} \; \sum_{k=0}^{\infty} \beta^k = \sqrt{\pi((h-\pi(h))^2)} \; / \, (1-\beta) < \infty\,.$$

Hence, the result follows from Corollary 32. ∎

Proof of Theorem 27. By Fact 10, there is $C < \infty$ and $\rho < 1$ with $|P^n f(x) - \pi(f)| \leq CV(x)\rho^n$ for $x \in \mathcal{X}$ and $f \leq V$, and furthermore $\pi(V) < \infty$. Let $g_k = P^k[h - \pi(h)]$ as in the proof of Corollary 32. Then by the Cauchy-Schwartz inequality, $(g_k)^2 = \big(P^k[h-\pi(h)]\big)^2 \leq$

$P^k\big([h-\pi(h)]^2\big)$. On the other hand, since $[h-\pi(h)]^2 \leq KV$, so $[h-\pi(h)]^2/K \leq V$, we have $(g_k)^2 \leq P^k\big([h-\pi(h)]^2\big) \leq CKV\rho^k$. This implies that $\pi((g_k)^2) \leq CK\rho^k\pi(V)$, so that

$$\sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)} \leq \sqrt{CK\pi((h-\pi(h))^2)} \sum_{k=0}^{\infty} \rho^{k/2} = \sqrt{CK\pi((h-\pi(h))^2)} \,/\, (1-\sqrt{\rho}) < \infty\,.$$

Hence, the result again follows from Corollary 32. ∎

5.4. Proof of Theorem 24 using Regenerations.

Here we use *regeneration theory* to give a reasonably direct proof of Theorem 24, following the outline of Hobert et al. [37], thereby avoiding the technicalities of the original proof of Ibragimov and Linnik [39].

We begin by noting from Fact 10 that since the chain is geometrically ergodic, there is a small set C and a drift function V satisfying (8) and (10).

In terms of this, we consider a *regeneration construction* for the chain (cf. [8], [4], [56], [37]). This is very similar to the coupling construction presented in Section 4, except now just for a *single* chain $\{X_n\}$. Thus, in the coupling construction we omit option 1, and merely update the single chain. More formally, given X_n, we proceed as follows. If $X_n \notin C$, then we simply choose $X_{n+1} \sim P(X_n, \cdot)$. Otherwise, if $X_n \in C$, then with probability ϵ we choose $X_{n+n_0} \sim \nu(\cdot)$, while with probability $1-\epsilon$ we choose $X_{n+n_0} \sim R(X_n, \cdot)$. [If $n_0 > 1$, we then fill in the missing values $X_{n+1}, \ldots, X_{n+n_0-1}$ as usual.]

We let $T_1, T_2, \ldots$ be the *regeneration times*, i.e. the times such that $X_{T_i} \sim \nu(\cdot)$ as above. Thus, the regeneration times occur with probability ϵ precisely n_0 iterations after each time the chain enters C (not counting those entries of C which are within n_0 of a previous regeneration attempt).

The benefit of regeneration times is that they break up sums like $\sum_{i=0}^{n}[h(X_i)-\pi(h)]$ into sums over *tours*, each of the form $\sum_{i=T_j}^{T_{j+1}-1}[h(X_i)-\pi(h)]$. Furthermore, since each subsequent tour begins from the same fixed distribution $\nu(\cdot)$, we see that *the different tours, after the first one, are independent and identically distributed (i.i.d.).*

More specifically, let $T_0 = 0$, and let $r(n) = \sup\{i \geq 0;\ T_i \leq n\}$. Then

$$\sum_{i=1}^{n}[h(X_i) - \pi(h)] \;=\; \sum_{j=1}^{r(n)} \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)] + E(n)\,, \tag{21}$$

where $E(n)$ is an error term which collects the terms corresponding to the incomplete final tour $X_{T_{r(n)+1}}, \ldots, X_n$, and also the first tour $X_0, \ldots, X_{T_1-1}$.

Now, the tours $\{\{X_{T_j}, X_{T_j+1}, \ldots, X_{T_{j+1}-1}\},\ \ j = 1, 2, \ldots\}$ are independent and identically distributed. Moreover, elementary renewal theory (see for example [4]) ensures that $r(n)/n \to \epsilon\pi(C)$ in probability. Hence, the classical central limit theorem (see e.g. [13], Theorem 27.1; or [82], p. 110) will prove Theorem 24, *provided* that each term has finite second moment, and that the error term $E(n)$ can be neglected.

To continue, we note that geometric ergodicity implies (as in the proof of Lemma 18) exponential tails on the return times to C. It then follows (cf. Theorem 2.5 of [93]) that there is $\beta > 1$ with

$$\mathbf{E}_\pi[\beta^{T_1}] < \infty\,, \quad \text{and} \quad \mathbf{E}[\beta^{T_{j+1}-T_j}] < \infty\,. \tag{22}$$

(This also follows from Theorem 15.0.1 of [53], together with a simple argument using probability generating functions.)

Now, it seems intuitively clear that $E(n)$ is $O_p(1)$ as $n \to \infty$, so when multiplied by $n^{-1/2}$, it will not contribute to the limit. Formally, this follows from (22), which implies by standard renewal theory that $E(n)$ has a limiting distribution as $n \to \infty$, which in turn implies that $E(n)$ is $O_p(1)$ as $n \to \infty$. Thus, the term $E(n)$ can be neglected without affecting the result.

Hence, it remains only to prove the finite second moments of each term in (21). Recalling that each tour begins in the distribution $\nu(\cdot)$, we see that the proof of Theorem 24 is completed by the following lemma:

Lemma 33. $\int_{x\in\mathcal{X}} \nu(dx)\, \mathbf{E}\big[\big(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)]\big)^2 \;\big|\; X_0 = x\big] \;<\; \infty$.

Proof. Note that

$$\pi(\cdot) \;=\; \int_{x\in\mathcal{X}} \pi(dx)\, P(x,\cdot) \;\geq\; \int_{x\in C} \pi(dx)\, P(x,\cdot) \;\geq\; \pi(C)\,\epsilon\,\nu(\cdot)\,,$$

so that $\nu(dx) \le \pi(dx) \,/\, \pi(C)\,\epsilon$. Hence, it suffices to prove the lemma with $\nu(dx)$ replaced by $\pi(dx)$ i.e. under the assumption that $X_0 \sim \pi(\cdot)$.

For notational simplicity, set $H_i = h(X_i) - \pi(h)$, and $\mathbf{E}_\pi[\cdots] = \int_{x\in\mathcal{X}} \mathbf{E}[\cdots \,|\, X_0 = x]\,\pi(dx)$. Note that $\Big(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\Big)^2 = \big(\sum_{i=0}^{\infty} \mathbf{1}_{i<T_1} H_i\big)^2$. Hence, by Cauchy-Schwartz,

$$\mathbf{E}_\pi\Big[\Big(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\Big)^2\Big] \;\le\; \Big(\sum_{i=0}^{\infty}\sqrt{\mathbf{E}_\pi\big[\mathbf{1}_{i<T_1}H_i^2\big]}\;\Big)^2. \tag{23}$$

To continue, let $p = 1+2/\delta$ and $q = 1+\delta/2$, so that $1/p+1/q = 1$. Then by *Hölder's inequality* (e.g. [13], p. 80),

$$\mathbf{E}_\pi[\mathbf{1}_{i<T_1}H_i^2] \;\le\; \mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{1/p}\,\mathbf{E}_\pi[|H_i|^{2q}]^{1/q}\,. \tag{24}$$

Now, since $X_0 \sim \pi(\cdot)$, therefore $\mathbf{E}_\pi[|H_i|^{2q}] \equiv K$ is a constant, independent of i, which is finite since $\pi(|h|^{2+\delta}) < \infty$.

Also, using (22), Markov's inequality then gives that $\mathbf{E}_\pi[\mathbf{1}_{0\le i<T_1}] \le \mathbf{E}_\pi[\mathbf{1}_{\beta^{T_1}>\beta^i}] \le \beta^{-i}\mathbf{E}_\pi[\beta^{T_1}]$. Hence, combining (23) and (24), we obtain that

$$\mathbf{E}_\pi\Big[\Big(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\Big)^2\Big] \;\le\; \Big(\sum_{i=0}^{\infty}\sqrt{\mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{1/p}\,\mathbf{E}_\pi[|H_i|^{2q}]^{1/q}}\Big)^2$$

$$\le \Big(K^{1/2q}\sum_{i=0}^{\infty}\sqrt{(\beta^{-i}\mathbf{E}_\pi[\beta^{T_1}])^{1/p}}\Big)^2 = \Big(K^{1/2q}\mathbf{E}_\pi[\beta^{T_1}]^{1/2p}\sum_{i=0}^{\infty}\beta^{-i/2}\Big)^2$$

$$= \Big(K^{1/2q}\mathbf{E}_\pi[\beta^{T_1}]^{1/2p}\,/\,(1-\beta^{-1/2})\Big)^2 < \infty\,. \qquad \blacksquare$$

It appears at first glance that Theorem 23 could be proved by similar regeneration arguments. However, we have been unable to do so.

Open Problem #3. *Can Theorem 23 be proved by direct regeneration arguments, similar to the above proof of Theorem 24?*

6. Optimal Scaling and Weak Convergence.

Finally, we briefly discuss another application of probability theory to MCMC, namely the *optimal scaling* problem. Our presentation here is quite brief; for further details see the review article [73].

Let $\pi_u : \mathbf{R}^d \to [0, \infty)$ be a continuous d-dimensional density (d large). Consider running a Metropolis-Hastings algorithm for π_u. The optimal scaling problem concerns the question of how we should choose the proposal distribution for this algorithm.

For concreteness, consider either the random-walk Metropolis (RWM) algorithm with proposal distribution given by $Q(x, \cdot) = N(x, \sigma^2 I_d)$, or the Langevin algorithm with proposal distribution given by $Q(x, \cdot) = N(x + \frac{\sigma^2}{2} \nabla \log \pi_u(x),\ \sigma^2 I_d)$. In either case, the question becomes, how should we choose σ^2?

If σ^2 is chosen to be too small, then by continuity the resulting Markov chain will nearly always accept its proposed value. However, the proposed value will usually be extremely close to the chain's previous state, so that the chain will move extremely slowly, leading to a very high acceptance rate, but very poor performance. On the other hand, if σ^2 is chosen to be too large, then the proposed values will usually be very far from the current state. Unless the chain gets very "lucky", then those proposed values will usually be rejected, so that the chain will tend to get "stuck" at the same state for large periods of time. This will lead to a very low acceptance rate, and again a very poorly performing algorithm. We conclude that proposal scalings satisfy a *Goldilocks Principle*: The choice of the proposal scaling σ^2 should be "just right", neither too small nor too large.

To prove theorems about this, assume for now that

$$\pi_u(\mathbf{x}) \ = \ \prod_{i=1}^{d} f(x_i)\,, \tag{25}$$

i.e. that the density π_u factors into i.i.d. components, each with (smooth) density f. (This assumption is obviously very restrictive, and is uninteresting in practice since then each coordinate can be simulated separately. However, it does allow us to develop some interesting theory, which may approximately apply in other cases as well.) Also, assume that chain begins in stationarity, i.e. that $X_0 \sim \pi(\cdot)$.

6.1. The Random Walk Metropolis (RWM) Case.

For RWM, let $I = \mathbf{E}[((\log f(Z))')^2]$ where $Z \sim f(z)\,dz$. Then it turns out, essentially, that under the assumption (25), as $d \to \infty$ it is optimal to choose $\sigma^2 \doteq (2.38)^2/Id$, leading to an asymptotic acceptance rate $\doteq 0.234$.

More precisely, set the proposal variance to be $\sigma_d^2 = \ell^2/d$, where $\ell > 0$ is to be chosen later. Let $\{X_n\}$ be the Random Walk Metropolis algorithm for $\pi(\cdot)$ on $\mathbf{R}^d$ with proposal variance σ_d^2. Also, let $\{N(t)\}_{t\geq 0}$ be a Poisson process with rate d which is independent of $\{X_n\}$. Finally, let

$$Z_t^d = X_{N(t)}^{(1)}\,, \qquad t \geq 0\,.$$

Thus, $\{Z_t^d\}_{t\geq 0}$ follows the first component of $\{X_n\}$, with time speeded up by a factor of d.

Then it is proved in [66] (see also [73]), using the theory from Ethier and Kurtz [26], that as $d \to \infty$, the process $\{Z_t^d\}_{t\geq 0}$ converges weakly to a diffusion process $\{Z_t\}_{t\geq 0}$ which satisfies the following stochastic differential equation:

$$dZ_t \;=\; h(\ell)^{1/2}\,dB_t + \frac{1}{2}\,h(\ell)\,\nabla \log \pi_u(Z_t)\;dt\,.$$

Here

$$h(\ell) \;=\; 2\,\ell^2\,\Phi\left(-\frac{\sqrt{I}\ell}{2}\right)$$

corresponds to the *speed* of the limiting diffusion, where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-s^2/2}ds$ is the cdf of a standard normal distribution.

We then compute numerically that the choice $\ell = \hat{\ell} \doteq 2.38/\sqrt{I}$ maximises the above speed function $h(\ell)$, and thus must be the choice leading to optimally fast mixing (at least, as $d \to \infty$). Furthermore, it is also proved in [66] that the asymptotic (i.e., expected value with respect to the stationary distribution) acceptance rate of the algorithm is given by the formula $A(\ell) = 2\,\Phi\left(-\frac{\sqrt{I}\ell}{2}\right)$, and we compute that $A(\hat{\ell}) \doteq 0.234$, thus giving the optimal asymptotic acceptance rate.

6.2. The Langevin Algorithm Case.

In the Langevin case, let $J = \mathbf{E}[(5((\log f(Z))'''))^2 - 3((\log f(Z))'')^3)/48]$ where again $Z \sim f(z)\,dz$. Then it turns out, essentially, that assuming (25), it is optimal as $d \to \infty$ to choose $\sigma^2 \doteq (0.825)^2/J^{1/2}d^{1/3}$, leading to an asymptotic acceptance rate $\doteq 0.574$.

More precisely, set $\sigma_d^2 = \ell^2/d^{1/3}$, let $\{X_n\}$ be the Langevin Algorithm for $\pi(\cdot)$ on $\mathbf{R}^d$ with proposal variance σ_d^2, let $\{N(t)\}_{t\geq 0}$ be a Poisson process with rate $d^{1/3}$ which is independent of $\{X_n\}$, and let

$$Z_t^d = X_{N(t)}^{(1)}\,,$$

so that $\{Z_t^d\}_{t\geq 0}$ follows the first component of $\{X_n\}$, with time speeded up by a factor of $d^{1/3}$. Then it is proved in [69] (see also [73]) that as $d \to \infty$, the process $\{Z_t^d\}_{t\geq 0}$ converges weakly to a diffusion process $\{Z_t\}_{t\geq 0}$ which satisfies the following stochastic differential equation:

$$dZ_t \;=\; g(\ell)^{1/2}\,dB_t + \frac{1}{2}\,g(\ell)\,\nabla \log \pi_u(Z_t)\;dt\,.$$

Here

$$g(\ell) \;=\; 2\,\ell^2\,\Phi\left(-J\ell^3\right)$$

represents the speed of the limiting diffusion. We then compute numerically that the choice $\ell = \hat{\ell} = 0.825/\sqrt{J}$ maximises $g(\ell)$, and thus must be the choice leading to optimally fast mixing (at least, as $d \to \infty$). Furthermore, it is proved in [69] that the asymptotic acceptance rate satisfies $A(\hat{\ell}) = 2\,\Phi(-J\hat{\ell}^3) \doteq 0.574$, thus giving the optimal asymptotic acceptance rate for the Langevin case.

6.3. Discussion of Optimal Scaling.

The above results show that for either the RWM or the Langevin algorithm, under the assumption (25), we can determine the optimal proposal scaling just in terms of universally optimal asymptotic acceptance rates (0.234 for RWM, 0.574 for Langevin). Such results are straightforward to apply in practice, since it is trivial for a computer to monitor the acceptance rate of the algorithm, and the user can modify σ^2 appropriately to achieve appropriate acceptance rates. Thus, these optimal scaling rates are often used in applied contexts (see e.g. Møller et al. [55]). (It may even be possible for the computer to adaptively

modify σ^2 to achieve the appropriate acceptance rates; see [5] and references therein. However it is important to recognise that adaptive strategies can violate the stationarity of π so they have to be carefully implemented; see for example [35].)

The above results also describe the *computational complexity* of these algorithms. Specifically, they say that as $d \to \infty$, the efficiency of RWM algorithms scales like d^{-1}, so its computational complexity is $O(d)$. Similarly, the efficiency of Langevin algorithms scales like $d^{-1/3}$, so its computational complexity is $O(d^{1/3})$ which is much lower order (i.e. better).

We note that for reasonable efficiency, we do not need the acceptance rate to be *exactly* 0.234 (or 0.574), just fairly close. Also, the dimension doesn't have to be *too* large before asymptotics approximately kick in; often 0.234 is approximately optimal in dimensions as low as 5 or 10. For further discussion of these issues, see the review article [73].

Now, the above results are only proved under the strong assumption (25). It is natural to ask what happens if this assumption is not satisfied. In that case, there are various extensions of the optimal-scaling results to cases of inhomogeneously-scaled components of the form $\pi_u(\mathbf{x}) = \prod_{i=1}^{d} C_i\, f(C_i x_i)$ (see [73]), to the discrete hypercube [64], and to finite-range homogeneous Markov random fields [14]; in particular, the optimal acceptance rate remains 0.234 (under appropriate assumptions) in all of these cases. On the other hand, surprising behaviour can result if we do not start in stationarity, i.e. if the assumption $X_0 \sim \pi(\cdot)$ is violated and the chain instead begins way out in the tails of $\pi(\cdot)$; see [16]. The true level of generality of these optimal scaling results is currently unknown, though investigations are ongoing [10]. In general this is an open problem:

Open Problem #4. *Determine the extent to which the above optimal scaling results continue to apply, even when assumption (25) is violated.*

APPENDIX: Proof of Lemma 17.

Lemma 17 above states (Meyn and Tweedie [53], Theorem 5.5.7) that for an aperiodic, ϕ-irreducible Markov chain, all petite sets are small sets.

To prove this, we require a lemma related to aperiodicity:

Lemma 34. *Consider an aperiodic Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Let $\nu(\cdot)$ be any probability measure on $\mathcal{X}$. Assume that $\nu(\cdot) \ll \pi(\cdot)$, and that for all $x \in \mathcal{X}$, there is $n = n(x) \in \mathbf{N}$ and $\delta = \delta(x) > 0$ such that $P^n(x, \cdot) \geq \delta\nu(\cdot)$ (for example, this always holds if $\nu(\cdot)$ is a minorisation measure for a small or petite set which is reachable from all states). Let $T = \{n \geq 1;\ \exists \delta_n > 0 \ s.t. \ \int \nu(dx)\, P^n(x, \cdot) \geq \delta_n \nu(\cdot)\}$, and assume that T is non-empty. Then there is $n_* \in \mathbf{N}$ with $T \supseteq \{n_*, n_* + 1, n_* + 2, \ldots\}$.*

Proof. Since $P^{(n(x))}(x, \cdot) \geq \delta(x)\, \nu(\cdot)$ for all $x \in \mathcal{X}$, it follows that T is non-empty.

Now, if $n, m \in T$, then since $\int_{x\in\mathcal{X}} \nu(dx)\, P^{n+m}(x, \cdot) = \int_{x\in\mathcal{X}} \int_{y\in\mathcal{X}} \nu(dx) P^n(x, dy) P^m(y, \cdot) \geq \int_{y\in\mathcal{X}} \delta_n \nu(dy) P^m(y, \cdot) \geq \delta_n \delta_m \nu(\cdot)$, we see that T is *additive*, i.e. if $n, m \in T$ then $n+m \in T$.

We shall prove below that $gcd(T) = 1$. It is then a standard and easy fact (e.g. [13], p. 541; or [82], p. 77) that if T is non-empty and additive, and $gcd(T) = 1$, then there is $n_* \in \mathbf{N}$ such that $T \supseteq \{n_*, n_* + 1, n_* + 2, \ldots\}$, as claimed.

We now proceed to prove that $gcd(T) = 1$. Indeed, suppose to the contrary that $gcd(T) = d > 1$. We will derive a contradiction.

For $1 \leq i \leq d$, let

$$\mathcal{X}_i \ = \ \{x \in \mathcal{X};\ \exists \ell \in \mathbf{N} \text{ and } \delta > 0 \text{ s.t. } P^{\ell d - i}(x, \cdot) \geq \delta\nu(\cdot)\}\,.$$

Then $\bigcup_{i=1}^d \mathcal{X}_i = \mathcal{X}$ by assumption. Now, let

$$S \ = \ \bigcup_{i\neq j} (\mathcal{X}_i \cap \mathcal{X}_j)\,,$$

let

$$\overline{S} \ = \ S \cup \{x \in \mathcal{X};\ \exists m \in \mathbf{N} \ s.t. \ P^m(x, S) > 0\}\,,$$

and let

$$\mathcal{X}_i' \ = \ \mathcal{X}_i \setminus \overline{S}\,.$$

Then $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ are disjoint by construction (since we have removed S). Also if $x \in \mathcal{X}_i'$, then $P(x, \overline{S}) = 0$, so that $P(x, \bigcup_{j=1}^d \mathcal{X}_j') = 1$ by construction. In fact we must have $P(x, \mathcal{X}_{i+1}') = 1$ in the case $i < d$ (with $P(x, \mathcal{X}_1') = 1$ for $i = d$), for if not then x would be in two different $\mathcal{X}_j'$ at once, contradicting their disjointedness.

We claim that for all $m \geq 0$, $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) = 0$ whenever $i \neq j$. Indeed, if we had $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) > 0$ for some $i \neq j$, then there would be $S' \subseteq \mathcal{X}$, $\ell_1, \ell_2 \in \mathbf{N}$, and $\delta > 0$ such that for all $x \in S'$, $P^{\ell_1 d+i}(x, \cdot) \geq \nu(\cdot)$ and $P^{\ell_2 > d+j}(x, \cdot) \geq \nu(\cdot)$, implying that $\ell_1 d + i + m \in T$ and $\ell_2 d + j + m \in T$, contradicting the fact that $\gcd(T) = d$.

It then follows (by sub-additivity of measures) that $\nu(\overline{S}) = 0$. Therefore, $\nu(\bigcup_{i=1}^d \mathcal{X}_i') = \nu(\bigcup_{i=1}^d \mathcal{X}_i) = \nu(\mathcal{X}) = 1$. Since $\nu \ll \pi$, we must have $\pi(\bigcup_{i=1}^d \mathcal{X}_i') > 0$.

We conclude from all of this that $\mathcal{X}_1', \ldots, \mathcal{X}_d'$ are subsets of positive π-measure, with respect to which the Markov chain is periodic (of period d), contradicting the assumption of aperiodicity. ∎

Proof of Lemma 17. Let R be $(n_0, \epsilon, \nu(\cdot))$-petite, so that $\sum_{i=1}^{n_0} P^i(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in R$. Let T be as in Lemma 34. Then $\sum_{i=1}^{n_0} \int_{x \in \mathcal{X}} \nu(dx)\, P^i(x, \cdot) \geq \epsilon\nu(\cdot)$, so we must have $i \in T$ for some $1 \leq i \leq n_0$, so that T is non-empty. Hence, from Lemma 34, we can find n_* and $\delta_n > 0$ such that $\int \nu(dx) P^n(x, \cdot) > \delta_n \nu(\cdot)$ for all $n \geq n_*$. Let $r = \min\Big\{\delta_n;\ n_* \leq n \leq n_* + n_0 - 1\Big\}$, and set $N = n_* + n_0$. Then for $x \in R$,

$$
\begin{aligned}
P^N(x, \cdot) &\geq \sum_{i=1}^{n_0} \int_{y \in \mathcal{X}} P^{N-i}(x, dy) P^i(y, \cdot) \\
&\geq \sum_{i=1}^{n_0} \int_{y \in R} r\nu(dy) P^i(y, \cdot) \\
&\geq \int_{y \in R} r\nu(dy) \epsilon\nu(\cdot) = r\epsilon\nu(\cdot)\,.
\end{aligned}
$$

Thus, R is $(N, r\epsilon, \nu(\cdot))$-small. ∎

Note added later: Olle Häggström has just produced a counter-example showing that the answer to our Open Problem #2 is no in general; see his paper "On the central limit theorem for geometrically ergodic Markov chains", available at

`http://www.math.chalmers.se/~olleh/papers.html`

Acknowledgements. The authors are sincerely grateful to Kun Zhang and to an anonymous referee for their extremely careful readings of this manuscript and their many insightful comments which lead to numerous improvements.

REFERENCES

[1] D.J. Aldous, (1983), Random walk on finite groups and rapidly mixing Markov chains. Séminaire de Probabilités *XVII*. Lecture Notes in Math. **986**, 243–297. Springer, New York.

[2] D.J. Aldous and P. Diaconis (1987), Strong stopping times and finite random walks. Adv. Appl. Math. **8**, 69–97.

[3] D.J. Aldous and H. Thorisson (1993), Shift-coupling. Stoch. Proc. Appl. **44**, 1–14.

[4] S. Asmussen (1987), Applied Probability and Queues. John Wiley & Sons, New York.

[5] Y.F. Atchadé and J.S. Rosenthal (2003), On Adaptive Markov Chain Monte Carlo Algorithms. Submitted.

[6] Y.F. Atchadé and J.S. Rosenthal (2003), Central Limit Theorems for Geometrically Ergodic Markov chains. Work in progress.

[7] K.B. Athreya, H. Doss, and J. Sethuraman (1996), On the Convergence of the Markov Chain Simulation Method. Ann. Stat. **24**, 69–100.

[8] K.B. Athreya and P. Ney (1978), A new approach to the limit theory of recurrent Markov chains. Trans. Amer. Math. Soc. **245**, 493–501.

[9] P.H. Baxendale (2003), Renewal theory and computable convergence rates for geometrically ergodic Markov chains. Preprint, University of Southern California.

[10] M. Bédard (2004), On the robustness of optimal scaling for Metropolis-Hastings algorithms. Ph.D. dissertation, University of Toronto. Work in progress.

[11] K. Berenhaut and R.B. Lund (2001), Geometric renewal convergence rates from hazard rates. J. Appl. Prob. **38**, 180–194.

[12] P. Billingsley (1961), The Lindeberg-Lévy theorem for martingales. Proc. Amer. Math. Soc. **12**, 788–792.

[13] P. Billingsley (1995), Probability and Measure, 3rd ed. John Wiley & Sons, New York.

[14] L. Breyer and G.O. Roberts (2000), From Metropolis to diffusions: Gibbs states and optimal scaling. Stoch. Proc. Appl. **90**, 181–206.

[15] K.S. Chan and C.J. Geyer (1994), Discussion to reference [92]. Ann. Stat. **22**, 1747–1758.

[16] O.F. Christensen, G.O. Roberts, and J.S. Rosenthal (2003), Scaling Limits for the Transient Phase of Local Metropolis-Hastings Algorithms. Submitted.

[17] R. Cogburn (1972), The central limit theorem for Markov processes. In Proc. Sixth Berkeley Symp. Math. Statist. Probab. **2**, 485–512.

[18] M.K. Cowles and B.P. Carlin (1996), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. J. Amer. Stat. Assoc. **91**, 883–904.

[19] M.K. Cowles, G.O. Roberts, and J.S. Rosenthal (1999), Possible biases induced by MCMC convergence diagnostics. J. Stat. Comp. Sim. **64**, 87–104.

[20] M.K. Cowles and J.S. Rosenthal (1998), A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statistics and Computing* **8**, 115–124.

[21] P. Diaconis (1988), Group representations in Probability and Statistics. Institute of Mathematical Statistics, Hayward, California.

[22] W. Doeblin (1938), Exposé de la théorie des châines simples constantes de Markov à un nombre fini d'états. Revue Mathematique de l'Union Interbalkanique **2**, 77–105.

[23] J.I. Doob (1953), Stochastic Processes. Wiley, New York.

[24] R. Douc, E. Moulines, and J.S. Rosenthal (2002), Quantitative bounds on convergence of time-inhomogeneous Markov Chains. Ann. Appl. Prob., to appear.

[25] R. Durrett (1991), Probability: theory and examples. Wadsworth, Pacific Grove, California.

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

[27] J.A. Fill, M. Machida, D.J. Murdoch, and J.S. Rosenthal (2000), Extension of Fill's perfect rejection sampling algorithm to general chains. Random Struct. Alg. **17**, 290–316.

[28] G. Fort (2003), Computable bounds for V-geometric ergodicity of Markov transition kernels. Preprint, Université Joseph Fourier, Grenoble, France.

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

[30] A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

[31] A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci., Vol. **7**, No. **4**, 457-472.

[32] S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.

[33] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, ed. (1996), Markov chain Monte Carlo in practice. Chapman and Hall, London.

[34] C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473-483.

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

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

[37] J.P. Hobert, G.L. Jones, B. Presnell, and J.S. Rosenthal (2002), On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo. *Biometrika* **89**, 731–743.

[38] I.A. Ibragimov (1963), A central limit theorem for a class of dependent random variables. Theory Prob. Appl. **8**, 83–89.

[39] I.A. Ibragimov and Y.V. Linnik (1971), Independent and Stationary Sequences of Random Variables. Wolters-Noordhoff, Groningen (English translation).

[40] N. Jain and B. Jamison (1967), Contributions to Doeblin's theory of Markov processes. Z. Wahrsch. Verw. Geb. **8**, 19–40.

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

[42] G.L. Jones (2004), On the Central Limit Theorem for Markov Chains (review

paper). Work in progress.

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

[44] 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.

[45] W. Kendall and J. Møller (2000), Perfect simulation using dominating processes on ordered state spaces, with application to locally stable point processes. Adv. Appl. Prob. **32**, 844–865.

[46] C. Kipnis and S.R.S. Varadhan (1986), Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Comm. Math. Phys. **104**, 1-19.

[47] T. Lindvall (1992), Lectures on the Coupling Method. Wiley & Sons, New York.

[48] R. Lund, S.P. Meyn, and R.L. Tweedie (1996), Rates of convergence of stochastically monotone Markov processes. Ann. App. Prob. **6**, 218–237.

[49] A.A. Markov (1906), Extension of the law of large numbers to dependent quantities (in Russian). Izv. Fiz.-Matem. Obsch. Kazan Univ. (2nd Ser) **15**, 135–156.

[50] P. Matthews (1993), A slowly mixing Markov chain with implications for Gibbs sampling. Stat. Prob. Lett. **17**, 231–236.

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

[52] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

[53] S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at http://decision.csl.uiuc.edu/∼meyn/pages/TOC.html.

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

[55] J. Møller, A.R. Syversveen, and R. Waagepetersen (1998), Log Gaussian Cox processes. Scand. J. Statist. **25**, 451–482.

[56] P.A. Mykland, L. Tierney, and B. Yu (1995), Regeneration in Markov chain samplers. J. Amer. Stat. Assoc. **90**, 233–241.

[57] R.M. Neal (2004), Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. Technical Report No. 0406, Dept. of Statistics, University of Toronto.

[58] E. Nummelin (1978), Uniform and ratio limit theorems for Markov renewal and semi-regenerative processes on a general state space. Ann. Inst. Henri Poincaré Series B **14**, 119–143.

[59] E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.

[60] S. Orey (1971), Lecture notes on limit theorems for Markov chain transition probabilities. Van Nostrand Reinhold, London.

[61] J.W. Pitman (1976), On coupling of Markov chains. Z. Wahrsch. verw. Gebiete **35**, 315–322.

[62] J.G. Propp and D.B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms **9**, 223–252.

[63] D. Randall (2003), Mixing, a tutorial on Markov chains. FOCS 2003.

[64] G.O. Roberts (1998), Optimal Metropolis algorithms for product measures on the vertices of a hypercube. Stochastics and Stochastic Reports **62**, 275–283.

[65] G.O. Roberts (1999), A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. J. Appl. Prob. **36**, 1210–1217.

[66] 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.

[67] G.O. Roberts and J.S. Rosenthal (1997), Shift-coupling and convergence rates of ergodic averages. Stochastic Models **13**, 147–165.

[68] G.O. Roberts and J.S. Rosenthal (1997), Geometric ergodicity and hybrid Markov chains. Electronic Comm. Prob. **2**, Paper no. 2, 13–25.

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

[70] G.O. Roberts and J.S. Rosenthal (1998), Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). Canadian J. Stat. **26**, 5–31.

[71] G.O. Roberts and J.S. Rosenthal (2001), Small and Pseudo-Small Sets for Markov

Chains. Stochastic Models **17**, 121–145.

[72] G.O. Roberts and J.S. Rosenthal (2001), Markov chains and de-initialising processes. Scandinavian Journal of Statistics **28**, 489–504.

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

[74] G.O. Roberts and J.S. Rosenthal (2003), Harris Recurrence of Metropolis-Within-Gibbs and Transdimensional MCMC Algorithms. In preparation.

[75] G.O. Roberts and R.L. Tweedie (1996), Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. Biometrika **83**, 95–110.

[76] G.O. Roberts and R.L. Tweedie (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229. See also the corrigendum, Stoch. Proc. Appl. **91** (2001), 337–338.

[77] J.S. Rosenthal (1993), Rates of convergence for data augmentation on finite sample spaces. Ann. Appl. Prob. **3**, 819–839.

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

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

[80] J.S. Rosenthal (1995), Convergence rates of Markov chains. SIAM Review **37**, 387–405.

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

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

[83] J.S. Rosenthal (2001), A review of asymptotic convergence for general state space Markov chains. Far East J. Theor. Stat. **5**, 37–50.

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

[85] J.S. Rosenthal (2003), Asymptotic Variance and Convergence Rates of Nearly-

Periodic MCMC Algorithms. J. Amer. Stat. Assoc. **98**, 169–177.

[86] J.S. Rosenthal (2003), Geometric Convergence Rates for Time-Sampled Markov Chains. J. Theor. Prob. **16**, 671–688.

[87] A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Prob., Comput. **1**, 351–370.

[88] A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3–24.

[89] C. Stein (1971), A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proc. Sixth Berkeley Symp. Math. Statist. Prob. **3**, 583–602. University of California Press.

[90] H. Thorisson (2000), Coupling, Stationarity, and Regeneration. Springer, New York.

[91] E. Thönnes, A Primer in Perfect Simulation. In Statistical Physics and Spatial Statistics (K.R. Mecke and D. Stoyan, ed.), Springer Lecture Notes in Physics, 349–378.

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

[93] P. Tuominen and R.L. Tweedie (1994), Subgeometric rates of convergence of f-ergodic Markov chains. Adv. Appl. Prob. **26**, 775–798.