PROBABILITY AND MCMC

Course notes for STA 4276H, Fall 2003.

by

Jeffrey S. Rosenthal

University of Toronto

jeff@math.toronto.edu

http://probability.ca/jeff/

Last updated December 1, 2003.

Based partially upon previous course notes written jointly with Gareth O. Roberts, Lancaster, July 2003.

1. Introduction.

Markov chain Monte Carlo (MCMC) algorithms – such as the Metropolis-Hastings algorithm ([45], [32]) and the Gibbs sampler ([29], [27]) – have become extremely popular in statistics, as a way of approximately sampling from complicated probability distributions in high dimensions (see for example the reviews [77], [74], [30], [59]). In particular, the existence of MCMC algorithms has transformed Bayesian inference, by allowing practitioners to sample from posterior distributions of complicated statistical models.

In addition to their importance in applied statistics, these algorithms also raise numerous questions related to probability theory and the mathematics of Markov chains. Specifically, 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 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 the long analytic technicalities of some 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.

1.1. The problem.

The problem addressed by MCMC algorithms is the following. We're given a (possibly un-normalised) density function π_u, on some state space $\mathcal{X}$. (Typically $\mathcal{X}$ is an open subset of $\mathbf{R}^d$, and the densities are taken with respect to Lebesgue measure, though other settings – such as 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.

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

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 variance of order $O(1/\sqrt{N})$. Furthermore, 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 its 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", 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 B is 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.

Remark. MCMC is, of course, not the only way to sample or estimate from complicated probability distributions. Other possible sampling algorithms include:

(1) "Rejection sampling": Suppose g is easy to sample from, and we can find $c < \infty$ such that $\pi_u(x) \leq c g(x)$ for all $x \in \mathcal{X}$. We then simulate $Z \sim g(x)dx$, and also simulate $U \sim \text{Uniform}[0,1]$. If $U \leq \pi_u(Z) \,/\, c\, g(Z)$ then we output Z. Otherwise, we *reject* Z, and begin again with fresh simulations of Z and U. This algorithm outputs an observation from π. However, it is usually difficult to find an appropriate function g such that the algorithm's rejection rate will be significantly less than 1.

(2) "Importance sampling": Given a sample $X_1, X_2, \ldots X_n \sim g(x)\,dx$, estimate $E_\pi[f(X)]$ by

$$\hat{\pi}(f) = \frac{\sum_{i=1}^n f(X_i) w(X_i)}{\sum_{i=1}^n w(X_i)}$$

where $w(x) = \frac{\pi_u(x)}{g(x)}$. However, $w(x)$ may be small, and estimation errors may be large.

1.2. Motivating example: Bayesian Statistics.

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 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 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.

2.1. A Very Simple Example.

Suppose $\mathcal{X} = \{1, 2, 3\}$, with:

$$\pi\{1\} = 1/6\,, \quad \pi\{2\} = 1/3\,, \quad \pi\{3\} = 1/2\,.$$

Let $P(1, \{2\}) = 1$, $P(2, \{1\}) = P(2, \{3\}) = 1/2$, and $P(3, \{2\}) = 1/3$, $P(3, \{3\}) = 2/3$. Then

$$\begin{aligned} \int \pi\{x\}\, P(x, \{2\}) &= \pi\{1\}\, P(1, \{2\}) + \pi\{3\}\, P(3, \{2\}) \\ &= (1/6)(1) + (1/2)(1/3) = 1/3 = \pi\{2\}\,. \end{aligned}$$

Similarly $\int \pi\{x\}\, P(x, \{1\}) = \pi\{1\}$ and $\int \pi\{x\}\, P(x, \{3\}) = \pi\{3\}$. Success!

So, for large n, probably have $\mathcal{L}(X_n) \approx \pi(\cdot)$. [Good!]

But how did we know? And, how would we proceed for a more *complicated* example? 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

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

The usefulness of reversibility is the following.

Proposition 1. *If 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.

2.2. The Metropolis-Hastings Algorithm.

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

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 $C = \int_{\mathcal{X}} \pi_u(x)\, dx$,

$$\pi(dx)\, P(x, dy) = [C^{-1} \pi_u(x)\, dx]\, [q(x, y)\, \alpha(x, y)\, dy]$$

$$= C^{-1} \pi_u(x)\, q(x, y)\, \min\left[1,\ \frac{\pi_u(y) q(y, x)}{\pi_u(x) q(x, y)}\right] dx\, dy$$

$$= C^{-1} \min[\pi_u(x)\, q(x, y),\ \pi_u(y) q(y, x)]\, dx\, dy\,,$$

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.

2.3. A Metropolis-Hastings Example.

Suppose $\mathcal{X} = \mathbf{R}^d$, with $\pi : \mathcal{X} \to (0, \infty)$ a complicated density function.

Let the proposal be $Q(x, \cdot) = N(x, 1)$, so that $q(x, y) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2} = q(y, x)$.

Then

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

Take $X_0 = 0$ (say).

Then, given X_n, we choose $Y_{n+1} \sim N(X_n, 1)$, and then set

$$X_{n+1} = \begin{cases} Y_{n+1}\,, & \text{probability } \alpha(X_n, Y_{n+1}) \\ X_n\,, & \text{probability } 1 - \alpha(X_n, Y_{n+1}) \end{cases}$$

This creates a sequence $X_0, X_1, X_2, \ldots$

Hopefully, for large n, the density of X_n is approximately equal to π.

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

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

- **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).

2.4. 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. [77], [57]), 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.

2.5. 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 $\mathbf{R}^d$. Write $\mathbf{x} = (x_1, \ldots, x_d)$, and $\mathbf{x}^{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

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 all 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.

2.6. 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 $\{Y_{ij}\}$ $(1 \leq i \leq K,\ 1 \leq j \leq J)$ 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:

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

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

$$\pi_u(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \propto$$

$$e^{-b_1/\sigma_\theta^2} \sigma_\theta^{2^{-a_1-1}} e^{-b_2/\sigma_e^2} \sigma_e^{2^{-a_2-1}} e^{-(\mu-\mu_0)^2/2\sigma_0^2}$$

$$\times \prod_{i=1}^{K} [e^{-(\theta_i-\mu)^2/2\sigma_\theta^2}/\sigma_\theta] \times \prod_{i=1}^{K}\prod_{j=1}^{J} [e^{-(Y_{ij}-\theta_i)^2/2\sigma_e^2}/\sigma_e] \,.$$

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 works well; this model was one of the original statistical applications of the Gibbs sampler [27], and versions of it have been used and studied often (see e.g. [65], [48], [67], [19], [38], [39]).

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 works well for this model, at least if the value of σ^2 is chosen appropriately. (This issue is discussed further herein.) 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. Convergence Issues.

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:

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

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? We shall measure this question in terms of total variation distance:

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) If $\pi(\cdot)$ is stationary for a Markov chain, then $\|P^n(x, \cdot) - \pi(\cdot)\|$ is non-increasing in n, i.e. $\|P^{n+1}(x, \cdot) - \pi(\cdot)\| \le \|P^n(x, \cdot) - \pi(\cdot)\|$.
(c) Let $t(n) = \sup_{x\in\mathcal{X}} \|P^n(x, \cdot) - \pi(\cdot)\|$, where $\pi(\cdot)$ is stationary. Then t is sub-multiplicative, i.e. $t(m + n) \le t(m)t(n)$ for $m, n \in \mathbf{N}$.

Proof. For (a), note that $|\nu_1(A) - \nu_2(A)| = |\int f d\nu_1 - \int f d\nu_2|$ where $f = \mathbf{1}_A$, thus proving that $\|\nu_1(\cdot) - \nu_2(\cdot)\| \leq \sup_{f:\mathcal{X}\to[0,1]} |\int f d\nu_1 - \int f d\nu_2|$. For the converse, let $\rho(\cdot)$ be any 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 $0 \leq f \leq 1$) when $f = 1$ on $\{g > h\}$ and $f = 0$ on $\{h > g\}$, in which case it equals $|\nu_1(A) - \nu_2(A)|$ for $A = \mathbf{1}_{g>h}$, thus proving the equivalence.

For part (b), we compute that

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

$$= |\int_{y\in A} P^n(x,dy)f(y) - \int_{y\in A} \pi(dy)f(y)| \leq \|P^n(x,\cdot) - \pi(\cdot)\|\,,$$

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

Part (c) follows since $t(n)$ is an L^∞ operator norm. 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

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

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

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

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 $g^* \leq t(m)$ by part (a). Now, if $g^* = 0$, then clearly $\hat{P}\hat{Q}f = 0$. Otherwise, we compute that

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

Since $0 \leq g/g^* \leq 1$, we have $(\hat{P}[g/g^*])(x) \leq \|P^n(x,\cdot) - \pi(\cdot)\|$ by part (a), so that $\sup_{x\in\mathcal{X}}(\hat{P}[g/g^*])(x) \leq t(n)$. The result then follows from part (a) together with (5). ∎

3.1. 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 $P(1, \{2\}) = P(2, \{1\}) = 1/2$, and $P(3, \{3\}) = 1$. Let $X_0 = 1$. Then $\pi(\cdot)$ is stationary. However, $X_n \in \{1, 2\}$ for all n, so $P(X_n = 3) = 0$ for all n, so $P(X_n = 3) \not\to \pi\{3\}$, and there is no convergence.

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 can eventually reach 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_* is eventually reachable from any state x. Thus, if a chain has any one state which is reachable from anywhere, 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, so that $\phi(\{x\}) = 0$ for all singleton sets, but such that all A with $\phi(A) > 0$ are eventually reachable from any $x \in \mathcal{X}$.

Even ϕ-irreducible chains might not converge, due to periodicity problems:

Example 2. Suppose again $\mathcal{X} = \{1, 2, 3\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$. Let $X_0 = 1$. Then $\pi(\cdot)$ is stationary, and the chain is ϕ-irreducible [e.g. with $\phi(\cdot) = \delta_1(\cdot)$]. However, $X_n = 1$ 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 stationarity.

To avoid this problem, we require aperiodicity:

Definition. A Markov chain 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 $\pi(\mathcal{X}_1) > 0$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$.

Now we can state the main asymptotic convergence theorem, whose proof is described in Section 4.

Theorem 4. *If a Markov chain 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, it follows that $\lim_{n\to\infty} P^n(x,A) = \pi(A)$ for all measurable $A \subseteq \mathcal{X}$. In fact, for any $h : \mathcal{X} \to \mathbf{R}$ with $\pi(|h|) < \infty$, a "strong law of large numbers" also holds (see e.g. [46]):

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

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 set of π-measure 0, and fail to converge there. Here is a simple example due to C. Geyer:

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 $\mathbf{P}(X_n = x + n$ 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 $x \geq 2$.

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 $\pi(\cdot)$ $[P(x, dy) = p(x, y)\, \pi(dy)]$, or for any Metropolis algorithm whose proposal distributions $Q(x, \cdot)$ are absolutely continuous with respect to $\pi(\cdot)$. 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: $\mathbf{P}[\exists n;\ X_n \in B \,|\, X_0 = x] = 1$. This condition is stronger than π-irreducibility; for discussions of this see e.g. [77], [62].

3.2. 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

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

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

The difference between uniform ergodicity and geometric ergodicity is that now the constant C does not depend on the initial state x. One equivalence of uniform ergodicity follows immediately from Proposition 3(c):

Proposition 5. *A Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{x \in \mathcal{X}} \|P^n(x, \cdot) - \pi(\cdot)\| < 1$ for some $n \in \mathbf{N}$.*

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:

$$P^{n_0}(x, \cdot) \ \geq\ \epsilon\, \nu(\cdot) \qquad x \in C\,, \tag{6}$$

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

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. (For background, see e.g. [21], [7], [50], [3], [46]; for applications to convergence rates see e.g. [47], [66], [67], [59], [64], [22], [70].) We note that if $\mathcal{X}$ is discrete, and if

$$\epsilon = \sum_{y \in \mathcal{X}} \inf_{x \in C} P^{n_0}(x, \{y\}) > 0\,, \tag{7}$$

then C is (n_0, ϵ, ν)-small where $\nu\{y\} = \epsilon^{-1} \inf_{x \in C} P^{n_0}(x, \{y\})$.

Remark. As observed in [60], 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 coupling (which includes all the bounds considered here). For ease of exposition, we do not pursue that here.

The main result guaranteeing uniform ergodicity is the following.

Theorem 6. *Consider a Markov chain with invariant probability distribution $\pi(\cdot)$. Suppose the minorisation condition (6) 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 $\|P^n(x, \cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \mathcal{X}$, where $\lfloor r \rfloor$ is the greatest integer not exceeding r.*

Theorem 6 is proved in Section 4. For example, on a discrete state space, we conclude using (7) that $\|P^n(x, \cdot) - \pi(\cdot)\| \leq (1 - \epsilon_{n_0})^{\lfloor n/n_0 \rfloor}$ where $\epsilon_{n_0} = \sum_{y \in \mathcal{X}} \inf_{x \in \mathcal{X}} P^{n_0}(x, \{y\})$. Furthermore, for an irreducible aperiodic chain on a *finite* state space, we always have $\epsilon_{n_0} > 0$ for sufficiently large n_0. Similarly, if the transition probabilities have densities, $P^{n_0}(x, dy) = p_{n_0}(x, y)\, dy$, then we can take $\epsilon_{n_0} = \int_{y \in \mathcal{X}} \big(\inf_{x \in \mathcal{X}} p(x, y)\big) dy$.

We note also that Theorem 6 provides a *quantitative* bound on the distance to stationarity $\|P^n(x, \cdot) - \pi(\cdot)\|$, namely that it must be $\leq (1 - \epsilon)^{\lfloor n/n_0 \rfloor}$. Thus, once n_0 and ϵ are known, we can find n_* so that, say, $\|P^{n_*}(x, \cdot) - \pi(\cdot)\| \leq 0.01$; we can then say that n_* iterations "suffices for convergence" of the Markov chain.

If a Markov chain is *not* uniformly ergodic (as few MCMC algorithms on unbounded state spaces are), then Theorem 6 cannot be applied. However, it is still of great importance to be able to find n_* so that $\|P^{n_*}(x,\cdot)-\pi(\cdot)\| \leq 0.01$. This issue is discussed in the following sections.

Example 4. Consider the Markov chain on $\mathcal{X} = \{1,2,3,4,5\}$ with transition matrix

$$P \;=\; \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 & 0 \\ 0.4 & 0 & 0.3 & 0.3 & 0 \\ 0.2 & 0.2 & 0.4 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.1 & 0.3 \\ 0.2 & 0 & 0.5 & 0.3 & 0 \end{pmatrix}$$

We see by inspection that the column minimums are $0.2, 0, 0.3, 0.1, 0$, respectively. We may let this be the measure $\epsilon\, \nu(\cdot)$. Thus, we may take $\epsilon = 0.2+0.3+0.1 = 0.6$, and $\nu(1) = 1/3$, $\nu(3) = 1/2$, and $\nu(4) = 1/6$, and immediately conclude that $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (0.4)^n$. ■

Example 5. Let $\mathcal{X} = [0,1]$ (the interval from 0 to 1), and set

$$P(x,dy) \;=\; \frac{1+x+y}{\frac{3}{2}+x}\; dy\,.$$

We see by inspection that $P(x,dy) \geq \frac{2}{3}\, dy$ for all x and y, so that we may take $\epsilon = \frac{2}{3}$ to conclude that $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1/3)^k$.

We can do even better by choosing

$$\epsilon \;=\; \int_0^1 \left(\inf_{0\leq x\leq 1} \frac{1+x+y}{\frac{3}{2}+x} \right) dy \;=\; \int_0^{\frac{1}{2}} \frac{2}{3}(1+y)dy \;+\; \int_{\frac{1}{2}}^1 \frac{2}{5}(2+y)dy \;=\; \frac{29}{30}\,,$$

and letting $\nu(\cdot)$ has density (with respect to dy) given by $\frac{30}{29}\,\frac{2}{3}(1+y)$ for $0 \leq y \leq \frac{1}{2}$, and by $\frac{30}{29}\,\frac{2}{5}(2+y)$ for $\frac{1}{2} < y \leq 1$. Hence, we actually have $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1/30)^k$. ■

3.3. Geometric ergodicity.

A slightly weaker condition than uniform ergodicity is geometric ergodicity:

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

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

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

Of course, if the state space $\mathcal{X}$ is *finite*, then all irreducible and aperiodic Markov chains are geometrically ergodic. However, for infinite $\mathcal{X}$ this is not the case. For example, it is shown in [44] (see also [63]) 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*, see [26] and [37].) Hence, we now discuss conditions which ensure geometric ergodicity.

Definition. Given Markov chain transition probabilities P on a state space $\mathcal{X}$, and a function $f : \mathcal{X} \to \mathbf{R}$, define the function $Pf : \mathcal{X} \to \mathbf{R}$ so 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

$$PV \leq \lambda V + b\mathbf{1}_C\,, \tag{8}$$

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 about geometric ergodicity is the following.

Theorem 7. *Consider a ϕ-irreducible, aperiodic Markov chain with invariant measure π. Suppose the minorisation condition (6) is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition (8) 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 7 is usually proved by complicated analytic arguments (see e.g. [50], [46], [6]). In Section 4, we describe a proof of Theorem 7 which uses direct coupling arguments. Note also that Theorem 7 provides no *quantitative* bounds on $C(x)$ or ρ, though this is remedied in Theorem 10 below.

Remark 8. In fact, it follows from Theorems 15.0.1 and 16.0.1 and 14.3.7 of [46], and Proposition 1 of [57], that the minorisation condition (6) and drift condition (8) of Theorem 7 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

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

where $\pi(f) = \int_{x \in \mathcal{X}} f(x) \, \pi(dx)$; furthermore, we always have $\pi(V) < \infty$.

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 6. ([59]) 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, with $\lambda = 5$, the sampler will usually return a mean value of about 0.8 instead of 1, and then occasionally return very large values 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.

Example 7. ("Witch's Hat", see e.g. [43]) 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 <u>miss</u> 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 8. Let $\mathcal{X} = \mathbf{R}$, and let $\pi_u(x) = 1/(1+x^2)$ be (unnormalised) Cauchy distribution. Then a random-walk Metropolis algorithm for π_1 (with, say, $Q(x,\cdot) = \text{Uniform}[x-1, x+1]$) is ergodic but is <u>not</u> geometrically ergodic. And, indeed, this sampler has very slow, poor convergence properties. On the other hand, let $\pi'_u(x) = \pi_u(x)\,\mathbf{1}(|x| \le 10^{100})$, i.e. π'_u corresponds to π_u truncated at $\pm$ one google. Then the same random-walk Metropolis algorithm for π'_u <u>is</u> geometrically ergodic, in fact uniformly ergodic. On the other hand, the two algorithms are <u>indistinguishable</u> when run for any remotely feasible amount of time. Thus, this is an example where geometric ergodicity does not in any way indicate improved performance of the algorithm.

Due to these limitations, it is desirable where possible to have *quantitative* bounds on convergence, not just *qualitative* bounds such as geometric ergodicity.

3.4. A Quantitative Bound.

In light of the above, we ideally want *quantitative* bounds on convergence rates, i.e. bounds of the form $\|P^n(x,\cdot) - \pi(\cdot)\| \le g(x,n)$ 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. [47], [66], [67], [42], [19], [64], [38], [39], [22], [10], [70], [25], [8], [71], [72].)

We here present a result from [70], which follows as a special case of [22]; it is based on the approach of [66] while also taking into account an improvement from [64].

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

$$\overline{P}h(x,y) \;\le\; h(x,y)\,/\,\alpha\,, \qquad (x,y) \notin C \times C \tag{9}$$

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

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

Of course, (9) is closely related to (8); for example we have the following (see also [66], and Proposition 2 of [72]):

Proposition 9. *Suppose the univariate drift condition (8) is satisfied for some $V : \mathcal{X} \to [1, \infty]$, $C \subseteq \mathcal{X}$, $\lambda < 1$, and $b < \infty$. Let $d = \inf_{x \in C} V(x)$. Then if $d > [b/(1-\lambda)] - 1$, then the bivariate drift condition (9) 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 $h(x, y) \geq \frac{1}{2}(1 + d)$, and $PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$. Then

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

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

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

Finally, we let

$$B = \max[1,\, \alpha^{n_0}(1-\epsilon) \sup_{C \times C} \overline{R}h]\,, \tag{10}$$

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 10. *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 $\mathcal{X}$, $\alpha > 1$, and $\epsilon > 0$, such that (6) and (9) hold. Define B by (10). 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*

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

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

Theorem 10 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. [67] and [39]. Thus, while it is admittedly hard work to apply Theorem 10 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 can be difficult or impossible to apply Theorem 10 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. [28], [17]). This technique often works well in practice, though it provides no rigorous guarantees and can sometimes be fooled into prematurely claiming convergence (see e.g. [43]) and can also introduce bias into the resulting estimates (see [18]). Overall, the "convergence time problem" remains largely unresolved for practical application of MCMC. (This is also the motivation for "perfect MCMC" algorithms, not reviewed here; see e.g. [53].)

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], [20], [73], [2], [75]). Here we focus on the method of *coupling*, which seems particularly well-suited to analysing MCMC algorithms. It is also particularly well-suited to encorporating small sets (though small sets can also be combined with *regeneration theory*, see e.g. [7], [3], [48], [33]). Some of the proofs below are new, and avoid many of the long analytic arguments of some previous proofs (e.g. [46]). Note that we prove the theorems in a different order from which we stated them; however, in the course of the proofs we only make use of results that we have already proved.

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

$$\begin{aligned} \|\mathcal{L}(X) - \mathcal{L}(Y)\| &= \sup_A |P(X \in A) - P(Y \in A)| \\ &= \sup_A |P(X \in A, X = Y) + P(X \in A, X \neq Y) \\ &\qquad - P(Y \in A, Y = X) - P(Y \in A, Y \neq X)| \qquad (11) \\ &= \sup_A |P(X \in A, X \neq Y) - P(Y \in A, Y \neq X)| \\ &\leq P(X \neq Y)\,. \end{aligned}$$

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. [52], [41], [76].

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 [49] and Athreya and Ney [7]; see also [50] and [46]. 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 $X'_0 \sim \pi(\cdot)$.

Given X_n and X'_n:

1. If $X_n = X'_n$, choose $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$, replace n by $n+1$, and repeat.
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. $1-\epsilon$, 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 for Y. In any case, replace n by $n + n_0$, and repeat.
3. Else, independently choose $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$, replace n by $n+1$, and repeat.

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.

The *coupling inequality* then says that $\|P^n(x,\cdot) - \pi(\cdot)\| \le \mathbf{P}[X_n \ne X'_n]$. 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 6.

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 $\mathbf{P}[X_n \ne X'_n] \le (1-\epsilon)^m$. Hence, from the coupling inequality, $\|P^n(x,\cdot) - \pi(\cdot)\| \le (1-\epsilon)^m = (1-\epsilon)^{n/n_0}$ in this case. It then follows from Proposition 3(b) that $\|P^n(x,\cdot) - \pi(\cdot)\| \le (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for any n. ∎

4.4. Proof of Theorem 10.

We follow the general outline of [70]. We again begin by assuming that $n_0 = 1$ in the minorisation condition for the small set C, and indicate at the end what changes are required if $n_0 > 1$.

Let

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

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 $1 \le j \le k$,

$$P[X_k \ne X'_k] \;=\; P[X_k \ne X'_k,\; N_{k-1} \ge j] \;+\; P[X_k \ne X'_k,\; N_{k-1} < j]\,. \tag{12}$$

Now, the event $\{X_k \ne X'_k,\; N_{k-1} \ge j\}$ is contained in the event that the first j coin flips all came up tails. Hence, $P[X_k \ne X'_k,\; N_{k-1} \ge j] \le (1-\epsilon)^j$. which bounds the first term in (12).

To bound the second term in (12), let

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

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

Lemma 11. *We have*

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

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

Proof. Since our coupling construction is Markovian,

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

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

$$E[M_{k+1} \,|\, X_k, X'_k, \{T > k\}] \;=\; \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k] \mathbf{1}(X_k \ne X'_k)$$

$$=\; M_k\, \alpha E[h(X_{k+1}, X'_{k+1}) \,|\, X_k, X'_k] \,/\, h(X_k, X'_k)$$

$$\leq \; M_k \, ,$$

by (9). 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

$$E[M_{k+1} \,|\, X_k, X'_k] \;=\; \alpha^{k+1} B^{-N_{k-1}-1} E[h(X_{k+1}, X'_{k+1}) \, \mathbf{1}(X_{k+1} \neq X'_{k+1}) \,|\, X_k, X'_k]$$

$$= \; \alpha^{k+1} B^{-N_{k-1}-1} (1-\epsilon)(\overline{R}h)(X_k, X'_k)$$

$$= \; M_k \, \alpha \, B^{-1} (1-\epsilon)(\overline{R}h)(X_k, X'_k)$$

$$\leq \; M_k \, ,$$

by (10). Hence, $\{M_k\}$ is a supermartingale. ∎

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

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

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

$$\leq \; B^{j-1} E[\mathbf{1}(X_k \neq X'_k) \, B^{-N_{k-1}}] \qquad \text{(by Markov's inequality)}$$

$$\leq \; B^{j-1} E[\mathbf{1}(X_k \neq X'_k) \, B^{-N_{k-1}} \, h(X_k, X'_k)] \qquad \text{(since } h \geq 1\text{)}$$

$$= \; \alpha^{-k} B^{j-1} E[M_k] \qquad \text{(by defn of } M_k\text{)}$$

$$\leq \; \alpha^{-k} B^{j-1} E[M_0] \qquad \text{(since } \{M_k\} \text{ is supermartingale)}$$

$$= \; \alpha^{-k} B^{j-1} E[h(X_0, X'_0)] \qquad \text{(by defn of } M_0\text{)} \, .$$

Theorem 10 now follows (in the case $n_0 = 1$), by combining these two bounds with (12) and (11).

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 Y] 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 (12) 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, usually $t(k) = k$, unless the joint chain visited $C \times C$ at some time between $k - n_0$ and $k - 1$.) [CHECK THIS BIT!] ∎

4.5. Proof of Theorem 7.

Here we give a direct coupling proof of Theorem 7, thereby avoiding the long technicalities of [46] (though admittedly with a slightly weaker conclusion; see Remark 8). Our proof requires one additional assumption (13) below, leading to the following question:

Open Question. *Can we prove Theorem 7 from direct coupling methods without assuming (13)?*

We now proceed to the proof of Theorem 7. Our approach shall be to make use of Theorem 10. To begin, set $h(x, y) = \frac{1}{2}[V(x) + V(y)]$. We shall assume that

$$\sup_{C \times C} \overline{R}h(x, y) \ < \ \infty \,, \tag{13}$$

which ensures that the quantity B of (10) is finite. Now, by (8), we see that (13) follows immediately if $\sup_C V < \infty$, and this condition holds in all practical problems that we are aware of. However, (13) might not hold in general.

Assuming (13), we will have $B < \infty$ in Theorem 10. To continue, let $d = \inf_C V$. Then we see from Proposition 9 that the bivariate drift condition (9) will hold, *provided* that $d > b/(1 - \lambda) - 1$. In that case, Theorem 7 follows immediately (in fact, in a *quantitative* version) by combining Proposition 9 with Theorem 10.

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 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 (6) still holds though perhaps for uncontrollably larger n_0 and smaller $\epsilon > 0$]. Theorem 7 will then follow from Proposition 9 and Theorem 10 as above. [Note that we will have no direct control over the new values of n_0 and C, 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 $\inf_{x \in C'} V(x) \geq d' > b/(1-\lambda) - 1$. Thus, Theorem 7 will follow from Proposition 9 and Theorem 10 if we can prove:

Lemma 12. *C' is a small set.*

To prove Lemma 12, we make use the notion of "petite set", following [46].

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

$$\sum_{i=1}^{n_0} P^i(x, \cdot) \geq \epsilon\, \nu(\cdot) \qquad x \in C\,. \tag{14}$$

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 small. 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 $x \in C$ cover $\epsilon\, \nu(\cdot)$ only at odd times, and others only at even times). Surprisingly, however, for an aperiodic ϕ-irreducible Markov chain, all petite sets are also small ([46], Theorem 5.5.7), and we shall use this to our advantage below.

Lemma 13. *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 (8) implies that W_n is a supermartingale. Hence, for $x \in S$, using Markov's inequality and the fact that $V \geq 1$,

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

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

$$\leq \lambda^N \, \mathbf{E}[Z_0 \,|\, X_0 = x] = \lambda^N \, V(x) \leq \lambda^N \, d\,,$$

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. ■

Lemma 14. *For an aperiodic, ϕ-irreducible Markov chain, all petite sets are small sets.*

To prove Lemma 14, we require a lemma related to aperiodicity:

Lemma 15. *Consider an aperiodic, ϕ-irreducible Markov chain on a state space $\mathcal{X}$. Let $\nu(\cdot)$ be any probability measure on $\mathcal{X}$. 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 $\int_{x\in\mathcal{X}} \nu(dx)\, P^{n+m}(x, \cdot) \geq \int_{x\in R} \nu(dx) P^n(x, dy) P^m(y, \cdot) \geq \delta_n\, P^m(y, \cdot)$, we see that T is *additive*, i.e. if $n, m \in T$ then $n + m \in T$.

Furthermore, we claim that $gcd(T) = 1$. Indeed, if $gcd(T) = d > 1$, then for $1 \leq i \leq d$, let $\mathcal{X}_i = \{x \in \mathcal{X};\ \exists \ell \in \mathbf{N},\ \int \nu(dx)\, P^{\ell d+i}(x, C) > 0\}$. Then $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ are disjoint subsets with respect to which the Markov chain is periodic (of period d), contradicting the assumption of aperiodicity.

Hence, $gcd(T) = 1$, and T is additive. It is then a standard and easy fact (cf. [12], p. 541; or [68], p. 77) that there is $n_* \in \mathbf{N}$ such that $T \supseteq \{n_*, n_* + 1, n_* + 2, \ldots\}$, as claimed. ■

Proof of Lemma 14. 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 15. 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 15, 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 $r = \min\left\{\delta_n;\ n_* \leq n \leq n_* + n_0 - 1\right\}$, and set $N = n_* + n_0$. Then for $x \in R$,

$$P^N(x, \cdot) \geq \sum_{i=1}^{n_0} \int_{y\in\mathcal{X}} P^{N-i}(x, dy) P^i(y, \cdot)$$

$$\geq \sum_{i=1}^{n_0} \int_{y\in R} r\nu(dy) P^i(y, \cdot)$$

$$\geq \int_{y\in R} r\nu(dy)\epsilon\nu(\cdot) = r\epsilon\nu(\cdot)\,.$$

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

Combining Lemmas 13 and 14, we see that C' must be small, proving Lemma 12, and hence proving Theorem 7. ∎

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 [36] (see also Orey [51]). We shall not prove it here; for modern proofs see e.g. [50], p. 16, or [46], Theorem 5.2.2. The key idea (cf. [46], 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 16. *(Jain and Jameson [36]) Every ϕ-irreducible Markov chain contains a small set $C \subseteq \mathcal{X}$ 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 [69]. We begin with a lemma about return probabilities:

Lemma 17. *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. Then we make the following claims (proved below):

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

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

Claim 2. Let $S = \sup\{k \geq 1;\ X_{k\ell} \in A\}$. Then for integers $1 \leq r \leq j$,

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

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

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

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

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

Proof of Claim 1. Since the conclusion does not hold, 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 $\eta = \#\{k \geq 1;\ X_{\ell_0 k} \in B_2\}$. 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

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

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

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

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

Proof of Claim 2. We compute using stationarity that

$$\begin{aligned}
&\int_{x \in \mathcal{X}} \pi(dx)\, \mathbf{P}_x[S = r,\ X_{jn} \notin A] \\
&= \int_{x \in \mathcal{X}} \pi(dx) \int_{y \in \mathcal{X}} P^{rn}(x, dy)\, \mathbf{P}_y[S = \infty,\ X_{(j-r)\ell} \notin A] \\
&= \int_{y \in \mathcal{X}} \int_{x \in \mathcal{X}} \pi(dx) P^{rn}(x, dy)\, \mathbf{P}_y[S = \infty,\ X_{(j-r)\ell} \notin A] \\
&= \int_{y \in \mathcal{X}} \pi(dy)\, \mathbf{P}_y[S = \infty,\ X_{(j-r)\ell} \notin A] \\
&\geq \int_{y \in B} \pi(dy)\, \mathbf{P}_y[S = \infty,\ X_{(j-r)\ell} \notin A] \\
&\geq \int_{y \in B} \pi(dy)\, \delta = \pi(B)\, \delta\,.
\end{aligned}$$ ∎

To proceed, we let C be a small set as in Theorem 16. 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 18. $\pi(\overline{G}) = 1$.

Proof. We first prove that $(\pi \times \pi)(G) = 1$. Indeed, since $\nu(C) > 0$ by Theorem 16, it follows from Lemma 15 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 17 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 $R \times R$ which must be absolutely continuous with respect to $\pi \times \pi$, and hence (again by Lemma 17) 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 $(\pi \times \pi)(G) = 1$. ∎

5. Central Limit Theorems.

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 error sum $n^{-1/2} \sum_{i=1}^{n} [h(\mathbf{X}_i) - \pi(h)]$ 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 (cf. [14]) that

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \mathbf{E}\Big[\Big(\sum_{i=1}^{n} [h(X_i) - \pi(h)]\Big)^2\Big], \tag{15}$$

and also $\sigma^2 = \tau\, \mathbf{Var}_\pi(h)$, where $\tau = \sum_{k\in\mathbf{Z}} \mathrm{Corr}(X_0, X_k)$ is the *integrated autocorrelation time*. (In the reversible case this is also related to spectral measures; see e.g. [40], [31], [57].) 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. [31], [77], [14], [46], [63], [33], [5]).

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 [55] 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 19. *Consider a reversible Markov chain, and let* $r(x) = \mathbf{P}(X_{n+1} = X_n \mid X_n = x)$. *Then if*

$$\lim_{n\to\infty} \; n\, \pi\big([h - \pi(h)]^2\, r^n\big) \;=\; \infty\,, \tag{16}$$

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

Proof. We compute directly from (15) that

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

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

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

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

by (16). Hence, a $\sqrt{n}$-CLT cannot exist. ∎

In particular, Theorem 19 is used in [55] 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 (cf. [12], Theorem 27.1; [68], p. 110). For *uniformly ergodic* chains, an identical result exists; it is shown in Corollary 4.2(ii) of [16] (cf. Theorem 5 of [77]) that:

Theorem 20. *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 [35] (see also Theorem 2 of [14], and Theorem 2 of [33]) that:

Theorem 21. *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 [57] that Theorem 21 can be strenghthened if the chain is *reversible*:

Theorem 22. *If the Markov chain is geometrically ergodic and reversible, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.*

Comparing Theorems 22 and 21 leads to the following open question (see [5]):

Open Question. *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?*

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.

If P is *reversible*, then it was proved by Kipnis and Varadhan [40] that finiteness of σ^2 is all that is required:

Theorem 23. *(Kipnis and Varadhan [40]) For a reversible Markov chain, a $\sqrt{n}$-CLT holds for h whenever $\sigma^2 < \infty$, where σ^2 is giving by (15).*

In a different direction, we have the following:

Theorem 24. *Suppose a Markov chain is geometrically ergodic, satisfying (8) for some $V : \mathcal{X} \to [1, \infty)$. 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.*

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 [11] and Ibragimov [34]; see e.g. p. 375 of [23].

Theorem 25. *(Billingsley [11] and Ibragimov [34]) Let $\{Z_n\}$ be a stationary ergodic sequence, with $\mathbf{E}[Z_n \,|\, Z_1, \ldots, Z_{n-1}] = 0$ 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 25, consider the *Poisson equation*: $h - \pi(h) = g - Pg$. A useful result is the following (cf. Theorem 17.4.4 of [46]):

Theorem 26. *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 $Z_n = g(X_n) - Pg(X_n)$. 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, $\mathbf{E}_\pi[Z_n^2] \leq 4\, \pi(g^2) < \infty$. Also, by the double-expectation formula,

$$\mathbf{E}_\pi[g(X_n) - Pg(X_n) \,|\, X_0, \ldots, X_{n-1}] = \mathbf{E}_\pi[g(X_n) - \mathbf{E}[X_{n+1} \,|\, X_n] \,|\, X_{n-1}]$$

$$= \mathbf{E}_\pi[g(X_n) \,|\, X_{n-1}] - \mathbf{E}_\pi[g(X_n) \,|\, X_{n-1}] = 0\,.$$

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 25, $n^{-1/2} \sum_{i=1}^n Z_i$ converges weakly to $N(0, \sigma^2)$. But

$$\sum_{i=1}^n Z_i = \sum_{i=1}^n [g(X_i) - Pg(X_i)] = \sum_{i=1}^n [h(X_i) - \pi(h)]\,.$$

The result follows. ∎

Corollary 27. *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

$$\begin{aligned}(g - Pg)(x) &= \sum_{k=0}^{\infty} g_k(x) - \sum_{k=0}^{\infty} P g_k(x) = \sum_{k=0}^{\infty} g_k(x) - \sum_{k=1}^{\infty} g_k(x) \\ &= g_0(x) = P^0 h(x) - \pi(h) = h(x) - \pi(h)\,.\end{aligned}$$

Hence, the result follows from Theorem 26, *provided* that $\pi(g^2) < \infty$. On the other hand, it is known (in fact, since $\mathbf{Cov}(X,Y) \le \sqrt{\mathbf{Var}(X)\,\mathbf{Var}(Y)}$) that the $L^2(\pi)$ norm satisfies the triangle inequality, so that

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

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

Proof of Theorem 22. Let

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

be the usual $L^2(\pi)$ operator norm for P (restricted to those functionals f with $\pi(f) = 0$ and $\pi(f^2) < \infty$). Then it is shown in Theorem 2 of [57] that reversible chains are geometrically ergodic if and only if they satisfy $\|P\|_{L^2(\pi)} < 1$, i.e. there is $\beta < 1$ with $\pi((Pf)^2) \le \beta^2 \pi(f^2)$ whenever $\pi(f) = 0$ and $\pi(f^2) < \infty$. Furthermore, reversibility implies self-adjointness of P in $L^2(\pi)$, so that $\|P^k\|_{L^2(\pi)} = \|P\|_{L^2(\pi)}^k$, and hence $\pi((P^k f)^2) \le \beta^{2k} \pi(f^2)$.

Let $g_k = P^k h - \pi(h)$ as in the proof of Corollary 27. Then this implies that $\pi((g_k)^2) \le \beta^{2k}\, \pi((h-\pi(h)^2)$, so that

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

Hence, the result follows from Corollary 27. ∎

Proof of Theorem 24. By Remark 8, 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 27. 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 27. ∎

5.4. Proof of Theorem 21 using Regenerations.

Here we use *regeneration theory* to give a reasonably direct proof of Theorem 21, following the outline of [33], thereby avoiding the technicalities of the original proof [35].

We begin by noting from Remark 8 that since the chain is geometrically ergodic, there is a small set C and a drift function V satisfying (6) and (8).

In terms of this, we consider a *regeneration construction* for the chain (cf. [7], [3], [48], [33]). 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

$$\sum_{i=1}^{n}[h(X_i) - \pi(h)] \;=\; \sum_{j=1}^{r(n)} \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)] + E(n)\,, \tag{17}$$

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 negatives of terms corresponding to the first tour $X_0, \ldots, X_{T_1-1}$. The point is 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.

Since the terms corresponding to different j in (17) are i.i.d., the classical central limit theorem (see e.g. [12], Theorem 27.1; [68], p. 110) will prove Theorem 21, *provided* that each term has finite second moment. Hence, recalling that each tour begins in the distribution $\nu(\cdot)$, we see that the proof of Theorem 21 is completed by the following lemma:

Lemma 28. $\int_{x\in\mathcal{X}} \nu(dx)\, \mathbf{E}\big[\big(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)]\big)^2 \;\big|\; X_0 = x\big] < \infty$.

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 $\nu(dx) \leq \pi(dx)\,/\,\pi(C)\,\epsilon$. 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 $\mathbf{E}_\pi[\cdots] = \int_{x\in\mathcal{X}} \mathbf{E}[\cdots \mid X_0 = x]\,\pi(dx)$. 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,

$$\mathbf{E}_\pi\Big[\Big(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)]\Big)^2\Big] \;\leq\; \Big(\sum_{i=0}^{\infty} \sqrt{\mathbf{E}_\pi\left[\mathbf{1}_{i<T_1} H_i^2\right]}\;\Big)^2. \tag{18}$$

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. [12], p. 80),

$$\mathbf{E}_\pi[\mathbf{1}_{i<T_1} H_i^2] \leq \mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{1/p}\, \mathbf{E}_\pi[|H_i|^{2q}]^{1/q}\,. \tag{19}$$

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, geometric ergodicity implies (as in the proof of Lemma 13) exponential tails on the return times to C. It then follows (cf. Theorem 2.5 of [78]) that $\mathbf{E}_\pi[\beta^{T_1}] < \infty$ for some $\beta > 1$. 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 (18) and (19), we obtain that

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

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

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

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 [61].

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 make a theory out of this, assume for now that

$$\pi_u(\mathbf{x}) \;=\; \prod_{i=1}^{d} f(x_i)\,, \tag{20}$$

i.e. that the density π_u factors into i.i.d. components, each with (smooth) density f. (This assumption is very restrictive, but it does allow us to develop some intersting theory.) 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 is proved in [56] (see also [61]) that under the assumpion (20), as $d \to \infty$ it is optimal to choose $\sigma^2 \doteq (2.38)^2/I\ d$, leading to an asymptotic acceptance rate $\doteq 0.234$.

More precisely, set $\sigma_d^2 = \ell^2/d$, and let

$$Z_t^d = X_{[dt]}^{(1)}\,.$$

[Thus, $\{Z_t\}$ follows the first component of $\{X_n\}$, with time speeded up by a factor of d.] Then as $d \to \infty$, the process $\{Z_t^d\}_{t\geq 0}$ converges weakly (see e.g. [24]) to a diffusion process $\{Z_t\}_{t\geq 0}$ which satisfies the following stochastic differential equation:

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

where

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

represents the *speed* of the diffusion. The choice $\ell = \hat{\ell} = 2.38/\sqrt{I}$ maximises $h(\ell)$. Furthermore, the asymptotic acceptance rate of the algorithm is given by $A(\ell) = 2\Phi\left(-\frac{\sqrt{I}\ell}{2}\right)$, so that $A(\hat{\ell}) \doteq 0.234$.

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 is proved in [58] that assuming (20), as $d \to \infty$ it is optimal 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 $\sigma_d^2 = \ell^2/d^{1/3}$, and let

$$Z_t^d = X^{(1)}_{[d^{1/3}t]}\,.$$

Then 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:

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

where

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

represents the speed of the diffusion. The choice $\ell = \hat{\ell} = 0.825/\sqrt{J}$ maximises $g(\ell)$, and the asymptotic acceptance rate satisfies $A(\hat{\ell}) = 2\,\Phi(-J\hat{\ell}^3) \doteq 0.574$.

6.3. Discussion of Optimal Scaling.

The above results show that for either the RWM or the Langevin algorithm, under the assumption [20], 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 extremely easy 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. (It may also be possible for the computer to adaptively modify σ^2 to achieve the appropriate acceptance rates; see [4] and references therein.)

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 (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 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 [61].

Now, the above results are only proved under the strong assumption (20). 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 compoenents of the form $\pi_u(\mathbf{x}) = \prod_{i=1}^d C_i\, f(C_i x_i)$ (see [61]), to the discrete hypercube [54], and to finite-range homogeneous Markov random fields [13]; 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 [15]. The true level of generality of these optimal scaling results is currently unknown, though investigations are ongoing [9].

REFERENCES

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

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

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

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

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

[7] 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.

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

[9] M. Bédard (2003), On the robustness of optimal scaling for Metropolis-Hastings algorithms. Ph.D. dissertation, University of Toronto. Work in progress.

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

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

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

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

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

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

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

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

[18] 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.

[19] 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.

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

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

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

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

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

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

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

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

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

[29] 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.

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

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

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

[33] 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.

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

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

[36] N. Jain and B. Jamison (1967), Contributions to Doeblin's theory of Markov

processes. Z. Wahrsch. Verw. Geb. **8**, 19–40.

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

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

[39] G.L. Jones and J.P. Hobert (2001), Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. Ann. Stat., to appear.

[40] 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.

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

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

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

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

[45] 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.

[46] S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London.

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

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

[49] 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.

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

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

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

[53] 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.

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

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

[56] 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.

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

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

[59] 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.

[60] G.O. Roberts and J.S. Rosenthal (2001), Small and Pseudo-Small Sets for Markov Chains. Stochastic Models **17**, 121–145.

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

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

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

[64] 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.

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

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

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

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

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

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

[71] J.S. Rosenthal (2003), Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms. J. Amer. Stat. Assoc. **98**, 169–177.

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

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

[74] 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.

[75] 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.

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

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

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