THE DRIFT AND MINORIZATION METHOD
FOR REVERSIBLE MARKOV CHAINS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Daniel Jerison

May 2016

Abstract

This thesis proves upper bounds on the convergence time to stationarity for reversible discrete time Markov chains on general state spaces. The method of proof is the well-established *drift and minorization* approach, which imposes a regenerative structure on the Markov chain according to a particular recipe. The resulting bounds are computable in terms of the ingredients of the recipe.

The convergence theorems in this thesis are developed using a new perspective on the interplay between regeneration and reversibility. They are more widely applicable than previous results and provide better numerical bounds on the time to stationarity in specific examples. Two Gibbs samplers from Bayesian statistics are treated in detail. When applied to these chains, the new bounds improve on earlier work but are still quite conservative compared with the true convergence rates.

For certain classes of finite chains, the drift and minorization method can give precise bounds on the mixing time. A striking result of this type is a sharp upper bound on the cutoff window for birth and death chains. In addition, an inductive argument shows that the spectral gap of the random walk on the hypercube can be recovered using drift and minorization up to a constant factor of 2.

The thesis also contains:

- An exposition of the drift and minorization method, explaining both the renewal theory approach of Meyn and Tweedie [MT93] and the coupling approach of Rosenthal [Ros95a], and showing how the bounds improve when the chain is reversible or stochastically monotone.
- A development of the properties of different types of regeneration times for general state space Markov chains. Defining a randomized stopping time usually requires enlarging the sample space to incorporate independent randomness. In full generality, this operation raises subtle questions of measurability, which are resolved using a new "compatibility condition."
- A brief consideration of quantile estimation in Markov chain Monte Carlo, focusing on concentration inequalities that provide finite-sample guarantees at specified confidence levels. The goal is to prove inequalities that rely as little as possible on theoretical convergence bounds (of the sort proved elsewhere in this thesis) and as much as possible on the empirical sample.

Acknowledgments

Many people have helped and supported me through the process of writing this thesis. Above all, I would like to thank my advisor, Persi Diaconis. He taught me everything I know about Markov chains and pointed me toward the topic of drift and minorization. His genial enthusiasm inspired me; his encyclopedic knowledge and deep understanding of the big picture led me in the right direction. He has been my strongest supporter, encouraging me with unfailing honesty and patience. While writing, I have guided myself by my conception of what Persi would consider "good work." If the result is worthwhile, he is ultimately responsible.

I wrote much of this thesis while at Cornell University. The probabilists at Cornell, Laurent Saloff-Coste and Lionel Levine, have been kind and hospitable. Their understanding and encouragement have given me the impetus to continue pushing forward until the work was complete.

I would also like to thank Amir Dembo, for suggesting the avenue that led to Theorem 1.3, and Aaron Smith and John Pike, for many wide-ranging conversations.

Finally, my wife Erin has stood by my side since the beginning. She sustains me and gives me strength; she is the constant and the light of my life. This work is dedicated to her.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Regeneration, reversibility, and convergence

This thesis is about convergence of Markov chains to their stationary distributions. The main goal is to find explicitly computable exponential convergence rates for discrete time chains on general state spaces.

The principal motivation comes from Markov chain Monte Carlo algorithms. Often in statistics, one would like to explore a probability distribution π on a subset of $\mathbf{R}^d$ (say, the posterior joint distribution of the parameters in a Bayesian statistical model). Sampling directly from π is intractable, but one can construct a discrete time Markov chain (X_t) whose stationary distribution is π. There are two related questions:

1. How many steps t suffice for the chain to "forget" its starting state and give an approximate sample from π?
2. Given a real-valued function f on the state space $\mathcal{X}$, how many steps t suffice for the empirical mean $\frac{1}{t}\sum_{s=1}^{t} f(X_s)$ to be near the average value $\pi(f) = \int_{\mathcal{X}} f(x)\pi(dx)$ with high probability?

Most of the thesis will be devoted to finding upper bounds for the first question. Chapter 6 provides upper bounds for the second question under certain conditions.

The most successful "off-the-shelf" technique to find an explicit convergence rate for a general state space Markov chain is the method of *drift and minorization*. This method gives a recipe for constructing a probability measure ν on $\mathcal{X}$ and a randomized stopping time T of the chain (X_t) such that the distribution of X_T is ν, and moreover, X_T is independent of the value of T. The random time T will be called a *strong ν time*, in analogy with strong stationary times, which satisfy the same definition except with the stationary distribution π in place of ν.

As part of the construction, one obtains an exponential bound on the tail of T: there exist a constant $0 < \lambda_* < 1$ and a function $F : \mathcal{X} \to \mathbf{R}$, both explicitly computable, such that

$$\mathbf{P}_x(T > t) \le F(x)\lambda_*^t \qquad \text{for all } x \in \mathcal{X} \text{ and } t \ge 0.$$

Here $\mathbf{P}_x$ denotes the probability for the Markov chain started from $X_0 = x$. In addition, there is a constant $B < \infty$ such that $\mathbf{P}_\nu(T > t) \le B\lambda_*^t$ also decays exponentially.

For intuition, consider the special case where $\mathcal{X}$ is countable, $\nu = \delta_c$ is concentrated at a single state $c \in \mathcal{X}$, and $T = \tau_c^+ = \min\{t \ge 1 : X_t = c\}$. (The notation τ_c^+ is to distinguish from $\tau_c = \min\{t \ge 0 : X_t = c\}$.) Given a sample path of the chain, let T_k be the time of the kth visit to c. The interarrival times $T_{k+1} - T_k$ are iid, as are the segments $(X_{T_k}, X_{T_k+1}, \dots, X_{T_{k+1}-1})$ of the sample path. The exponential bound on the tail of T controls the length of each segment, including the initial segment $(X_0, \dots, X_{T_1-1})$, whose distribution is different.

The general case is nearly equivalent. By resetting the timer every time the chain "reaches ν," one defines a sequence of strong ν times $T_1 < T_2 < \cdots$ such that the interarrival times $T_{k+1} - T_k$ are iid with exponentially decaying tail. Interestingly, the segments $(X_{T_k}, X_{T_k+1}, \dots, X_{T_{k+1}-1})$ are *not* necessarily iid, but they are 1-dependent; see Chapter 3.

The method of drift and minorization imposes a regenerative structure on the Markov chain. This is not enough to prove convergence to stationarity in the sense of Question 1 above: a periodic chain could regenerate frequently and never converge. Therefore some aperiodicity condition is needed. In this thesis, the standing assumption will be that the chain is reversible with nonnegative eigenvalues. Formally, the transition kernel $P(x, dy)$ of the chain induces an operator $f \mapsto Pf$ on the space

$$L^2(\pi) = \{f : \mathcal{X} \to \mathbf{R} : \langle f, f\rangle_\pi < \infty\}, \qquad \text{where } \langle f, g\rangle_\pi = \int_{\mathcal{X}} f(x)g(x)\pi(dx),$$

given by $(Pf)(x) = \int_{\mathcal{X}} f(y)P(x, dy)$. The chain is reversible with nonnegative eigenvalues if the operator P is self-adjoint with nonnegative spectrum. This assumption leads to straightforward proofs of convergence with explicit rates that substantially improve on the existing literature. As well, many chains frequently used in MCMC, such as Gibbs samplers and Metropolis–Hastings chains, are reversible with nonnegative eigenvalues or can be made that way with minor modifications.

It should be noted that periodicity does not present an obstacle when finding upper bounds for differences of the form $|\frac{1}{t}\sum_{s=1}^t f(X_s) - \pi(f)|$. Accordingly, the results in Chapter 6 rely only on regeneration and do not require reversibility or aperiodicity.

1.2 A few results

This section presents three results that will be proved in later chapters. The first two are general statements showing that if (X_t) is a reversible Markov chain with nonnegative eigenvalues and T is a strong ν time, then the distance from stationarity after t steps is controlled directly by the tail of T. The third result illustrates the strength of the general method by deriving a sharp mixing time bound for birth and death chains.

Let μ_1, μ_2 be probability measures on the state space $\mathcal{X}$. Their total variation distance is

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \sup_{A \subseteq \mathcal{X}} |\mu_1(A) - \mu_2(A)| = \frac{1}{2} \int_{\mathcal{X}} |\mu_1 - \mu_2|(dx),$$

where the supremum is over all measurable subsets $A \subseteq \mathcal{X}$. Their $L^2(\pi)$ distance is

$$\|\mu_1 - \mu_2\|_{L^2(\pi)}^2 = \int_{\mathcal{X}} \left[\frac{d(\mu_1 - \mu_2)}{d\pi}(x) \right]^2 \pi(dx)$$

if $\mu_1 - \mu_2$ is absolutely continuous with respect to π (so that the Radon-Nikodym derivative $d(\mu_1 - \mu_2)/d\pi$ is defined) and $\|\mu_1 - \mu_2\|_{L^2(\pi)}^2 = \infty$ otherwise. It is well-known that $\|\mu_1 - \mu_2\|_{\mathrm{TV}} \leq \frac{1}{2}\|\mu_1 - \mu_2\|_{L^2(\pi)}$. Let $P^t(\mu, \cdot) = \mathbf{P}_\mu(X_t \in \cdot)$ denote the law of X_t given that $X_0 \sim \mu$.

Theorem 1.1. *Let (X_t) be a Markov chain on a state space $\mathcal{X}$, reversible with respect to a stationary distribution π and having nonnegative eigenvalues. Suppose that ν is a probability measure on $\mathcal{X}$ and T is a strong ν time for (X_t) such that $\mathbf{E}_\nu[T] < \infty$ and $\mathbf{P}_\mu(1 \leq T < \infty) = 1$ for all probability measures μ on $\mathcal{X}$. Then for all $t \geq 0$, $P^t(\nu, \cdot)$ is absolutely continuous with respect to π and*

$$\|P^t(\nu, \cdot) - \pi\|_{L^2(\pi)}^2 \leq \sum_{n=2t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

Suppose the tail of T decays exponentially, $\mathbf{P}_\nu(T > t) \leq B\lambda_*^t$ for some $B < \infty$ and $\lambda_* < 1$, as happens when T is constructed using drift and minorization. It follows that

$$\|P^t(\nu, \cdot) - \pi\|_{L^2(\pi)} \leq \sqrt{\frac{B\lambda_*}{1 - \lambda_*}} \cdot \lambda_*^t,$$

so the chain converges to π at the same exponential rate λ_* governing the tail of T, or faster.

Theorem 1.1 immediately gives an upper bound on the total variation distance $\|P^t(\nu, \cdot) - \pi\|_{\mathrm{TV}}$. This bound, unlike the original $L^2(\pi)$ bound, easily extends to the case where the chain is started from a fixed state $X_0 = x$ rather than $X_0 \sim \nu$; see Theorem 4.4.

The second result concerns the special case where $\nu = \delta_c$ is concentrated at a state $c \in \mathcal{X}$. Its flavor is similar to Theorem 1.1, but the proof relies on a hidden stochastic monotonicity.

Theorem 1.2. *Let (X_t) be a Markov chain on a state space $\mathcal{X}$, reversible with respect to a stationary distribution π and having nonnegative eigenvalues. For fixed $c \in \mathcal{X}$, suppose that $\mathbf{E}_c[\tau_c^+] < \infty$ and $\mathbf{P}_x(\tau_c < \infty) = 1$ for all $x \in \mathcal{X}$. Then for all $t \geq 0$,*

$$\|P^t(c, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbf{P}_\pi(\tau_c > t).$$

As before, this bound extends easily to starting states $x \neq c$; see Theorem 4.5. The assumption that $\mathbf{E}_c[\tau_c^+] < \infty$ implies that $\pi(\{c\}) > 0$, so Theorem 1.2 requires a stationary distribution that assigns positive mass to at least one singleton element of the state space. This condition is automatically satisfied when $\mathcal{X}$ is countable and (X_t) is positive recurrent.

Finding a regenerative structure is one of the only ways to prove exponential convergence to stationarity for general state space Markov chains. By contrast, if the state space is finite, exponential convergence is guaranteed as long as the chain is irreducible and aperiodic. The modern theory of finite Markov chains considers quantities like the mixing time. For a chain with transition matrix P and stationary distribution π, the total variation mixing time with parameter $0 < \varepsilon < 1$ is

$$t_{\mathrm{mix}}(\varepsilon) = \min\{t \geq 0 : \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \varepsilon \text{ for all } x \in \mathcal{X}\}.$$

Theorem 1.2 leads to an upper bound on mixing time in terms of the hitting time τ_c for any fixed state $c \in \mathcal{X}$. For many finite chains, the bound is far from sharp no matter which state c is chosen. As an example, consider the lazy simple random walk on the hypercube $\{0,1\}^n$, which takes a step from a state $(x_1, \ldots, x_n) \in \{0,1\}^n$ by choosing a coordinate $1 \leq i \leq n$ uniformly at random and replacing x_i with either 0 or 1, each with probability 1/2. For any $0 < \varepsilon < 1$, the mixing time $t_{\mathrm{mix}}(\varepsilon)$ is asymptotic to $\frac{1}{2} n \log n$ as $n \to \infty$. Meanwhile, the expected hitting times $\min\{\mathbf{E}_x[\tau_c] : x \neq c\}$ and $\max\{\mathbf{E}_x[\tau_c] : x \neq c\}$ are both asymptotic to 2^{n+1} ([LPW09], Theorem 18.3 and Exercise 10.5).

Birth and death chains are a family of examples where Theorem 1.2 gives sharp bounds. A birth and death chain on $\{0, \ldots, n\}$ is a Markov chain whose transition matrix satisfies $P(i,j) = 0$ if $|i - j| \geq 2$. Every irreducible birth and death chain is reversible with respect to its stationary distribution π. The state $0 \leq m \leq n$ is called a *median state* if $\pi(\{0, \ldots, m\}) \geq 1/2$ and $\pi(\{m, \ldots, n\}) \geq 1/2$. Usually there is only one median state, but if $\pi(\{0, \ldots, k\}) = 1/2$ for some k, then both $m = k$ and $m = k + 1$ are median states.

It was shown by [DLP10] that the mixing time of a lazy irreducible birth and death chain is approximately equal to the time for the chain to reach the median when started from one of the endpoints.

(“Lazy” means that $P(i,i) \geq 1/2$ for all states i.) Specifically, let m be a median state and define

$$t_{\text{hit}} = \max\{\mathbf{E}_0[\tau_m], \mathbf{E}_n[\tau_m]\}.$$

Also define the relaxation time $t_{\text{rel}} = 1/\gamma$, where γ is the spectral gap of the transition matrix (that is, the difference between 1 and the next-largest eigenvalue). One can show that $t_{\text{rel}} \leq t_{\text{hit}}$. The work of [DLP10] implies that there exist functions $F(\varepsilon)$ and $G(\varepsilon)$, with no dependence on n, such that any lazy irreducible birth and death chain on $\{0,\dots,n\}$ satisfies

$$t_{\text{hit}} + F(\varepsilon)\sqrt{t_{\text{hit}} \cdot t_{\text{rel}}} \leq t_{\text{mix}}(\varepsilon) \leq t_{\text{hit}} + G(\varepsilon)\sqrt{t_{\text{hit}} \cdot t_{\text{rel}}} \qquad \text{for all } 0 < \varepsilon < 1. \tag{1.1}$$

A version of the upper bound in (1.1) that is more precise than the corresponding result in [DLP10] can be obtained as a consequence of Theorem 1.2.

Theorem 1.3. *Let P be the transition matrix for a lazy irreducible birth and death chain on $\{0,\dots,n\}$ with stationary distribution π. For any $\delta > 0$,*

$$t_{\text{mix}}(\varepsilon) \leq (1+\delta)t_{\text{hit}} + \left(1 + \frac{1}{\delta}\right) 2t_{\text{rel}} \log(2/\varepsilon) \qquad \textit{for all } 0 < \varepsilon < 1.$$

Optimizing in δ yields

$$t_{\text{mix}}(\varepsilon) \leq t_{\text{hit}} + 2\sqrt{2t_{\text{hit}}t_{\text{rel}}\log(2/\varepsilon)} + 2t_{\text{rel}}\log(2/\varepsilon), \tag{1.2}$$

which has the form $t_{\text{mix}}(\varepsilon) \leq t_{\text{hit}} + G(\varepsilon)\sqrt{t_{\text{hit}} \cdot t_{\text{rel}}}$ since $t_{\text{rel}} \leq \sqrt{t_{\text{hit}} \cdot t_{\text{rel}}}$. Unlike the analogous bound in [DLP10], the inequality (1.2) has the correct dependence on ε in the sense that the right side of (1.2) is bounded above by a fixed constant multiple of $t_{\text{mix}}(\varepsilon)$ as $\varepsilon \to 0$.

1.3 Chapter summaries

Following is a brief summary of each chapter in this thesis.

Chapter 1: Introduction. Section 1.4 provides the foundational definitions for Markov chains on general state spaces. The theory is developed in the greatest possible generality; it is not assumed that the σ-algebra of measurable subsets of the state space is countably generated. For this reason, enlarging the sample space (say, to add an independent source of randomness for a randomized stopping time) leads to subtle questions of measurability. A novel “compatibility condition” is proposed to resolve these questions. Finally, Section 1.5 collects frequently used definitions and notation.

Chapter 2: Drift and minorization. This chapter describes the method of drift and minorization in detail. It presents the historical development from the original proof of exponential convergence due to Nummelin and Tuominen [NT82] through the first quantitative bounds proved by Meyn and Tweedie [MT94] along with the bivariate drift approach of Rosenthal [Ros95a]. It is explained why the extra assumptions of monotonicity and reversibility lead to faster convergence rates. The chapter has few formal proofs; the focus is on building intuition while placing the new results in context.

Chapter 3: Regeneration times. Section 1.1 asserted that if one applies the method of drift and minorization to a Markov chain, one obtains a strong ν time encoding a regenerative structure. This chapter carefully develops the theory of strong ν times and other types of regeneration times in the general setting defined in Section 1.4. Although the statements in this chapter are well motivated, the proofs are technical and relegated to the Appendix.

Chapter 4: Convergence results. This chapter states and proves the main convergence theorems. First, if a Markov chain satisfies a drift and minorization condition, it has a strong ν time whose tail decays exponentially. This is already well-known, but the new bound in Theorem 4.9 improves on the best previous result in [RT99]. Second and more crucially, if the chain is reversible with nonnegative eigenvalues, the tail of the strong ν time directly controls the convergence rate. This is the content of Theorems 1.1 and Theorem 1.2, which are proved as Theorems 4.16 and 4.19, respectively. Putting together the two steps yields explicit convergence bounds in terms of the drift and minorization data, which are given in Theorems 4.4 and 4.5. An interesting highlight is the proof of Theorem 4.19, which relies on a stochastic monotonicity discovered initially by [LZK06].

Chapter 5: Examples. Two example chains are analyzed using the convergence theorems from Chapter 4. Both chains are two-variable Gibbs samplers designed to converge to the joint posterior distribution of the parameters in a Bayesian statistical model. The chapter proposes a way to optimize many aspects of the drift and minorization computation. Carrying out the process yields explicit convergence bounds that are quite conservative compared with the chains' actual behavior but still represent a significant improvement over the best bounds previously available.

Chapter 6: MCMC estimation of quantiles. This chapter considers Question 2 from Section 1.1, the estimation of $\pi(f) = \int_{\mathcal{X}} f(x)\pi(dx)$ by $\frac{1}{t}\sum_{s=1}^{t} f(X_s)$ for some function $f : \mathcal{X} \to \mathbf{R}$. Most of the attention is given to the case where $\theta : \mathcal{X} \to \mathbf{R}$ is a parameter of interest and $f(x) = f_c(x) = \mathbf{1}\{\theta(x) \leq c\}$ for some $c \in \mathbf{R}$, since accurate estimates of $\pi(f_c)$ lead to bounds on the quantiles $\theta_q = \inf\{r \in \mathbf{R} : \pi(\theta \leq r) \geq q\}$ of θ. Suppose $0 < q < 1/2$ and $0 < \delta < 1$. A procedure is set forth to find an interval $[c, d]$ that contains the "$100(1-2q)\%$ credible interval" for θ, namely $[\theta_q, \theta_{1-q}]$, with confidence at least $1 - \delta$. This procedure is carried out for one of the examples from Chapter 5. Although many steps of the Markov chain are needed to ensure the prescribed confidence level, the method—based on so-called *empirical Bernstein inequalities*—is quite promising.

Chapter 7: Finite chains. Section 7.1 defines the notion of cutoff for finite Markov chains and discusses the proof by [DLP10] of (1.1), showing that a sequence of lazy irreducible birth and death chains has cutoff if it satisfies the Peres condition [Per04] that $t_{\mathrm{rel}} = o(t_{\mathrm{mix}})$. Theorem 1.3, which gives a tighter version of the [DLP10] upper bound, is proved (as Theorem 7.3) using results from Chapter 4. Section 7.2 considers the lazy simple random walk on the hypercube as a more difficult test case. Although the bound from Theorem 1.2 is very bad, another approach based on Chapter 4 finds the spectral gap to within a factor of 2.

1.4 Foundations

This section defines the fundamental objects to be studied in this thesis: transition kernels, Markov chains, and so forth. It is assumed that the reader is already familiar with the basic theory of Markov chains on general state spaces; for a detailed exposition, see Chapter 3 of [MT93] or Chapter 1 of [Rev84]. Since a Markov chain on a state space $\mathcal{X}$ is determined by its transition kernel and its initial distribution, each fixed transition kernel P induces a family of Markov chains on $\mathcal{X}$ indexed by the possible initial distributions μ. A *Markov scheme* simultaneously defines all these chains on a single sample space Ω using a family of probability measures $\mathbf{P}_\mu$. This terminology is nonstandard. The definition in this section imposes a compatibility condition on the measures $\mathbf{P}_\mu$ that resolves questions of measurability while providing freedom to enlarge the sample space to include new sources of randomness.

> A *Markov chain* with transition kernel P has a specified initial distribution and is given by a single probability measure on the sample space.
> A *Markov scheme* with transition kernel P is a family of probability measures on the sample space, each one corresponding to the chain started from a different initial distribution.

Let $(\mathcal{X}, \mathcal{E})$ be a measurable space, where $\mathcal{E}$ is the collection of measurable subsets of $\mathcal{X}$. Denote the space of probability measures on $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$. Endow $[0,1]$ with the Borel σ-algebra. A *transition kernel* $P = P(x, dy)$ on $\mathcal{X}$ is a function $P : \mathcal{X} \times \mathcal{E} \to [0,1]$ such that for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $\mathcal{X}$, and for each $A \in \mathcal{E}$, the function $x \mapsto P(x, A)$ is measurable.

Let $\mathcal{X}^\infty = \{(x_0, x_1, \ldots) : x_j \in \mathcal{X}\}$ be the space of infinite sequences on $\mathcal{X}$, equipped with the product σ-algebra $\mathcal{E}^\infty$. A *Markov chain* on $\mathcal{X}$ with transition kernel P consists of a measurable space $(\Omega, \mathcal{F})$, a measurable function $X : \Omega \to \mathcal{X}^\infty$ taking each $\omega \in \Omega$ to a sequence $(X_0(\omega), X_1(\omega), \ldots)$, a measure $\mu \in \mathcal{P}(\mathcal{X})$, and a measure $\mathbf{P}_\mu \in \mathcal{P}(\Omega)$ satisfying

$$\mathbf{P}_\mu(X_0 \in A_0, \ldots, X_n \in A_n) = \int_{x_0 \in A_0} \cdots \int_{x_n \in A_n} \mu(dx_0) P(x_0, dx_1) \cdots P(x_{n-1}, dx_n) \tag{1.3}$$

for all $n \geq 0$ and $A_0, \ldots, A_n \in \mathcal{E}$. $\mathcal{X}$ is called the *state space*, Ω is called the *sample space*, and μ is called the *initial distribution*. From (1.3) one can verify the *Markov property*: under $\mathbf{P}_\mu$, $(X_{n+1}, X_{n+2}, \ldots)$ is conditionally independent of $(X_0, \ldots, X_{n-1})$ given X_n. If the initial distribution is the delta measure

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases}$$

one can write $\mathbf{P}_x$ instead of $\mathbf{P}_{\delta_x}$.

Suppose P is a transition kernel on $\mathcal{X}$. A standard construction (see [Rev84]) simultaneously defines on the same sample space all the Markov chains with transition kernel P and different initial distributions. The sample space is $\Omega = \mathcal{X}^\infty$, and the function $X : \Omega \to \mathcal{X}^\infty$ is the identity map. For each $\mu \in \mathcal{P}(\mathcal{X})$ there is a measure $\mathbf{P}_\mu \in \mathcal{P}(\mathcal{X}^\infty)$ satisfying (1.3). In addition, for each measurable subset $E \in \mathcal{E}^\infty$ (called an *event*), the function $x \mapsto \mathbf{P}_x(E)$ is measurable; and for every $\mu \in \mathcal{P}(\mathcal{X})$,

$$\mathbf{P}_\mu(E) = \int_{\mathcal{X}} \mathbf{P}_x(E)\mu(dx). \tag{1.4}$$

These properties uniquely determine the family $\{\mathbf{P}_\mu\}_{\mu \in \mathcal{P}(\mathcal{X})}$, making the construction canonical.

In this thesis, it will be necessary to incorporate sources of randomness beyond the sample path $(X_0, X_1, \ldots)$ taken by the Markov chain. This will require the use of sample spaces larger than $\mathcal{X}^\infty$.

Definition 1.4. Let P be a Markov transition kernel on the state space $(\mathcal{X}, \mathcal{E})$. A *Markov scheme* on $\mathcal{X}$ with transition kernel P consists of a measurable space $(\Omega, \mathcal{F})$, a measurable function $X : \Omega \to \mathcal{X}^\infty$, and a family $\{\mathbf{P}_\mu\}_{\mu \in \mathcal{P}(\mathcal{X})}$ of probability measures on Ω such that:

1. For each $\mu \in \mathcal{P}(\mathcal{X})$, the measure $\mathbf{P}_\mu$ satisfies (1.3), so that it defines a Markov chain on $\mathcal{X}$ with transition kernel P and initial distribution μ.

2. For any event $E \in X^{-1}(\mathcal{E}^\infty)$, the function $x \mapsto \mathbf{P}_x(E)$ is measurable, and (1.4) holds for every $\mu \in \mathcal{P}(\mathcal{X})$. (Here, $X^{-1}(\mathcal{E}^\infty)$ is the collection of preimages of sets in $\mathcal{E}^\infty$ under X, which is a sub-σ-algebra of $\mathcal{F}$.)

3. If $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ and μ' is absolutely continuous with respect to μ, so that the Radon-Nikodym derivative $\frac{d\mu'}{d\mu}$ exists, then for every bounded measurable function $f : \Omega \to \mathbf{R}$,

$$\mathbf{E}_{\mu'}[f(\omega)] = \mathbf{E}_\mu \left[\frac{d\mu'}{d\mu}(X_0) f(\omega) \right]. \tag{1.5}$$

Condition 3 is called the *compatibility condition*. Note that if (1.5) holds for all functions

$$f(\omega) = \mathbf{1}\{\omega \in E\} = \begin{cases} 1 & \text{if } \omega \in E, \\ 0 & \text{otherwise,} \end{cases}$$

for $E \in \mathcal{F}$, then it holds for all bounded measurable functions f.

The compatibility condition is weaker than (1.4). If (1.4) holds, suppose that μ' is absolutely continuous with respect to μ and that $f(\omega) = \mathbf{1}\{\omega \in E\}$ is given. Then

$$\begin{aligned} \mathbf{E}_{\mu'}[f(\omega)] = \mathbf{P}_{\mu'}(E) = \int_\Omega \mathbf{P}_x(E)\mu'(dx) &= \int_\Omega \mathbf{P}_x(E)\frac{d\mu'}{d\mu}(x)\mu(dx) \\ = \int_\Omega \mathbf{E}_x\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega \in E\}\right]\mu(dx) &= \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)f(\omega)\right]. \end{aligned}$$

An alternative definition of a Markov scheme would require each function $x \mapsto \mathbf{P}_x(E)$ to be measurable and (1.4) to hold for all $E \in \mathcal{F}$. This definition is fine when the σ-algebra $\mathcal{E}$ is countably generated. Without that hypothesis, the constructions of regeneration times in Chapter 3 succeed under Definition 1.4 but would fail under the alternative definition.

Definition 1.5. Consider a Markov scheme consisting of a sample space $(\Omega, \mathcal{F})$, a function $X : \Omega \to \mathcal{X}^\infty$, and a family $\{\mathbf{P}_\mu\}_{\mu \in \mathcal{P}(\mathcal{X})}$ of probability measures on Ω. An *extension* of the Markov scheme is a measurable space $(\bar{\Omega}, \bar{\mathcal{F}})$ along with a surjective measurable function $p : \bar{\Omega} \to \Omega$ and a family $\{\bar{\mathbf{P}}_\mu\}_{\mu \in \mathcal{P}(\mathcal{X})}$ of probability measures on $\bar{\Omega}$ such that $\mathbf{P}_\mu(E) = \bar{\mathbf{P}}_\mu(p^{-1}(E))$ for all $\mu \in \mathcal{P}(\mathcal{X})$ and $E \in \mathcal{F}$, and the space $(\bar{\Omega}, \bar{\mathcal{F}})$ together with the function $X \circ p$ and the family $\{\bar{\mathbf{P}}_\mu\}_{\mu \in \mathcal{P}(\mathcal{X})}$ is itself a Markov scheme. (Conditions 1 and 2 of Definition 1.4 are automatically satisfied, but the compatibility condition must be checked.)

It will be convenient to describe both Markov chains and Markov schemes using the notation (X_t). In both cases the sample space Ω is suppressed, and in the former case the initial distribution μ is also suppressed. Whether (X_t) refers to a chain or scheme should be clear from context. Note also that specifying an initial distribution for a scheme yields a Markov chain.

1.5 Notation and definitions

This section collects various notations and definitions.

A *transition kernel* from a measurable space $(\mathcal{X}, \mathcal{E})$ to a measurable space $(\mathcal{Y}, \mathcal{F})$ is a function $P : \mathcal{X} \times \mathcal{F} \to [0, 1]$ (with the Borel σ-algebra) such that for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $(\mathcal{Y}, \mathcal{F})$, and for each $A \in \mathcal{F}$, the function $x \mapsto P(x, A)$ is $(\mathcal{X}, \mathcal{E})$-measurable. When

$(\mathcal{X}, \mathcal{E}) = (\mathcal{Y}, \mathcal{F})$, one says that P is a transition kernel on $(\mathcal{X}, \mathcal{E})$ or simply on $\mathcal{X}$.

Suppose $P(x, dy)$ is a transition kernel on $(\mathcal{X}, \mathcal{E})$. For any $\mu \in \mathcal{P}(\mathcal{X})$, define the measure $P(\mu, \cdot)$ by

$$P(\mu, A) = \int_{\mathcal{X}} \mu(dx) P(x, A).$$

For integers $t \geq 0$, define the *t-step transition kernels* $P^t(x, \cdot)$ inductively by $P^0(x, \cdot) = \delta_x(\cdot)$ and

$$P^{t+1}(x, A) = \int_{\mathcal{X}} P(y, A) P^t(x, dy),$$

so that in particular $P^1(x, \cdot) = P(x, \cdot)$. Analogously define $P^t(\mu, \cdot)$. The notation $\mu P^t(\cdot)$ means the same thing as $P^t(\mu, \cdot)$.

A probability measure $\pi \in \mathcal{P}(\mathcal{X})$ is a *stationary distribution* for P (or for any Markov chain having transition kernel P) if for all $A \in \mathcal{E}$,

$$\pi(A) = \int_{\mathcal{X}} \pi(dx) P(x, A). \tag{1.6}$$

A (possibly signed) measure μ, with $\mu(\mathcal{X})$ not required to equal 1, is called an *invariant measure* for P or for the Markov chain if it satisfies (1.6). The term "stationary distribution" is reserved for probability measures.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $X : \Omega \to \mathbf{R}$ be a random variable. If $E \in \mathcal{F}$ is an event with $\mathbf{P}(E) > 0$, the *conditional expectation* of X given E is

$$\mathbf{E}[X \mid E] = \frac{\mathbf{E}[X, E]}{\mathbf{P}(E)},$$

where

$$\mathbf{E}[X, E] = \mathbf{E}[X \mathbf{1}_E] = \int_E X(\omega)\, \mathbf{P}(d\omega).$$

Here $\mathbf{1}_E$ is the indicator function on Ω, $\mathbf{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbf{1}_E(\omega) = 0$ otherwise. It can be shown that

$$E \mapsto \begin{cases} \mathbf{E}[X \mid E] & \text{if } \mathbf{P}(E) > 0, \\ 0 & \text{if } \mathbf{P}(E) = 0, \end{cases}$$

is a signed measure on $(\Omega, \mathcal{F})$, denoted $\mathbf{E}[X \mid \cdot\,]$, that is absolutely continuous with respect to $\mathbf{P}$. If $\mathcal{F}' \subseteq \mathcal{F}$ is a sub-σ-algebra, the conditional expectation of X given $\mathcal{F}'$ is the Radon-Nikodym derivative

$$\mathbf{E}[X \mid \mathcal{F}'] = \frac{d\,\mathbf{E}[X \mid \cdot\,]|_{\mathcal{F}'}}{d\,\mathbf{P}|_{\mathcal{F}'}},$$

where the measures in the numerator and denominator are restricted to $\mathcal{F}'$. The conditional expectation is defined up to $\mathbf{P}|_{\mathcal{F}'}$-almost everywhere equivalence.

If $(\Omega_1, \mathcal{F}_1)$ is another measurable space and $Y : \Omega \to \Omega_1$ is a random variable, then $\mathbf{P}$ induces a probability measure $\mathbf{P}(Y \in \cdot\,) = \mathbf{P}(Y^{-1}(\cdot))$ on $(\Omega_1, \mathcal{F}_1)$. The conditional expectation $\mathbf{E}[X \mid Y]$ is the same as $\mathbf{E}[X \mid \sigma(Y)]$, where $\sigma(Y) = Y^{-1}(\mathcal{F}_1)$ is the sub-σ-algebra of $\mathcal{F}$ generated by Y. The conditional expectation can also be written as a function of Y: if $y \in \Omega_1$, then

$$\mathbf{E}[X \mid Y = y] = \frac{d\,\mathbf{E}[X \mid Y^{-1}(\cdot)]}{d\,\mathbf{P}(Y^{-1}(\cdot))}(y),$$

defined $\mathbf{P}(Y \in \cdot\,)$-almost everywhere.

Given an event $E \in \mathcal{F}$, the *conditional probability* $\mathbf{P}(E \mid \mathcal{F}')$ is defined to be $\mathbf{E}[\mathbf{1}_E \mid \mathcal{F}']$, and likewise for conditioning on events or random variables. For example, the statement "$\mathbf{P}(E \mid Y = y) = c$" means that the function $y \mapsto \mathbf{E}[\mathbf{1}_E \mid Y = y]$ on Ω_1 is $\mathbf{P}(Y \in \cdot\,)$-a.e. equal to the constant function c.

Two events $E_1, E_2 \in \mathcal{F}$ are *conditionally independent* given an event $E \in \mathcal{F}$ with $\mathbf{P}(E) > 0$ if $\mathbf{P}(E_1, E_2 \mid E) = \mathbf{P}(E_1 \mid E)\,\mathbf{P}(E_2 \mid E)$. E_1 and E_2 are conditionally independent given the σ-algebra $\mathcal{F}' \subseteq \mathcal{F}$ if

$$\mathbf{P}(E_1, E_2 \mid \mathcal{F}') = \mathbf{P}(E_1 \mid \mathcal{F}')\,\mathbf{P}(E_2 \mid \mathcal{F}'),$$

where the equality is $\mathbf{P}|_{\mathcal{F}'}$-a.e. equivalence. Two random variables X_1, X_2 on Ω are conditionally independent given E or $\mathcal{F}'$ if the corresponding equation above holds for every $E_1 \in \sigma(X_1)$ and $E_2 \in \sigma(X_2)$. Conditional independence given a random variable Y is the same as conditional independence given $\sigma(Y)$. The definitions extend to more than two events or random variables in the natural way.

Let (X_t) be a Markov chain or scheme with sample space Ω. A *random time* is a measurable function $T : \Omega \to \bar{\mathbf{N}} = \{0, 1, 2, \ldots\} \cup \{\infty\}$. The random time T is a *stopping time* for (X_t) if for each $n \geq 0$, the event $\{T = n\}$ depends only on the values of $X_0, \ldots, X_n$. That is, for every $(x_0, \ldots, x_n) \in \mathcal{X}^{n+1}$, the set $\{\omega \in \Omega \,:\, X_0(\omega) = x_0, \ldots, X_n(\omega) = x_n\}$ is either contained in the event $\{T = n\}$ or disjoint from it.

An example of a stopping time is the *hitting time* for a subset $C \in \mathcal{E}$. Define

$$\begin{aligned} \tau_C &= \min\{t \geq 0 \,:\, X_t \in C\}, \\ \tau_C^+ &= \min\{t \geq 1 \,:\, X_t \in C\}. \end{aligned}$$

If $C = \{c\}$ is a single element, one can write τ_c, τ_c^+ instead of $\tau_{\{c\}}, \tau_{\{c\}}^+$.

A *randomized stopping time* for (X_t) is a slight generalization of the notion of a stopping time. The event $\{T = n\}$ is allowed to depend not just deterministically on the sequence $(X_0, \ldots, X_n)$ but also on a source of randomness independent of the future path $(X_{n+1}, X_{n+2}, \ldots)$. Formally, the random time T is a randomized stopping time for the Markov chain (X_t) with initial distribution μ if for every $n \geq 0$, the event $\{T = n\}$ is conditionally independent of $(X_{n+1}, X_{n+2}, \ldots)$ under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_n)$. T is a randomized stopping time for the Markov scheme (X_t) if the conditional independence holds under $\mathbf{P}_\mu$ for every $\mu \in \mathcal{P}(\mathcal{X})$.

Let $f : \mathcal{X} \to \mathbf{R}$ be a measurable function, and let $\mu \in \mathcal{P}(\mathcal{X})$. Define

$$\mu(f) = \int_{\mathcal{X}} f(x)\mu(dx).$$

This definition makes sense whenever $f \geq 0$, and also for general f with $\mu(|f|) < \infty$. If P is a transition kernel on $\mathcal{X}$ and $t \geq 0$, then

$$P^t(\mu, f) = \int_{\mathcal{X}} f(x)P^t(\mu, dx).$$

It is said that $f \in L^p(\mu)$, for $1 \leq p < \infty$, if

$$\int_{\mathcal{X}} |f(x)|^p \mu(dx) < \infty.$$

If π is a stationary distribution for the transition kernel P, then one can define an operator on $L^p(\pi)$ for each p by $f \mapsto Pf$, where

$$(Pf)(x) = \int_{\mathcal{X}} f(y)P(x, dy).$$

The formula for $(Pf)(x)$ also makes sense if $f \geq 0$ even if it is not in any $L^p(\pi)$.

When $p = 2$, $L^2(\pi)$ is a Hilbert space with the inner product

$$\langle f, g\rangle_\pi = \int_{\mathcal{X}} f(x)g(x)\pi(dx).$$

P is *reversible* with respect to π if the operator $f \mapsto Pf$ on $L^2(\pi)$ is self-adjoint; this is equivalent to the condition that

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

as measures on $(\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{E})$. If P is reversible, its $L^2(\pi)$ spectrum $\sigma(P)$ is contained in the interval $[-1, 1]$. It will be said that P has *nonnegative eigenvalues* if $\sigma(P) \subseteq [0, 1]$. The $L^2(\pi)$ spectral gap of P is $1 - \sup\{r < 1 : r \in \sigma(P)\}$. When P is a finite-dimensional matrix this is simply the difference between 1 and the second-highest eigenvalue.

If P is reversible, an equivalent condition to $\sigma(P) \subseteq [0, 1]$ is that $\langle Pf, f\rangle_\pi \geq 0$ for all $f \in L^2(\pi)$. In fact, if P is reversible with $\sigma(P) \subseteq [0, 1]$, then for every $f \in L^2(\pi)$ the sequence $\langle P^n f, f\rangle_\pi$ is nonnegative and nonincreasing.

The norm of a function $f \in L^2(\pi)$ is

$$\|f\|_{L^2(\pi)}^2 = \int_{\mathcal{X}} |f(x)|^2 \pi(dx) = \langle f, f \rangle_\pi.$$

The $L^2(\pi)$ norm of a signed measure $\mu \in \mathcal{P}(\mathcal{X})$ that is absolutely continuous with respect to π is defined using the Radon-Nikodym derivative,

$$\|\mu\|_{L^2(\pi)} = \left\| \frac{d\mu}{d\pi} \right\|_{L^2(\pi)}.$$

If μ is not absolutely continuous with respect to π, one says that $\|\mu\|_{L^2(\pi)} = \infty$. For $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$, their $L^2(\pi)$ distance is $\|\mu_1 - \mu_2\|_{L^2(\pi)}$, interpreted as above.

Other distances between measures include the total variation distance, the L^1 distance, and the separation distance. The L^1 norm of a signed measure is

$$\|\mu\|_1 = \int_{\mathcal{X}} |\mu|(dx),$$

and the L^1 distance between measures $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ is given by $\|\mu_1 - \mu_2\|_1$. The *total variation distance* is

$$\|\mu_1 - \mu_2\|_{\mathrm{TV}} = \frac{1}{2}\|\mu_1 - \mu_2\|_1 = \sup_{A \in \mathcal{E}} |\mu_1(A) - \mu_2(A)|.$$

The *separation distance*, which is not symmetric, is given by

$$d_{\mathrm{sep}}(\mu_1, \mu_2) = \inf\{0 \le \beta \le 1 : \mu_1(A) \ge (1-\beta)\mu_2(A) \text{ for all } A \in \mathcal{E}\}.$$

It is well-known that $\|\mu_1 - \mu_2\|_{\mathrm{TV}} \le d_{\mathrm{sep}}(\mu_1, \mu_2)$.

For a Markov chain with transition matrix P and stationary distribution π on a finite state space $\mathcal{X}$, the *total variation mixing time* with parameter $0 < \varepsilon < 1$ is

$$t_{\mathrm{mix}}(\varepsilon) = \min\{t \ge 0 : \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \le \varepsilon \text{ for all } x \in \mathcal{X}\}.$$

By convention, the expression t_{mix} without any specified ε refers to $t_{\mathrm{mix}}(1/4)$.

Given vectors $v, w \in \mathbf{R}^d$, denote their standard inner product by (v, w) and their Euclidean norms by $\|v\|_2, \|w\|_2$.

Several well-known distributions of random variables will be considered, most frequently the multivariate normal distribution and the gamma distribution. A multivariate normal random variable in k dimensions with mean $m \in \mathbf{R}^k$ and $k \times k$ covariance matrix Σ will be denoted by $X \sim N_k(m, \Sigma)$.

A gamma random variable with shape parameter α and rate parameter β, whose density function is

$$G_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

will be denoted by $X \sim G(\alpha, \beta)$.

Chapter 2

Drift and minorization

The main object of study in this thesis will be a Markov transition kernel satisfying a drift and minorization condition. Any aperiodic chain whose transition kernel satisfies such a condition must converge to its stationary distribution at an exponential rate. This chapter is devoted to exploring this fundamental convergence result. The goal is to build intuition while presenting the historical development of the subject and placing the new results of this thesis in context.

Section 2.1 gives precise definitions and explains the original proof of exponential convergence, due to Nummelin and Tuominen [NT82]. Section 2.2 discusses quantitative convergence bounds and introduces a second method of proof developed by Rosenthal [Ros95a; Ros02]. Section 2.3 explains how the quantitative bounds improve when the Markov chain is reversible. Here there are two approaches, one taken by Baxendale [Bax05] and the other to be developed in Chapters 3 and 4 of this work. Section 2.4 talks about the various norms under which one can demonstrate exponential convergence and the relationships between them.

2.1 Exponential convergence

A drift and minorization condition consists of two pieces: a drift function with respect to a subset C of the state space $\mathcal{X}$, and a minorization property that holds for the same subset C.

Definition 2.1. Let P be a transition kernel on the state space $(\mathcal{X}, \mathcal{E})$, and let $C \in \mathcal{E}$. A *drift function* for P with respect to C is a measurable function $V : \mathcal{X} \to [1, \infty)$ along with constants

$\lambda < 1$ and $K < \infty$ such that for all $x \in \mathcal{X}$,

$$PV(x) \leq \begin{cases} \lambda V(x) & \text{if } x \notin C, \\ K & \text{if } x \in C. \end{cases}$$

Drift functions are also frequently called *Lyapunov functions*.

Definition 2.2. Let P be a transition kernel on $(\mathcal{X}, \mathcal{E})$, and let $C \in \mathcal{E}$. A *minorization property* for C is a probability measure $\nu \in \mathcal{P}(\mathcal{X})$, called the *minorization measure*, along with an integer $m \geq 1$ and a constant $\varepsilon > 0$, such that

$$P^m(x, \cdot) \geq \varepsilon \nu(\cdot) \qquad \text{for all } x \in C.$$

A set C that has a minorization property is called a *small set* for P. If the value of m is important, one can speak of a small set with m-step minorization.

Definition 2.3. A transition kernel P on $\mathcal{X}$ is said to satisfy a *drift and minorization condition* if P has a drift function with respect to a small set.

There are a few variants of these definitions in the literature, but they are all essentially equivalent. It is well-known, though certainly not immediate, that any transition kernel satisfying a drift and minorization condition has a unique stationary distribution π. The next definition formalizes the idea of exponential convergence to stationarity.

Definition 2.4. A transition kernel P on $\mathcal{X}$ having stationary distribution π is *geometrically ergodic* if there are a constant $\gamma < 1$ and a measurable function $B : \mathcal{X} \to \mathbf{R}$ such that for all $x \in \mathcal{X}$,

$$\|P^t(x, \cdot) - \pi\|_{\text{TV}} \leq B(x)\gamma^t \qquad \text{for all } t \geq 0. \tag{2.1}$$

If (X_t) is a Markov chain or scheme with transition kernel P, one says that (X_t) satisfies a drift and minorization condition or is geometrically ergodic if the appropriate properties hold for P.

One key point of geometric ergodicity is that the exponential rate γ is independent of the starting location x. It turns out that if the transition kernel exhibits exponential convergence, in the sense that there are functions $B : \mathcal{X} \to \mathbf{R}$ and $\gamma : \mathcal{X} \to [0, 1)$ such that

$$\|P^t(x, \cdot) - \pi\|_{\text{TV}} \leq B(x)\gamma(x)^t \qquad \text{for } \pi\text{-almost every } x \in \mathcal{X} \text{ and all } t \geq 0, \tag{2.2}$$

then there is a fixed constant $\gamma < 1$ such that (2.1) holds for π-almost every $x \in \mathcal{X}$. This result is due to Vere-Jones [VJ62] in the case where $\mathcal{X}$ is countable, and Nummelin and Tweedie [NT78] for general $\mathcal{X}$ (subject only to the requirement that $\mathcal{E}$ is countably generated).

If (2.2) holds for every $x \in \mathcal{X}$ rather than just π-almost every x, it is still possible for (2.1) to fail on a set of π-measure zero. Consider the following transition kernel on the nonnegative integers: $P(0,0) = 1$, and for $j > 0$, $P(j,0) = 1/j$ while $P(j,j) = 1 - 1/j$. The stationary distribution π is concentrated at zero, which is an absorbing state, and for $j > 0$,

$$\|P^t(j,\cdot) - \pi\|_{\mathrm{TV}} = \left(1 - \frac{1}{j}\right)^t.$$

Sometimes in the literature (e.g. [NT82; RR04]), geometric ergodicity is defined so that (2.1) is only required to be true for π-almost every $x \in \mathcal{X}$. This definition has the advantage of working well with the theorem of Vere-Jones and Nummelin–Tweedie, and the disadvantage that the Markov chain is allowed to behave badly when started from a set of π-measure zero, as in the example given above. Definition 2.4, which requires (2.1) for every $x \in \mathcal{X}$, is also standard (e.g. [MT93; Bax05]), and will be used henceforth.

The idea of drift and minorization was devised, by Popov [Pop77] in the case of countable state space and by Nummelin and Tuominen [NT82] in the general case, as a sufficient condition for geometric ergodicity. Nummelin and Tuominen essentially proved the following:

Theorem 2.5. *Suppose the transition kernel P is aperiodic and satisfies a drift and minorization condition. Then P is geometrically ergodic.*

A full proof of Theorem 2.5 will not be given here, but two strategies will be outlined.

Roughly speaking, a transition kernel is aperiodic if there is not a partition of the state space $\mathcal{X} = D_1 \cup D_2 \cup \cdots \cup D_k$ such that for all $x \in D_i$, $P(x, D_{i+1}) = 1$ (taking the indices mod k). This definition elides subtleties having to do with sets of measure zero; for a rigorous development, see Chapter 5 of [MT93]. If P satisfies a drift and minorization condition with small set C, then aperiodicity is equivalent to the statement that $\gcd\{t \geq 0 : P^t(x, C) > 0\} = 1$ for all $x \in C$.

The assumption of aperiodicity in Theorem 2.5 is very important. Considering what goes wrong when the Markov chain is periodic or almost periodic will provide motivation for the main results in this thesis. Here is an instructive example: the state space is the discrete circle $\mathcal{X} = \mathbf{Z}/N\mathbf{Z}$, and the transition kernel is given by $P(k, k-1) = 1$ for all $k \in \mathcal{X}$. A chain with transition kernel P will keep going around the circle forever and never converge to stationarity. However, it is straightforward to define a drift and minorization condition for P. For minorization, set $m = 1$, $\varepsilon = 1$, and $\nu = \delta_{N-1}$ to be the appropriate delta measure. (In general, no minorization is needed when $C = \{c\}$ is a single element, but one can always let $\nu(\cdot) = P(c, \cdot)$, $m = 1$, and $\varepsilon = 1$ if Definition 2.2 must be satisfied.) For drift, fix any $0 < \lambda < 1$ and set $V(k) = \lambda^{-k}$ for $0 \leq k \leq N-1$. Then for $k \neq 0$, $PV(k) = \lambda V(k)$, and $PV(0) = \lambda^{-(N-1)}$.

A slightly modified version of this example will be discussed in Section 2.2 where the Markov chain is not periodic and has a well-behaved drift and minorization condition, but nevertheless converges quite slowly. At that point it will be useful to make N large while keeping the upper bound $K \geq PV(0)$ constant. This can be accomplished by choosing $\lambda = 1 - 1/N$, so that

$$PV(0) = \left(1 + \frac{1}{N-1}\right)^{N-1} \leq e.$$

Here is the first strategy for proving Theorem 2.5. This is roughly the original method of Nummelin and Tuominen, as well as the approach taken later in the book of Meyn and Tweedie [MT93].

The drift property says that if $x \notin C$, then $PV(x) \leq \lambda V(x)$. Suppose the Markov chain (X_t) with transition kernel P is started from $X_0 = x$. With each time step, if the chain has not yet reached C, one expects the value of $V(X_t)$ to drop by a factor of λ. This cannot keep going forever because of the requirement that $V \geq 1$ at all points in $\mathcal{X}$. Therefore, the value of $V(x)$ controls the law of the hitting time τ_C for the Markov chain started from x. This control is captured in the following well-known lemma, which will be proved in Chapter 4 as Lemma 4.7.

Lemma 2.6. *Let (X_t) be a Markov scheme with transition kernel P on $(\mathcal{X}, \mathcal{E})$, and let $C \in \mathcal{E}$ be a subset of $\mathcal{X}$. Suppose the measurable function $V : \mathcal{X} \to [1, \infty)$ satisfies $PV(x) \leq \lambda V(x)$ for $x \notin C$, where $\lambda < 1$ is fixed. Then for all $x \in \mathcal{X}$,*

$$\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x).$$

Note that the function $V_0(x) = \mathbf{E}_x[\lambda^{-\tau_C}]$ itself satisfies $PV_0(x) = \lambda V_0(x)$ for $x \notin C$, and $V_0 \geq 1$ on all of $\mathcal{X}$. Thus $V_0(x)$ is the minimal drift function for P given the subset C and the constant λ.

By convexity, Lemma 2.6 implies that

$$\mathbf{E}_x[\tau_C] \leq \frac{\log V(x)}{\log \lambda^{-1}}. \tag{2.3}$$

If the value of $V(X_t)$ dropped by exactly a factor of λ at each time step, $\log V(x)/\log \lambda^{-1}$ is the number of steps it would take to reach 1 from a starting point of $V(x)$. So, (2.3) should not be too surprising. The statement of Lemma 2.6 is much more powerful, because it controls an exponential moment of the hitting time τ_C rather than just the expected value.

The next ingredient in the proof of Theorem 2.5 has to do with the minorization. Assume for the moment that the small set C has 1-step minorization, that is, $m = 1$ in Definition 2.2. If $X_t \in C$, the next step of the Markov chain can be described by the following algorithm: first, flip a coin with probabilities $\varepsilon, 1 - \varepsilon$. If the coin shows ε, choose $X_{t+1} \sim \nu$. If the coin shows $1 - \varepsilon$, choose $X_{t+1} \sim \frac{1}{1-\varepsilon}[P(X_t, \cdot) - \varepsilon\nu(\cdot)]$. In both cases, continue running the chain afterwards, flipping an independent $\varepsilon, 1 - \varepsilon$ coin every time t that $X_t \in C$.

Denote the random times that the coin-flip shows ε by $T_1 - 1, T_2 - 1, \ldots$, so that at each time T_j, $X_{T_j} \sim \nu$. The T_j are called *regeneration times* for the chain (X_t): X_{T_j} is still distributed according to ν even if the value of T_j and the entire sample path $(X_0, \ldots, X_{T_j - 1})$ are known. Therefore, the behavior of (X_t) after time T_j is completely independent of what came before.

The observation that such a sequence of regeneration times can be defined for any Markov chain having a small set with 1-step minorization is due independently to Athreya and Ney [AN78] and to Nummelin [Num78]. To describe how [AN78] and [Num78] used this construction to analyze the convergence properties of the Markov chain requires a short detour into the subject of discrete renewal theory. (For a textbook treatment, see Chapter 2 of [BL08] and the references therein.)

Suppose that Q is the transition kernel for a Markov chain (Y_t) on a countable state space $\mathcal{W}$ starting from a distinguished state $c \in \mathcal{W}$. Let $u_n = Q^n(c, c)$. Then $u_0 = 1$, and for $n \geq 1$, u_n satisfies the recurrence

$$u_n = \sum_{k=1}^{n} u_{n-k} \, \mathbf{P}_c(\tau_c^+ = k).$$

A sequence (u_n) with $u_0 = 1$ and

$$u_n = \sum_{k=1}^{n} u_{n-k} b_k \qquad \text{for } n \geq 1, \tag{2.4}$$

where each $b_k \geq 0$ and $\sum_{k=1}^{\infty} b_k = 1$, is called a *renewal sequence*. The sequence (b_n) is called the *increment sequence* associated with (u_n).

The sample path $(Y_0, Y_1, \ldots)$ can be partitioned into independent tours between consecutive visits to c. Assume that the expected tour length $\mathbf{E}_c[\tau_c^+]$ is finite. The distribution of Y_t is given by the formula

$$\mathbf{P}_c(Y_t \in A) = \sum_{k=0}^{\infty} u_{t-k} \, \mathbf{P}_c(Y_k \in A, \tau_c^+ > k), \tag{2.5}$$

where u_n is taken to be zero when $n < 0$. Since u_n is the probability that a tour begins at time n, if $\lim_{n\to\infty} u_n$ exists, it must equal $1/\,\mathbf{E}_c[\tau_c^+]$. Then, dominated convergence shows that

$$\lim_{t\to\infty} \mathbf{P}_c(Y_t \in A) = \sum_{k=0}^{\infty} \frac{1}{\mathbf{E}_c[\tau_c^+]} \, \mathbf{P}_c(Y_k \in A, \tau_c^+ > k). \tag{2.6}$$

The right side is necessarily a stationary distribution for (Y_t); this can also be checked directly without much trouble (see e.g. Theorem 6.5.2 of [Dur10]). Denote this distribution by $\pi(A)$. Subtracting (2.6) from (2.5),

$$|Q^t(c, A) - \pi(A)| \leq \sum_{k=0}^{\infty} \left| u_{t-k} - \frac{1}{\mathbf{E}_c[\tau_c^+]} \right| \mathbf{P}_c(\tau_c^+ > k).$$

Therefore, control over the rate of convergence of u_n to its limit, along with control over the tail of the increment sequence b_n, leads to a bound on $\|Q^t(c,\cdot) - \pi\|_{\mathrm{TV}}$. For a starting point $x \neq c$, one obtains a similar result as long as one can control $\mathbf{P}_x(\tau_c > k)$.

The regeneration time construction of [AN78] and [Num78] allowed them to apply this same analysis to the general state space transition kernel P. The role of the distinguished state c is played by the minorization measure ν. Suppose that the Markov chain (X_t) with transition kernel P starts from $X_0 \sim \nu$, and let $\mathbf{T} = \{0, T_1, T_2, \ldots\}$ be the set of regeneration times, with $T = T_1$. Then one can define $u_n = \mathbf{P}_\nu(n \in \mathbf{T})$ and $b_n = \mathbf{P}_\nu(T = n)$. By the regeneration property, the recurrence (2.4) holds. The same logic as before leads to the conclusion

$$|P^t(\nu, A) - \pi(A)| \leq \sum_{k=0}^{\infty} \left| u_{t-k} - \frac{1}{\mathbf{E}_\nu[T]} \right| \mathbf{P}_\nu(T > k). \tag{2.7}$$

With all this preparation, it is now possible to summarize the proof of Theorem 2.5. Suppose P is an aperiodic transition kernel on $\mathcal{X}$ satisfying a drift and minorization condition with respect to a small set C. Fix $x \in \mathcal{X}$ and let (X_t) be a Markov chain started from x with transition kernel P. First, assume that C has 1-step minorization.

1. Lemma 2.6 bounds $\mathbf{E}_x[\lambda^{-\tau_C}]$, so the law of τ_C decays exponentially. Every time t that the Markov chain enters C, there is a probability ε that the first regeneration time T equals $t+1$ (assuming that $T > t$). Therefore, the law of T also decays exponentially.

2. A lemma of Kendall [Ken59] says the following. Let (u_n) be a renewal sequence, and (b_n) be the associated increment sequence. If $b_n \to 0$ at an exponential rate, then u_n converges to its limit at an exponential rate, provided that the aperiodicity condition $\gcd\{n : b_n > 0\} = 1$ is satisfied.

3. Let $u_n = \mathbf{P}_\nu(n \in \mathbf{T})$ and $b_n = \mathbf{P}_\nu(T = n)$. By step 1, $b_n \to 0$ at an exponential rate. The gcd condition follows from aperiodicity of the transition kernel P. Thus Kendall's Lemma applies, and both pieces of the right side of (2.7) decay exponentially.

4. For any $x \in \mathcal{X}$, an exponential bound on $\mathbf{P}_x(T > t)$ combined with the exponential bound on $\|P^t(\nu,\cdot) - \pi\|_{\mathrm{TV}}$ leads to an exponential bound on $\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}}$.

If the small set has m-step minorization for $m > 1$, apply the previous argument to the Markov chain $(X_{mt}) = (X_0, X_m, X_{2m}, \ldots)$. The only issue is the exponential bound on the law of τ_C. Set $\tau_C^{(m)} = \min\{t \geq 0 : X_{mt} \in C\}$. An argument is given based on aperiodicity of P that if the law of τ_C decays exponentially, so does the law of $\tau_C^{(m)}$. The rest of the argument goes through, giving a bound on $\|P^{mt}(x,\cdot) - \pi\|_{\mathrm{TV}}$. Because $\|P^s(x,\cdot) - \pi\|_{\mathrm{TV}}$ is nonincreasing in s, the bound on

$\|P^{mt}(x,\cdot) - \pi\|_{\mathrm{TV}}$ also gives a bound on $\|P^{s}(x,\cdot) - \pi\|_{\mathrm{TV}}$ for general s (i.e. not just the multiples of m). This finishes the proof.

The converse of Theorem 2.5 is also true. The following result was first stated explicitly by Roberts and Rosenthal [RR97], though it is implicit in earlier works such as [MT93].

Theorem 2.7. *Suppose P is a geometrically ergodic transition kernel on a state space $(\mathcal{X}, \mathcal{E})$ where $\mathcal{E}$ is countably generated. Then P is aperiodic and satisfies a drift and minorization condition.*

It should be noted that [RR97] proved Theorem 2.7 using the "almost everywhere" definition of geometric ergodicity, with the drift function in the conclusion allowed to be infinite on a set of measure zero. If one strengthens the assumption to the "everywhere" definition of geometric ergodicity, one obtains a drift function which is everywhere finite.

The proof of Theorem 2.7 is nonconstructive. Its interest is that it confirms the power of the drift and minorization technique: if a Markov chain is in fact geometrically ergodic, it can be proved by drift and minorization.

2.2 Quantitative bounds and bivariate drift

With the advent of Markov chain Monte Carlo, researchers started giving more attention to quantitative convergence bounds for general state space Markov chains. Theorem 2.5 is "soft" in that it provides no such quantitative bounds. The reason is that aperiodicity is a soft assumption. The periodic example in the previous section, with state space $\mathcal{X} = \mathbf{Z}/N\mathbf{Z}$, can be perturbed so that it is aperiodic but converges arbitrarily slowly.

Meyn and Tweedie [MT94] obtained the first quantitative version of Theorem 2.5 under the following aperiodicity assumption: in the minorization $P^m(x,\cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$, one has $m = 1$ and $\nu(C) \geq \delta/\varepsilon$ for some constant $\delta > 0$. This assumption is called "strong aperiodicity" in the literature. Baxendale [Bax05] and Bednorz [Bed13] get better quantitative bounds starting from the same assumption. (Their formulas are quite involved but fully explicit.) When applied to toy examples from MCMC, all of these results are several orders of magnitude too conservative. A modification of the periodic example gives one reason why.

With $\mathcal{X} = \mathbf{Z}/N\mathbf{Z}$, define the transition kernel by $P(k, k-1) = 1$ when $k \neq 0$. When $k = 0$, let $P(0,0) = 1/2$ and $P(0, N-1) = 1/2$. As before, let $C = \{0\}$, and keep the drift function $V(k) = \lambda^{-k}$ where $\lambda = 1 - 1/N$. One can take $K = (1+e)/2$. For minorization, let $\varepsilon = 1$, $m = 1$, and $\nu(0) = \nu(N-1) = 1/2$.

Imagine a random walker moving around the circle $\mathbf{Z}/N\mathbf{Z}$ according to P. Every time it reaches zero, it pauses for a random number of time steps before continuing around. The amount of time that the walker pauses at zero is a geometric random variable with parameter $1/2$. In order for $\|P^t(x,\cdot)-\pi\|_{\text{TV}}$ to be small, the total amount of time paused at zero, essentially a sum of independent Geometric($1/2$) random variables, must have standard deviation of at least order N. This will not happen until the random walker has taken order N^2 trips around the circle, so t must be at least order N^3. Baxendale [Bax05] considers exactly this example at the end of Section 3.1 (though in different language). He confirms that the best possible γ in $\|P^t(x,\cdot)-\pi\|_{\text{TV}} \le B(x)\gamma^t$ satisfies $1-\gamma = O(1/N^3)$.

To summarize, the minorization data are $m = 1$, $\varepsilon = 1$, and $\delta = 1/2$; and the drift data are $\lambda = 1 - 1/N$ and $K = (1+e)/2$. The Markov chain regenerates roughly every N steps, but it needs order $N^3 = 1/(1-\lambda)^3$ steps to mix. Any quantitative convergence bound starting from the same assumptions as [MT94; Bax05; Bed13] is stuck with this problem: the mixing time may be as large as the cube of the number of steps needed to regenerate. The unavoidable factor of $1/(1-\lambda)^3$ appears in the formulas of [Bax05] and [Bed13]. (The bound in [MT94] had a factor of $1/(1-\lambda)^4$.)

A different approach that usually gives better bounds for chains used in MCMC is the bivariate drift method of Rosenthal [Ros95a; Ros02], which uses the technique of coupling.

Definition 2.8. Let (X_t) and (X'_t) be two Markov chains with the same transition kernel on the same state space $(\mathcal{X},\mathcal{E})$, having starting distributions μ and μ' respectively. A *coupling* of (X_t) and (X'_t) is a single probability measure $\mathbf{P}_{\mu,\mu'}$ on a sample space Ω, together with maps $X, X' : \Omega \to \mathcal{X}^\infty$, such that the law of $X(\omega)$ under $\mathbf{P}_{\mu,\mu'}$ is the same as the law of (X_t), and the law of $X'(\omega)$ under $\mathbf{P}_{\mu,\mu'}$ is the same as the law of (X'_t). Define the *coupling time*

$$T_{\text{couple}} = \min\{t \ge 0 \,:\, X_t = X'_t\}.$$

The coupling is *faithful* if

$$\mathbf{P}_{\mu,\mu'}(X_t = X'_t \text{ for all } t \ge T_{\text{couple}}) = 1.$$

The well-known Coupling Inequality ([LPW09], Theorem 5.2) relates faithful couplings to total variation convergence.

Proposition 2.9. *Let $\mathbf{P}_{\mu,\mu'}$ be a faithful coupling of the Markov chains (X_t) and (X'_t), both with transition kernel P, as in Definition 2.8. Then*

$$\|P^t(\mu,\cdot) - P^t(\mu',\cdot)\|_{\text{TV}} \le \mathbf{P}_{\mu,\mu'}(T_{\text{couple}} > t).$$

Usually one sets $\mu' = \pi$ to get a bound on $\|P^t(\mu,\cdot)-\pi\|_{\text{TV}}$.

Definition 2.10. Let (X_t) and (X'_t) be Markov chains on $(\mathcal{X},\mathcal{E})$ with transition kernel P and initial distributions μ,μ'. Suppose $\bar{P}$ is a transition kernel on $\mathcal{X}\times\mathcal{X}$ (with the product σ-algebra $\mathcal{E}\otimes\mathcal{E}$)

such that for all $x, x' \in \mathcal{X}$ and $A, A' \in \mathcal{E}$,

$$\begin{aligned} \bar{P}((x,x'), A \times \mathcal{X}) &= P(x, A), \\ \bar{P}((x,x'), \mathcal{X} \times A') &= P(x', A'). \end{aligned} \tag{2.8}$$

Also suppose that $\bar{\mu}$ is a probability measure on $\mathcal{X} \times \mathcal{X}$ such that $\bar{\mu}(A \times \mathcal{X}) = \mu(A)$ and $\bar{\mu}(\mathcal{X} \times A') = \mu'(A')$ for all $A, A' \in \mathcal{E}$. A Markov chain on $(\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{E})$ with transition kernel $\bar{P}$ and initial distribution $\bar{\mu}$ is called a *Markovian coupling* of (X_t) and (X'_t). Any Markovian coupling satisfies Definition 2.8.

Let $\Delta = \{(x, x') \in \mathcal{X} \times \mathcal{X} \,:\, x = x'\}$. If $\bar{P}((x,x), \Delta) = 1$ for all $x \in \mathcal{X}$, the Markovian coupling is faithful. In order for this to be true, $\bar{P}((x,x), \cdot)$ must be given by

$$\bar{P}((x,x), A \times A') = P(x, A \cap A') \tag{2.9}$$

for all $x \in \mathcal{X}$ and $A, A' \in \mathcal{E}$.

Most couplings used in practice are Markovian. The construction below will use a coupling that is non-Markovian, but only slightly.

Definition 2.11. Let P be a transition kernel on $(\mathcal{X}, \mathcal{E})$. A *bivariate drift and minorization condition* for P consists of:

- A small set $C \in \mathcal{E}$, with $P^m(x, \cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$;
- A transition kernel $\bar{P}$ on $(\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{E})$ that satisfies (2.8) and (2.9);
- A measurable function $\bar{V} : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, called the *bivariate drift function*, together with a constant $\bar{\lambda} < 1$, such that $\bar{P}\bar{V}(x, x') \leq \bar{\lambda}\bar{V}(x, x')$ for all $(x, x') \notin C \times C$;
- A finite constant $\bar{K}$ such that $\bar{P}\bar{V}(x, x) \leq \bar{K}$ for all $x \in C$;
- If $\varepsilon < 1$, a remainder kernel $\bar{R}$ assigning to each $(x, x') \in C \times C$ a probability measure on $\mathcal{X} \times \mathcal{X}$ such that for all $A, A' \in \mathcal{E}$,

  $$\begin{aligned} \varepsilon\nu(A) + (1-\varepsilon)\bar{R}((x,x'), A \times \mathcal{X}) &= P^m(x, A), \\ \varepsilon\nu(A') + (1-\varepsilon)\bar{R}((x,x'), \mathcal{X} \times A') &= P^m(x', A'), \end{aligned}$$

  and a finite constant $\bar{L}$ for which

  $$\sup_{(x,x') \in C \times C} \bar{R}\bar{V}(x, x') \leq \bar{L}.$$

Note that the function $V(x) = \bar{V}(x, x)$ is automatically a (univariate) drift function for P with respect to C, because $PV(x) = \bar{P}\bar{V}(x, x)$.

Here is Rosenthal's construction (from [RR04]; see also [Ros02]) for a faithful coupling of two chains (X_t) and (X'_t) whose transition kernel P satisfies a bivariate drift and minorization condition. Start with $X_0 \sim \mu$ and $X'_0 \sim \mu'$; the joint distribution of (X_0, X'_0) is allowed to be anything as long as the marginal distributions of X_0 and X'_0 are μ and μ', respectively. Choose $(X_1, X'_1), (X_2, X'_2), \ldots$ according to $\bar{P}$ until the first time $S_1 \geq 0$ that $(X_{S_1}, X'_{S_1}) \in C \times C$. Now, flip an $\varepsilon, 1 - \varepsilon$ coin.

If the coin shows ε, set $X_{S_1+m} = X'_{S_1+m}$, both distributed according to ν. For $t > S_1+m$, keep $X_t = X'_t$, distributed according to $P(X_{t-1}, \cdot)$. Finally, fill in the missing values $(X_{S_1+1}, \ldots, X_{S_1+m-1})$ according to their conditional distribution under $\mathbf{P}_\mu$ given X_{S_1} and X_{S_1+m}, conditionally independent of everything else; and fill in the values $(X'_{S_1+1}, \ldots, X'_{S_1+m-1})$ according to their conditional distribution under $\mathbf{P}_{\mu'}$ given X'_{S_1} and X'_{S_1+m}, conditionally independent of everything else.

If the coin shows $1 - \varepsilon$, choose (X_{S_1+m}, X'_{S_1+m}) according to $\bar{R}((X_{S_1}, X'_{S_1}), \cdot)$. From there, choose $(X_{S_1+m+1}, X'_{S_1+m+1}), (X_{S_1+m+2}, X'_{S_1+m+2}), \ldots$ according to $\bar{P}$ until the first time $S_2 \geq S_1 + m$ that $(X_{S_2}, X'_{S_2}) \in C \times C$. Then flip another independent $\varepsilon, 1 - \varepsilon$ coin. If the coin shows ε, set $X_{S_2+m} = X'_{S_2+m}$, both distributed according to ν, and proceed as in the previous paragraph. If the coin shows $1 - \varepsilon$, choose (X_{S_2+m}, X'_{S_2+m}) according to $\bar{R}((X_{S_2}, X'_{S_2}), \cdot)$ and continue to S_3. Keep going until the first time the coin shows ε. At that point, there will be a number of missing sequences. Fill in each sequence $(X_{S_j+1}, \ldots, X_{S_j+m-1})$ according to its conditional distribution under $\mathbf{P}_\mu$ given X_{S_j} and X_{S_j+m}, conditionally independent of everything else, and likewise with the sequences $(X'_{S_j+1}, \ldots, X'_{S_j+m-1})$.

When $m = 1$, there are no missing sequences, and the coupling is Markovian. When $m > 1$, the missing sequence construction means that the coupling may not be Markovian.

The advantage of the bivariate drift approach is that it is relatively straightforward to obtain a bound on T_{couple} given the data in Definition 2.11. The computation is exactly the same as the one that bounds the law of the regeneration time (or strong ν time) T in the case of univariate drift and minorization. By the Coupling Inequality, taking $\mu' = \pi$, one immediately has an exponential bound on the total variation distance $\|P^t(\mu, \cdot) - \pi\|_{\text{TV}}$.

Theorem 2.12. *Suppose the transition kernel P satisfies a bivariate drift and minorization condition. Let $\bar{\mu}$ be a measure on $\mathcal{X} \times \mathcal{X}$ with $\bar{\mu}(\bar{V}) < \infty$, and define the marginal measures μ, μ' by $\mu(A) = \bar{\mu}(A \times \mathcal{X})$ and $\mu'(A') = \bar{\mu}(\mathcal{X} \times A')$. Then there are explicit constants B and $\gamma < 1$ such that*

$$\|P^t(\mu, \cdot) - P^t(\mu', \cdot)\|_{\text{TV}} \leq B\gamma^t.$$

The value of γ depends only on the bivariate drift and minorization data and not on $\bar{\mu}$.

Let π be the stationary distribution for P. For $x \in \mathcal{X}$, the measure $\delta_x \otimes \pi$ on $\mathcal{X} \times \mathcal{X}$ is given by

$$(\delta_x \otimes \pi)(A \times A') = \begin{cases} \pi(A') & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

If $(\delta_x \otimes \pi)(\bar{V}) < \infty$ for every $x \in \mathcal{X}$, then P is geometrically ergodic, with an explicit upper bound on the rate of convergence.

Theorem 2.12 is first stated in [Ros95a]. A simple proof along the lines described above is given in [Ros02], at least for the case $m = 1$; the extension to $m > 1$ is routine. Improved formulas for B and γ will be given in Theorem 4.13.

The disadvantage of the bivariate drift approach is that one needs to find a bivariate drift function $\bar{V}$ along with the kernels $\bar{P}$ and $\bar{R}$. There are two situations where this is no more difficult than finding a univariate drift function. The first is when the transition kernel P satisfies a monotonicity property, and the second is when the small set C covers a large fraction of the state space. In both cases the finiteness condition $(\delta_x \otimes \pi)(\bar{V}) < \infty$ is automatically satisfied, so Theorem 2.12 gives an explicit form of geometric ergodicity.

Here is a discussion of the monotonicity situation. The theory was developed by Lund and Tweedie [LT96] and Roberts and Tweedie [RT00]. Suppose the transition kernel P has a drift function V with respect to a small set of the form $C = \{x \in \mathcal{X} : V(x) \leq d\}$. Imagine that (X_t) and (X'_t) are two random walkers on $\mathcal{X}$, starting from distributions μ and μ' respectively and each moving according to P. Now, suppose there is a coupling of (X_t) with (X'_t) so that if $V(X_0) \leq V(X'_0)$, then $V(X_t) \leq V(X'_t)$ for all t; and if $V(X_0) \geq V(X'_0)$, then $V(X_t) \geq V(X'_t)$ for all t. This is called a *monotone coupling*. Assume without loss of generality that $V(X_0) \leq V(X'_0)$. Whenever $X'_t \in C$, $X_t \in C$ also, because $V(X_t) \leq V(X'_t)$. Therefore one can proceed as follows. Run the coupled chain until $X'_t \in C$, and flip the $\varepsilon, 1-\varepsilon$ coin. If the coin shows ε, make $X_{t+m} = X'_{t+m}$. If the coin shows $1-\varepsilon$, it can still be arranged that $V(X_{t+m}) \leq V(X'_{t+m})$, so the coupling can proceed from there.

Effectively, if $V(X_0) \leq V(X'_0)$, the coupling time is determined only by the path $(X'_0, X'_1, \ldots)$ and the coin flips, because $V(X_t)$ is "trapped" below $V(X'_t)$. This fits into the framework of bivariate drift and minorization if one chooses $\bar{V}(x, x') = \max\{V(x), V(x')\}$.

Definition 2.13. Let P be a transition kernel on $\mathcal{X}$, and let $f : \mathcal{X} \to \mathbf{R}$ be a measurable function. P is *stochastically monotone with respect to f* if whenever $f(x) \leq f(x')$, then for every $r \in \mathbf{R}$,

$$\mathbf{P}_x(f(X_1) > r) \leq \mathbf{P}_{x'}(f(X_1) > r).$$

If $\mathcal{X} \subseteq \mathbf{R}$ and f is the identity function, P is simply *stochastically monotone*.

Proposition 2.14. *Let the transition kernel P have a drift function V with respect to a small set*

$C = \{x \in \mathcal{X} : V(x) \leq d\}$. If P is stochastically monotone with respect to V, then P satisfies a bivariate drift and minorization condition with $\bar{V}(x, x') = \max\{V(x), V(x')\}$ and $\bar{\lambda} = \lambda$.

In [LT96] it is shown that the quantitative convergence bounds arising from Proposition 2.14 are sharp for a certain family of chains on the state space $\mathcal{X} = [0, \infty)$. Note that if the transition kernel P of a chain (X_t) is stochastically monotone with respect to V and $V(x) = V(x')$, then the law of $V(X_1)$ given that $X_0 = x$ must be equal to the law of $V(X_1)$ given that $X_0 = x'$. Thus, the projection $(V(X_t))$ is itself a Markov chain on $[1, \infty)$. In this sense, (X_t) is essentially one-dimensional.

Monotonicity enabled the bivariate drift function $\bar{V}$ to be constructed directly from the univariate drift function V. This is possible in the general (non-monotone) case if the small set C is large enough. The following proposition is a slightly modified version of Theorem 12 in [Ros95a].

Proposition 2.15. *Let the transition kernel P have a drift function V with respect to a small set $C = \{x \in \mathcal{X} : V(x) \leq d\}$. If $d > (K-1)/(1-\lambda)$, then P satisfies a bivariate drift and minorization condition with $\bar{V}(x, x') = [V(x) + V(x')]/2$ and $\bar{\lambda} = \lambda + (K - \lambda)/(d + 1)$.*

Proof. Let $\bar{P}((x, x'), A \times A') = P(x, A)P(x', A')$ when $x \neq x'$, the so-called *independent coupling*, and $\bar{P}((x, x), A \times A') = P(x, A \cap A')$. Let

$$\bar{R}((x, x'), A \times A') = \left(\frac{1}{1-\varepsilon}[P(x, A) - \varepsilon\nu(A)]\right)\left(\frac{1}{1-\varepsilon}[P(x', A') - \varepsilon\nu(A')]\right).$$

If $x \notin C$ and $x' \notin C$, then $\bar{P}\bar{V}(x, x') \leq \lambda\bar{V}(x, x')$. The key computation is that if $x \notin C$ but $x' \in C$,

$$\bar{P}\bar{V}(x, x') \leq \frac{1}{2}\lambda V(x) + \frac{1}{2}K = \lambda\bar{V}(x, x') + \frac{1}{2}[K - \lambda V(x')].$$

Since $V(x) \geq d$ and $V(x') \geq 1$,

$$\frac{K - \lambda V(x')}{V(x) + V(x')} \leq \frac{K - \lambda}{d + 1}.$$

It follows that

$$\bar{P}\bar{V}(x, x') \leq \left(\lambda + \frac{K - \lambda}{d + 1}\right)\bar{V}(x, x').$$

The last step is to check that $\sup_{(x,x') \in C \times C} \bar{R}\bar{V}(x, x') < \infty$. From the definitions of $\bar{R}$ and $\bar{V}$,

$$\sup_{(x,x') \in C \times C} \bar{R}\bar{V}(x, x') = \sup_{x \in C} \frac{1}{1-\varepsilon}\left[P^m(x, V) - \varepsilon\nu(V)\right].$$

An upper bound for this last quantity will be given in Corollary 4.11. □

In order for Propositions 2.14 and 2.15 to give a quantitative form of geometric ergodicity via Theorem 2.12, the bivariate drift functions must satisfy the finiteness condition $(\delta_x \otimes \pi)(\bar{V}) < \infty$

for all $x \in \mathcal{X}$. This follows from the general fact that if V is a (univariate) drift function for P with respect to C, and π is a stationary distribution for P,

$$\pi(V) \leq \frac{K-\lambda}{1-\lambda}\pi(C). \tag{2.10}$$

If $\bar{V}(x,x') = \max\{V(x), V(x')\}$, then $(\delta_x \otimes \pi)(\bar{V}) \leq V(x) + \pi(V)$, while if $\bar{V}(x,x') = [V(x)+V(x')]/2$, then $(\delta_x \otimes \pi)(\bar{V}) = [V(x)+\pi(V)]/2$. The inequality (2.10) will be proved in Lemma 4.8.

Proposition 2.15 provides another route to proving Theorem 2.5. Suppose P has a drift function V with respect to a small set C. If C contains $\{x \in \mathcal{X} : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$, then P satisfies a bivariate drift and minorization condition, leading immediately to geometric ergodicity via Theorem 2.12. If C is too small, it is argued in [RR04] that for any finite d, the set $C_d = C \cup \{x \in \mathcal{X} : V(x) \leq d\}$ is also a small set for P, as long as P is aperiodic. In the expansion from C to C_d, one loses quantitative control over the constants ε_d and m_d in the statement

$$P^{m_d}(x,\cdot) \geq \varepsilon_d \nu_d(\cdot) \qquad \text{for all } x \in C_d.$$

This is inevitable since aperiodicity is a soft condition; but even under the quantitative "strong aperiodicity" assumption of [MT94], it seems difficult to get good bounds on ε_d and m_d. In any case, the almost periodic example on $\mathcal{X} = \mathbf{Z}/N\mathbf{Z}$ precludes any drastic improvement over the results of [Bax05] and [Bed13].

Here is the overall situation. Suppose P has a drift function V with respect to a small set C, and P satisfies a quantitative aperiodicity condition along the lines of the strong aperiodicity of [MT94]. If C does not contain $\{x \in \mathcal{X} : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$, then the renewal theory approach of [MT94; Bax05; Bed13] yields explicit formulas for $B(x)$ and γ such that $\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t$. However, the numerical bounds are often too conservative for practical use, partly because of the almost periodic examples. The bivariate drift approach of [Ros95a; Ros02; RR04] yields no explicit formulas. If P is stochastically monotone with respect to V, the bivariate drift approach as applied by [LT96; RT00] gives very good numerical bounds.

If the small set C does contain $\{x \in \mathcal{X} : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$, both the renewal theory approach and the bivariate drift approach give explicit formulas for $B(x)$ and γ. In this situation, the bounds from the bivariate drift approach tend to be orders of magnitude better. Indeed, the bivariate drift approach has been successfully applied to some toy examples from MCMC, with encouraging results: see e.g. [Ros95a; JH01; JH04]. Unlike in the stochastically monotone case, the estimates of convergence rate are not sharp, but they are good enough to be useful.

Given that the numerical bounds are much better when C contains $\{x \in \mathcal{X} : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$, an important question is how restrictive this condition really is. The next result, originally due to Partha Dey (unpublished personal communication) in a slightly weaker form, shows that the condition is quite restrictive: the “small set” must encompass at least half the state space!

Proposition 2.16. *Suppose P has a drift function V with respect to a small set C that contains $\{x \in \mathcal{X} : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$. If π is a stationary distribution for P, then $\pi(C) > 1/2$.*

Proof. By (2.10), $\pi(V) < \infty$. Therefore,

$$\begin{aligned} 0 = (\pi P)(V) - \pi(V) &= \int_{\mathcal{X}} [PV(x) - V(x)]\pi(dx) \\ &= \int_C [PV(x) - V(x)]\pi(dx) - \int_{\mathcal{X}\setminus C} [V(x) - PV(x)]\pi(dx). \end{aligned}$$

Let $a = \sup_{x\in C}[PV(x) - V(x)]$ and $b = \inf_{x\notin C}[V(x) - PV(x)]$; note that $a \geq 0$ and $b > 0$. Then $0 \leq a\pi(C) - b[1-\pi(C)]$, which implies that

$$\frac{1-\pi(C)}{\pi(C)} \leq \frac{a}{b}. \tag{2.11}$$

The key step in the proof of Proposition 2.15 was the construction of a constant $\bar{\lambda} < 1$ such that

$$PV(x) + PV(x') \leq \bar{\lambda}[V(x) + V(x')] \qquad \text{for all } x \in C,\ x' \notin C. \tag{2.12}$$

This allowed the independent coupling to give rise to a bivariate drift and minorization condition. Note that if (2.12) is satisfied, then for all $x \in C$ and $x' \notin C$,

$$PV(x) + PV(x') + 2(1-\bar{\lambda}) \leq PV(x) + PV(x') + (1-\bar{\lambda})[V(x) + V(x')] \leq V(x) + V(x'),$$

which implies that $a + 2(1-\bar{\lambda}) \leq b$. Therefore, if the independent coupling gives rise to a bivariate drift and minorization condition, then $a < b$. It follows from (2.11) that $\pi(C) > 1/2$.

The requirement that $C \supseteq \{x : V(x) \leq d\}$ for some $d > (K-1)/(1-\lambda)$ was engineered so that the argument in Proposition 2.15 would work. Indeed, if the requirement is satisfied then

$$b = \inf_{x\notin C}[V(x) - PV(x)] \geq \inf_{x\notin C}(1-\lambda)V(x) \geq (1-\lambda)d > K-1,$$

and it is always true that $a = \sup_{x\in C}[PV(x) - V(x)] \leq K-1$. Thus the requirement implies that $a < b$, which immediately gives $\pi(C) > 1/2$. □

There are various alternative forms of the drift function criterion in the literature. For each one there is a corresponding version of Proposition 2.15. In all cases the bivariate drift construction relies on (2.12), so the argument above leads to the same conclusion that $\pi(C) > 1/2$.

Suppose (X_t) is a random walk on a graph, or more generally that (X_t) makes only local moves on the state space $\mathcal{X}$. In order to obtain an explicit convergence bound using Proposition 2.15, $\pi(C)$ must be greater than $1/2$. This in turn requires the number of minorization steps m to be relatively large, so that $P^m(x, \cdot)$ and $P^m(x', \cdot)$ can overlap even if x, x' are at "opposite ends" of C.

2.3 Reversibility

This section returns to the renewal theory approach of [MT93; Bax05; Bed13]. As discussed in Section 2.1, this approach converts drift and minorization data for a Markov chain (X_t) into explicit bounds on the chain's convergence to stationarity, assuming an aperiodicity condition is met. In practice, the bounds are usually orders of magnitude too conservative. This is partly because the aperiodicity condition still allows almost periodic chains that regenerate frequently but converge slowly.

The main idea in this section is that if (X_t) is reversible with nonnegative eigenvalues, it cannot exhibit almost periodic behavior. This is good because a very large number of Markov chains are reversible. (In MCMC, the Metropolis–Hastings algorithm and the random scan Gibbs sampler are reversible. For the sequential scan Gibbs sampler in two variables, each variable considered alone is a reversible chain [Bax05]. For finite chains, any simple random walk on an undirected and possibly weighted graph is reversible; a special case is the random walk on a group driven by a symmetric increment distribution. There are many more examples. Some of these automatically have nonnegative eigenvalues. If not, one can replace the transition kernel P by the *lazy* version $P_L(x, \cdot) = \frac{1}{2}[P(x, \cdot) + \delta_x(\cdot)]$, which necessarily has nonnegative eigenvalues.) Since there is no almost periodic behavior, it is no longer possible for (X_t) to regenerate frequently but converge slowly.

Let P be a transition kernel on $\mathcal{X}$ satisfying a drift and minorization condition. Assume for the moment that the number of minorization steps is $m = 1$, that is, $P(x, \cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$. In this situation, if (X_t) is a Markov scheme with transition kernel P, Section 2.1 described (following [AN78; Num78]) how a regeneration time T can be defined for (X_t). Let

$$\lambda_* = \inf\{\gamma \geq 0 : \text{for some finite function } F(x),\ \mathbf{P}_x(T > t) \leq F(x)\gamma^t \text{ for all } t \geq 0\}, \tag{2.13}$$

$$\rho_{\mathrm{TV}} = \inf\{\gamma \geq 0 : \text{for some finite function } B(x),\ \|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t \text{ for all } t \geq 0\}. \tag{2.14}$$

With no further assumptions, P may be periodic, in which case there is no convergence to stationarity and $\rho_{\mathrm{TV}} = 1$. With a standard aperiodicity assumption, $\rho_{\mathrm{TV}} < 1$ but may be much larger than λ_*. If P is stochastically monotone with respect to the drift function V, the regeneration time T can serve as a coupling time between the chain started from x and the chain started from π. Therefore,

$\rho_{\mathrm{TV}} \leq \lambda_*$. (For a proof of this statement along the lines of the discussion in Section 2.2, see [RT00]).

The first main result in this section is due to Baxendale [Bax05].

Theorem 2.17. *Let P be a transition kernel satisfying a drift and minorization condition, and assume the number of minorization steps is $m = 1$. Suppose P is reversible with nonnegative eigenvalues. If $\lambda_*, \rho_{\mathrm{TV}}$ are defined as in (2.13) and (2.14), then $\rho_{\mathrm{TV}} \leq \lambda_*$.*

In other words, the assumption that P is reversible with nonnegative eigenvalues gives nearly the same speed-up in convergence rate as the assumption that P is stochastically monotone with respect to V. In both cases, the time to stationarity is comparable to the time until the first regeneration.

Probably the most important theoretical result in this thesis is a generalization of Theorem 2.17 to the case of $m \geq 1$. When $m > 1$, the regeneration time construction of [AN78] and [Num78] yields instead a strong ν time T (as discussed in Section 1.1). The new result is this:

Theorem 2.18. *Theorem 2.17 holds in the general case $m \geq 1$.*

There are three ways in which Theorem 2.18 differs from Theorem 2.17. First, the method of proof for Theorem 2.18 is different and more probabilistic. In [Bax05], Theorem 2.17 is proved by looking at the generating functions

$$u(z) = \sum_{n=0}^{\infty} u_n z^n, \qquad b(z) = \sum_{n=1}^{\infty} b_n z^n,$$

where $u_n = \mathbf{P}_\nu(n \in \mathbf{T})$ and $b_n = \mathbf{P}_\nu(T = n)$ as in Section 2.1. These functions are related by the *renewal equation*

$$u(z) = \frac{1}{1 - b(z)}.$$

Let $u_\infty = \lim_{n\to\infty} u_n$ and

$$\rho_u = \inf\{\gamma \geq 0 : \text{there exists } D \text{ such that } |u_n - u_\infty| \leq D\gamma^n \text{ for all } n \geq 0\}.$$

It is not hard to show that $\rho_{\mathrm{TV}} \leq \rho_u$, so it suffices to prove $\rho_u \leq \lambda_*$.

The assumption that P is reversible with nonnegative eigenvalues is shown to imply that all the poles of $u(z)$ lie on the positive real axis. The definition of λ_* means that $b(z)$ has radius of convergence at least λ_*^{-1}. Since $b(1) = 1$ and $b(z)$ is increasing when $0 \leq z < \lambda_*^{-1}$, $b(z)$ cannot equal 1 anywhere else in that interval, so $u(z)$ has no poles inside the circle of radius λ_*^{-1} besides one at $z = 1$. This in turn means that $\rho_u \leq \lambda_*$, completing the proof.

The proof of Theorem 2.18 uses no generating functions. It starts with a version of the recurrence associated with the regeneration: if f is a measurable function on the state space $\mathcal{X}$,

$$\mathbf{E}_\nu[f(X_n), T \leq n] = \sum_{j=0}^{n} \mathbf{P}_\nu(T = n - j)\, \mathbf{E}_\nu[f(X_j)], \tag{2.15}$$

assuming that all the expectations are finite or that $f \geq 0$. A function f is chosen so that the sequence $\mathbf{E}_\nu[f(X_n)]$ controls the convergence of $P^n(\nu, \cdot)$ to π in the $L^2(\pi)$ distance; the exact equation is

$$\|P^n(\nu, \cdot) - \pi\|_{L^2(\pi)}^2 = \mathbf{E}_\nu[f(X_{2n})] - 1.$$

Using that P is reversible with nonnegative eigenvalues, the sequence $\mathbf{E}_\nu[f(X_n)]$ is shown to be decreasing in n. Theorem 2.18 follows from doing summation by parts on (2.15). The details are given in Section 4.4.

The second way that Theorems 2.17 and 2.18 differ is in the generalization to the case $m > 1$. The generating function approach described above for Theorem 2.17 does not work when $m > 1$ because it is possible to have $\rho_u > \lambda_*$. (Roughly speaking, this is because the definition of T may add artificial periodicity.) Therefore it is impossible to prove that $\rho_{\mathrm{TV}} \leq \lambda_*$ using $\rho_{\mathrm{TV}} \leq \rho_u$ as an intermediate step.

For chains used in MCMC, the case $m = 1$ has historically been the most important, due to the difficulty of getting quantitative m-step minorization conditions when $m > 1$. But higher values of m are essential for the analysis in Section 7.2 of a random walk on a high-dimensional finite graph. If $m = 1$ was required, the small set C would need to have diameter at most 2. Typically in high dimensions, the hitting time for such a set C is very large compared to the mixing time of the random walk. Since the Markov chain cannot regenerate until it reaches C for the first time, any drift-and-minorization bound for the mixing time of the chain will be far from sharp.

The analysis in Section 7.2 is of the lazy random walk on the hypercube, and uses a small set C which is half the state space. By sending $m \to \infty$, the drift-and-minorization bound gets the spectral gap to within a factor of 2.

The third difference between Theorems 2.17 and 2.18 has to do with the explicit bounds

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t.$$

There are quantitative versions of both theorems that provide explicit formulas for $B(x)$ and γ. When $m = 1$, the bounds can be compared. A standard operational definition of the time to stationarity for the Markov chain started from x is the least integer t such that $B(x)\gamma^t < 0.01$. For the MCMC examples treated in Chapter 5, the value of t given by Theorem 2.17 is 2 to 5 times the value of t given by Theorem 2.18.

The last topic in this section is a special result when the transition kernel P has a drift function with respect to a single-element small set $C = \{c\}$. Note that in this case the regeneration time T is simply the hitting time τ_c^+. If λ is the drift parameter, then $\lambda_* \le \lambda$ by Lemma 2.6. Assume P is reversible with nonnegative eigenvalues. Let (X_t) be a Markov chain with transition kernel P started from c, and let (X'_t) be a chain with transition kernel P started from π. An argument to be given in Section 4.6 implies that there is a faithful coupling of (X_t) with (X'_t) such that whenever $X'_t = c$, also $X_t = c$. Therefore, the first time that $X'_t = c$ is a coupling time for the two chains, giving a bound on total variation distance. The key ingredient is a theorem of Lund, Zhao, and Kiessler [LZK06] about the *hazard rates* of the renewal sequence $u_n = P^n(c, c)$. Let $b_n = \mathbf{P}_c(\tau_c^+ = n)$ be the increment sequence associated with u_n, and let $B_n = \sum_{k=n+1}^{\infty} b_k$. The hazard rate h_n is defined by

$$h_n = \frac{b_n}{B_{n-1}} = \mathbf{P}_c(\tau_c^+ = n \mid \tau_c^+ \ge n).$$

It is proved in [LZK06] that if P is reversible with nonnegative eigenvalues, the sequence (h_n) is decreasing. In other words, the more recently the Markov chain left state c, the more likely it is to jump back at a given time. This is why the coupling described above is possible. The result is a clean bound on the convergence of (X_t) to stationarity.

Theorem 2.19. *Suppose the transition kernel P has a drift function with respect to a single-element set $C = \{c\}$. Let λ be the drift parameter. If P is reversible with nonnegative eigenvalues, then:*

$$\begin{aligned} &\|P^t(c,\cdot) - \pi\|_{\mathrm{TV}} \le \lambda^{t+1}, \\ &\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le 2V(x)\lambda^{t+1} \textit{ for all } x \in \mathcal{X}. \end{aligned}$$

This theorem will be used in Chapter 7 to provide a sharp upper bound on the cutoff window for lazy irreducible birth and death chains.

2.4 Norms of convergence

Convergence results based on drift and minorization are often stated with respect to a stronger norm than total variation, the so-called *V-norm.*

Definition 2.20. Let $(\mathcal{X}, \mathcal{E})$ be a measurable space. Fix a measurable function $F : \mathcal{X} \to (0, \infty)$. The *$F$-norm* for measurable functions $f : \mathcal{X} \to \mathbf{R}$ is given by

$$\|f\|_F = \sup_{x \in \mathcal{X}} \frac{|f(x)|}{F(x)}.$$

Define the space $L_F^\infty = \{f : \mathcal{X} \to \mathbf{R} : \|f\|_F < \infty\}$.

The F-norm for signed measures η on $\mathcal{X}$ is

$$\|\eta\|_F = \sup_{\substack{f:\mathcal{X}\to\mathbf{R} \\ \|f\|_F \le 1}} \eta(f).$$

If μ, μ' are probability measures on $\mathcal{X}$, the F-distance between μ and μ' is $\|\mu - \mu'\|_F$.

If $F \equiv 1$, the F-norm for functions is the L^∞ norm (setting aside issues of almost-everywhere equivalence). The F-distance between measures is the same as the L^1 distance, which is double the total variation distance.

Suppose the transition kernel P satisfies a drift and minorization condition with drift function V and associated constant λ. Whenever this is used to obtain an explicit bound $\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le B(x)\gamma^t$, one has $\gamma \ge \lambda$, and $B(x)$ can always be chosen so that $B(x) \le AV(x)$ for some constant A not depending on x. This last observation is due to Chan [Cha89].

Lemma 4.18, to be proved in Section 4.5, strengthens bounds of the form

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le AV(x)\gamma^t,$$

where $\gamma > \lambda$, into bounds of the form

$$\|P^t(x,\cdot) - \pi\|_V \le A'V(x)\gamma^t. \tag{2.16}$$

Here the drift function V plays the role of F in Definition 2.20. (If $\gamma = \lambda$ the right side of (2.16) has an extra factor of t, but this is irrelevant to the present discussion.)

The bound (2.16) can be rewritten in terms of the L_V^∞ operator norm (or V-operator norm), defined for linear operators $Q : L_V^\infty \to L_V^\infty$ by

$$|||Q|||_V = \sup_{\substack{f:\mathcal{X}\to\mathbf{R} \\ \|f\|_V = 1}} \|Qf\|_V.$$

Define the operator Π on $L^1(\pi)$ by $(\Pi f)(x) = \pi(f)$, so that every function f is mapped by Π to a constant function. Since $\pi(V) < \infty$ by (2.10), one has $L_V^\infty \subseteq L^1(\pi)$. With these definitions, (2.16) is equivalent to $|||P^t - \Pi|||_V \le A'\gamma^t$. If this is true, noting that $P^t - \Pi = (P - \Pi)^t$, the L_V^∞-spectral radius ρ_V of $P - \Pi$ is at most γ. Indeed,

$$\rho_V = \inf\{\gamma \ge 0 \,:\, \text{there exists } A' \text{ such that (2.16) holds}\}.$$

Theorem 2.21. *Suppose P is a Markov transition kernel on $(\mathcal{X}, \mathcal{E})$, where $\mathcal{E}$ is countably generated. Then P is geometrically ergodic with stationary distribution π if and only if there is a measurable* *function $V \geq 1$ with $\pi(V) < \infty$ such that the operator $P - \Pi$ has L^∞_V-spectral radius $\rho_V < 1$. If ρ_{TV} is defined as in (2.14), then $\rho_{\mathrm{TV}} \leq \rho_V$.*

The "only if" direction follows from Theorem 2.7 combined with the discussion in the previous paragraphs. The "if" direction is easier: for any $\gamma > \rho_V$ there is A such that $\left\|\left|P^t - \Pi\right|\right\|_V \leq A\gamma^t$. Since $V \geq 1$,

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \frac{1}{2}\|P^t(x,\cdot) - \pi\|_V \leq \frac{1}{2}AV(x)\gamma^t.$$

This also proves that $\rho_{\mathrm{TV}} \leq \rho_V$.

The interpretation of drift and minorization bounds for general state space Markov chains in terms of the V-operator norm is due to Meyn and Tweedie [MT92; MT93]. (Related results for countable state space chains were proved in [Spi90; HS92].) Theorem 2.21, or at least an "almost everywhere" version, is proved in [RR97]. In [KM12] (see also [KM03]) it is shown further that $\rho_V < 1$ if and only if P is aperiodic and has an L^∞_V-spectral gap.

The drift condition $PV(x) \leq \lambda V(x)$ for $x \notin C$ can also be written in terms of the V-operator norm as $\left\|\left|P\mathbf{1}_{\mathcal{X}\setminus C}\right|\right\|_V \leq \lambda$, where $(\mathbf{1}_{\mathcal{X}\setminus C}f)(x) = \mathbf{1}\{x \in \mathcal{X} \setminus C\}f(x)$. This indicates that it may be possible to give operator-theoretic proofs of convergence results such as Theorem 2.5; but see the remark at the end of this section.

If P is reversible with respect to π, one can also consider the $L^2(\pi)$ spectrum of P, which is a subset of the interval $[-1, 1]$. Let ρ_2 be the $L^2(\pi)$-spectral radius of $P - \Pi$. For any $\mu \in \mathcal{P}(\mathcal{X})$ with $\|\mu\|_{L^2(\pi)} < \infty$,

$$\|P^t(\mu,\cdot) - \pi\|_{L^2(\pi)} \leq \|\mu - \pi\|_{L^2(\pi)}\rho_2^t.$$

Theorem 2.22. *Suppose P is a Markov transition kernel on $(\mathcal{X}, \mathcal{E})$, where $\mathcal{E}$ is countably generated. If P is reversible with respect to the stationary distribution π, then $P - \Pi$ has $L^2(\pi)$-spectral radius $\rho_2 < 1$ if and only if P is π-almost everywhere geometrically ergodic, that is, if*

$$\tilde{\rho}_{\mathrm{TV}} = \inf\{\gamma \geq 0 \,:\, \textit{for some } \pi\textit{-a.e. finite } B(x),\ \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t \textit{ for all } t \geq 0\} < 1.$$

In addition, $\rho_2 \leq \tilde{\rho}_{\mathrm{TV}}$.

The "if" direction, along with the inequality $\rho_2 \leq \tilde{\rho}_{\mathrm{TV}}$, is due to [RR97]. The "only if" direction is due to [RT01b].

Combining Theorem 2.22 with the "almost everywhere" version of Theorem 2.21 gives the following result. If P is reversible with respect to π and $\mathcal{E}$ is countably generated, then $\rho_2 < 1$ if and only if there is a measurable function $V \geq 1$, finite π-almost everywhere and with $\pi(V) < \infty$, such that $\rho_V < 1$. In that case, $\rho_2 \leq \rho_V$.

If P is not reversible, one can still consider the $L^2(\pi)$ spectrum and the radius ρ_2, but it is not necessarily true that $\rho_2 \leq \rho_V$. In fact, [KM12] provides an example where $\rho_2 = 1$ but $\rho_V < 1$.

A curious feature of these operator-norm results is that for the most part, they are not proved using operator theory. Rather, the statements about spectra are translated into statements about Markov chains with drift and minorization. The main arguments are made at the level of the Markov chains. At the end, the results are translated back into the language of operators. It seems in principle that one should be able to work directly with the operator P on the spaces $L^p(\pi)$ and L_V^∞, but so far this approach has not been successful.

Chapter 3

Regeneration times

3.1 Introduction and history

Let (X_t) be a Markov chain. Loosely, a regeneration time for (X_t) is a randomized stopping time T such that X_T is distributed according to a particular measure ν, called the regeneration measure.

The traditional definition of a regeneration time (as in [Num84; MTY95]) has an extra requirement: given that $T = k$, the value X_k (and therefore also the future sample path $(X_{k+1}, X_{k+2}, \ldots)$) must be independent of the history $(X_0, \ldots, X_{k-1})$. If this condition is met, the Markov chain at and after time T has absolutely no dependence on what came before, which is why the chain is said to regenerate.

The full force of this extra requirement is not necessary for the main convergence results in this thesis. Here are two alternative notions, called strong and weak ν times.

Definition 3.1. Let (X_t) be a Markov chain on the state space $(\mathcal{X}, \mathcal{E})$ with initial distribution μ, and let T be a randomized stopping time for (X_t). Fix a probability measure ν on $\mathcal{X}$. T is called a *weak ν time* for (X_t) if for all $A \in \mathcal{E}$,

$$\mathbf{P}_\mu(X_T \in A) = \nu(A). \tag{3.1}$$

T is called a *strong ν time* for (X_t) if for all $A \in \mathcal{E}$ and all $k \geq 0$,

$$\mathbf{P}_\mu(X_k \in A \mid T = k) = \nu(A). \tag{3.2}$$

T is called a *ν-regeneration time* for (X_t) if for all $A \in \mathcal{E}$ and all $k \geq 0$,

$$\mathbf{P}_\mu(X_k \in A \mid T = k, X_0, \dots, X_{k-1}) = \nu(A). \tag{3.3}$$

If (X_t) is a Markov scheme, a weak ν time for (X_t) (respectively, strong ν time, ν-regeneration time) is a randomized stopping time for (X_t) that satisfies (3.1) (respectively, (3.2), (3.3)) for all $\mu \in \mathcal{P}(\mathcal{X})$.

Any regeneration time is a strong ν time, and any strong ν time is a weak ν time. The name "strong ν time" was chosen to follow the very similar notion of a *strong stationary time*, which is a randomized stopping time T satisfying (3.2) for a measure $\nu = \pi$ that is a stationary distribution for the Markov chain. Likewise, a *stationary time* is a randomized stopping time T satisfying (3.1) for $\nu = \pi$.

To illustrate the differences between these definitions, consider the simple random walk (X_t) on the directed graph in Figure 3.1. Define stopping times and probability measures:

$$\begin{aligned}
T^{(1)} &= \min\{t \geq \tau_a : X_t \in \{b, e\}\} &\qquad \nu_1(b) &= \nu_1(e) = \frac{1}{2},\ \nu_1(\{a, c, d\}) = 0 \\
T^{(2)} &= \min\{t \geq \tau_a : X_t \in \{d, e\}\} &\qquad \nu_2(d) &= \nu_2(e) = \frac{1}{2},\ \nu_2(\{a, b, c\}) = 0 \\
T^{(3)} &= \min\{t \geq \tau_a : X_t \in \{b, c\}\} &\qquad \nu_3(b) &= \nu_3(c) = \frac{1}{2},\ \nu_3(\{a, d, e\}) = 0
\end{aligned}$$

With these definitions, $T^{(1)}$ is a weak ν_1 time but not a strong ν_1 time; $T^{(2)}$ is a strong ν_2 time but not a ν_2-regeneration time; and $T^{(3)}$ is a ν_3-regeneration time.

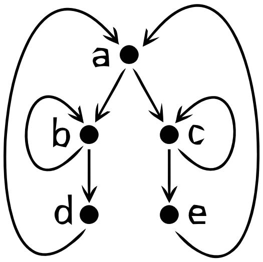

Figure 3.1: The simple random walk on this directed graph is used as an example for constructing different types of regeneration times.

There is some literature on strong stationary times, beginning with papers of Aldous and Diaconis [AD86; AD87] and Diaconis and Fill [DF90]. The connection with the regeneration theory of general state space Markov chains was noted in [AD87]. Aldous, Lovasz, and Winkler [ALW97; LW98] defined many types of randomized stopping times for Markov chains, some of which are classified as weak ν times using this chapter's terminology. The precise notion of strong ν time given above appears very rarely in the literature. The Ph.D. thesis of Pak [Pak97] calls it a time-invariant stopping time and uses it as a tool to construct strong stationary times for random walks on groups. See Section 3.5 of [Pak97] for details. Miclo [Mic10] calls it a strong random time; he constructs a particular sequence of such times in order to give a stochastic interpretation of the eigenvalues for reversible chains with an absorbing state.

Athreya and Ney [AN78] and Nummelin [Num78] used regeneration times to prove convergence results about so-called Harris recurrent Markov chains.

Definition 3.2. Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$. (X_t) is called *Harris recurrent* if there is a σ-finite positive measure φ on $\mathcal{X}$, called the *irreducibility measure*, such that for any $x \in \mathcal{X}$ and any $A \in \mathcal{E}$ with $\varphi(A) > 0$, $\mathbf{P}_x(\tau_A < \infty) = 1$.

The Harris condition is a natural extension of the usual definition of recurrence for countable state space Markov chains to the context of general state space. One of the earliest structural results about Harris recurrent chains, due to Jain and Jamison [JJ67], is the existence of small sets.

Theorem 3.3. *Let (X_t) be a Markov scheme with transition kernel P on $(\mathcal{X}, \mathcal{E})$, where $\mathcal{E}$ is countably generated. If (X_t) is Harris recurrent, then there exists a subset $C \in \mathcal{E}$ that is a small set for P.*

A proof of Theorem 3.3 is given in Chapter 5 of [MT93]. Note that the number of minorization steps, that is, the value of m in the minorization $P^m(x, \cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$, may be arbitrarily large.

The contribution of [AN78] and [Num78] was a construction of a regeneration time for any Markov chain having a small set with 1-step minorization. The version presented here is from [Num78] (see also [Num84]). It is often called the "split chain" or "Nummelin splitting."

Let (X_t) be a Markov chain on $(\mathcal{X}, \mathcal{E})$ with transition kernel P, and suppose $C \in \mathcal{E}$ is a small set for P with 1-step minorization. The Nummelin splitting defines random variables $Y_t \in \{0, 1\}$ such that if $X_t \in C$, then with probability ε, $X_{t+1} \sim \nu$ and $Y_{t+1} = 1$, while with probability $1 - \varepsilon$, $X_{t+1} \sim \frac{1}{1-\varepsilon}[P(X_t, \cdot) - \varepsilon\nu(\cdot)]$ and $Y_{t+1} = 0$. If $X_t \notin C$, then $Y_{t+1} = 0$ with probability 1. Thus, the chain (X_t) regenerates whenever $Y_t = 1$.

In fact, (X_t, Y_t) can be defined as a Markov chain on the state space $\mathcal{X} \times \{0, 1\}$. Its transition kernel $\tilde{P}$ is given as follows. If $x \in \mathcal{X} \backslash C$ and $A \in \mathcal{E}$, let $\tilde{P}((x, y), A \times \{0\}) = P(x, A)$ and $\tilde{P}((x, y), A \times \{1\}) = 0$. If $x \in C$, let $\tilde{P}((x, y), A \times \{0\}) = P(x, A) - \varepsilon\nu(A)$ and $\tilde{P}((x, y), A \times \{1\}) = \varepsilon\nu(A)$.

Let $\mathbf{T} = \{t \geq 0 : Y_t = 1\}$, and label its elements in increasing order as $T = T_1 < T_2 < T_3 < \cdots$. Each T_j is a ν-regeneration time for (X_t), and the tours $\mathbf{W}_j = (X_{T_j}, X_{T_j+1}, \ldots, X_{T_{j+1}-1})$ are independent and identically distributed.

One of the principal aims of this chapter is to extend the Nummelin splitting construction to the case $m > 1$. The goal is to construct random variables $Y_t \in \{0, 1\}$ such that if $X_t \in C$ and t is a "valid arrival time" (to be defined in the next paragraph), then with probability ε, $X_{t+m} \sim \nu$ and $Y_{t+m} = 1$, while with probability $1 - \varepsilon$, $X_{t+m} \sim \frac{1}{1-\varepsilon}[P^m(X_t, \cdot) - \varepsilon\nu(\cdot)]$ and $Y_{t+m} = 0$. If $X_t \notin C$, or $X_t \in C$ but t is not a valid arrival time, then $Y_{t+m} = 0$ with probability 1. This construction will be performed in Section 3.3.

For fixed $m \geq 1$ and $C \in \mathcal{E}$, the set $\mathbf{S} = \{S_1, S_2, \ldots\}$ of valid arrival times to C is defined inductively as follows. The first one is $S_1 = \tau_C$, and for $k \geq 1$, $S_{k+1} = \min\{t \geq S_k + m : X_t \in C\}$. The idea is that at each time S_k, an $\varepsilon, 1 - \varepsilon$ coin is flipped that determines the value of Y_{S_k+m} and affects the trajectory $(X_{S_k+1}, \ldots, X_{S_k+m})$. Any time $S_k < t < S_k + m$ for which $X_t \in C$ is ignored.

When $m > 1$, (X_t, Y_t) may no longer be a Markov chain. The times T_j are not necessarily ν-regeneration times, and the tours $\mathbf{W}_j$ are not necessarily independent. The issue is this. Suppose $Y_t = 1$, so that $X_{t-m} \in C$ and $X_t \sim \nu$. Conditioned on $Y_t = 1$, X_t is independent of X_{t-m} and of X_s for all $s < t - m$, but X_t may still depend on $(X_{t-m+1}, \ldots, X_{t-1})$. What can be said is that the T_j are strong ν times, and the tours $\mathbf{W}_j$ are 1-dependent: $(\mathbf{W}_1, \ldots, \mathbf{W}_{j-1})$ is independent of $(\mathbf{W}_{j+1}, \mathbf{W}_{j+2}, \ldots)$. For more information on 1-dependent processes and pointers to the literature, see [BDF10] and [Val94].

For a concrete example, let (X_t) be the simple random walk on the directed graph in Figure 3.1. Let $C = \{a\}$ be the small set, $m = 2$ be the number of minorization steps, $\nu = \nu_2$ be the minorization measure, and $\varepsilon = 1/2$ be the minorization mass. The strong ν times $T_1, T_2, \ldots$ are defined as follows:

$$T_j = \min\{t \geq 0 : t - 2 \geq T_{j-1}, X_{t-2} = a, X_t \in \{d, e\}\},$$

where T_0 is taken to be 0. The T_j are strong ν times but not ν-regeneration times, and the tours $\mathbf{W}_j$ are 1-dependent but not independent. For instance, if the tour $\mathbf{W}_{j-1}$ ends with $X_{T_j - 1} = b$, then the tour $\mathbf{W}_j$ begins with $X_{T_j} = d$. This example is very similar to one given in [BLL08].

The rest of this chapter is organized as follows. Section 3.2 begins with the construction of the sequence $\mathbf{T} = \{T_1, T_2, \ldots\}$ in the general setting of a Markov chain with a weak ν time (or strong ν time, or ν-regeneration time) T. This is followed by a discussion of the differences between the variants of regeneration times, as reflected in the level of independence of the tours $\mathbf{W}_j$ between consecutive regenerations. Finally, it is proved that a Markov chain having a weak ν time with finite expectation has unique stationary distribution given by the normalized occupation measure

$$\pi(A) = \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(X_n \in A, T > n).$$

Section 3.3 generalizes the Nummelin splitting construction to the case $m > 1$.

The proofs of the results in this chapter are quite technical, so they are given in the Appendix.

3.2 Properties of chains with regeneration

Suppose (X_t) is a Markov scheme, and T is a weak ν time for (X_t) (which may also be a strong ν time or ν-regeneration time). Assume that the following condition is satisfied.

Condition 3.4. The weak ν time (or strong ν time, or ν-regeneration time) T satisfies $\mathbf{P}_\mu(T < \infty) = 1$ for all $\mu \in \mathcal{P}(\mathcal{X})$, and $\mathbf{P}_\nu(T > 0) = 1$.

Here is the intuitive idea of the construction of the set $\mathbf{T} = \{T_1, T_2, \ldots\}$. Start with $T_1 = T$. Inductively, suppose that T_j has been defined, and view the sample path $(X_{T_j}, X_{T_j+1}, \ldots)$ as an instance of the Markov chain started from initial measure ν. T_{j+1} will be defined so that the law of $T_{j+1} - T_j$ is the same as the law of T given the sample path $(X_{T_j}, X_{T_j+1}, \ldots)$.

To make the construction rigorous, for each $\ell \geq 0$, define functions f_ℓ and r_ℓ on $\mathcal{X}^{\ell+1}$ to satisfy

$$\mathbf{P}_\nu(T = \ell \mid X_0, \ldots, X_\ell) = f_\ell(X_0, \ldots, X_\ell),$$
$$\mathbf{P}_\nu(T = \ell \mid T \geq \ell, X_0, \ldots, X_\ell) = r_\ell(X_0, \ldots, X_\ell).$$

(The notation is taken from Section 3.3 of [Num84].) One has

$$f_\ell(x_0, \ldots, x_\ell) = \left(\prod_{i=0}^{\ell-1} \Big[1 - r_i(x_0, \ldots, x_i) \Big] \right) r_\ell(x_0, \ldots, x_\ell). \tag{3.4}$$

It will be possible to construct a sequence (T_j) that meets the following requirements.

Requirements 3.5. Each T_j is a randomized stopping time for (X_t), and $T_1 = T$. For any $\mu \in \mathcal{P}(\mathcal{X})$ and any k such that $\mathbf{P}_\mu(T_j = k) > 0$, the value of $T_{j+1} - T_j$ is conditionally independent of $(X_0, \ldots, X_{k-1})$ under $\mathbf{P}_\mu$ given that $T_j = k$ and given the future path $(X_k, X_{k+1}, \ldots)$. Its law is

$$\mathbf{P}_\mu(T_{j+1} - T_j = \ell \mid T_j = k, X_k, X_{k+1}, \ldots) = f_\ell(X_k, X_{k+1}, \ldots, X_{k+\ell}).$$

Recall from Definition 1.5 that an extension of a Markov scheme is an enlargement of the sample space to encompass extra randomness.

Proposition 3.6. *Suppose the Markov scheme (X_t) has a weak ν time T satisfying Condition 3.4. Then there is an extension of (X_t) with a sequence of times (T_j) satisfying Requirements 3.5.*

The next proposition is the main result in this section. If the sequence (T_j) is constructed from T, the extent to which the tours are independent of each other depends on whether T is a weak ν time, strong ν time, or ν-regeneration time.

Proposition 3.7. *Suppose the Markov scheme (X_t) on $(\mathcal{X}, \mathcal{E})$ has a weak ν time T satisfying Condition 3.4, and a sequence of randomized stopping times (T_j) satisfying Requirements 3.5. Define the tours $\mathbf{W}_1, \mathbf{W}_2, \ldots$ by $\mathbf{W}_j = (X_{T_j}, \ldots, X_{T_{j+1}-1})$. For each initial distribution μ:*

(i) Each T_j is a weak ν time for the chain (X_t) started from μ. The sequence $(\mathbf{W}_j)$ is stationary and 1-dependent: each $(\mathbf{W}_j, \mathbf{W}_{j+1}, \ldots)$ has the same distribution as $(\mathbf{W}_1, \mathbf{W}_2, \ldots)$, and for all j, $(\mathbf{W}_1, \ldots, \mathbf{W}_{j-1})$ is independent of $(\mathbf{W}_{j+1}, \mathbf{W}_{j+2}, \ldots)$. In addition, the law of $(\mathbf{W}_j)$ has no dependence on μ.

(ii) If T is a strong ν time, then each T_j is a strong ν time for the chain (X_t) started from μ, and the sequence of tour lengths $(T_{j+1} - T_j)$ is independent.

(iii) If T is a ν-regeneration time, then each T_j is a ν-regeneration time for the chain (X_t) started from μ, and the sequence $(\mathbf{W}_j)$ is independent.

To understand Proposition 3.7, suppose T is a weak ν time and $T_j = k$. By the construction, $X_{T_{j+1}}$ is conditionally distributed as ν. If the value of X_k is revealed, $X_{T_{j+1}}$ is still conditionally distributed as ν. Therefore, the conditional distribution of $(X_{T_{j+1}}, X_{T_{j+1}+1}, \ldots)$ given the value of T_j and the sequence $(X_0, \ldots, X_{T_j})$ is the same as the unconditional distribution of $(X_{T_{j+1}}, X_{T_{j+1}+1}, \ldots)$. This implies that the sequence $(\mathbf{W}_j)$ is 1-dependent.

Now, suppose the tour length $T_{j+1} - T_j$ is revealed. In general, $X_{T_{j+1}}$ may no longer be conditionally distributed as ν. But if T is a strong ν time, $X_{T_{j+1}} \sim \nu$ still. Therefore, the behavior of the chain from time T_{j+1} onward is independent of $T_{j+1} - T_j$, meaning that the sequence $(T_{j+1} - T_j)$ is independent.

Finally, suppose the sample path $(X_{T_j+1}, \ldots, X_{T_{j+1}-1})$ is revealed. Again, the conditional distribution of $X_{T_{j+1}}$ may not be ν in general, but it is ν if T is a ν-regeneration time. In that case, the conditional distribution of $(X_{T_{j+1}}, X_{T_{j+1}+1}, \ldots)$ given the values of T_j, T_{j+1} and the sequence $(X_0, \ldots, X_{T_{j+1}-1})$ is the same as the unconditional distribution of $(X_{T_{j+1}}, X_{T_{j+1}+1}, \ldots)$. Hence the sequence $(\mathbf{W}_j)$ is independent.

Even when T is a ν-regeneration time, the independence of the $\mathbf{W}_j$ is still somewhat delicate. It holds when $\mathbf{W}_j$ is defined to be $(X_{T_j}, \ldots, X_{T_{j+1}-1})$. Under the alternative definition $\tilde{\mathbf{W}}_j = (X_{T_j+1}, \ldots, X_{T_{j+1}})$, the $\tilde{\mathbf{W}}_j$ would not necessarily be independent.

As an example, consider the simple random walk (X_t) on the directed graph in Figure 3.1. Use the weak ν_1 time $T^{(1)}$ to generate the sequence $T_j^{(1)}$ defined by $T_1^{(1)} = T^{(1)}$ and

$$T_{j+1}^{(1)} = \min\{t \geq \tau_a(j) : X_t \in \{b, e\}\} \qquad \text{for all } j \geq 1,$$

where $\tau_a(j) = \min\{t \geq T_j^{(1)} : X_t = a\}$. This is the sequence of weak ν_1 times given by Proposition 3.6. It can be observed directly that the sequence of tours $(\mathbf{W}_j^{(1)})$ as in Proposition 3.7 is stationary and 1-dependent, but neither $(\mathbf{W}_j^{(1)})$ nor $(T_{j+1}^{(1)} - T_j^{(1)})$ is an independent sequence.

If instead the strong ν_2 time $T^{(2)}$ is used to generate the times $T_j^{(2)}$ and tours $\mathbf{W}_j^{(2)}$, then the sequence $(T_{j+1}^{(2)} - T_j^{(2)})$ is iid while the sequence $(\mathbf{W}_j^{(2)})$ is stationary and 1-dependent but not independent. If the ν_3-regeneration time $T^{(3)}$ is used to generate the times $T_j^{(3)}$ and tours $\mathbf{W}_j^{(3)}$, then both sequences $(T_{j+1}^{(3)} - T_j^{(3)})$ and $(\mathbf{W}_j^{(3)})$ are iid. Note that the tours $\tilde{\mathbf{W}}_j^{(3)} = (X_{T_j^{(3)}+1}, \ldots, X_{T_{j+1}^{(3)}})$ are not independent.

The last topic in this section is the existence and uniqueness of the stationary distribution, which holds whenever there is an almost surely finite weak ν time T such that $\mathbf{E}_\nu[T] < \infty$.

Condition 3.8. The weak ν time (or strong ν time, or ν-regeneration time) T satisfies $\mathbf{P}_\mu(T < \infty) = 1$ for all $\mu \in \mathcal{P}(\mathcal{X})$, $\mathbf{P}_\nu(T > 0) = 1$, and $\mathbf{E}_\nu[T] < \infty$.

Proposition 3.9. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Suppose (X_t) has a weak ν time T satisfying Condition 3.8. Then (X_t) has a unique stationary distribution π given by*

$$\pi(A) = \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(X_n \in A, T > n). \tag{3.5}$$

In the special case where $\nu = \delta_c$ is concentrated at a single state and $T = \tau_c^+$, Proposition 3.9 reduces to the well-known formula

$$\pi(A) = \frac{1}{\mathbf{E}_c[\tau_c^+]} \sum_{n=0}^{\infty} \mathbf{P}_c(X_n \in A, \tau_c^+ > n) \tag{3.6}$$

previously mentioned in equation (2.6). Plugging in $A = \{c\}$ yields the celebrated identity of Kac [Kac47] that $\pi(c) = 1/\mathbf{E}_c[\tau_c^+]$. Sometimes (3.6) is also attributed to Kac as a natural consequence of his original result. Proposition 3.9 extends (3.6) to the more general setting of weak ν times.

3.3 Strong ν times from small sets

This section extends the Nummelin splitting construction to the setting where (X_t) is a Markov scheme whose transition kernel P has a small set with m-step minorization. It will follow that a randomized stopping time T, which is a strong ν time when $m > 1$ and a ν-regeneration time when $m = 1$, can be defined on an extension of (X_t). By Propositions 3.6 and 3.7, one obtains a full sequence of strong ν times $T = T_1 < T_2 < T_3 < \cdots$. As seen in Section 3.1, when $m > 1$ the strong ν time may not be a ν-regeneration time, and the tours $\mathbf{W}_j$ between successive strong ν times may not be independent.

The strong ν time construction involves a sequence of $\varepsilon, 1-\varepsilon$ coins being flipped at specific times S_k, where $S_1 = \tau_C$ and $S_{k+1} = \min\{t \geq S_k + m : X_t \in C\}$. For a given sample path $(X_0, X_1, \ldots)$, define the sequence (S_k) by this rule, and let $\mathbf{S} = \{S_1, S_2, \ldots\}$. By construction, consecutive elements of $\mathbf{S}$ differ by at least m.

The goal is to introduce a sequence (Y_t) of $\{0,1\}$-valued random variables so that $Y_t = 1$ if and only if both $t - m \in \mathbf{S}$ and the coin that was flipped at time $t - m$ showed ε. That way, the strong ν time T can be defined as the first time t that $Y_t = 1$, and the sequence $\mathbf{T} = \{T = T_1, T_2, \ldots\}$ of successive regenerations can be defined as $\mathbf{T} = \{t \geq 0 : Y_t = 1\}$.

Proposition 3.10. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Suppose C is a small set for P. That is, there are a probability measure ν, an integer $m \geq 1$, and a constant $\varepsilon > 0$ such that*

$$P^m(x, \cdot) \geq \varepsilon \nu(\cdot) \quad \textit{for all } x \in C.$$

Also assume that for all $x \in \mathcal{X}$, $\mathbf{P}_x(\tau_C < \infty) = 1$. Then a sequence $(Y_0, Y_1, \ldots)$ of $\{0,1\}$-valued random variables can be defined on an extension of (X_t), such that if the elements of the set $\mathbf{T} = \{t \geq 0 : Y_t = 1\}$ are listed in increasing order as $T = T_1 < T_2 < T_3 < \cdots$, the following properties are satisfied:

1. *T is a strong ν time for (X_t) satisfying Condition 3.4. If $m = 1$, T is a ν-regeneration time for (X_t).*
2. *The sequence (T_j) satisfies Requirements 3.5, meaning that Proposition 3.7 applies. In particular, each T_j is a strong ν time for (X_t). Under every $\mathbf{P}_\mu$, the tour lengths $T_{j+1} - T_j$ are independent and the tours $\mathbf{W}_j = (X_{T_j}, \ldots, X_{T_{j+1}-1})$ are 1-dependent. If $m = 1$, each T_j is a ν-regeneration time for (X_t), and the tours $\mathbf{W}_j$ are independent under every $\mathbf{P}_\mu$.*
3. *If $\varepsilon = 1$ then $T = \tau_C + m$. If $\varepsilon < 1$, suppose $\mu \in \mathcal{P}(\mathcal{X})$ is fixed and $\mathbf{P}_\mu(\tau_C = s) > 0$. Then*

$$\mathbf{P}_\mu(T < s + m \mid \tau_C = s) = 0,$$

and for all $t \geq s + m$,

$$\mathbf{P}_\mu(T > t \mid \tau_C = s) = (1 - \varepsilon)\, \mathbf{P}_{\mu_s}(T > t - s - m),$$

where μ_s is the remainder measure

$$\mu_s(A) = \frac{1}{1-\varepsilon} \left[\mathbf{P}_\mu(X_{s+m} \in A \mid \tau_C = s) - \varepsilon \nu(A) \right].$$

Properties 1 and 3 encode all the desired behavior of T. If $\tau_C = s$, then $T \geq s + m$ with probability 1, and $T = s + m$ with probability ε. In case $T = s + m$, then $X_{s+m} \sim \nu$. In case $T > s + m$, then $X_{s+m} \sim \mu_s$, and the law of $T - (s+m)$ in this situation is the same as the law of T for the chain started from μ_s. This is a recursive way to write the rule "keep returning to C and flipping independent $\varepsilon, 1-\varepsilon$ coins until one of them shows ε."

Chapter 4

Convergence results

4.1 Summary

This chapter states and proves convergence results of the following form: if a reversible Markov chain with nonnegative eigenvalues satisfies a drift and minorization condition, its distance from stationarity after t steps is bounded above by an explicit function that decays exponentially in t. These are the main new theoretical results in this thesis.

Here are three versions of the drift and minorization condition. The first is a repeat of Definition 2.3.

Definition 4.1. The transition kernel P on the state space $(\mathcal{X}, \mathcal{E})$ satisfies a *general drift and minorization condition* if there are a subset $C \in \mathcal{E}$, a measurable function $V : \mathcal{X} \to [1, \infty)$, and constants $\lambda < 1$ and K such that for all $x \in \mathcal{X}$,

$$PV(x) \leq \begin{cases} \lambda V(x) & \text{if } x \notin C, \\ K & \text{if } x \in C, \end{cases}$$

and if there are a positive integer m, a constant $\varepsilon > 0$, and a probability measure ν on $\mathcal{X}$ such that

$$P^m(x, \cdot) \geq \varepsilon \nu(\cdot) \qquad \text{for all } x \in C.$$

Definition 4.2. The transition kernel P on $(\mathcal{X}, \mathcal{E})$ satisfies a *uniform drift and minorization condition* if there are a subset $C \in \mathcal{E}$, a measurable function $V : \mathcal{X} \to [1, \infty)$, and constants $\lambda < 1$ and M such that $V(x) \leq M$ for all $x \in \mathcal{X}$ and $PV(x) \leq \lambda V(x)$ for all $x \notin C$, and if there are a positive integer m, a constant $\varepsilon > 0$, and a probability measure ν on $\mathcal{X}$ such that

$$P^m(x, \cdot) \geq \varepsilon\nu(\cdot) \qquad \text{for all } x \in C.$$

Definition 4.3. The transition kernel P on $(\mathcal{X}, \mathcal{E})$ satisfies a *single element drift condition* if there are an element $c \in \mathcal{X}$, a measurable function $V : \mathcal{X} \to [1, \infty)$, and constants $\lambda < 1$ and K such that $PV(x) \leq \lambda V(x)$ for all $x \in \mathcal{X} \setminus \{c\}$ and $PV(c) \leq K$.

Clearly Definitions 4.2 and 4.3 are stronger than Definition 4.1.

The next two theorems collect the various bounds that will be proved in this chapter. As discussed in Section 2.3, they improve upon the similar results of Baxendale [Bax05] in two ways. First, the bounds below hold for all values of m, while the statements in [Bax05] are proved only for $m = 1$. Second, when $m = 1$ the new bounds are numerically smaller than those of [Bax05] in examples. The present results are proved using probabilistic arguments, while the proofs of [Bax05] are analytic.

Theorem 4.4. *Let P be a Markov transition kernel on the state space $(\mathcal{X}, \mathcal{E})$. Assume that P is reversible with respect to a stationary measure π and that P has nonnegative eigenvalues. Suppose P satisfies a general or uniform drift and minorization condition. Then π is the unique stationary distribution for P, and:*

$$\begin{aligned}
\|P^t(\nu, \cdot) - \pi\|_{L^2(\pi)} &\leq B_1 \lambda_*^t && \textit{for all } t \geq 0 \\
\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} &\leq B_2(x,t) \lambda_*^t && \textit{for all } t \geq 0, x \in \mathcal{X} \\
\|P^t(\nu, \cdot) - \pi\|_{V} &\leq B_3 \lambda_*^t + B_4(t)\lambda^t && \textit{for all } t \geq 0 \\
\|P^t(x, \cdot) - \pi\|_{V} &\leq B_5(x,t) \lambda_*^t + B_6(x,t)\lambda^t && \textit{for all } t \geq 0, x \in \mathcal{X}
\end{aligned}$$

where the formulas for λ_* and the B_i are given at the end of this section. The dependence on t of B_2, B_4, B_5, B_6 is either linear or quadratic. For example, $B_6(x,t) = \blacksquare + \blacksquare t + \blacksquare t^2$, where the $\blacksquare$ terms depend only on x and the drift-minorization data. The dependence on x of B_2, B_5, B_6 is through the value of $V(x)$ and at most linear in that value. For example, $B_6(x,t) \leq \blacksquare V(x)$, where the $\blacksquare$ depends only on the drift-minorization data and quadratically on t.

Theorem 4.4 will be used to find upper bounds for the time to stationarity of two Gibbs samplers in Chapter 5. A direct comparison with the results of [Ros95a] and [Bax05] shows that the bounds from Theorem 4.4 are significantly better but still far from optimal.

Recall that the value of $V(x)$ controls how long it will take for the Markov chain started at x to reach C, by the inequality $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x)$. (This was stated as Lemma 2.6 and will be proved in Section 4.2 as Lemma 4.7.) Therefore, it is no surprise that B_2, B_5, B_6 depend on x through $V(x)$. A consequence is that if P satisfies a uniform drift and minorization condition, then B_2, B_5, B_6 are uniformly bounded in x, which is a property called *uniform ergodicity*.

If P satisfies a single element drift condition, the following bound holds.

Theorem 4.5. *Let P be a Markov transition kernel on the state space $(\mathcal{X}, \mathcal{E})$. Assume that P is reversible with respect to a stationary measure π and that P has nonnegative eigenvalues. Suppose P satisfies a single element drift condition with respect to an element $c \in \mathcal{X}$. Then π is the unique stationary distribution for P, and:*

$$\begin{aligned} \|P^t(c,\cdot) - \pi\|_{\mathrm{TV}} &\le \lambda^{t+1} && \textit{for all } t \ge 0 \\ \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} &\le 2V(x)\lambda^{t+1} && \textit{for all } t \ge 0, x \in \mathcal{X} \\ \|P^t(x,\cdot) - \pi\|_V &\le \left[4KV(x)t + V(x) + \frac{K-\lambda}{1-\lambda}\right]\lambda^t && \textit{for all } t \ge 0, x \in \mathcal{X}. \end{aligned}$$

Theorem 4.5 will be used to find a sharp upper bound on the cutoff window for birth and death chains in Section 7.1.

The proofs of Theorem 4.4 and Theorem 4.5 are quite different. Theorem 4.4 will be proved using the strong ν time construction from Chapter 3. Suppose that T is a strong ν time for a Markov chain (X_t) with transition kernel P started from $X_0 \sim \nu$. If $f \ge 0$ is a measurable function on $\mathcal{X}$, one has the recurrence

$$\mathbf{E}_\nu[f(X_n), T \le n] = \sum_{j=0}^{n} \mathbf{P}_\nu(T = n-j)\,\mathbf{E}_\nu[f(X_j)],$$

which was stated as (2.15). Once the appropriate function f is chosen, Theorem 4.4 will follow from a summation by parts.

Theorem 4.5 will be proved by coupling. Let (X_t) be a Markov chain with transition kernel P started from $X_0 = c$. Under the hypotheses, Lund, Zhao, and Kiessler [LZK06] showed that the sequence of hazard rates

$$h_n = \mathbf{P}_c(\tau_c^+ = n \mid \tau_c^+ \ge n)$$

is decreasing. Define the *age process* (A_t) by

$$A_t = t - \max\{s \ge 0 \,:\, X_s = c\}.$$

In other words, A_t is the amount of time since the last visit to c. The property of decreasing hazard rates means that the age process is stochastically monotone. The monotone coupling provides a bound on the convergence of the age process, which translates into a bound on the convergence of (X_t).

Perhaps the most interesting aspect of the proof is the link between reversibility (with nonnegative eigenvalues) and monotonicity. It was discussed in Section 2.3 that the assumptions of reversibility and monotonicity are somewhat interchangeable in the context of drift-minorization theorems. When (X_t) satisfies a single element drift condition, the age process provides a concrete connection.

Returning to the context of Theorem 4.4, suppose that for large m one has the inequality $P^m(x, \cdot) \geq \varepsilon_m \nu_m(\cdot)$ for all $x \in C$, and the minorization masses ε_m approach 1 at an exponential rate: $\varepsilon_m \geq 1 - A\beta^m$ for some $\beta < 1$. Then one can send $m \to \infty$ in Theorem 4.4 to get a bound on the spectral gap of the chain.

Theorem 4.6. *Under the hypotheses of Theorem 4.4, suppose that for sufficiently large m,*

$$P^m(x, \cdot) \geq (1 - A\beta^m)\nu_m(\cdot) \qquad \text{for all } x \in C, \tag{4.1}$$

where A and $\beta < 1$ are fixed constants and the ν_m are probability measures. In the case of general drift and minorization, define

$$\rho = \exp\left(\frac{\log\beta \log\lambda}{\log\beta + \log\lambda}\right).$$

In the case of uniform drift and minorization, define $\rho = \max\{\beta, \lambda\}$. Then the $L^2(\pi)$ spectral gap of P is at least $1 - \rho$.

Theorem 4.6 will be used to find the spectral gap of the lazy simple random walk on the hypercube to within a factor of 2 in Section 7.2. Outside the context of drift and minorization, results that use coupling to get lower bounds on the spectral gap are discussed in [BK00]; an early argument of this kind is in [Hol85]. To prove Theorem 4.6, the heavy lifting is done by the "if" direction of Theorem 2.22, which was shown by [RR97].

A special case of (4.1) is when there is a uniform bound on separation distance

$$d_{\text{sep}}(P^t(x, \cdot), \pi) \leq A\beta^t \qquad \text{for all } x \in C. \tag{4.2}$$

Then one may choose $\varepsilon_m = 1 - A\beta^m$ and $\nu_m = \pi$. In that case the strong ν_m time T is actually a strong stationary time, so Theorem 4.9 (to be proved in Section 4.3) leads directly to an exponential bound on $d_{\text{sep}}(P^t(x, \cdot), \pi)$ for $x \notin C$. The exponential rate is the same as in Theorem 4.6.

The rest of this chapter is devoted to the proof of these theorems. Section 4.2 proves some preliminary lemmas. Sections 4.3–4.5 are devoted to the proof of Theorem 4.4. If P satisfies a drift and minorization condition, the construction in Chapter 3 produces a strong ν time T. Section 4.3 shows how the drift function gives an exponential bound on the law of T. As a consequence, explicit convergence bounds are given for chains satisfying a bivariate drift and minorization condition. Section 4.4 contains the main step to prove Theorem 4.4: if P is reversible with nonnegative eigenvalues, a strong ν time controls its rate of convergence to stationarity. Section 4.5 finishes the proof of Theorem 4.4. Section 4.6 proves Theorem 4.5 using monotonicity of the age process, and Section 4.7 proves Theorem 4.6.

Formulas for Theorem 4.4

In the case of general drift and minorization, define

$$B = \frac{1-\lambda^m}{1-\lambda}(K-\lambda) + \lambda^m$$

and $J = [B-(1-\varepsilon)]/\varepsilon$. If $\varepsilon = 1$, let $\lambda_* = \lambda$. If $\varepsilon < 1$, let $L = (B-\varepsilon)/(1-\varepsilon)$ and

$$\lambda_* = \max\left\{\lambda, \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log L}\right)\right\} < 1. \tag{4.3}$$

In the case of uniform drift and minorization, let $J = L = M$. If $\varepsilon = 1$, let $\lambda_* = \lambda$, and if $\varepsilon < 1$, define λ_* using (4.3).

In all cases, let $r = \log\lambda_*/\log\lambda$, so that $0 \le r \le 1$. Next, define

$$D = \sqrt{\frac{J^r\lambda_*^{2-m}}{1-\lambda_*}}.$$

The formulas for B_1 and B_2 are $B_1 = D$ and

$$B_2(x,t) = \max\left\{\frac{D}{2}, 1\right\}\lambda_*^{-m}[1+(1-\lambda_*)V(x)^r t].$$

The formulas for B_3 through B_6 look different depending on whether $\lambda = \lambda_*$ or $\lambda < \lambda_*$. They are given below for the case of general drift and minorization. In the case of uniform drift and minorization, the exact same formulas can be used, but with M in place of K.

First, formulas for B_3 and B_4. When $\lambda = \lambda_*$,

$$B_3 = 0, \qquad B_4(t) = KD\lambda^{-1}t + J + \frac{K-\lambda}{1-\lambda}.$$

When $\lambda < \lambda_*$,

$$B_3 = \frac{KD}{\lambda_*-\lambda}, \qquad B_4(t) = -\frac{KD}{\lambda_*-\lambda} + J + \frac{K-\lambda}{1-\lambda}.$$

Last, formulas for B_5 and B_6. When $\lambda = \lambda_*$,

$$\begin{aligned} B_5(x,t) &= 0, \\ B_6(x,t) &= 2K \max\left\{\frac{D}{2}, 1\right\} \lambda^{-m} \left[\lambda^{-1} t + \frac{1-\lambda}{\lambda} V(x)^r \frac{t^2 - t}{2}\right] + V(x) + \frac{K-\lambda}{1-\lambda}. \end{aligned}$$

When $\lambda < \lambda_*$,

$$\begin{aligned} B_5(x,t) &= 2K \max\left\{\frac{D}{2}, 1\right\} \lambda_*^{-m} \left(\frac{1}{\lambda_* - \lambda} + (1-\lambda_*) V(x)^r \left[\frac{t}{\lambda_* - \lambda} - \frac{\lambda_*}{(\lambda_* - \lambda)^2}\right]\right), \\ B_6(x,t) &= 2K \max\left\{\frac{D}{2}, 1\right\} \lambda_*^{-m} \left[-\frac{1}{\lambda_* - \lambda} + (1-\lambda_*) V(x)^r \frac{\lambda_*}{(\lambda_* - \lambda)^2}\right] + V(x) + \frac{K-\lambda}{1-\lambda}. \end{aligned}$$

4.2 Preliminary lemmas

This section proves two well-known results that follow immediately from the drift condition. The first was already stated as Lemma 2.6.

Lemma 4.7. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P, and let $C \in \mathcal{E}$ be a subset. Suppose the measurable function $V : \mathcal{X} \to [1, \infty)$ satisfies $PV(x) \le \lambda V(x)$ for $x \notin C$, where $\lambda < 1$ is fixed. Then for all $x \in \mathcal{X}$,*

$$\mathbf{E}_x[\lambda^{-\tau_C}] \le V(x).$$

It follows immediately by integration that for any probability measure μ on $\mathcal{X}$, $\mathbf{E}_\mu[\lambda^{-\tau_C}] \le \mu(V)$.

The second lemma was previously stated as the inequality (2.10).

Lemma 4.8. *Let P be a transition kernel on $(\mathcal{X}, \mathcal{E})$, and suppose the measurable function $V : \mathcal{X} \to [1, \infty)$ satisfies*

$$PV(x) \le \begin{cases} \lambda V(x) & \text{if } x \notin C, \\ K & \text{if } x \in C, \end{cases}$$

for a particular subset $C \in \mathcal{E}$ and constants $\lambda < 1$ and K. If π is a stationary distribution for P,

$$\pi(V) \le \frac{K - \lambda}{1 - \lambda} \pi(C).$$

Proof of Lemma 4.7. The first step is to show by induction on t that for every $t \ge 0$ and $x \notin C$,

$$V(x) \ge \lambda^{-t}\, \mathbf{E}_x[V(X_t), \tau_C > t] + \sum_{s=1}^{t} \lambda^{-s}\, \mathbf{P}_x(\tau_C = s). \tag{4.4}$$

The base case $t = 0$ is trivial. For the inductive step, assume (4.4) holds for t. When $x \notin C$,

$$\begin{aligned}
&\mathbf{E}_x[V(X_t), \tau_C > t] \\
&\geq \lambda^{-1}\,\mathbf{E}_x[PV(X_t), \tau_C > t] \\
&= \lambda^{-1}\,\mathbf{E}_x[V(X_{t+1}), \tau_C > t] \\
&= \lambda^{-1}\Big(\mathbf{E}_x[V(X_{t+1}), \tau_C > t+1] + \mathbf{E}_x[V(X_{t+1}), \tau_C = t+1]\Big) \\
&\geq \lambda^{-1}\,\mathbf{E}_x[V(X_{t+1}), \tau_C > t+1] + \lambda^{-1}\,\mathbf{P}_x(\tau_C = t+1).
\end{aligned}$$

Substituting this inequality into (4.4) for t yields (4.4) for $t+1$. This completes the induction.

It follows that for every $x \notin C$ and $t \geq 0$,

$$V(x) \geq \lambda^{-t}\,\mathbf{P}_x(\tau_C > t) + \sum_{s=1}^{t} \lambda^{-s}\,\mathbf{P}_x(\tau_C = s), \tag{4.5}$$

so sending $t \to \infty$,

$$V(x) \geq \sum_{s=1}^{\infty} \lambda^{-s}\,\mathbf{P}_x(\tau_C = s).$$

The right side is equal to $\mathbf{E}_x[\lambda^{-\tau_C}]$ as long as $\mathbf{P}_x(\tau_C = \infty) = 0$. This is true because (4.5) implies that

$$V(x) \geq \lambda^{-t}\,\mathbf{P}_x(\tau_C = \infty).$$

Since $V(x)$ is finite and $0 < \lambda < 1$, necessarily $\mathbf{P}_x(\tau_C = \infty) = 0$, finishing the proof. □

Proof of Lemma 4.8. Under the assumption that $\pi(V)$ is finite, Lemma 4.8 can be proved simply by integrating the drift condition over $\mathcal{X}$. This proof, which is new, uses a truncation argument to avoid having to make that assumption.

Fix a positive integer N. Let $S_N = \{x \in \mathcal{X} : V(x) \leq N\}$, and let $U_N = \mathcal{X} \setminus S_N$. Then

$$\int_{S_N} V(y)\pi(dy) = \int_{S_N} V(y)(\pi P)(dy) = \int_{S_N}\int_{S_N} V(y)P(x,dy)\pi(dx) + \int_{U_N}\int_{S_N} V(y)P(x,dy)\pi(dx).$$

The second term satisfies

$$\begin{aligned}\int_{U_N}\int_{S_N} V(y)P(x,dy)\pi(dx) &\leq N\int_{U_N} P(x,S_N)\pi(dx)\\ &= N\left[\pi(S_N) - \int_{S_N} P(x,S_N)\pi(dx)\right]\\ &= N\int_{S_N} P(x,U_N)\pi(dx)\\ &\leq \int_{S_N}\int_{U_N} V(y)P(x,dy)\pi(dx).\end{aligned}$$

Therefore,

$$\begin{aligned}\int_{S_N} V(y)\pi(dy) &\leq \int_{S_N}\int_{S_N} V(y)P(x,dy)\pi(dx) + \int_{S_N}\int_{U_N} V(y)P(x,dy)\pi(dx)\\ &= \int_{S_N}\int_{\mathcal{X}} V(y)P(x,dy)\pi(dx)\\ &= \int_{S_N} PV(x)\pi(dx)\\ &\leq \int_{S_N\setminus C} \lambda V(x)\pi(dx) + \int_{S_N\cap C} K\pi(dx)\\ &= \int_{S_N} \lambda V(x)\pi(dx) + \int_{S_N\cap C} (K - \lambda V(x))\pi(dx)\\ &\leq \lambda\int_{S_N} V(x)\pi(dx) + (K-\lambda)\pi(C).\end{aligned}$$

Since

$$\lambda\int_{S_N} V(x)\pi(dx) \leq \lambda N\pi(S_N) < \infty,$$

that quantity can be subtracted from both sides, getting

$$\int_{S_N} V(x)\pi(dx) \leq \frac{K-\lambda}{1-\lambda}\pi(C).$$

It follows by monotone convergence that

$$\pi(V) = \int_{\mathcal{X}} V(x)\pi(dx) = \lim_{N\to\infty}\int_{S_N} V(x)\pi(dx) \leq \frac{K-\lambda}{1-\lambda}\pi(C). \qquad \square$$

4.3 Tail bound

Suppose that (X_t) is a Markov scheme whose transition kernel P satisfies a general or uniform drift and minorization condition. Lemma 4.7 says that for all $x \in \mathcal{X}$, $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x) < \infty$. This means in particular that $\mathbf{P}_x(\tau_C < \infty) = 1$, so by Proposition 3.10, the small set leads to the construction of a strong ν time T. The main result of this section is that the law of T decays exponentially.

Theorem 4.9. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Assume P satisfies a drift and minorization condition, and let T be a strong ν time for (X_t) satisfying the conclusions of Proposition 3.10. Then there are explicit constants $\lambda_* < 1$ and $r \leq 1$ such that for any initial distribution μ with $\mu(V) < \infty$,*

$$\mathbf{P}_\mu(T > t) \leq \mu(V)^r \lambda_*^{t+1-m} \qquad \text{for all } t \geq 0. \tag{4.6}$$

Here are the formulas for λ_* and r. If $\varepsilon = 1$, let $\lambda_* = \lambda$ and $r = 1$. If $\varepsilon < 1$, in the case of general drift and minorization, define

$$B = \frac{1 - \lambda^m}{1 - \lambda}(K - \lambda) + \lambda^m$$

and $L = (B - \varepsilon)/(1 - \varepsilon)$. Note that when $m = 1$, $B = K$; in general, $1 \leq B \leq mK$ and also

$$B \leq \frac{\max\{1, K - \lambda\}}{1 - \lambda}.$$

In the case of uniform drift and minorization, define $L = M$. Then set

$$\lambda_* = \max\left\{\lambda, \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log L}\right)\right\} < 1$$

and $r = \log\lambda_* / \log\lambda$.

In the case $m = 1$, the constant λ_* appears for the first time in [RT99], under the notation $\beta_{RT} = \lambda_*^{-1}$. (See also the corrigendum [RT01a].) Theorem 5.1 in [RT99] is very similar to (4.6), but the result is not quite as tight: while (4.6) takes the form $\mathbf{P}_\mu(T > t) \leq (\text{const})\lambda_*^t$, Theorem 5.1 takes the form $\mathbf{P}_\mu(T > t) \leq (\text{const}) t \lambda_*^t$. The generalization to the case $m > 1$ is new (but relatively routine).

The significance of B and L in the statement of Theorem 4.9 is the following.

Lemma 4.10. *Suppose P satisfies a general drift and minorization condition with respect to a small set C. Define B as in the statement of Theorem 4.9, and let μ be a probability measure supported on C. If $\eta(\cdot) = P^m(\mu, \cdot)$, then $\eta(V) \leq B$. In addition, one has the upper bounds $B \leq mK$ and*

$$B \leq \frac{\max\{1, K - \lambda\}}{1 - \lambda}.$$

Corollary 4.11. *Under the conditions of Lemma 4.10, suppose that $\varepsilon < 1$. Define the remainder measure $\bar{\mu}$ to satisfy*

$$P^m(\mu, \cdot) = \varepsilon\nu(\cdot) + (1 - \varepsilon)\bar{\mu}(\cdot).$$

Then

$$\nu(V) \le \frac{B - (1-\varepsilon)}{\varepsilon}, \qquad \bar{\mu}(V) \le \frac{B-\varepsilon}{1-\varepsilon} = L.$$

If instead P satisfies a uniform drift and minorization condition with $\varepsilon < 1$, and $\bar{\mu}$ is defined in the same way, both $\nu(V)$ and $\bar{\mu}(V)$ are bounded above by $M = L$.

A consequence of Lemma 4.10 and Corollary 4.11 is that $\nu(V) < \infty$ always. (If $\varepsilon = 1$ then ν is the same as η in Lemma 4.10, and if $\varepsilon < 1$ then Corollary 4.11 applies.) Therefore, Theorem 4.9 with initial distribution $\mu = \nu$ implies that $\mathbf{E}_\nu[T] < \infty$, and Condition 3.8 is satisfied.

Proposition 4.12. *Under the conditions of Theorem 4.9, the strong ν time T satisfies Condition 3.8. Thus, Proposition 3.9 ensures that (X_t) has unique stationary distribution given by the normalized occupation measure (3.5).*

Theorem 4.9 can also be used to bound the coupling time for a Markov chain satisfying a bivariate drift and minorization condition.

Theorem 4.13. *In the setting of Theorem 2.12, set*

$$\bar{\lambda}_* = \max\left\{\bar{\lambda}, \exp\left(\frac{-\log(1-\varepsilon)\log\bar{\lambda}}{-m\log\bar{\lambda} + \log\bar{L}}\right)\right\}$$

and $\bar{r} = \log\bar{\lambda}_* / \log\bar{\lambda}$, assuming $\varepsilon < 1$. (If $\varepsilon = 1$ then set $\bar{\lambda}_* = \bar{\lambda}$ and $\bar{r} = 1$.) Let $\bar{\mathbf{P}}_{\bar{\mu}}$ be the probability for the coupling described in Section 2.2, and let $\bar{T}$ be the coupling time. Then

$$\bar{\mathbf{P}}_{\bar{\mu}}(\bar{T} > t) \le \bar{\mu}(\bar{V})^{\bar{r}} \bar{\lambda}_*^{t+1-m} \qquad \textit{for all } t \ge 0.$$

In Theorem 2.12, one can take $\gamma = \bar{\lambda}_*$ and $B = \bar{\mu}(\bar{V})^{\bar{r}}\bar{\lambda}_*^{1-m}$.

Suppose P satisfies a (univariate) drift and minorization condition. Section 2.2 discussed two situations in which a bivariate drift function can be generated automatically: when P is stochastically monotone with respect to the univariate drift function V, and when the small set is of the form $C = \{x \in \mathcal{X} : V(x) \le d\}$ for large enough d. Theorem 4.13 combined with Propositions 2.14 and 2.15, along with the bounds on $(\delta_x \otimes \pi)(\bar{V})$ given in Section 2.2, leads to the following results.

Corollary 4.14. *Let the transition kernel P have a drift function V with respect to a small set $C = \{x \in \mathcal{X} : V(x) \le d\}$. If P is stochastically monotone with respect to V, then*

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le \left[V(x) + \frac{K-\lambda}{1-\lambda}\right]^r \lambda_*^{t+1-m} \qquad \textit{for all } t \ge 0, x \in \mathcal{X},$$

where λ_* and r are defined as in Theorem 4.9.

Corollary 4.15. *Let the transition kernel P have a drift function V with respect to a small set $C = \{x \in \mathcal{X} : V(x) \leq d\}$. If $d > (K-1)/(1-\lambda)$, then*

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \left(\frac{1}{2}\left[V(x) + \frac{K-\lambda}{1-\lambda}\right]\right)^{\bar{r}} \bar{\lambda}_*^{t+1-m} \qquad \text{for all } t \geq 0, x \in \mathcal{X},$$

where $\bar{\lambda}_*$ and $\bar{r}$ are defined as in Theorem 4.13 using $\bar{\lambda} = \lambda + (K-\lambda)/(d+1)$ and $\bar{L} = L$ from Theorem 4.9.

Proofs

No proof for Proposition 4.12 is required, since as discussed above, it follows immediately from Lemma 4.10, Corollary 4.11, and Theorem 4.9. Theorem 4.13 is proved in exactly the same way as the main Theorem 4.9, so its proof is also omitted. Corollaries 4.14 and 4.15 are a direct consequence of Theorem 4.13 combined with the discussion in Section 2.2.

Proof of Lemma 4.10. Because

$$\eta(V) = \mathbf{E}_\eta[V(X_0)] = \mathbf{E}_\mu[V(X_m)],$$

the desired statement is

$$\mathbf{E}_\mu[V(X_m)] \leq \frac{1-\lambda^m}{1-\lambda}(K-\lambda) + \lambda^m.$$

This is proved by induction on m. For $m = 1$, the right side is exactly K, and certainly $\mathbf{E}_\mu[V(X_1)] \leq K$. For the inductive step,

$$\begin{aligned}
\mathbf{E}_\mu[V(X_{m+1})] &= \mathbf{E}_\mu[V(X_{m+1}), X_m \in C] + \mathbf{E}_\mu[V(X_{m+1}), X_m \notin C] \\
&\leq K\,\mathbf{P}_\mu(X_m \in C) + \mathbf{E}_\mu[\lambda V(X_m), X_m \notin C] \\
&= K\,\mathbf{P}_\mu(X_m \in C) + \lambda\,\mathbf{E}_\mu[V(X_m)] - \lambda\,\mathbf{E}_\mu[V(X_m), X_m \in C] \\
&\leq \lambda\,\mathbf{E}_\mu[V(X_m)] + (K-\lambda)\,\mathbf{P}_\mu(X_m \in C) \\
&\leq \lambda\,\mathbf{E}_\mu[V(X_m)] + (K-\lambda).
\end{aligned}$$

If

$$\mathbf{E}_\mu[V(X_m)] \leq \frac{1-\lambda^m}{1-\lambda}(K-\lambda) + \lambda^m,$$

then

$$\begin{aligned}
\mathbf{E}_\mu[V(X_{m+1})] &\leq \left(\lambda \cdot \frac{1-\lambda^m}{1-\lambda} + 1\right)(K-\lambda) + \lambda^{m+1} \\
&= \frac{1-\lambda^{m+1}}{1-\lambda}(K-\lambda) + \lambda^{m+1}.
\end{aligned}$$

This completes the induction. To show that $B \leq mK$, since $\lambda < 1$,

$$\frac{1-\lambda^m}{1-\lambda}(K-\lambda) + \lambda^m = (1+\lambda+\cdots+\lambda^{m-1})(K-\lambda) + \lambda^m \leq mK - m\lambda + \lambda^m.$$

For all $m \geq 1$, $m\lambda \geq \lambda^m$. It follows that $B \leq mK$. One also has

$$B = (1+\lambda+\cdots+\lambda^{m-1})(K-\lambda) + \lambda^m \leq (1+\lambda+\cdots+\lambda^m)\max\{1, K-\lambda\} \leq \frac{\max\{1,K-\lambda\}}{1-\lambda}. \quad \square$$

Proof of Corollary 4.11. The statement for uniform drift and minorization is clear. In the case of general drift and minorization, setting $\eta(\cdot) = P^m(\mu,\cdot)$,

$$\eta(V) = \varepsilon\nu(V) + (1-\varepsilon)\bar{\mu}(V). \tag{4.7}$$

By Lemma 4.10, $\eta(V) \leq B$. Using that $\bar{\mu}(V) \geq 1$, it follows that

$$\nu(V) \leq \frac{B-(1-\varepsilon)}{\varepsilon}.$$

Using instead that $\nu(V) \geq 1$, it follows from (4.7) that

$$\bar{\mu}(V) \leq \frac{B-\varepsilon}{1-\varepsilon}. \qquad \square$$

Proof of Theorem 4.9. First suppose $\varepsilon = 1$, so that $T = \tau_C + m$. By Lemma 4.7, $\mathbf{E}_\mu[\lambda^{-\tau_C}] \leq \mu(V)$. Using Markov's inequality,

$$\mathbf{P}_\mu(T > t) \leq \lambda^{t+1}\,\mathbf{E}_\mu[\lambda^{-T}] = \lambda^{t+1-m}\,\mathbf{E}_\mu[\lambda^{-\tau_C}] \leq \mu(V)\lambda^{t+1-m}.$$

For the rest of the proof, assume $\varepsilon < 1$. Let

$$\mathcal{D} = \left\{\frac{1}{1-\varepsilon}[P^m(\mu,\cdot) - \varepsilon\nu(\cdot)] \,:\, \mu \text{ is a probability measure supported on } C\right\}.$$

By the minorization property, all the elements of $\mathcal{D}$ are probability measures. If $\bar{\mu} \in \mathcal{D}$, then by Corollary 4.11, $\bar{\mu}(V) \leq L$.

The only relevant properties of λ_* are that it is less than 1 and that for all $\bar{\mu} \in \mathcal{D}$,

$$\mathbf{E}_{\bar{\mu}}[\lambda_*^{-\tau_C - m}] \leq \frac{1}{1-\varepsilon}. \tag{4.8}$$

To prove (4.8), using that the function $f(x) = x^r$ is concave and Lemma 4.7,

$$\begin{aligned}\mathbf{E}_{\bar{\mu}}[\lambda_*^{-\tau_C - m}] = \lambda_*^{-m}\,\mathbf{E}_{\bar{\mu}}[(\lambda^{-\tau_C})^r] \le \lambda_*^{-m}\,\mathbf{E}_{\bar{\mu}}[\lambda^{-\tau_C}]^r \le \lambda_*^{-m}\bar{\mu}(V)^r \\ \le \lambda_*^{-m} L^r = \exp\left(-m\log\lambda_* + \frac{\log\lambda_*}{\log\lambda}\cdot\log L\right).\end{aligned}$$

This quantity is less than or equal to $\frac{1}{1-\varepsilon}$ exactly when

$$\lambda_* \ge \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log L}\right).$$

This proves (4.8).

The next step is to prove by induction on t that for all $\bar{\mu} \in \mathcal{D}$,

$$\mathbf{P}_{\bar{\mu}}(T > t) \le \frac{1}{1-\varepsilon}\lambda_*^{t+1}.$$

Fix $t \ge 0$. Suppose $0 \le s \le t - m$. Then define the remainder measure $\bar{\mu}_s \in \mathcal{D}$ to satisfy

$$\mathbf{P}_{\bar{\mu}}(X_{s+m} \in \cdot \mid \tau_C = s) = \varepsilon\nu(\cdot) + (1-\varepsilon)\bar{\mu}_s(\cdot).$$

To bound $\mathbf{P}_{\bar{\mu}}(T > t)$, first write

$$\mathbf{P}_{\bar{\mu}}(T > t) = \sum_{s=0}^{t-m} \mathbf{P}_{\bar{\mu}}(T > t \mid \tau_C = s)\,\mathbf{P}_{\bar{\mu}}(\tau_C = s) + \mathbf{P}_{\bar{\mu}}(\tau_C > t - m).$$

Using property 3 of Proposition 3.10 and the inductive hypothesis,

$$\mathbf{P}_{\bar{\mu}}(T > t \mid \tau_C = s) = (1-\varepsilon)\,\mathbf{P}_{\bar{\mu}_s}(T > t - m - s) \le \lambda_*^{t-m-s+1}.$$

Hence

$$\begin{aligned}\mathbf{P}_{\bar{\mu}}(T > t) &\le \lambda_*^{t+1}\sum_{s=0}^{t-m}\lambda_*^{-s-m}\,\mathbf{P}_{\bar{\mu}}(\tau_C = s) + \mathbf{P}_{\bar{\mu}}(\tau_C > t - m) \\ &= \lambda_*^{t+1}\left[\mathbf{E}_{\bar{\mu}}[\lambda_*^{-\tau_C - m}] - \sum_{s=t-m+1}^{\infty}\lambda_*^{-s-m}\,\mathbf{P}_{\bar{\mu}}(\tau_C = s)\right] + \sum_{s=t-m+1}^{\infty}\mathbf{P}_{\bar{\mu}}(\tau_C = s) \\ &\le \frac{1}{1-\varepsilon}\lambda_*^{t+1} - \sum_{s=t-m+1}^{\infty}(\lambda_*^{t-m+1-s} - 1)\,\mathbf{P}_{\bar{\mu}}(\tau_C = s) \\ &\le \frac{1}{1-\varepsilon}\lambda_*^{t+1}.\end{aligned}$$

Note that when $0 \le t \le m - 1$, the sum from $s = 0$ to $t - m$ is empty, so the inductive hypothesis is never invoked. Thus the preceding computation proves the base case as well as the inductive step.

Suppose now that μ is any probability measure on $\mathcal{X}$ with $\mu(V) < \infty$. As before, for $s \geq 0$, define the remainder measure $\mu_s \in \mathcal{D}$ to satisfy

$$\mathbf{P}_\mu(X_{s+m} \in \cdot \mid \tau_C = s) = \varepsilon\nu(\cdot) + (1-\varepsilon)\mu_s(\cdot).$$

Then if $s \leq t - m$, again using property 3 from Proposition 3.10,

$$\mathbf{P}_\mu(T > t \mid \tau_C = s) = (1-\varepsilon)\,\mathbf{P}_{\mu_s}(T > t - m - s) \leq \lambda_*^{t-m-s+1}.$$

To bound $\mathbf{P}_\mu(T > t)$, run the same computation that bounded $\mathbf{P}_{\bar{\mu}}(T > t)$. At the point in the argument that used

$$\mathbf{E}_{\bar{\mu}}[\lambda_*^{-\tau_C - m}] \leq \frac{1}{1-\varepsilon},$$

use instead

$$\mathbf{E}_\mu[\lambda_*^{-\tau_C - m}] = \lambda_*^{-m}\,\mathbf{E}_\mu[(\lambda^{-\tau_C})^r] \leq \lambda_*^{-m}\,\mathbf{E}_\mu[\lambda^{-\tau_C}]^r \leq \lambda_*^{-m}\mu(V)^r.$$

The result is

$$\mathbf{P}_\mu(T > t) \leq \mu(V)^r \lambda_*^{t+1-m}. \qquad \square$$

4.4 L^2 convergence from strong ν times

If P is reversible with nonnegative eigenvalues, a strong ν time provides direct control over the rate of convergence to stationarity in L^2 distance.

Theorem 4.16. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Assume that P is reversible with respect to a stationary measure π and that P has nonnegative eigenvalues. Let ν be any probability measure on $\mathcal{X}$, and let T be a strong ν time for (X_t) satisfying Condition 3.8. Then for all $t \geq 0$, $P^t(\nu, \cdot)$ is absolutely continuous with respect to π and*

$$\|P^t(\nu,\cdot) - \pi\|_{L^2(\pi)}^2 \leq \sum_{n=2t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

The core of the proof of Theorem 4.16 is in the following lemma.

Lemma 4.17. *Let ν be a probability measure on $\mathcal{X}$ and T be a strong ν time for the Markov chain (X_t) started from $X_0 \sim \nu$. Suppose $\mathbf{P}_\nu(T > 0) = 1$ and $\mathbf{E}_\nu[T] < \infty$. Let $f \geq 0$ be a measurable function on $\mathcal{X}$ such that $\mathbf{E}_\nu[f(X_0)]$ is finite and the sequence $\mathbf{E}_\nu[f(X_t)]$ is nonincreasing in t. Denote* *the limit of the sequence by* $\mathbf{E}_\nu[f(X_\infty)]$. *Then for all* $t \geq 0$,

$$\mathbf{E}_\nu[f(X_t)] - \mathbf{E}_\nu[f(X_\infty)] \leq \mathbf{E}_\nu[f(X_\infty)] \sum_{n=t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

Proof. Fix a positive integer n. Since T is a strong ν time,

$$\mathbf{E}_\nu[f(X_n), T \leq n] = \sum_{j=0}^{n-1} \mathbf{E}_\nu[f(X_j)]\, \mathbf{P}_\nu(T = n - j).$$

Using that $\mathbf{P}_\nu(T = n - j) = \mathbf{P}_\nu(T > n - j - 1) - \mathbf{P}_\nu(T > n - j)$, apply summation by parts. The result is

$$\begin{aligned} \sum_{j=1}^{n} \Big(\mathbf{E}_\nu[f(X_{j-1})] - \mathbf{E}_\nu[f(X_j)] \Big) \mathbf{P}_\nu(T > n - j) &= \mathbf{E}_\nu[f(X_0)]\, \mathbf{P}_\nu(T > n) - \mathbf{E}_\nu[f(X_n), T > n] \\ &\leq \mathbf{E}_\nu[f(X_0)]\, \mathbf{P}_\nu(T > n). \end{aligned} \tag{4.9}$$

Each term $\mathbf{E}_\nu[f(X_{j-1})] - \mathbf{E}_\nu[f(X_j)]$ is nonnegative. Summing (4.9) from $n = 1$ to ∞ gives

$$\Big(\mathbf{E}_\nu[f(X_0)] - \mathbf{E}_\nu[f(X_\infty)] \Big) \mathbf{E}_\nu[T] \leq \mathbf{E}_\nu[f(X_0)](\mathbf{E}_\nu[T] - 1),$$

which means that $\mathbf{E}_\nu[f(X_0)] \leq \mathbf{E}_\nu[f(X_\infty)]\, \mathbf{E}_\nu[T]$.

Fix $t \geq 0$. Summing (4.9) from $n = t + 1$ to ∞, the left side is

$$\sum_{j=1}^{\infty} \Big(\mathbf{E}_\nu[f(X_{j-1})] - \mathbf{E}_\nu[f(X_j)] \Big) \sum_{n=\max\{j, t+1\}}^{\infty} \mathbf{P}_\nu(T > n - j),$$

which is greater than or equal to

$$\sum_{j=t+1}^{\infty} \Big(\mathbf{E}_\nu[f(X_{j-1})] - \mathbf{E}_\nu[f(X_j)] \Big) \sum_{n=j}^{\infty} \mathbf{P}_\nu(T > n - j) = \Big(\mathbf{E}_\nu[f(X_t)] - \mathbf{E}_\nu[f(X_\infty)] \Big) \mathbf{E}_\nu[T].$$

The right side of the sum of (4.9) from $n = t + 1$ to ∞ is

$$\mathbf{E}_\nu[f(X_0)] \sum_{n=t+1}^{\infty} \mathbf{P}_\nu(T > n) \leq \mathbf{E}_\nu[f(X_\infty)]\, \mathbf{E}_\nu[T] \sum_{n=t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

Hence,

$$\mathbf{E}_\nu[f(X_t)] - \mathbf{E}_\nu[f(X_\infty)] \leq \mathbf{E}_\nu[f(X_\infty)] \sum_{n=t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

□

Proof of Theorem 4.16. By Proposition 3.9, for $A \in \mathcal{E}$,

$$\pi(A) = \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(X_n \in A, T > n) \geq \frac{1}{\mathbf{E}_\nu[T]} \mathbf{P}_\nu(X_0 \in A, T > 0) = \frac{\nu(A)}{\mathbf{E}_\nu[T]}.$$

Thus one can take the Radon-Nikodym derivative $\frac{d\nu}{d\pi}(x)$ to be less than or equal to $\mathbf{E}_\nu[T]$ for all $x \in \mathcal{X}$. In particular, $\frac{d\nu}{d\pi} \in L^2(\pi)$.

Fix $t \geq 0$. For any $f \in L^2(\pi)$,

$$\int_{\mathcal{X}} f(x) P^t(\nu, dx) = \int_{\mathcal{X}} (P^t f)(x) \frac{d\nu}{d\pi}(x) \pi(dx) = \int_{\mathcal{X}} f(x) \left(P^t \frac{d\nu}{d\pi} \right)(x) \pi(dx),$$

where the second equality used reversibility of P. This means precisely that $P^t(\nu, \cdot)$ is absolutely continuous with respect to π and

$$\frac{dP^t(\nu, \cdot)}{d\pi}(x) = \left(P^t \frac{d\nu}{d\pi} \right)(x).$$

(If P were not reversible, the absolute continuity would still be true, but the formula for the Radon-Nikodym derivative would not hold.) Now,

$$\begin{aligned} \|P^t(\nu, \cdot) - \pi\|_{L^2(\pi)}^2 &= \left\langle P^t \frac{d\nu}{d\pi} - 1, P^t \frac{d\nu}{d\pi} - 1 \right\rangle_\pi = \left\langle P^t \frac{d\nu}{d\pi}, P^t \frac{d\nu}{d\pi} \right\rangle_\pi - 1 \\ &= \left\langle P^{2t} \frac{d\nu}{d\pi}, \frac{d\nu}{d\pi} \right\rangle_\pi - 1, \end{aligned}$$

using reversibility of P in the last equality.

Consider the sequence

$$\left\langle P^t \frac{d\nu}{d\pi}, \frac{d\nu}{d\pi} \right\rangle_\pi = \int_{\mathcal{X}} \left(P^t \frac{d\nu}{d\pi} \right)(x) \nu(dx) = \mathbf{E}_\nu \left[\frac{d\nu}{d\pi}(X_t) \right].$$

Since P is reversible with nonnegative eigenvalues, this sequence is nonincreasing. Let

$$a = \lim_{t \to \infty} \mathbf{E}_\nu \left[\frac{d\nu}{d\pi}(X_t) \right].$$

The following argument shows that $a = 1$. For any $t \geq 0$,

$$\mathbf{E}_\pi \left[\frac{d\nu}{d\pi}(X_t) \right] = \mathbf{E}_\pi \left[\frac{d\nu}{d\pi}(X_0) \right] = \int_{\mathcal{X}} \frac{d\nu}{d\pi}(x) \pi(dx) = \int_{\mathcal{X}} \nu(dx) = 1.$$

As well,

$$\mathbf{E}_\pi\left[\frac{d\nu}{d\pi}(X_t)\right] = \mathbf{E}_\pi\left[\frac{d\nu}{d\pi}(X_t),\ T > t\right] + \sum_{s=0}^{\infty} \mathbf{P}_\pi(T = s)\,\mathbf{E}_\nu\left[\frac{d\nu}{d\pi}(X_{t-s})\right] \tag{4.10}$$

where $\mathbf{E}_\nu\left[\frac{d\nu}{d\pi}(X_{t-s})\right]$ is taken to be zero when $t - s < 0$. Take the limit as $t \to \infty$ of (4.10). The left side is 1. For the first part of the right side,

$$\lim_{t\to\infty} \mathbf{E}_\pi\left[\frac{d\nu}{d\pi}(X_t),\ T > t\right] \le \lim_{t\to\infty} \mathbf{E}_\nu[T]\,\mathbf{P}_\pi(T > t) = 0.$$

For the second part of the right side, use dominated convergence to interchange the sum and the limit. This is legal because for all t,

$$\mathbf{P}_\pi(T = s)\,\mathbf{E}_\nu\left[\frac{d\nu}{d\pi}(X_{t-s})\right] \le \mathbf{E}_\nu[T]\,\mathbf{P}_\pi(T = s),$$

and

$$\sum_{s=0}^{\infty} \mathbf{E}_\nu[T]\,\mathbf{P}_\pi(T = s) = \mathbf{E}_\nu[T] < \infty.$$

Hence

$$\begin{aligned}\lim_{t\to\infty} \sum_{s=0}^{\infty} \mathbf{P}_\pi(T = s)\,\mathbf{E}_\nu\left[\frac{d\nu}{d\pi}(X_{t-s})\right] &= \sum_{s=0}^{\infty} \lim_{t\to\infty} \mathbf{P}_\pi(T = s)\,\mathbf{E}_\nu\left[\frac{d\nu}{d\pi}(X_{t-s})\right] \\ &= \sum_{s=0}^{\infty} \mathbf{P}_\pi(T = s) \cdot a = a.\end{aligned}$$

So taking the limit as $t \to \infty$ of (4.10) yields $1 = 0 + a$.

Lemma 4.17 with $f = \frac{d\nu}{d\pi}$ gives

$$\left\langle P^t \frac{d\nu}{d\pi}, \frac{d\nu}{d\pi} \right\rangle_\pi - 1 \le \sum_{n=t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

It follows that

$$\|P^t(\nu,\cdot) - \pi\|^2_{L^2(\pi)} = \left\langle P^{2t} \frac{d\nu}{d\pi}, \frac{d\nu}{d\pi} \right\rangle_\pi - 1 \le \sum_{n=2t+1}^{\infty} \mathbf{P}_\nu(T > n). \qquad \square$$

4.5 Proof of Theorem 4.4

The L^2 bound in Theorem 4.4 is a combination of Theorem 4.16 with Theorem 4.9. The total variation and V-norm bounds will follow from the L^2 bound.

Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Since P satisfies a general or uniform drift and minorization condition, an extension of (X_t) has a strong ν time T satisfying Condition 3.8 and the tail bound of Theorem 4.9. Define

$$J = \frac{B - (1 - \varepsilon)}{\varepsilon}$$

in the case of general drift and minorization, and $J = M$ in the case of uniform drift and minorization, so that $\nu(V) \leq J$ by Corollary 4.11. Using Theorem 4.16 followed by Theorem 4.9,

$$\begin{aligned}\|P^t(\nu, \cdot) - \pi\|^2_{L^2(\pi)} &\leq \sum_{n=2t+1}^{\infty} \mathbf{P}_\nu(T > n) \\ &\leq \sum_{n=2t+1}^{\infty} J^r \lambda_*^{n+1-m} \\ &= J^r \lambda_*^{2t+2-m} \frac{1}{1 - \lambda_*} \\ &= D^2 \lambda_*^{2t}.\end{aligned}$$

This proves the L^2 bound.

For the total variation bound, first note that for any $x \in \mathcal{X}$,

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \sum_{n=0}^{t} \mathbf{P}_x(T = n) \|P^{t-n}(\nu, \cdot) - \pi\|_{\mathrm{TV}} + \mathbf{P}_x(T > t). \tag{4.11}$$

The proof of (4.11) begins with the definition

$$\|P^t(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} = \sup_{A \in \mathcal{E}} \Big[P^t(x, A) - \pi(A)\Big].$$

Because T is a strong ν time,

$$\begin{aligned}P^t(x, A) &= \sum_{n=0}^{t} \mathbf{P}_x(T = n) P^{t-n}(\nu, A) + \mathbf{P}_x(T > t)\, \mathbf{P}_x(X_t \in A \mid T > t), \\ \pi(A) &= \sum_{n=0}^{t} \mathbf{P}_x(T = n) \pi(A) + \mathbf{P}_x(T > t) \pi(A).\end{aligned}$$

Hence,

$$P^t(x, A) - \pi(A) \leq \sum_{n=0}^{t} \mathbf{P}_x(T = n) \sup_{A' \subseteq \mathcal{X}} \Big[P^{t-n}(\nu, A') - \pi(A')\Big] + \mathbf{P}_x(T > t).$$

Taking the supremum over all $A \in \mathcal{E}$ gives (4.11).

The L^2 bound combined with (4.11) yields

$$\begin{aligned}\|P^t(x,\cdot)-\pi\|_{\mathrm{TV}} &\le \sum_{n=0}^{t} \mathbf{P}_x(T=n)\|P^{t-n}(\nu,\cdot)-\pi\|_{\mathrm{TV}} + \mathbf{P}_x(T>t)\\ &\le \frac{D}{2}\lambda_*^t \sum_{n=0}^{t} \lambda_*^{-n}\,\mathbf{P}_x(T=n) + \mathbf{P}_x(T>t).\end{aligned}$$

Using summation by parts,

$$\sum_{n=0}^{t} \lambda_*^{-n}\,\mathbf{P}_x(T=n) = 1 + (\lambda_*^{-1}-1)\sum_{n=0}^{t-1} \lambda_*^{-n}\,\mathbf{P}_x(T>n) - \lambda_*^{-t}\,\mathbf{P}_x(T>t).$$

Then,

$$\begin{aligned}&\|P^t(x,\cdot)-\pi\|_{\mathrm{TV}}\\ &\le \max\left\{\frac{D}{2},1\right\}\lambda_*^t \sum_{n=0}^{t} \lambda_*^{-n}\,\mathbf{P}_x(T=n) + \mathbf{P}_x(T>t)\\ &= \max\left\{\frac{D}{2},1\right\}\lambda_*^t \left[1 + (\lambda_*^{-1}-1)\sum_{n=0}^{t-1} \lambda_*^{-n}\,\mathbf{P}_x(T>n) - \lambda_*^{-t}\,\mathbf{P}_x(T>t)\right] + \mathbf{P}_x(T>t)\\ &\le \max\left\{\frac{D}{2},1\right\}\lambda_*^t \left[1 + (\lambda_*^{-1}-1)\sum_{n=0}^{t-1} \lambda_*^{-n}\,\mathbf{P}_x(T>n)\right]\\ &\le \max\left\{\frac{D}{2},1\right\}\lambda_*^t \left[1 + (\lambda_*^{-1}-1)\sum_{n=0}^{t-1} \lambda_*^{-n} V(x)^r \lambda_*^{n+1-m}\right]\\ &= \max\left\{\frac{D}{2},1\right\}\left[1+(1-\lambda_*)V(x)^r t\right]\lambda_*^{-m}\cdot\lambda_*^t.\end{aligned}$$

This proves the total variation bound.

For the V-norm bounds, the following lemma will be useful.

Lemma 4.18. *Suppose the transition kernel P on $(\mathcal{X},\mathcal{E})$ has a drift function V with respect to a subset $C\in\mathcal{E}$. That is, $PV(x)\le\lambda V(x)$ if $x\notin C$ and $PV(x)\le K$ if $x\in C$. For any measures $\mu,\mu'\in\mathcal{P}(\mathcal{X})$ and any $t\ge 0$,*

$$\|P^t(\mu,\cdot)-P^t(\mu',\cdot)\|_V \le 2K\sum_{n=1}^{t}\lambda^{n-1}\|P^{t-n}(\mu,\cdot)-P^{t-n}(\mu',\cdot)\|_{\mathrm{TV}} + [\mu(V)+\mu'(V)]\lambda^t.$$

Proof. Let (X_t) be a Markov scheme with transition kernel P. First note that for any $\mu\in\mathcal{P}(\mathcal{X})$ and any $t\ge 0$,

$$\mathbf{E}_\mu[V(X_t),\tau_C\ge t]\le \mu(V)\lambda^t.$$

This is very similar to (4.4), and the proof is essentially the same. When $t = 0$ it is true. For $t \geq 1$,

$$\mathbf{E}_\mu[V(X_t), \tau_C \geq t] = \mathbf{E}_\mu[PV(X_{t-1}), \tau_C \geq t] \leq \lambda\, \mathbf{E}_\mu[V(X_{t-1}), \tau_C \geq t] \leq \lambda\, \mathbf{E}_\mu[V(X_{t-1}), \tau_C \geq t-1],$$

which finishes the inductive proof. Similarly, if $x \in C$ and $t \geq 1$,

$$\mathbf{E}_x[V(X_t), \tau_C^+ \geq t] \leq K\lambda^{t-1}.$$

Now,

$$\|P^t(\mu,\cdot) - P^t(\mu',\cdot)\|_V = \sup_{|f|\leq V} |P^t(\mu,f) - P^t(\mu',f)|,$$

where

$$P^t(\mu,f) = \mathbf{E}_\mu[f(X_t)] = \sum_{n=1}^{t} \int_C P^{t-n}(\mu,dx)\, \mathbf{E}_x[f(X_n), \tau_C^+ \geq n] + \mathbf{E}_\mu[f(X_t), \tau_C \geq t].$$

If $|f| \leq V$, then

$$\begin{aligned}
|P^t(\mu,f) - P^t(\mu',f)| &\leq \sum_{n=1}^{t} \int_C |P^{t-n}(\mu,\cdot) - P^{t-n}(\mu',\cdot)|(dx)\, \mathbf{E}_x[V(X_n), \tau_C^+ \geq n] \\
&\quad + \mathbf{E}_\mu[V(X_t), \tau_C \geq t] + \mathbf{E}_{\mu'}[V(X_t), \tau_C \geq t] \\
&\leq \sum_{n=1}^{t} K\lambda^{n-1} \int_{\mathcal{X}} |P^{t-n}(\mu,\cdot) - P^{t-n}(\mu',\cdot)|(dx) + [\mu(V) + \mu'(V)]\lambda^t \\
&= 2K \sum_{n=1}^{t} \lambda^{n-1} \|P^{t-n}(\mu,\cdot) - P^{t-n}(\mu',\cdot)\|_{\mathrm{TV}} + [\mu(V) + \mu'(V)]\lambda^t. \qquad \square
\end{aligned}$$

The V-norm bounds in Theorem 4.4 come from Lemma 4.18 (using $\mu' = \pi$) and the previously proved total variation bounds. First, using that

$$\|P^t(\nu,\cdot) - \pi\|_{\mathrm{TV}} \leq \frac{D}{2}\lambda_*^t,$$

one has

$$\begin{aligned}
\|P^t(\nu,\cdot) - \pi\|_V &\leq KD \sum_{n=1}^{t} \lambda^{n-1}\lambda_*^{t-n} + [\nu(V) + \pi(V)]\lambda^t \\
&\leq KD \sum_{n=1}^{t} \lambda^{n-1}\lambda_*^{t-n} + \left[J + \frac{K-\lambda}{1-\lambda}\right]\lambda^t.
\end{aligned}$$

The formula for the sum looks different in the two cases $\lambda = \lambda_*$ and $\lambda < \lambda_*$. If $\lambda = \lambda_*$,

$$\sum_{n=1}^{t} \lambda^{n-1} \lambda_*^{t-n} = t\lambda^{t-1}.$$

If $\lambda < \lambda_*$,

$$\sum_{n=1}^{t} \lambda^{n-1} \lambda_*^{t-n} = \frac{\lambda_*^t - \lambda^t}{\lambda_* - \lambda}.$$

This proves the bound on $\|P^t(\nu, \cdot) - \pi\|_V$.

For the chain started from $x \in \mathcal{X}$, the total variation bound is

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \max\left\{\frac{D}{2}, 1\right\} \lambda_*^{-m}[1 + (1 - \lambda_*)V(x)^r t]\lambda_*^t.$$

Therefore,

$$\begin{aligned}\|P^t(x, \cdot) - \pi\|_V \leq 2K \max\left\{\frac{D}{2}, 1\right\} \lambda_*^{-m} \left[\sum_{n=1}^{t} \lambda^{n-1}\lambda_*^{t-n} + (1 - \lambda_*)V(x)^r \sum_{n=1}^{t} \lambda^{n-1}(t - n)\lambda_*^{t-n}\right] \\ + \left[V(x) + \frac{K - \lambda}{1 - \lambda}\right] \lambda^t.\end{aligned}$$

The closed form of the first sum has already been given. For the second sum, when $\lambda = \lambda_*$,

$$\sum_{n=1}^{t} \lambda^{n-1}(t - n)\lambda_*^{t-n} = \frac{t^2 - t}{2}\lambda^{t-1},$$

and when $\lambda < \lambda_*$,

$$\sum_{n=1}^{t} \lambda^{n-1}(t - n)\lambda_*^{t-n} = \frac{t\lambda_*^t}{\lambda_* - \lambda} - \frac{\lambda_*(\lambda_*^t - \lambda^t)}{(\lambda_* - \lambda)^2}.$$

This proves the bound on $\|P^t(x, \cdot) - \pi\|_V$, finishing the proof of Theorem 4.4. □

4.6 Decreasing hazard rates

The object of this section is to prove Theorem 4.5. The main step is showing the following result, which may be of independent interest.

Theorem 4.19. *Let (X_t) be a Markov scheme on $(\mathcal{X}, \mathcal{E})$ with transition kernel P. Assume that P is reversible with respect to a stationary distribution π and has nonnegative eigenvalues. For fixed*

$c \in \mathcal{X}$, suppose that $\mathbf{E}_c[\tau_c^+] < \infty$ and $\mathbf{P}_x(\tau_c < \infty) = 1$ for all $x \in \mathcal{X}$. Then for all $t \geq 0$,

$$\|P^t(c, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbf{P}_\pi(\tau_c > t).$$

Proof. First observe that the stationary distribution π is necessarily unique by Proposition 3.9 applied to the stopping time τ_c^+. Suppose now that the chain (X_t) is started from $X_0 = c$. Let $b_n = \mathbf{P}_c(\tau_c^+ = n)$ for $n \geq 1$, and let $B_n = \sum_{k=n+1}^{\infty} b_k = \mathbf{P}_c(\tau_c^+ > n)$ for $n \geq 0$. The sequence of hazard rates associated with (b_n) is defined by

$$h_n = \frac{b_n}{B_{n-1}} = \mathbf{P}_c(\tau_c^+ = n \mid \tau_c^+ \geq n) \tag{4.12}$$

for $n \geq 1$. If $B_{n-1} = 0$, one can take $h_n = 1$.

The following result is due to Lund, Zhao, and Kiessler [LZK06], who proved it under the assumption that the state space is finite or countable. For completeness, a proof of the general case will be given at the end of this section.

Theorem 4.20. *Let P be a Markov transition kernel on $(\mathcal{X}, \mathcal{E})$, reversible with respect to its unique stationary distribution π and having nonnegative eigenvalues. Fix $c \in \mathcal{X}$ and let (X_t) be a Markov chain with transition kernel P started at $X_0 = c$. Suppose $\mathbf{E}_c[\tau_c^+] < \infty$. Then the sequence (h_n) defined by (4.12) is nonincreasing.*

Theorem 4.20 can be interpreted in terms of the age process (A_t) associated with (X_t):

$$A_t = t - \max\{s \geq 0 : X_s = c\}.$$

The value of A_t is always an element of the set $\mathbf{A} = \{n \geq 0 : h_k \neq 1 \text{ for all } 1 \leq k \leq n\}$. In general one could have $\mathbf{A} = \{0, 1, \ldots, a\}$. In the present situation, when (h_n) is a decreasing sequence, either $\mathbf{A} = \{0\}$ or $\mathbf{A}$ is the set of all nonnegative integers. If $\mathbf{A} = \{0\}$ then $P(c, c) = 1$; by uniqueness of π, it follows that $\pi(c) = 1$, so the desired statement is trivial. In all nontrivial cases, $\mathbf{A}$ is the set of nonnegative integers.

It is easily seen that (A_t) is itself a Markov chain on $\mathbf{A}$, with $A_0 = 0$ and transition matrix Q given by $Q(i, 0) = h_{i+1}$, $Q(i, i+1) = 1 - h_{i+1}$, and $Q(i, j) = 0$ for all other values of j.

Saying that (h_n) is nonincreasing is equivalent to saying that the transition matrix Q is stochastically monotone. That is, if $i \leq j$ then for all $r \geq 0$,

$$\sum_{k \geq r,\, k \in \mathbf{A}} Q(i, k) \leq \sum_{k \geq r,\, k \in \mathbf{A}} Q(j, k).$$

Since Q is stochastically monotone, there is an associated monotone (Markovian) coupling. In this case the transition matrix $\bar{Q}$ for the coupled chain, on the state space $\mathbf{A} \times \mathbf{A}$, is given as follows. For $i \le j$,

$$\bar{Q}((i,j),(0,0)) = h_{j+1}, \qquad \bar{Q}((i,j),(0,j+1)) = h_{i+1} - h_{j+1}, \qquad \bar{Q}((i,j),(i+1,j+1)) = 1 - h_{i+1}.$$

The monotone coupling can be used to determine how fast the age process converges to its stationary distribution. This will lead to a bound on the convergence of (X_t) to π.

The connection between (X_t) and (A_t) is an *intertwining relation*. For each $i \in \mathbf{A}$, define the probability measure $\Lambda(i,\cdot)$ on $\mathcal{X}$ by

$$\Lambda(i,\cdot) = \mathbf{P}_c(X_i \in \cdot \mid \tau_c^+ > i). \tag{4.13}$$

(Note that $\mathbf{P}_c(\tau_c^+ > i) > 0$ automatically.) Λ is a transition kernel from $\mathbf{A}$ to $\mathcal{X}$, also called a *link*. For any measure μ on $\mathbf{A}$, define the measure $\mu\Lambda$ on $\mathcal{X}$ by

$$(\mu\Lambda)(\cdot) = \sum_{i \in \mathbf{A}} \mu(i)\Lambda(i,\cdot).$$

Similarly, for any transition kernels K on $\mathcal{X}$ and L on $\mathbf{A}$, define the transition kernels ΛK and $L\Lambda$ from $\mathbf{A}$ to $\mathcal{X}$ by

$$(\Lambda K)(i,\cdot) = \int_{\mathcal{X}} K(x,\cdot)\Lambda(i,dx), \qquad (L\Lambda)(i,\cdot) = \sum_{j \in \mathbf{A}} L(i,j)\Lambda(j,\cdot).$$

The intertwining between P and Q is expressed by the equation

$$\Lambda P = Q\Lambda, \tag{4.14}$$

which follows from the definitions of Q and Λ. This immediately implies that $\Lambda P^t = Q^t \Lambda$ for all $t \ge 0$. In addition, the starting distributions $A_0 \sim \delta_0$ and $X_0 \sim \delta_c$ are related by $\delta_0 \Lambda = \delta_c$. Hence, if $\mathbf{Q}_0$ is the probability for the chain (A_t), the laws of A_t and X_t are related by

$$\mathbf{P}_c(X_t \in \cdot) = \sum_{i \in \mathbf{A}} \mathbf{Q}_0(A_t = i)\Lambda(i,\cdot) \qquad \text{for all } t \ge 0.$$

For more information on intertwining relations, see [DF90].

The next step is to characterize the stationary distribution $\tilde{\pi}$ of Q. Suppose $\tilde{\pi}$ is a measure on $\mathbf{A}$ (not necessarily a probability measure) for which $\tilde{\pi} Q = \tilde{\pi}$. Using the definition of Q, for each $i \ge 1$,

$\tilde{\pi}(i) = (1-h_i)\tilde{\pi}(i-1)$. Therefore,

$$\tilde{\pi}(i) = \left(\prod_{k=1}^{i}(1-h_k)\right)\tilde{\pi}(0),$$

meaning that

$$\tilde{\pi}(\mathbf{A}) = \left(\sum_{i\in\mathbf{A}}\prod_{k=1}^{i}(1-h_k)\right)\tilde{\pi}(0).$$

Set

$$Z = \sum_{i\in\mathbf{A}}\prod_{k=1}^{i}(1-h_k).$$

If $Z = \infty$, then Q has no stationary distribution. If $Z < \infty$, then Q has a unique stationary distribution given by

$$\tilde{\pi}(i) = \frac{1}{Z}\prod_{k=1}^{i}(1-h_k). \tag{4.15}$$

The quantity $\prod_{k=1}^{i}(1-h_k)$ has a natural interpretation in terms of the chain (X_t):

$$\mathbf{P}_c(\tau_c^+ > i) = \prod_{k=1}^{i}(1-h_k).$$

Hence

$$Z = \sum_{i\in\mathbf{A}}\mathbf{P}_c(\tau_c^+ > i) = \mathbf{E}_c[\tau_c^+] < \infty.$$

It follows that $\tilde{\pi}$ defined as in (4.15) is the unique stationary distribution for Q.

Due to the intertwining (4.14), $\tilde{\pi}\Lambda$ is a stationary distribution for P. By uniqueness, $\tilde{\pi}\Lambda = \pi$. It is also straightforward to check that $\tilde{\pi}\Lambda$ is precisely the normalized occupation measure (3.5) with $\nu = \delta_c$ and $T = \tau_c^+$.

Another consequence of the intertwining is that

$$\|P^t(c,\cdot) - \pi\|_{\mathrm{TV}} \le \|Q^t(0,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} \tag{4.16}$$

for all $t \ge 0$. This is proved as follows. Let $\|\mu\|_1$ denote the total mass of a signed measure μ. If $\tilde{\mu}$ is a signed measure on $\mathbf{A}$, with positive and negative parts $\tilde{\mu}_+$ and $\tilde{\mu}_-$, then $\|\tilde{\mu}\|_1 = \tilde{\mu}_+(\mathbf{A}) + \tilde{\mu}_-(\mathbf{A})$. The signed measure $\mu = \tilde{\mu}\Lambda$ on $\mathcal{X}$ satisfies $\mu = \tilde{\mu}_+\Lambda - \tilde{\mu}_-\Lambda$, so $\|\mu\|_1 \le (\tilde{\mu}_+\Lambda)(\mathcal{X}) + (\tilde{\mu}_-\Lambda)(\mathcal{X}) = \tilde{\mu}_+(\mathbf{A}) + \tilde{\mu}_-(\mathbf{A}) = \|\tilde{\mu}\|_1$. By the intertwining, if $\tilde{\mu}(\cdot) = Q^t(0,\cdot) - \tilde{\pi}$, then $\mu(\cdot) = P^t(c,\cdot) - \pi$. Hence

$$\|P^t(c,\cdot) - \pi\|_1 \le \|Q^t(0,\cdot) - \tilde{\pi}\|_1,$$

and the total variation norm is exactly one-half the total mass norm.

Let (A_t) and (A'_t) be two copies of the age process, one started from $A_0 = 0$ and the other from $A'_0 \sim \tilde{\pi}$. The proof will use the monotone coupling for (A_t) and (A'_t) to bound $\|Q^t(0,\cdot) - \tilde{\pi}\|_{\mathrm{TV}}$, followed by (4.16). It would also be possible to convert the coupling of (A_t) and (A'_t) into a coupling of two copies of the original Markov chain, (X_t) started from c and (X'_t) started from π, using the link Λ. This method would circumvent (4.16). It is carried out in [DF90] in the context of finite state space; but due to the technicalities associated with general state spaces, the argument below will proceed using (4.16).

The monotone coupling of (A_t) and (A'_t) is a Markov chain (A_t, A'_t) on $\mathbf{A} \times \mathbf{A}$ with transition matrix $\bar{Q}$ and initial distribution $\delta_0 \otimes \tilde{\pi}$. Let $\bar{\mathbf{P}}$ denote the probability associated with this chain. The coupling time is

$$T = \tau'_0 = \min\{t \geq 0 \,:\, A'_t = 0\},$$

since by monotonicity, $A'_t = 0$ implies that $A_t = 0$; and it is impossible for A_t and A'_t to be equal before A'_t has reached 0. By the Coupling Inequality (Proposition 2.9),

$$\|Q^t(0,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} \leq \bar{\mathbf{P}}(T > t) = \mathbf{Q}_{\tilde{\pi}}(\tau'_0 > t), \tag{4.17}$$

where $\mathbf{Q}_{\tilde{\pi}}$ is the probability for the chain (A'_t). Actually, equality holds in (4.17) since the monotone coupling is optimal, but that will not be necessary for the rest of the proof.

Let (X'_t) be a Markov chain with transition kernel P started from π, and let $\tau'_c = \min\{t \geq 0 : X'_t = c\}$. It will be shown that

$$\mathbf{Q}_{\tilde{\pi}}(\tau'_0 > t) = \mathbf{P}_{\pi}(\tau'_c > t). \tag{4.18}$$

The combination of (4.16), (4.17), and (4.18) gives the statement of the theorem. Therefore it suffices to prove (4.18).

Define $\mathbf{A}_1 = \mathbf{A} \setminus \{0\}$ and $\mathcal{X}_1 = \mathcal{X} \setminus \{c\}$. Let P_1 be the restriction of P to $\mathcal{X}_1$ and let Q_1 be the restriction of Q to $\mathbf{A}_1$. That is, $P_1(x, B) = P(x, B)$ for $x \in \mathcal{X}_1$ and $B \subseteq \mathcal{X}_1$, while $Q_1(i,j) = Q(i,j)$ for $i, j \in \mathbf{A}_1$. Both P_1 and Q_1 are sub-stochastic transition kernels. Let $\tilde{\pi}_1$ and π_1 be the restrictions of $\tilde{\pi}$ and π to $\mathbf{A}_1$ and $\mathcal{X}_1$, so that in particular $\tilde{\pi}_1(\mathbf{A}_1) = 1 - \tilde{\pi}(0)$ and $\pi_1(\mathcal{X}_1) = 1 - \pi(\{c\})$. Then

$$\mathbf{Q}_{\tilde{\pi}}(\tau'_0 > t) = \sum_{i_0, \ldots, i_t \in \mathbf{A}_1} \tilde{\pi}(i_0) Q(i_0, i_1) \cdots Q(i_{t-1}, i_t) = (\tilde{\pi}_1 Q_1^t)(\mathbf{A}_1),$$

and similarly $\mathbf{P}_{\pi}(\tau'_c > t) = (\pi_1 P_1^t)(\mathcal{X}_1)$.

It follows from the definition (4.13) that $\Lambda(0, \{c\}) = 1$ and $\Lambda(i, \{c\}) = 0$ for all $i \geq 1$. Therefore one can define a link Λ_1 from $\mathbf{A}_1$ to $\mathcal{X}_1$ by $\Lambda_1(i, B) = \Lambda(i, B)$ for $i \in \mathbf{A}_1$ and $B \subseteq \mathcal{X}_1$. Each row of Λ_1 has total mass 1, that is, $\Lambda_1(i, \mathcal{X}_1) = 1$ for all $i \in \mathbf{A}_1$. As before, one has the intertwining

$\Lambda_1 P_1 = Q_1 \Lambda_1$, which immediately implies that $\Lambda_1 P_1^t = Q_1^t \Lambda_1$ for all $t \geq 0$. Since also $\tilde{\pi}_1 \Lambda_1 = \pi_1$,

$$\begin{aligned}(\pi_1 P_1^t)(\mathcal{X}_1) &= (\tilde{\pi}_1 \Lambda_1 P_1^t)(\mathcal{X}_1) = (\tilde{\pi}_1 Q_1^t \Lambda_1)(\mathcal{X}_1) \\ &= \sum_{i \in \mathbf{A}_1} (\tilde{\pi}_1 Q_1^t)(i) \Lambda_1(i, \mathcal{X}_1) = \sum_{i \in \mathbf{A}_1} (\tilde{\pi}_1 Q_1^t)(i) = (\tilde{\pi}_1 Q_1^t)(\mathbf{A}_1).\end{aligned}$$

Thus (4.18) holds, and the proof is complete. □

Proof of Theorem 4.5. Let (X_t) be a Markov scheme with transition kernel P. By Lemma 4.7, $\mathbf{P}_x(\tau_c < \infty) = 1$ for all $x \in \mathcal{X}$, and $\mathbf{E}_c[\lambda^{-\tau_c^+}] \leq \lambda^{-1} K < \infty$. Therefore the conditions of Theorem 4.19 are met. Using notation from the proof of that theorem,

$$\|P^t(c, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbf{P}_\pi(\tau_c > t) = \mathbf{Q}_{\tilde{\pi}}(\tau_0' > t),$$

where $\mathbf{Q}_{\tilde{\pi}}$ is the probability for the age process (A_t') started from $A_0' \sim \tilde{\pi}$. In terms of the hazard rates h_n,

$$\mathbf{Q}_{\tilde{\pi}}(\tau_0' > t) = \sum_{i=1}^{\infty} \tilde{\pi}(i) \prod_{k=1}^{t} (1 - h_{i+k}), \tag{4.19}$$

where the empty product is taken to be 1 when $t = 0$.

Because (h_n) is nonincreasing, it has a limit $h_\infty = \lim_{n\to\infty} h_n$. The drift function provides a lower bound on h_∞. Specifically, since $\mathbf{E}_c[\lambda^{-\tau_c^+}] \leq \lambda^{-1} K$, it follows by Markov's inequality that

$$\prod_{n=1}^{t} (1 - h_n) = \mathbf{P}_c(\tau_c^+ > t) \leq K \lambda^t,$$

which yields

$$\frac{1}{t} \sum_{n=1}^{t} \log(1 - h_n) \leq \frac{\log K}{t} + \log \lambda.$$

Taking the limit as $t \to \infty$ gives $\log(1 - h_\infty) \leq \log \lambda$, meaning that $1 - h_\infty \leq \lambda$. Therefore,

$$\|P^t(c, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbf{Q}_{\tilde{\pi}}(\tau_0' > t) \leq \sum_{i=1}^{\infty} \tilde{\pi}(i) \prod_{k=1}^{t} (1 - h_\infty) \leq [1 - \tilde{\pi}(0)] \lambda^t.$$

For every $i \geq 0$, $Q(i, 0) = h_{i+1} \geq h_\infty \geq 1 - \lambda$. Thus $\tilde{\pi}(0) \geq 1 - \lambda$, and $1 - \tilde{\pi}(0) \leq \lambda$. This proves the first inequality in Theorem 4.5.

At this point a comparison can be drawn with a similar result in [LZK06]. Using that $\tilde{\pi}(0) = \pi(c)$, the previous argument showed that

$$\|P^t(c, \cdot) - \pi\|_{\mathrm{TV}} \leq [1 - \pi(c)](1 - h_\infty)^t. \tag{4.20}$$

Theorem 4.3 of [LZK06] is proved using the decreasing hazard rate property but without the interpretation in terms of the age process. It says that

$$\|P^t(c,\cdot)-\pi\|_{\mathrm{TV}} \le \left(\sqrt{\frac{1-h_\infty}{4\pi(c)}}\right)(1-h_\infty)^t.$$

Since $\pi(c) \ge h_\infty$,

$$\frac{1-h_\infty}{1-\pi(c)} \ge 1 \ge 4\pi(c)[1-\pi(c)],$$

which implies that

$$\sqrt{\frac{1-h_\infty}{4\pi(c)}} \ge 1-\pi(c).$$

Hence the bound (4.20) is always better than Theorem 4.3 of [LZK06]. In the setting of a drift function with λ close to 1, Theorem 4.3 of [LZK06] is worse by about a factor of $(1-\lambda)^{-1/2}$. This might correspond to a log factor in a typical finite Markov chain mixing time problem.

Returning to the proof of Theorem 4.5, the analogous version of (4.11) is

$$\|P^t(x,\cdot)-\pi\|_{\mathrm{TV}} \le \sum_{n=0}^{t} \mathbf{P}_x(\tau_c=n)\|P^{t-n}(c,\cdot)-\pi\|_{\mathrm{TV}} + \mathbf{P}_x(\tau_c>t).$$

By Markov's inequality applied to Lemma 4.7, $\mathbf{P}_x(\tau_c > t) \le V(x)\lambda^{t+1}$. As well,

$$\begin{aligned}\sum_{n=0}^{t} \mathbf{P}_x(\tau_c=n)\|P^{t-n}(c,\cdot)-\pi\|_{\mathrm{TV}} &\le \sum_{n=0}^{t} \mathbf{P}_x(\tau_c=n)\lambda^{t-n+1} \\ &\le \lambda^{t+1}\sum_{n=0}^{\infty} \mathbf{P}_x(\tau_c=n)\lambda^{-n} \\ &= \lambda^{t+1}\,\mathbf{E}_x[\lambda^{-\tau_c}] \le V(x)\lambda^{t+1}.\end{aligned}$$

This proves the second inequality in Theorem 4.5.

The third inequality in Theorem 4.5 follows from the second inequality and Lemma 4.18. Indeed,

$$\begin{aligned}\|P^t(x,\cdot)-\pi\|_V &\le 2K\sum_{n=1}^{t}\lambda^{n-1}\|P^{t-n}(x,\cdot)-\pi\|_{\mathrm{TV}} + [V(x)+\pi(V)]\lambda^t \\ &\le 2K\sum_{n=1}^{t}\lambda^{n-1}2V(x)\lambda^{t-n+1} + \left[V(x)+\frac{K-\lambda}{1-\lambda}\right]\lambda^t \\ &= 4KV(x)t\lambda^t + \left[V(x)+\frac{K-\lambda}{1-\lambda}\right]\lambda^t.\end{aligned}$$

This completes the proof. □

Proof of Theorem 4.20. First, note that

$$\pi(A) = \frac{1}{\mathbf{E}_c[\tau_c^+]} \sum_{n=0}^{\infty} \mathbf{P}_c(X_n \in A, \tau_c^+ > n). \tag{4.21}$$

This is because the right side of (4.21) is a stationary distribution for P (by the computation in the proof of Proposition 3.9), and π is assumed to be unique. It follows that $\pi(c) = 1/\mathbf{E}_c[\tau_c^+] > 0$.

Define

$$b_n = \mathbf{P}_c(\tau_c^+ = n), \qquad B_n = \mathbf{P}_c(\tau_c^+ > n) = \sum_{k=n+1}^{\infty} b_k, \qquad h_n = \mathbf{P}_c(\tau_c^+ = n \mid \tau_c^+ \geq n) = \frac{b_n}{B_{n-1}}.$$

Also define $u_n = \mathbf{P}_n(X_n = c)$ to be the renewal sequence associated with the increment sequence (b_n). Thus $u_0 = 1$ and for $n \geq 1$,

$$u_n = \sum_{k=1}^{n} u_{n-k} b_k. \tag{4.22}$$

Using summation by parts on (4.22) leads to the alternate form

$$B_n = \sum_{k=1}^{n} (u_{k-1} - u_k) B_{n-k}$$

for $n \geq 1$.

The goal is to show that the sequence (h_n) is nonincreasing. Since

$$h_n = \frac{B_{n-1} - B_n}{B_{n-1}} = 1 - \frac{B_n}{B_{n-1}},$$

it will suffice to show that

$$B_n^2 \leq B_{n-1} B_{n+1}$$

for all $n \geq 1$. This property of the sequence (B_n) is called *log-convexity.*

A classical result of de Bruijn and Erdős [BE53] says the following. Let $(c_k : k \geq 1)$ be a sequence of nonnegative real numbers, and define $(a_n : n \geq 0)$ by $a_0 = 1$ and

$$a_n = \sum_{k=1}^{n} c_k a_{n-k}$$

for $n \geq 1$. If (c_k) is log-convex, so is (a_n).

Using $c_k = u_{k-1} - u_k$ and $a_n = B_n$, it will suffice to show that (c_k) is nonnegative and log-convex.

Consider the function $\mathbf{1}_c$ on $\mathcal{X}$ given by

$$\mathbf{1}_c(x) = \begin{cases} 1 & \text{if } x = c, \\ 0 & \text{if } x \neq c. \end{cases}$$

One has $\langle P^k \mathbf{1}_c, \mathbf{1}_c \rangle_\pi = \pi(c) P^k(c,c) = \pi(c) u_k$. Let μ be the spectral measure associated with P and the function $\mathbf{1}_c$. This means that μ is a (positive) measure on $[0,1]$, with total mass $\pi(c)$, and

$$\langle P^k \mathbf{1}_c, \mathbf{1}_c \rangle_\pi = \int_{[0,1]} x^k \mu(dx).$$

It follows that

$$c_k = \frac{1}{\pi(c)} \int_{[0,1]} (1-x) x^{k-1} \mu(dx) \geq 0.$$

For log-convexity, the goal is to show that $c_k^2 \leq c_{k-1} c_{k+1}$ for $k \geq 2$. The left side is

$$c_k^2 = \frac{1}{\pi(c)^2} \int_{[0,1]} \int_{[0,1]} (1-x)(1-y) x^{k-1} y^{k-1} \mu(dx) \mu(dy).$$

The right side is

$$\begin{aligned} c_{k-1} c_{k+1} &= \frac{1}{\pi(c)^2} \int_{[0,1]} \int_{[0,1]} (1-x)(1-y) x^{k-2} y^k \mu(dx) \mu(dy) \\ &= \frac{1}{\pi(c)^2} \int_{[0,1]} \int_{[0,1]} (1-x)(1-y) x^k y^{k-2} \mu(dx) \mu(dy). \end{aligned}$$

Therefore,

$$\begin{aligned} c_{k-1} c_{k+1} - c_k^2 &= \frac{1}{\pi(c)^2} \int_{[0,1]} \int_{[0,1]} (1-x)(1-y) \left(\frac{1}{2} x^{k-2} y^k + \frac{1}{2} x^k y^{k-2} - x^{k-1} y^{k-1} \right) \mu(dx) \mu(dy) \\ &= \frac{1}{\pi(c)^2} \int_{[0,1]} \int_{[0,1]} (1-x)(1-y) \frac{1}{2} x^{k-2} y^{k-2} (x-y)^2 \mu(dx) \mu(dy) \\ &\geq 0. \end{aligned}$$

This proves that (c_k) is nonnegative and log-convex, so (B_n) is log-convex, and (h_n) is decreasing. □

4.7 Proof of Theorem 4.6

Let ρ_2 be the $L^2(\pi)$ spectral radius of P, and define

$$\tilde{\rho}_{\mathrm{TV}} = \inf\{\gamma \geq 0 \, : \, \text{for some } \pi\text{-a.e. finite } B(x), \; \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t \text{ for all } t \geq 0\},$$
$$\rho_{\mathrm{TV}} = \inf\{\gamma \geq 0 \, : \, \text{for some everywhere finite } B(x), \; \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq B(x)\gamma^t \text{ for all } t \geq 0\}.$$

Theorem 2.22 states that $\rho_2 \leq \tilde{\rho}_{\mathrm{TV}}$. (The theorem assumes that the σ-algebra $\mathcal{E}$ is countably generated; but that hypothesis is not necessary to show that $\rho_2 \leq \tilde{\rho}_{\mathrm{TV}}$. There is a proof in [RR97].) Also, $\tilde{\rho}_{\mathrm{TV}} \leq \rho_{\mathrm{TV}}$ directly from the definitions. The proof below will bound ρ_{TV} using Theorem 4.4.

First consider the case of general drift and minorization. The exponential rate of convergence in Theorem 4.4 is

$$\lambda_* = \max\left\{\lambda, \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log L}\right)\right\}, \tag{4.23}$$

where $L = (B-\varepsilon)/(1-\varepsilon)$ and $B \leq K/(1-\lambda)$ (as stated in Theorem 4.9). Therefore,

$$\lambda_* \leq \max\left\{\lambda, \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log\frac{K}{1-\lambda} - \log(1-\varepsilon)}\right)\right\}.$$

Using $\varepsilon = 1 - A\beta^m$, a straightforward computation gives

$$\lim_{m\to\infty} \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log\frac{K}{1-\lambda} - \log(1-\varepsilon)}\right) = \exp\left(\frac{\log\beta\log\lambda}{\log\beta + \log\lambda}\right).$$

Because

$$\exp\left(\frac{\log\beta\log\lambda}{\log\beta + \log\lambda}\right) = \lambda^{\log\beta/(\log\beta+\log\lambda)} > \lambda,$$

the exponential rate of convergence gets arbitrarily close to

$$\rho = \exp\left(\frac{\log\beta\log\lambda}{\log\beta + \log\lambda}\right)$$

as $m \to \infty$. This gives the desired bound on ρ_2.

Next, consider the case of uniform drift and minorization. The exponential rate of convergence is again given by (4.23), but this time $L = M$ has no dependence on ε. Using $\varepsilon = 1 - A\beta^m$,

$$\lim_{m\to\infty} \exp\left(\frac{-\log(1-\varepsilon)\log\lambda}{-m\log\lambda + \log M}\right) = \beta.$$

Hence the exponential rate of convergence gets arbitrarily close to $\rho = \max\{\lambda, \beta\}$ as $m \to \infty$. This completes the proof. □

Chapter 5

Examples

This chapter considers two examples from Bayesian statistics. Both are *two-variable Gibbs samplers.* The results of Chapter 4 provide theoretical upper bounds on the time to convergence. It will be seen that these bounds are quite conservative compared with the actual behavior of the Markov chains, but still reasonable enough to be used in practice.

The Gibbs sampler was named by Geman and Geman [GG84], though the idea is implicit in earlier works such as [Met+53]. See [CG92] for an introduction and [DKSC08] for a brief history along with many references. It is used in the following very common situation. One would like to sample from a probability measure π on a multidimensional space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Sampling directly from π is intractable; but if $1 \leq j \leq d$ and $a_1 \in \mathcal{X}_1, \ldots, a_{j-1} \in \mathcal{X}_{j-1}, a_{j+1} \in \mathcal{X}_{j+1}, \ldots, a_d \in \mathcal{X}_d$ are given, one can sample from the one-variable conditional distribution $\pi_j(dx_j \mid x_1 = a_1, \ldots, x_{j-1} = a_{j-1}, x_{j+1} = a_{j+1}, \ldots, x_d = a_d)$. The Gibbs sampler is a Markov chain (X_t) on $\mathcal{X}$ with the following transition rule. Given that $X_t = (a_1, \ldots, a_d)$, sample b_1 from $\pi_1(dx_1 \mid x_2 = a_2, \ldots, x_d = a_d)$. This is referred to as "updating the first coordinate." Now sample b_2 from $\pi_2(dx_2 \mid x_1 = b_1, x_3 = a_3, \ldots, x_d = a_d)$. Continue in this fashion until b_d has been chosen from $\pi_d(dx_d \mid x_1 = b_1, \ldots, x_{d-1} = b_{d-1})$. Then set $X_{t+1} = (b_1, \ldots, b_d)$. The chain (X_t) is called the *Gibbs sampler.* It has π as a stationary distribution, and is ergodic under conditions that are easily verified in practical examples.

Variations of the method above are possible. For example, rather than iterating through the dimensions in order, at each step one may update a coordinate $1 \leq j \leq d$ chosen uniformly at random. This is called the *random scan Gibbs sampler*, in contrast to the *deterministic scan Gibbs sampler* described in the previous paragraph. Another variation, called *Metropolis-within-Gibbs*, is useful when it is impractical to sample directly even from the one-variable conditional distributions. In that case one replaces each step "choose b_j from the distribution $\pi_j(dx_j \mid x_1 = b_1, \ldots, x_{j-1} = b_{j-1}, x_{j+1} = a_{j+1}, \ldots, x_d = a_d)$" with an instance of the Metropolis–Hastings algorithm [Met+53].

This is actually the way that Metropolis et al introduced their algorithm in the aforementioned paper.

For the examples in this chapter, it is possible to sample directly from the one-variable conditional distributions, so Metropolis-within-Gibbs is unnecessary. In addition, the number of variables can be chosen to be $d = 2$. For instance, Section 5.1 considers a measure π on $\mathbf{R}^{11}$ where the last 10 coordinates $x_2, \dots, x_{11}$ are conditionally independent given x_1. Therefore, if $y_1 = x_1$ and $y_2 = (x_2, \dots, x_{11})$, sequentially updating the coordinates $x_2, \dots, x_{11}$ is the same as updating y_2 according to its conditional density on $\mathbf{R}^{10}$ given y_1. It follows that the 11-variable deterministic scan Gibbs sampler (X_t) on $\mathbf{R}^{11}$ is identical to the 2-variable deterministic scan Gibbs sampler (Y_t) on $\mathbf{R} \times \mathbf{R}^{10}$. As discussed below, reducing the number of variables to two greatly simplifies the analysis.

In order for Gibbs sampling to be useful, the convergence to the stationary distribution π must be reasonably quick. In many papers it is proved that particular Gibbs samplers are geometrically ergodic, from which it follows that the convergence happens at an exponential rate with non-explicit constants. The Ph.D. thesis of Johnson [Joh09] covers the topic in depth, and the survey [JJN13] has many references. Since the publication of [JJN13], papers that prove geometric ergodicity for Gibbs samplers include [KH13; PK14; Fit14; LDM15; RH15; ALV15; JJ15; JB15]. A smaller number of papers have found explicit convergence rates: see [Ros95a; Ros95b; JH01; JH04; DKSC08].

The papers cited above deal mostly with two-variable Gibbs samplers, which are easier to analyze than the general case of three or more variables. This is mainly due to the following observation. If $(X_t) = (A_t, B_t)$ is a two-variable (deterministic scan) Gibbs sampler, the updating proceeds as follows:

$$\cdots \to (A_t, B_t) \to (A_{t+1}, B_t) \to (A_{t+1}, B_{t+1}) \to (A_{t+2}, B_{t+1}) \to \cdots$$

By construction, A_{t+1} depends on $X_t = (A_t, B_t)$ only through B_t, and then B_{t+1} depends on (A_{t+1}, B_t) only through A_{t+1}. Therefore, the projections (A_t) and (B_t) are themselves Markov chains. If $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ is the state space, define transition kernels $Q_1(b, dx_1) = \pi_1(dx_1 \mid x_2 = b)$ and $Q_2(a, dx_2) = \pi_2(dx_2 \mid x_1 = a)$. Then Q_2Q_1 is the transition kernel for (A_t), and Q_1Q_2 is the transition kernel for (B_t).

The convergence rate of (X_t) to stationarity is essentially the same as the convergence rates of (A_t) and (B_t). Specifically, suppose μ is the initial distribution for (X_t), with projections μ_1, μ_2 onto the two coordinates. It can be checked by coupling that for each $t \geq 1$,

$$\begin{gathered}\|P^t(\mu, \cdot) - \pi\|_{\mathrm{TV}} = \|(Q_2Q_1)^{t-1}(\mu_2 Q_1, \cdot) - \pi_1\|_{\mathrm{TV}}, \\ \|(Q_1Q_2)^{t-1}(\mu_2, \cdot) - \pi_2\|_{\mathrm{TV}} \geq \|P^t(\mu, \cdot) - \pi\|_{\mathrm{TV}} \geq \|(Q_1Q_2)^t(\mu_2, \cdot) - \pi_2\|_{\mathrm{TV}}.\end{gathered}$$

Another feature of deterministic scan two-variable Gibbs samplers is that the projection chains in the first and second variable ((A_t) and (B_t) above) are reversible with nonnegative eigenvalues [Bax05]. That means Theorem 4.4 can find explicit convergence rates for deterministic scan two-variable Gibbs samplers. In three or more variables, all that can be said is that the random scan Gibbs sampler is reversible with nonnegative eigenvalues [LWK95].

5.1 Nuclear pump example

This is a well-studied toy example dating to the 1980s. Table 5.1 shows pump failures at the Farley 1 nuclear plant in Columbia, AL. Each pump failed a certain number of times over a certain number of hours of operation.

Pump	# Failures	# Hours
1	5	94,320
2	1	15,720
3	5	62,880
4	14	125,760
5	3	5,240
6	19	31,440
7	1	1,048
8	1	1,048
9	4	2,096
10	22	10,480

Table 5.1: Failure data for pumps at the Farley 1 nuclear plant.

These data were first analyzed by [GO87]. The specific model used here was first suggested by [GS90] and later used by [Tie94; MTY95; Ros95a]. It is a Bayesian hierarchical model with three stages. At the first stage, each pump's failures are taken to be independent Poisson processes, with rate $\theta^{(j)}$ for pump j. (The units for $\theta^{(j)}$ are failures per thousand hours.) At the second stage, the rates $\theta^{(j)}$ are drawn independently from a Gamma distribution $G(\alpha,\beta)$. (The density function is $G_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.) At the third stage, $\alpha = 1.802$ is fixed while β is drawn from a $G(0.01,1)$ distribution. Let s_j be the number of failures of pump j, and let t_j be its length of operation in thousands of hours, as given in the table. Also define $\theta = (\theta^{(1)},\dots,\theta^{(10)})$. The model is:

$$\begin{aligned} s_j \mid \theta,\beta &\sim \text{Poisson}(t_j\theta^{(j)}), \text{ independently for each } j \\ \theta^{(j)} \mid \beta &\sim G(1.802,\beta), \text{ independently for each } j \\ \beta &\sim G(0.01,1) \end{aligned}$$

The choice $\alpha = 1.802$ comes from the method of moments estimator.

There are eleven unknown parameters: β and the $\theta^{(j)}$. Let $S = \sum_{j=1}^{10} \theta^{(j)}$. For convenience, let

$d = (s_1, \ldots, s_{10}; t_1, \ldots, t_{10})$ denote the pump failure data. The one-variable conditional densities (see e.g. [Tie94; Ros95a]) are:

$$\beta \mid \theta, d \sim G(0.01 + 10\alpha, 1 + S)$$
$$\theta^{(j)} \mid \beta, d \sim G(\alpha + s_j, \beta + t_j), \text{ independently for each } j$$

Define a Gibbs sampler based on these conditional densities, with the updating rule

$$\cdots \to (\beta_t, \theta_t) \to (\beta_{t+1}, \theta_t) \to (\beta_{t+1}, \theta_{t+1}) \to (\beta_{t+2}, \theta_{t+1}) \to \cdots$$

Since the coordinates of θ are conditionally independent given β, it makes no difference whether they are updated sequentially or all at once. Let $(X_t) = (\beta_t, \theta_t)$, with stationary distribution π and transition kernel P. The goal is to find upper bounds for $\|P^t(\mu, \cdot) - \pi\|_{\text{TV}}$.

The choice to fix $\alpha = 1.802$ while drawing β from the "hyperprior" distribution $G(0.01, 1)$ may seem arbitrary. Why not also draw α from a hyperprior distribution? In that case one would be faced with a three-variable Gibbs sampler, complicating the analysis. On the other hand, if both α and β were fixed, no Gibbs sampling would be necessary. The model described above was deliberately chosen to be as simple as possible while remaining nontrivial.

As discussed earlier, both (β_t) and (θ_t) are Markov chains. In this example, a further simplification is possible. Since β_{t+1} depends on θ_t only through the value of $S_t = \sum_{j=1}^{10} \theta_t^{(j)}$, it follows that (S_t) is a Markov chain (reversible, with nonnegative eigenvalues). Let (S_t) have transition kernel Q_S and stationary distribution π_S. If μ_S is the projection onto the S variable of the initial distribution μ, for all $t \geq 1$,

$$\|Q_S^{t-1}(\mu_S, \cdot) - \pi_S\|_{\text{TV}} \geq \|P^t(\mu, \cdot) - \pi\|_{\text{TV}} \geq \|Q_S^t(\mu_S, \cdot) - \pi_S\|_{\text{TV}}.$$

Rosenthal [Ros95a] obtains numerical bounds on the convergence of (X_t) by showing that (S_t) satisfies a drift and minorization condition. (Specifically, he uses the bivariate drift approach described in Chapter 2.) He proves the following:

Theorem 5.1 (Theorem 11 of [Ros95a]). *Let P be the transition kernel for the nuclear pump Gibbs sampler, and let π be its stationary distribution. For any initial distribution μ on $\mathbf{R}^{11}$,*

$$\|P^t(\mu, \cdot) - \pi\|_{\text{TV}} \leq (0.976)^t + (0.951)^t(6.2 + \mathbf{E}_\mu[(S_0 - 6.5)^2]).$$

In particular, if the initial distribution μ is set so that $S_0 = 6.5$,

$$\|P^t(\mu, \cdot) - \pi\|_{\text{TV}} \leq (0.976)^t + 6.2 \cdot (0.951)^t.$$

The time until the total variation distance drops below 0.01 is $t_{0.01} \leq 192$.

Theorem 4.4 in this thesis gives the following improved bound.

Theorem 5.2. *Let P be the transition kernel for the nuclear pump Gibbs sampler, and let π be its stationary distribution. For any initial distribution μ on $\mathbf{R}^{11}$,*

$$\|P^{t+1}(\mu, \cdot) - \pi\|_{\mathrm{TV}} \leq [2.16 + 0.19t(1 + \mathbf{E}_\mu[(S_0 - 6.5)^2])^{0.19}] \cdot (0.914)^t.$$

In particular, if the initial distribution μ is set so that $S_0 = 6.5$,

$$\|P^{t+1}(\mu, \cdot) - \pi\|_{\mathrm{TV}} \leq (2.16 + 0.19t) \cdot (0.914)^t.$$

The time until the total variation distance drops below 0.01 is $t_{0.01} \leq 85$.

Although Theorem 5.2 improves on Theorem 5.1 by more than a factor of two, it does not come close to the actual convergence rate of the Markov chain. A non-rigorous approach described below indicates that if the (S_t) chain is started from $S_0 = 6.5$, the number of steps until the total variation distance from stationarity drops below 0.01 is likely $t_{0.01} = 2$. Thus the corresponding number of steps for the Gibbs sampler is either $t_{0.01} = 2$ or $t_{0.01} = 3$. Possible reasons for this disparity will be discussed at the end of this section.

Proof of Theorem 5.2. The argument is very similar to the proof of Theorem 5.1 in [Ros95a]. This is by design, so that the difference in final estimates ($t_{0.01} \leq 192$ versus $t_{0.01} \leq 85$) is due to improvements in the theoretical machinery rather than better estimates for this specific example.

The main idea is to show that the chain (S_t) satisfies a drift and minorization condition with explicit constants. In [Ros95a] it is observed that the stationary distribution for (S_t) is concentrated near 6.5. For this reason, Rosenthal chooses the bivariate drift function $W(x, y) = 1 + (x - 6.5)^2 + (y - 6.5)^2$. The corresponding univariate drift function $V(x) = 1 + (x - 6.5)^2$ will be used in this proof.

Let $Q = Q_S$ be the transition kernel for (S_t). Once $V(x)$ has been chosen, the function $QV(x)$ can be computed numerically. (See [Ros95a] for details.) Figure 5.1 shows the curves $y = QV(x)$ along with $y = \lambda V(x)$ for values $\lambda = 0.8, 0.6, 0.4$. It should be noted that the behavior of $QV(x)$ as $x \to \infty$ is also known: $\lim_{x\to\infty} QV(x) \approx 43.2$.

For each value of λ, the optimal small set C is the region of x-values for which $QV(x) > \lambda V(x)$. When $\lambda = 0.8$ this is roughly the interval $[5, 8]$, and when $\lambda = 0.6$ it is a slightly wider interval. When $\lambda = 0.4$ the optimal set C is the union of two intervals: one centered at $x = 6.5$ (roughly $[4, 9]$) and another very small one near $x = 0$ (roughly $[0, 0.1]$).

Once the set C has been identified, the minorization mass can be computed according to the method

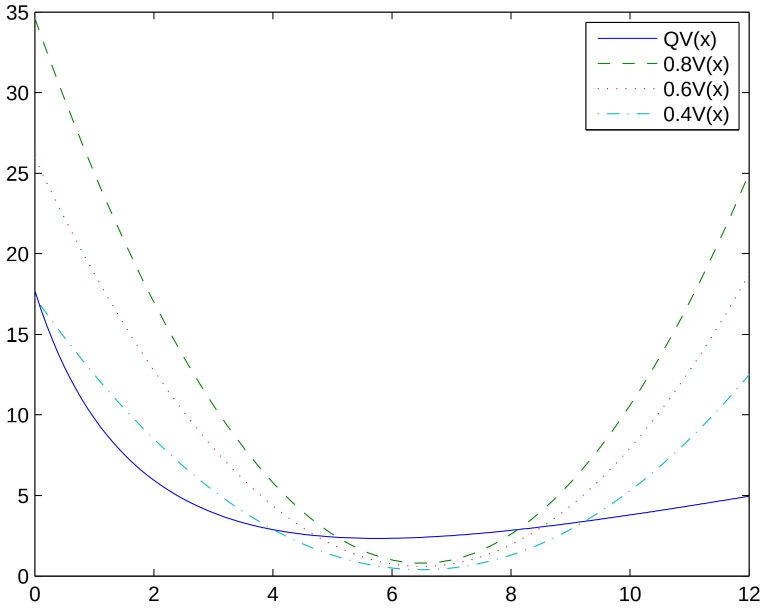

Figure 5.1: Graphs of $y = QV(x)$ and $y = \lambda V(x)$ where $\lambda = 0.8, 0.6, 0.4$ for the nuclear pump chain. Here Q is the transition kernel for the chain (S_t) and $V(x) = 1 + (x - 6.5)^2$.

used by [Ros95a]. Since the Gibbs sampler updates in the following order:

$$\cdots \to S_t \to \beta_{t+1} \to S_{t+1} \to \beta_{t+2} \to S_{t+2} \to \cdots$$

one can find a lower bound for the minorization mass using only the transition $S_t \to \beta_{t+1}$. Specifically, for $c \in \mathbf{R}$ let $f_c(x)$ be the probability density function for β_{t+1} given that $S_t = c$, and let $g_c(x)$ be the probability density function for S_{t+1} given that $S_t = c$. The optimal minorization mass for the small set C is

$$\varepsilon_{\text{opt}} = \int_0^\infty \left(\inf_{c \in C} g_c(x) \right) dx.$$

Unfortunately, ε_{opt} is somewhat difficult to compute. For this reason, [Ros95a] uses instead the minorization mass for β_{t+1}, namely

$$\varepsilon = \int_0^\infty \left(\inf_{c \in C} f_c(x) \right) dx. \tag{5.1}$$

It is immediate from a coupling argument that $\varepsilon \le \varepsilon_{\text{opt}}$. As well, ε can be computed quickly with the help of the following observation. One has

$$f_c(x) = G_{18.03,1+c}(x) = \frac{(1+c)^{18.03}}{\Gamma(18.03)} x^{17.03} e^{-(1+c)x}. \tag{5.2}$$

Holding x constant, this is log-concave as a function of c. Thus if $c_{\min} = \inf(C)$ and $c_{\max} = \sup(C)$, for all x,

$$\inf_{c \in C} f_c(x) = \min\{f_{c_{\min}}(x), f_{c_{\max}}(x)\}.$$

This equality is illustrated by Figure 5.2, which shows $f_c(x)$ for different values of c. The lowest curve at each x-value is either the curve corresponding to the smallest value of c or the curve corresponding to the largest value of c.

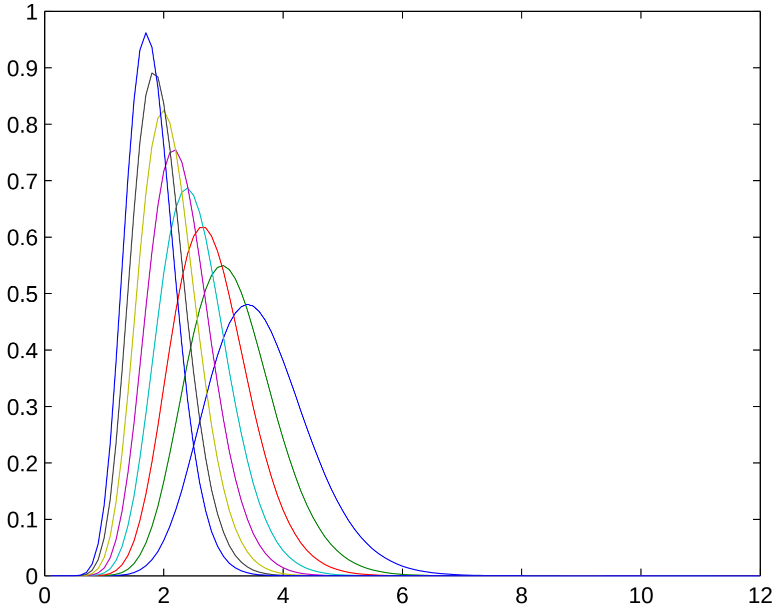

Figure 5.2: Graphs of $y = f_c(x)$ as defined in (5.2) for values of c between 4 and 9.

Having made one arbitrary choice, namely the drift function $V(x) = 1+(x-6.5)^2$, and one simplifying bound, namely the use of ε instead of ε_{opt}, the rest of the computation can be optimized. For each choice of the drift constant λ one can find the ideal small set C. Then, one computes the constant $K = \sup_{c \in C} QV(c)$ and the minorization mass ε. Since the transition kernel Q satisfies a drift and minorization condition, and Q is reversible with nonnegative eigenvalues, Theorem 4.4 gives explicit convergence bounds. The exponential convergence rate λ_* will vary depending on the choice of λ, as shown in Figure 5.3. The horizontal axis is λ, and the vertical axis is the values of ε and λ_*

determined by λ. Note the steep drop in the value of ε as λ decreases past 0.4, reading from right to left. This is because the small set C starts to include values of x near zero, as shown in Figure 5.1. (Though Figure 5.3 seems to show a sharp linear drop, it is actually a discontinuity.)

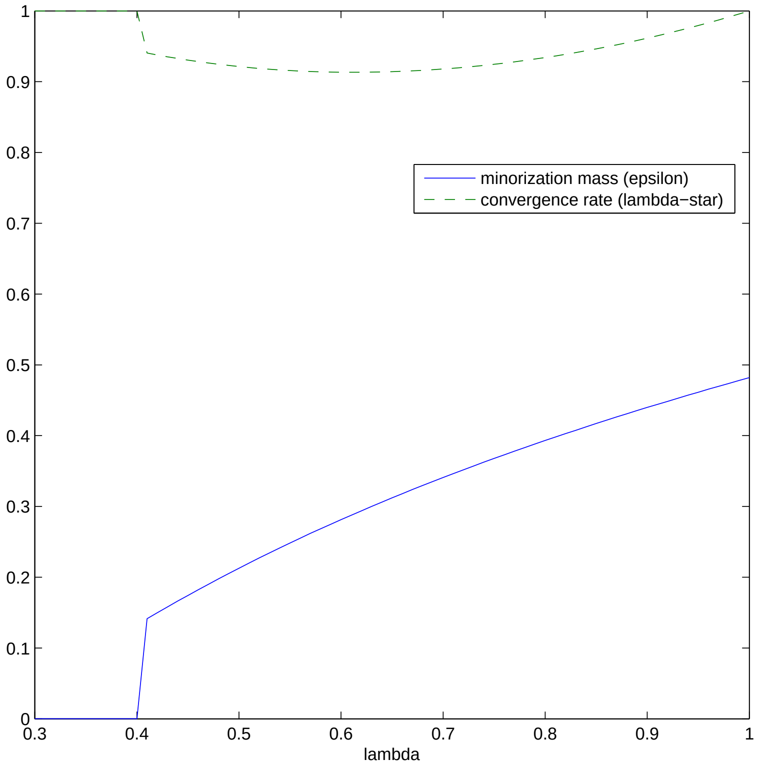

Figure 5.3: Graphs of the minorization mass ε and convergence rate λ_* as functions of the drift parameter λ for the nuclear pump chain. For each λ, the optimal small set C associated with λ and the drift function $V(x) = 1 + (x - 6.5)^2$ is computed as in Figure 5.1. Then ε is computed using (5.1) and λ_* is computed using Theorem 4.4.

Choosing $\lambda = 0.61$ minimizes λ_*, which gives the best convergence bounds. The method described above yields the following drift and minorization data:

$$C = [4.74, 8.50], \qquad K = 3.05, \qquad m = 1, \qquad \varepsilon = 0.287, \qquad \lambda_* = 0.914. \tag{5.3}$$

Then Theorem 4.4 gives the bound

$$\|Q^t(\mu_S, \cdot) - \pi_S\|_{\mathrm{TV}} \le [2.16 + 0.19t(1 + \mathbf{E}_\mu[(S_0 - 6.5)^2])^{0.19}] \cdot (0.914)^t.$$

Since it was already seen that $\|P^{t+1}(\mu,\cdot)-\pi\|_{\mathrm{TV}} \leq \|Q^t(\mu_S,\cdot)-\pi_S\|_{\mathrm{TV}}$, this completes the proof. □

When started at $S_0 = 6.5$, the chain (S_t) satisfies $t_{0.01} \leq 84$. For comparison, plugging the drift and minorization data (5.3) into Theorem 1.3 of [Bax05] yields

$$\|Q^t(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} \leq 102651 \cdot (0.927)^t,$$

which gives $t_{0.01} \leq 211$. Therefore Theorem 4.4 outperforms the analogous results of both [Ros95a] and [Bax05].

While the bound of 84 steps is a significant improvement, it is still an overestimate compared with the true behavior of the Markov chain. The following non-rigorous argument indicates that actually, $t_{0.01} = 2$.

Let $F_t(x)$ be the cumulative distribution function for the law of S_t given that $S_0 = 6.5$. Also, let $F(x)$ be the cumulative distribution function for the stationary distribution π_S of (S_t). The total variation distance $\|Q^t(6.5,\cdot)-\pi_S\|_{\mathrm{TV}}$ equals one-half the total variation of the function $F_t(x)-F(x)$.

To estimate $F_t(x)$, $N = 10^8$ independent instances of the Markov chain (S_t) started at $S_0 = 6.5$ were simulated. Let $\hat{F}_t(x)$ be the empirical cumulative distribution function determined by the simulation. The DKW inequality (originally proved in [DKW56], and with optimal constant in [Mas90]) says that for each t,

$$\mathbf{P}\left(\sup_{x\in\mathbf{R}} |\hat{F}_t(x) - F_t(x)| > 2 \cdot 10^{-4}\right) \leq 2e^{-8}. \tag{5.4}$$

Therefore the functions $F_t(x)$ can be well-estimated pointwise. The function $F(x)$ will be estimated using $F_T(x)$ for $T = 140$, since it has been proved that

$$\|Q^{140}(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} \leq 10^{-4}.$$

(It would also be possible to estimate $F(x)$ using perfect sampling techniques. Example 6.2.1 in the recent book of Huber [Hub16] describes how to implement the "multigamma coupler" of Murdoch and Green [MG98] in this exact example.)

The reason this approach is non-rigorous is that pointwise control over the functions $F_t(x)$ says nothing in principle about their total variation. In theory they could have tiny sharp oscillations. In practice this does not seem to be the case. Figure 5.4 shows the curves $y = \hat{F}_t(x) - \hat{F}_{140}(x)$ for $t = 1, 2, 3$.

Each curve appears smooth, with clearly defined minima and maxima. For instance, the curve $y = \hat{F}_1(x) - \hat{F}_{140}(x)$ seems to decrease from 0 to a minimum of about -0.024. It then rises to a maximum of about 0.006 before decreasing back to 0. The total variation is estimated to be

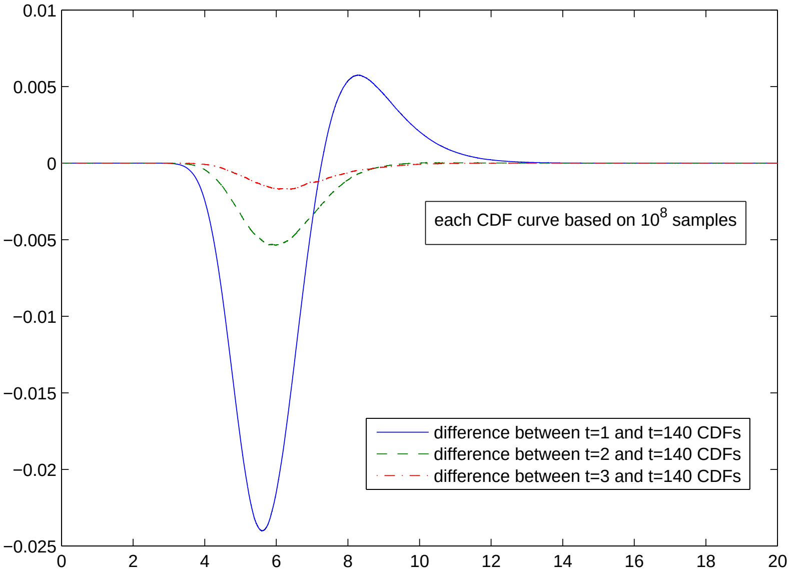

Figure 5.4: Graphs of $y = \hat{F}_t(x) - \hat{F}_{140}(x)$ for $t = 1, 2, 3$. If $(S_t^{(j)})$ for $1 \leq j \leq 10^8$ are independent instances of the chain (S_t) started at $S_0 = 6.5$, then $\hat{F}_{t_0}(x) = 10^{-8} \times \#\{1 \leq j \leq 10^8 : S_{t_0}^{(j)} \leq x\}$.

$(0 - -0.024) + (0.006 - -0.024) + (0.006 - 0) = 0.060$. Assume that the actual difference of CDF curves $y = F_1(x) - F_{140}(x)$ follows the same pattern: decrease to a minimum, increase to a maximum, decrease back to zero. The DKW inequality (5.4) says that with probability at least $1 - 4e^{-8}$, the minimum value is in the range $-0.024 \pm 4 \cdot 10^{-4}$, and the maximum value is in the range $0.006 \pm 4 \cdot 10^{-4}$. Therefore the total variation is $0.060 \pm 16 \cdot 10^{-4}$, which means that

$$\|Q^1(6.5, \cdot) - Q^{140}(6.5, \cdot)\|_{\mathrm{TV}} = 0.030 \pm 8 \cdot 10^{-4}$$

with probability at least $1 - 4e^{-8}$. Since $\|Q^{140}(6.5, \cdot) - \pi_S\|_{\mathrm{TV}} \leq 1 \cdot 10^{-4}$,

$$\|Q^1(6.5, \cdot) - \pi_S\|_{\mathrm{TV}} = 0.030 \pm 9 \cdot 10^{-4}.$$

The error $9 \cdot 10^{-4}$ is slightly less than 0.001. This is the reasoning behind the first estimate below; the other estimates are made by the same method, and also hold with probability at least $1 - 4e^{-8}$

(under the same smoothness assumptions).

$$\begin{aligned}\|Q^1(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} &= 0.030 \pm 0.001,\\ \|Q^2(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} &= 0.005 \pm 0.001,\\ \|Q^3(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} &= 0.002 \pm 0.001.\end{aligned}$$

This strongly indicates that the minimum number of steps t for which $\|Q^t(6.5,\cdot) - \pi_S\|_{\mathrm{TV}} \leq 0.01$ is $t_{0.01} = 2$.

It is worth considering why the method of drift and minorization produces such conservative estimates. There are three potential reasons:

1. Poor choice of the drift function $V(x)$. This was carried over from the analysis of [Ros95a] but is certainly not optimal. A different approach would be to fix the small set C and try to find a good drift function $V(x)$. Lemma 4.7 shows that the best possible drift function is $V(x) = \mathbf{E}_x[\lambda^{-\tau_C}]$, where λ is chosen as small as possible so that the expectation is finite. (It is known that for fixed λ, the quantity $\mathbf{E}_x[\lambda^{-\tau_C}]$ is either finite for all $x \notin C$ or infinite for all such x.) This leaves open the questions of how to choose C and how to determine rigorously whether $\mathbf{E}_x[\lambda^{-\tau_C}]$ is finite for particular choices of C and λ.

2. Underestimation of the minorization mass ε. As discussed in the proof of the convergence bound, the true minorization mass $\varepsilon_{\mathrm{opt}}$ may be much larger than the computed minorization mass. Better computation of this quantity would probably improve the overall bound significantly.

3. Weakness in the general drift-minorization paradigm. The overall idea is to find a strong ν time T and then observe that the Markov chain converges to stationarity on the same time scale as it reaches T. There is no guarantee that this approach will come near the true convergence rate for any given Markov chain, especially since the form of the strong ν time is constrained.

5.2 HMO premium example

This is a data set of the premiums charged by 341 American health plans in the early 1990s. Each plan was located in a particular US state. In [Hod98], Hodges proposed a hierarchical model to predict the premium for each plan based on two variables: the average cost of a hospital admission in its state, and an indicator for whether the plan was in New England. (These variables were chosen from a slightly longer list; see [Hod98] for details.) The convergence of the resulting Gibbs sampler has been studied by Johnson and Jones [JJ10], whose notation is followed here.

Figure 5.5 shows the data along with two best fit lines, produced by least squares using the following model:

$$\text{(premium)} = c_0 + c_1\text{(hospital cost in state)} + c_2\text{(New England indicator variable)} + \text{(error)}$$

Thus the lines are constrained to be parallel.

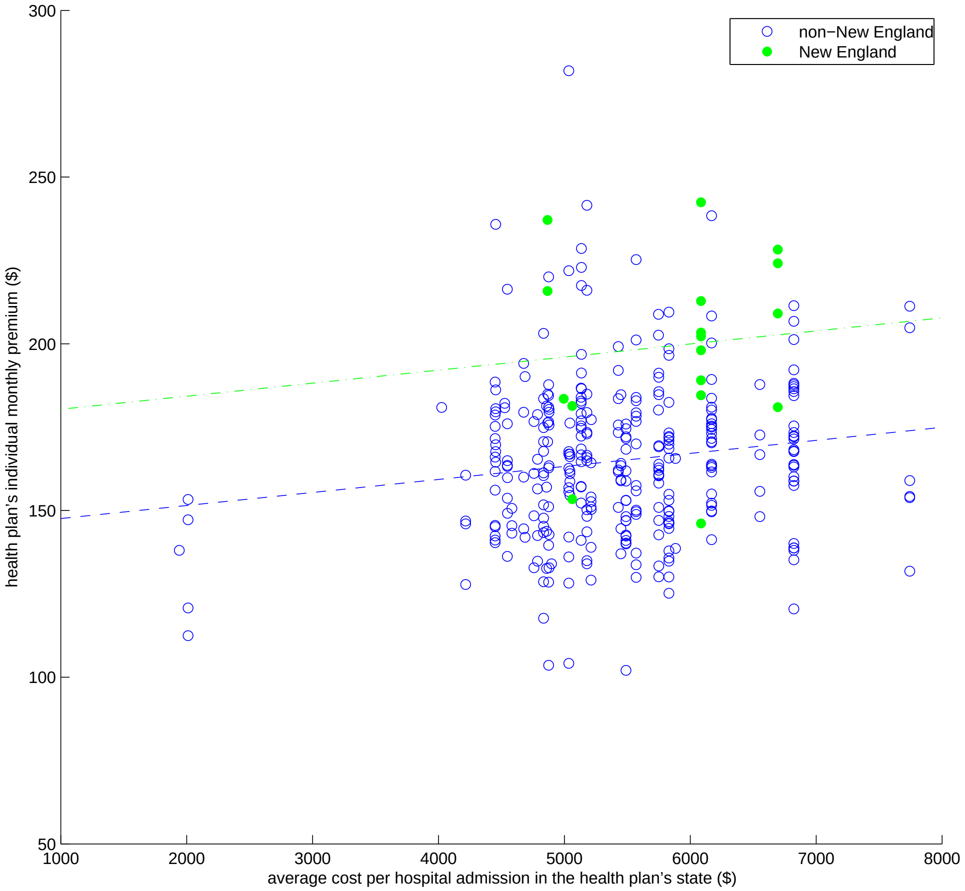

Figure 5.5: HMO premium data. Each circle represents a health plan. Its location on the horizontal axis is the average cost per hospital admission in the plan's state, and its location on the vertical axis is the plan's individual monthly premium. Hollow blue circles represent plans outside of New England, and solid green circles represent plans in New England.

Set $N = 341$. Let y be an $N \times 1$ vector whose jth entry is the individual monthly premium of health plan number j. Let X be an $N \times 3$ matrix with columns numbered $0, 1, 2$, defined as follows: $X(j, 0) = 1$; $X(j, 1)$ is a normalized version of the average cost per hospital admission in the state where health plan j is located; $X(j,2) = 1$ if health plan j is in New England and $X(j,2) = 0$ otherwise. The exact formula for $X(j,1)$ is this: if $\bar{x}_1$ is the national average cost per hospital admission, and $\tilde{x}_{j1}$ is the average cost per hospital admission in the state of plan j, then $X(j,1) = (\tilde{x}_{j1} - \bar{x}_1)/1000$. This is all consistent with [JJ10]. A model for these data is:

$$y_j = \beta_0 + \beta_1 X(j,1) + \beta_2 X(j,2) + \varepsilon_j, \tag{5.5}$$

where the ε_j are independent mean-zero normal variables with variance λ_R^{-1}. Let $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ be the 3×1 vector of parameters. A least-squares regression would minimize the L^2 distance $\|y - X\beta\|_2$. Following [Hod98], Johnson and Jones [JJ10] propose a Bayesian hierarchical version of the model:

$$\begin{aligned} y \mid \beta, \lambda_R &\sim N_N(X\beta, \lambda_R^{-1} I_N) \\ \beta \mid \lambda_R &\sim N_3(b, B^{-1}) \\ \lambda_R &\sim G(r_1, r_2) \end{aligned}$$

where N_k represents the multivariable Gaussian distribution in k dimensions, I_k is the $k \times k$ identity matrix, and b, B^{-1}, r_1, r_2 are hyperparameters taking the values

$$b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad B^{-1} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}, \qquad r_1 = 3.122 \times 10^{-6}, \qquad r_2 = 0.00177.$$

The values of r_1 and r_2 are chosen using the method of moments; see [JJ10].

There are four unknown parameters: λ_R and the three coordinates of β. As computed by [JJ10], here are the conditional laws given the data:

$$\begin{aligned} &\lambda_R \mid \beta, y \sim G\left(r_1 + \frac{1}{2}N, r_2 + \frac{1}{2}\|y - X\beta\|_2^2\right) \\ &\beta \mid \lambda_R, y \sim N_3(m, \Sigma^{-1}) \end{aligned}$$

where

$$\Sigma^{-1} = (\lambda_R X^T X + B)^{-1}, \qquad m = \Sigma^{-1}(\lambda_R X^T y + Bb).$$

The Gibbs sampler proposed by [JJ10] updates first λ_R and then β according to the laws above. As in the nuclear pump example, λ_R depends on β only through the value $E = \|y - X\beta\|_2^2$. Therefore the Markov chain (E_t) controls the convergence of the full Gibbs sampler $(X_t) = ((\lambda_R)_t, \beta_t)$. As well, (E_t) is reversible with nonnegative eigenvalues. Let (X_t) have transition kernel P and stationary distribution π; let (E_t) have transition kernel Q and stationary distribution π_E. If μ_E is the projection onto the E variable of the initial distribution μ of (X_t), then for all $t \geq 1$,

$$\|Q^{t-1}(\mu_E, \cdot) - \pi_E\|_{\mathrm{TV}} \geq \|P^t(\mu, \cdot) - \pi\|_{\mathrm{TV}} \geq \|Q^t(\mu_E, \cdot) - \pi_E\|_{\mathrm{TV}}.$$

The smallest possible value of E is reached when $\beta = \hat{\beta}$, the least-squares estimate in (5.5). This gives a value of $E_{\min} \approx 191240$. The state space of the chain (E_t) can be taken as $[E_{\min}, \infty)$.

It is proved in [JJ10] that the Gibbs sampler (X_t) is geometrically ergodic, but without rigorous quantitative control on its convergence rate. This section applies Theorem 4.4 to the (E_t) chain in order to prove the following bound.

Theorem 5.3. *Let P be the transition kernel for the HMO data Gibbs sampler, and let π be its stationary distribution. For any initial distribution μ on $\mathbf{R}^4$,*

$$\|P^{t+1}(\mu, \cdot) - \pi\|_{\mathrm{TV}} \leq [2.65 + 0.25t(1 + \mathbf{E}_\mu[E_0 - E_{\min}])^{0.10}] \cdot (0.909)^t.$$

In particular, if the initial distribution μ is set so that $E_0 = E_{\min}$,

$$\|P^{t+1}(\mu, \cdot) - \pi\|_{\mathrm{TV}} \leq (2.65 + 0.25t) \cdot (0.909)^t.$$

The time until the total variation distance drops below 0.01 is $t_{0.01} \leq 83$.

As in the nuclear pump example, this bound may be very conservative. A non-rigorous analysis of the actual convergence rate, like for the nuclear pump chain, suggests that the actual mixing time for the (E_t) chain started at $E_0 = E_{\min}$ is likely $t_{0.01} = 1$ or 2.

Along with the reasons discussed in the previous section for this disparity, another one comes into play here. The main convergence bound, Theorem 4.4, is proved using Lemma 4.7, which states that $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x)$. In other words, the value of the drift function at x controls the tail of the hitting time τ_C for the Markov chain started at x. The proof relies on the fact that $V(X_t)$ tends to decrease as long as $X_t \notin C$, and it cannot drop below 1. In this example, the optimal small set C is roughly the set $\{x : V(x) \leq 14600\}$. Therefore if $V(X_t)$ drops below 14600, it is guaranteed that $X_t \in C$ and so $\tau_C \leq t$. A tightened version of Lemma 4.7 taking this extra information into account would probably lead to an overall improvement in the convergence bound produced by Theorem 4.4.

Proof of Theorem 5.3. To prove a drift and minorization condition for the chain (E_t), the drift function $V(x) = 1 + x - E_{\min}$ will be used. Although it may seem that this drift function is linear, as compared to the quadratic drift function in the nuclear pump example, keep in mind that the quantity $E_t = \|y - X\beta_t\|_2^2$ is itself quadratic.

The value $QV(x)$ can be computed numerically. As a first step, the conditional distribution of $y - X\beta$ given λ_R and y is multivariate normal with mean $y - Xm$ and covariance matrix $X\Sigma^{-1}X^T$. Therefore,

$$\mathbf{E}\left[\|y - X\beta\|_2^2 \,\middle|\, \lambda_R, y\right] = \|y - Xm\|_2^2 + \operatorname{tr}(X\Sigma^{-1}X^T).$$

Note that both m and Σ^{-1} depend on λ_R. Since $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, one has

$$y - Xm = (I_N - \lambda_R X\Sigma^{-1}X^T)y.$$

Thus both terms $\|y - Xm\|_2^2$ and $\operatorname{tr}(X\Sigma^{-1}X^T)$ can be computed by diagonalizing $X\Sigma^{-1}X^T$, which is an $N \times N$ positive semidefinite symmetric matrix of rank 3. To do this, recall that $B = \frac{1}{100}I_3$ is a scalar multiple of the identity matrix and that $\Sigma^{-1} = (\lambda_R X^T X + B)^{-1}$. Let $X^T X$ have orthonormal eigenvectors v_1, v_2, v_3 with eigenvalues ρ_1, ρ_2, ρ_3. Then Σ^{-1} has eigenvectors $\{v_i\}$ with eigenvalues $\{(\lambda_R\rho_i + \frac{1}{100})^{-1}\}$. It follows that $X\Sigma^{-1}X^T$ has eigenvectors $\{Xv_i\}$ with eigenvalues $\{\rho_i(\lambda_R\rho_i + \frac{1}{100})^{-1}\}$, and an $(N-3)$-dimensional nullspace which is the orthogonal complement of the linear span of the vectors Xv_i. Hence

$$\operatorname{tr}(X\Sigma^{-1}X^T) = \sum_{i=1}^{3} \frac{\rho_i}{\lambda_R\rho_i + \frac{1}{100}}.$$

To compute $\|y - Xm\|_2^2$, let $c_i = \frac{1}{\rho_i}(y, Xv_i)$ for $i = 1, 2, 3$. The factor of $1/\rho_i$ is because $\|Xv_i\|_2^2 = \rho_i$, so one has the orthogonal decomposition

$$y = \tilde{y} + \sum_{i=1}^{3} c_i X v_i,$$

where $(X\Sigma^{-1}X^T)\tilde{y} = 0$. The least-squares estimate $\hat{\beta} = \sum_{i=1}^{3} c_i v_i$, and the remainder $\tilde{y}$ satisfies $\|\tilde{y}\|_2^2 = E_{\min}$. Now,

$$\begin{aligned} y - Xm &= (I_N - \lambda_R X\Sigma^{-1}X^T)\left(\tilde{y} + \sum_{i=1}^{3} c_i X v_i\right) \\ &= \tilde{y} + \sum_{i=1}^{3}\left(1 - \frac{\lambda_R\rho_i}{\lambda_R\rho_i + \frac{1}{100}}\right) c_i X v_i, \end{aligned}$$

which yields

$$\|y - Xm\|_2^2 = \|\tilde{y}\|_2^2 + \sum_{i=1}^{3}\left(1 - \frac{\lambda_R\rho_i}{\lambda_R\rho_i + \frac{1}{100}}\right)^2 c_i^2\rho_i = E_{\min} + \sum_{i=1}^{3}\left(\frac{\frac{1}{100}}{\lambda_R\rho_i + \frac{1}{100}}\right)^2 c_i^2\rho_i.$$

This completes the computation of the conditional expectation of $\|y - X\beta\|_2^2$ given λ_R and y. Now,

$$\begin{aligned}
QV(x) &= \mathbf{E}[V(E_{t+1}) \mid E_t = x] \\
&= 1 - E_{\min} + \int_0^\infty \mathbf{E}[E_{t+1} \mid (\lambda_R)_{t+1} = s]\, \mathbf{P}((\lambda_R)_{t+1} \in ds \mid E_t = x) \\
&= 1 - E_{\min} + \int_0^\infty \left[E_{\min} + \sum_{i=1}^{3} \left(\frac{\frac{1}{100}}{s\rho_i + \frac{1}{100}} \right)^2 c_i^2 \rho_i + \sum_{i=1}^{3} \frac{\rho_i}{s\rho_i + \frac{1}{100}} \right] G_{r_1+N/2, r_2+x/2}(s) ds \\
&= 1 + \int_0^\infty \sum_{i=1}^{3} \left[\left(\frac{\frac{1}{100}}{s\rho_i + \frac{1}{100}} \right)^2 c_i^2 \rho_i + \frac{\rho_i}{s\rho_i + \frac{1}{100}} \right] G_{r_1+N/2, r_2+x/2}(s) ds.
\end{aligned}$$

This integral can be computed numerically. Figure 5.6 shows the curve $y = QV(x)$ along with the lines $y = \lambda V(x)$ for $\lambda = 0.6, 0.4, 0.2$. The first graph has x-values from $x = E_{\min} \approx 191240$ up to $x = 250000$. The second graph has x-values from $x = E_{\min}$ up to $x = 6 \cdot 10^7$. In the first graph it appears that the curve $y = QV(x)$ is nearly flat. In the second graph, the curve $y = QV(x)$ grows close to linearly before flattening out. It can be computed that $\lim_{x\to\infty} QV(x) \approx 9.74 \cdot 10^6$, so the "flattening out" continues for larger x-values. (Already $QV(6 \cdot 10^7)$ is between $6 \cdot 10^6$ and $7 \cdot 10^6$.)

For $\lambda = 0.6$ and $\lambda = 0.4$, the optimal small set C is a relatively short interval close to $x = E_{\min}$. For $\lambda = 0.2$, the optimal C is larger by about three orders of magnitude. There is a discontinuity at the value of λ just above 0.2 for which the line $y = \lambda V(x)$ is tangent to the curve $y = QV(x)$.

To compute the minorization based on the optimal set C, the same method used in the nuclear pump example can be used. The conditional law of λ_R given that $E = x$ has probability density function $f_x(s) = G_{r_1+N/2, r_2+x/2}(s)$, which is log-concave as a function of x when s is held constant. If $c_{\min} = \inf(C)$ and $c_{\max} = \sup(C)$, the quantity

$$\varepsilon = \int_0^\infty \min\left\{ f_{c_{\min}}(s), f_{c_{\max}}(s) \right\} ds \tag{5.6}$$

is a lower bound for the minorization mass associated with C.

For each value of λ, the preceding discussion shows how to compute all the constants associated with the drift and minorization condition: the small set C, the value $K = \sup_{c\in C} QV(c)$, and the minorization mass ε. As in the nuclear pump example, C and K are optimal given the choice of drift function $V(x)$ and the value of λ, while ε is not optimal. Next, Theorem 4.4 gives convergence bounds for the Markov chain (E_t), including the exponential rate λ_*. Figure 5.7 shows the computed values of ε and λ_* for each chosen λ. The discontinuity just above $\lambda = 0.2$ is at the value of λ where the line $y = \lambda V(x)$ is tangent to the curve $y = QV(x)$, as shown in Figure 5.6.

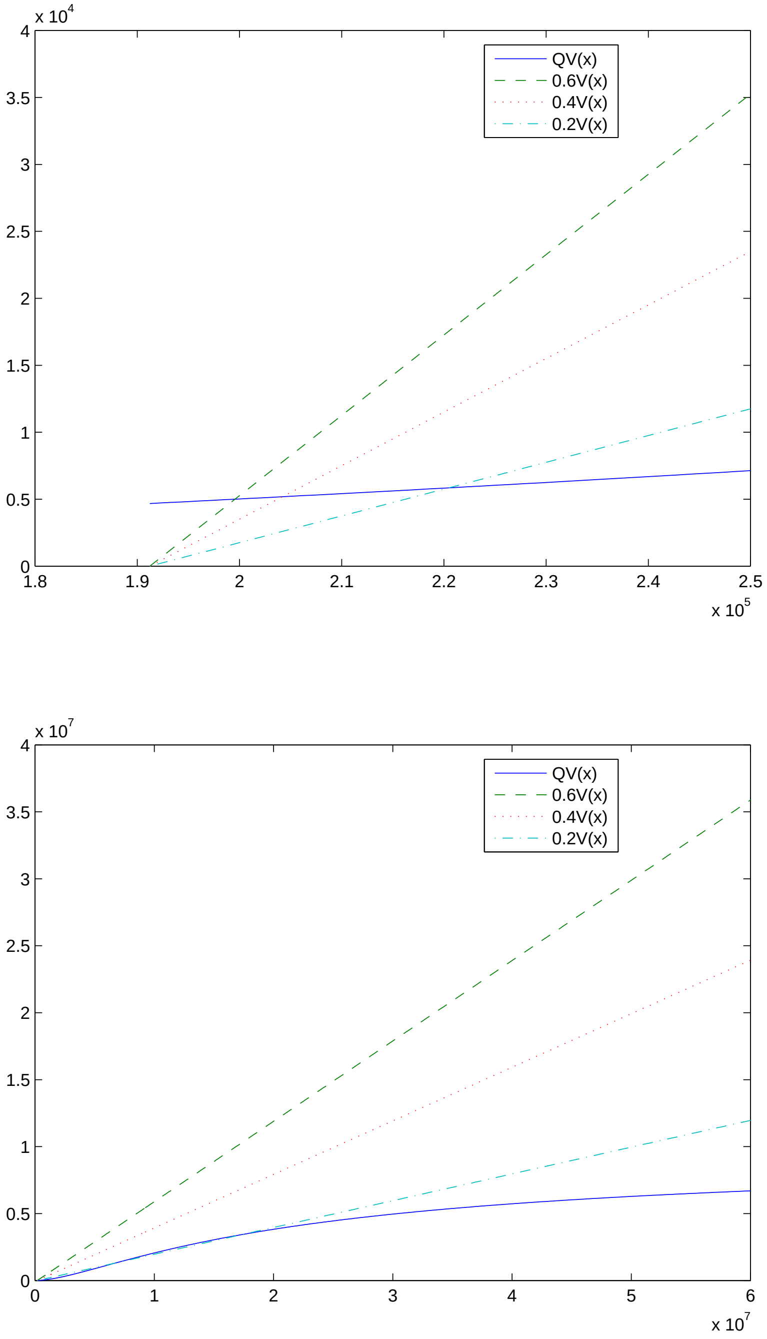

Figure 5.6: Graphs of $y = QV(x)$ and $y = \lambda V(x)$ where $\lambda = 0.6, 0.4, 0.2$ for the HMO premium chain. Here Q is the transition kernel for the chain (E_t) and $V(x) = 1 + x - E_{\min}$. The upper graph shows values $x \in [E_{\min}, 250000]$. The lower graph shows values $x \in [E_{\min}, 6 \cdot 10^7]$.

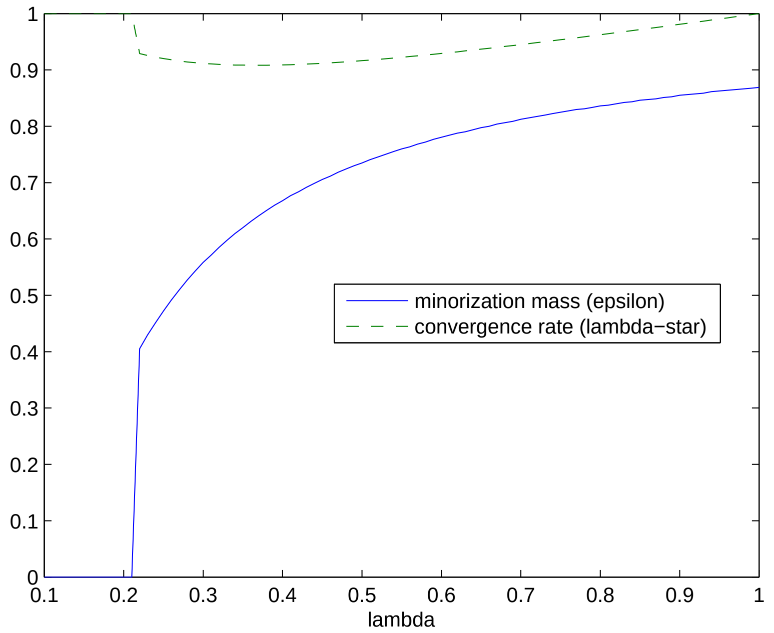

Figure 5.7: Graphs of the minorization mass ε and convergence rate λ_* as functions of the drift parameter λ for the HMO premium chain. For each λ, the optimal small set C associated with λ and the drift function $V(x) = 1 + x - E_{\min}$ is computed as in Figure 5.6. Then ε is computed using (5.6) and λ_* is computed using Theorem 4.4.

The smallest value of λ_* occurs at $\lambda = 0.36$. The drift and minorization data are

$$C = [191240, 205840], \qquad K = 5250, \qquad m = 1, \qquad \varepsilon = 0.631, \qquad \lambda_* = 0.909. \tag{5.7}$$

Theorem 4.4 yields

$$\|Q^t(\mu_E, \cdot) - \pi_E\|_{\mathrm{TV}} \le [2.65 + 0.25t(1 + \mathbf{E}_\mu[E_0 - E_{\min}])^{0.10}] \cdot (0.909)^t,$$

which combined with the bound $\|P^{t+1}(\mu, \cdot) - \pi\|_{\mathrm{TV}} \le \|Q^t(\mu_E, \cdot) - \pi_E\|_{\mathrm{TV}}$ completes the proof. □

If the (E_t) chain is started from $E_0 = E_{\min}$, the argument above gives $t_{0.01} \le 82$. This result can be compared with Theorem 1.3 of [Bax05], which when applied to the drift and minorization data (5.7) yields

$$\|Q^t(E_{\min}, \cdot) - \pi_E\|_{\mathrm{TV}} \le (6.44 \cdot 10^{11}) \cdot (0.916)^t$$

and $t_{0.01} \le 362$. Again, Theorem 4.4 gives tighter bounds. The high value $6.44 \cdot 10^{11}$ is partly due to a suboptimal dependence on K in the formulas of [Bax05].

Chapter 6

MCMC estimation of quantiles

6.1 Discussion of the problem

The main purpose of this thesis is to bound Markov chain convergence rates. This chapter discusses a closely related problem. Let (X_t) be a Markov chain on $\mathcal{X}$ with stationary distribution π, and let $\theta : \mathcal{X} \to \mathbf{R}$ be a (measurable) function. The problem is to explore the law of θ under π, that is, the probability measure $\pi(\theta \in \cdot) = \pi(\{x : \theta(x) \in \cdot\})$ on $\mathbf{R}$. For instance, what is the expectation $\pi(\theta) = \int_{\mathcal{X}} \theta(x)\pi(dx)$? What are the quantiles $\theta_q = \inf\{r \in \mathbf{R} : \pi(\theta \leq r) \geq q\}$?

The Markov chain provides estimates for these quantities. Given a positive integer n, define

$$\hat{\pi}_n(\theta) = \frac{1}{n}\sum_{t=0}^{n-1} \theta(X_t).$$

If (X_t) satisfies appropriate irreducibility assumptions and $\pi(|\theta|) < \infty$, it follows that $\hat{\pi}_n(\theta) \to \pi(\theta)$ with probability 1 as $n \to \infty$ ([RR04], Section 3.2; [MT93], Chapter 17). That is, $\hat{\pi}_n(\theta)$ is a strongly consistent estimator for $\pi(\theta)$.

To estimate the quantile θ_q for $0 < q < 1$, fix $n > 0$ and let $Y_1 \leq \cdots \leq Y_n$ be the order statistics of $\theta(X_0), \theta(X_1), \ldots, \theta(X_{n-1})$. Then

$$\hat{\theta}_{n,q} = Y_{\lceil nq \rceil} \tag{6.1}$$

is a strongly consistent estimator for θ_q.

Under conditions that are often easy to verify, the errors $\hat{\pi}_n(\theta) - \pi(\theta)$ and $\hat{\theta}_{n,q} - \theta_q$ satisfy central limit theorems; see [RR04; Jon04; Dos+14]. Nonasymptotic bounds of the form $\mathbf{P}_\mu(|\hat{\pi}_n(\theta) - \pi(\theta)| \geq \alpha) \leq \delta$ have been proved under various assumptions; see [Lez98; JO10; LMN13; AB15] and the detailed literature summary in [LMN13]. All these results could be adapted to provide nonasymptotic bounds for quantile estimates. In addition, the paper [Dos+14] provides a direct result of the form $\mathbf{P}_\mu(|\hat{\theta}_{n,q} - \theta_q| \geq \alpha) \leq \delta$ (but in terms of quantities that may be difficult to compute).

Applying these nonasymptotic bounds in concrete examples such as the nuclear pump and HMO chains from Chapter 5 is challenging. Consider for example the following theorem.

Theorem 6.1 (Theorem 28 of [RR04]; Theorems 3.1 and 4.2 of [LMN13])**.** *Suppose the transition kernel P of the Markov chain (X_t) on $\mathcal{X}$ with initial distribution μ and stationary distribution π satisfies a one-step general drift and minorization condition (that is, $m = 1$ in Definition 4.1). Let $\theta : \mathcal{X} \to \mathbf{R}$ be a function such that*

$$M = \sup_{x \in \mathcal{X}} \frac{|\theta(x)|^2}{V(x)} < \infty,$$

where V is the drift function. There exists σ^2 such that the following central limit theorem holds:

$$\sqrt{n}\,[\hat{\pi}_n(\theta) - \pi(\theta)] \to N(0, \sigma^2) \quad \textit{in distribution},$$

which implies immediately that

$$\lim_{n\to\infty} n\,\mathbf{E}_\mu[(\hat{\pi}_n(\theta) - \pi(\theta))^2] = \sigma^2.$$

In addition, there exists a constant B with an explicit formula in terms of M and the drift and minorization data such that for all $n \geq 1$,

$$\sqrt{\mathbf{E}_\mu[(\hat{\pi}_n(\theta) - \pi(\theta))^2]} \leq \frac{\sigma}{\sqrt{n}} + \frac{B}{n}. \tag{6.2}$$

This result is satisfying in that the leading term on the right side of (6.2) is asymptotically correct. Even if the drift and minorization condition does not capture the true convergence rate of the Markov chain, so that B is larger than it potentially could be, the effect is limited to a lower-order term.

The challenge in applying (6.2) is that the asymptotic variance σ^2 is usually unknown. In [LMN13] there is an upper bound on σ^2 in terms of M and the drift and minorization data; but in examples this bound can be conservative to the point where the numerical results are not practically useful.

This chapter aims to show that results like the above can potentially be improved by replacing the unknown value σ^2 with an empirical estimate $\hat{\sigma}_n^2$ calculated from the sample path of the Markov chain. One cannot simply replace σ with $\hat{\sigma}_n$ in (6.2), since $\hat{\sigma}_n$ is a random quantity. However, applying Chebyshev's inequality to (6.2) gives that for all $\alpha > 0$,

$$\mathbf{P}_\mu(|\hat{\pi}_n(\theta) - \pi(\theta)| \geq \alpha) \leq \frac{1}{\alpha^2}\left(\frac{\sigma}{\sqrt{n}} + \frac{B}{n}\right)^2.$$

Equivalently, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$|\hat{\pi}_n(\theta) - \pi(\theta)| < \frac{1}{\sqrt{\delta}}\left(\frac{\sigma}{\sqrt{n}} + \frac{B}{n}\right), \tag{6.3}$$

and one could replace σ with $\hat{\sigma}_n$ in (6.3). This chapter's main theorem has the same flavor but applies to quantile estimation in a somewhat different setting.

Suppose as above that the transition kernel P satisfies a one-step general drift and minorization condition with small set C, minorization mass ε, and minorization measure ν. By Proposition 3.10, the Markov chain (X_t) has a sequence of regeneration times. Formally, one can define a sequence $(Y_0, Y_1, \ldots)$ of $\{0, 1\}$-valued random variables such that the chain regenerates precisely at the times t for which $Y_t = 1$.

It was first observed by [MTY95] that the following procedure gives a random sample from the joint distribution of $(X_0, X_1, \ldots; Y_0, Y_1, \ldots)$.

1. Draw X_0 according to μ and let $Y_0 = 0$.
2. Given X_n, draw X_{n+1} according to $P(X_n, \cdot)$.
3. If $X_n \notin C$, let $Y_{n+1} = 0$. If $X_n \in C$, set

$$p = \varepsilon \cdot \frac{d\nu}{dP(X_n, \cdot)}(X_{n+1}). \tag{6.4}$$

   (Since $P(X_n, \cdot) \geq \varepsilon\nu(\cdot)$, p can be interpreted as a probability.) Then let $Y_{n+1} = 1$ with probability p and $Y_{n+1} = 0$ with probability $1 - p$.

In the general setting of Chapter 2, there is no guarantee that the Radon-Nikodym derivative in (6.4) is measurable as a function of X_n and X_{n+1}; but for real-world examples where the one-step minorization condition can be verified at all, this is not an issue. Thus one can usually implement this procedure to get a sample path of the Markov chain together with the regeneration times $T_1, T_2, \ldots$.

Suppose now that one wants to estimate $\pi(f)$ for some function $f : \mathcal{X} \to \mathbf{R}$. The method of *regenerative simulation* says to run the chain for some fixed number $N + 1$ of regenerations and use the estimator

$$\hat{\pi}_N^{\text{reg}}(f) = \frac{1}{T_{N+1} - T_1} \sum_{t=T_1}^{T_{N+1}-1} f(X_t) = \frac{\sum_{j=1}^N \Xi_j}{\sum_{j=1}^N \tau_j},$$

where

$$\Xi_j = \sum_{t=T_j}^{T_{j+1}-1} f(X_t), \qquad \tau_j = T_{j+1} - T_j \tag{6.5}$$

are the contribution of f over each tour and the length of each tour, respectively. By the regenerative property, the Ξ_j are independent and identically distributed (say, as Ξ), and the τ_j are also independent and identically distributed (say, as τ). The drift condition implies that $\mathbf{E}[\tau] < \infty$; if also $\mathbf{E}[|\Xi|] < \infty$, then with probability 1 as $N \to \infty$,

$$\hat{\pi}_N^{\text{reg}}(f) = \frac{\frac{1}{N}\sum_{j=1}^N \Xi_j}{\frac{1}{N}\sum_{j=1}^N \tau_j} \to \frac{\mathbf{E}[\Xi]}{\mathbf{E}[\tau]} = \pi(f).$$

Therefore the estimator is strongly consistent. The errors

$$\frac{1}{N}\sum_{j=1}^N \Xi_j - \mathbf{E}[\Xi], \qquad \frac{1}{N}\sum_{j=1}^N \tau_j - \mathbf{E}[\tau] \tag{6.6}$$

can be bounded using standard techniques for iid random variables, giving control over the error $\hat{\pi}_N^{\text{reg}}(f) - \pi(f)$.

In order to use this technique to estimate quantiles of some $\theta : \mathcal{X} \to \mathbf{R}$, consider the family of functions $\{f_c : c \in \mathbf{R}\}$ where $f_c(x) = \mathbf{1}\{\theta(x) \le c\}$. If $f = f_c$, then $0 \le \Xi_j \le \tau_j$ for each j. The drift function implies that the law of τ decays exponentially; see Theorem 4.9. The same bound applies to Ξ. Therefore a classical concentration inequality for sums of independent subexponential random variables, such as Bernstein's inequality (Theorem 6.3 below), will control both the errors (6.6).

Unfortunately, this approach has the same problem as Theorem 6.1. The bounds on the quantities (6.6) are given in terms of the variances $\mathrm{Var}(\Xi)$ and $\mathrm{Var}(\tau)$, which are themselves unknown. Both $\mathrm{Var}(\Xi)$ and $\mathrm{Var}(\tau)$ are at most $\mathbf{E}[\tau^2]$, which can be controlled using the tail bound from Theorem 4.9. But in examples, this bound is far too conservative, to the point where the numerical results are not useful in practice.

This chapter proposes instead to use a so-called *empirical Bernstein inequality*: a version of Bernstein's inequality that uses the sample variances of $(\Xi_1, \dots, \Xi_N)$ and $(\tau_1, \dots, \tau_N)$ in place of $\mathrm{Var}(\Xi)$ and $\mathrm{Var}(\tau)$. This controls the errors (6.6) using quantities that are computable from the sample path of the Markov chain.

The remaining sections are organized as follows. Section 6.2 discusses empirical Bernstein inequalities and states the main theorem of the chapter, which is an inequality of this type for subexponential random variables. Section 6.3 applies the main theorem to get nonasymptotic quantile estimates in the HMO chain example from Chapter 5. These results are compared against asymptotic estimates computed using a method proposed by [Dos+14]. Finally, Section 6.4 provides some ways this chapter's results could be improved with further research.

6.2 Empirical Bernstein inequalities

The idea of an empirical Bernstein inequality comes from the machine learning community. The first such inequality was proved by Audibert, Munos, and Szepesvári ([AMS09]; see also [MSA08]). This chapter will use a slightly improved version due to Maurer and Pontil [MP09]. A subsequent generalization to U-statistics has been made by [PAR10].

The starting point is the classical Bernstein inequality. The following version is usually called Bennett's inequality [Ben62]. The particular form of the statement is taken from [MP09].

Theorem 6.2. *Let $Z_1, \ldots, Z_N$ be iid random variables distributed as Z, taking values in the interval $[0,1]$ with probability 1. For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}[Z] \leq \sqrt{\frac{2\operatorname{Var}(Z)\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{3N}. \tag{6.7}$$

The same upper bound applies to the other tail $\mathbf{E}[Z] - \frac{1}{N}\sum_{i=1}^{N} Z_i$, as can be seen by replacing each Z_i with $1 - Z_i$.

Another variant of the classical Bernstein inequality applies to subexponential random variables.

Theorem 6.3 (Proposition 2.9 of [Mas07])**.** *Let $Z_1, \ldots, Z_N$ be iid random variables distributed as Z. Assume that there exist constants V and c such that $\mathbf{E}[Z^2] \leq V$ and*

$$\mathbf{E}[(Z_+)^k] \leq \frac{k!}{2} V c^{k-2} \quad \textit{for all } k \geq 3,$$

where $Z_+ = \max\{Z, 0\}$. For any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}[Z] \leq \sqrt{\frac{2V\log(1/\delta)}{N}} + \frac{c\log(1/\delta)}{N}.$$

Note that Theorem 6.3 implies Theorem 6.2. Suppose $Z \in [0,1]$ with probability 1 and $Z_1, \ldots, Z_N$ are iid random variables distributed as Z. The variables $\bar{Z}_i = Z_i - \mathbf{E}[Z]$ are distributed as $\bar{Z} = Z - \mathbf{E}[Z]$, and $\bar{Z}_+ \in [0,1]$ with probability 1. Therefore, for $k \geq 3$,

$$\mathbf{E}[(\bar{Z}_+)^k] \leq \mathbf{E}[(\bar{Z}_+)^2] \leq \mathbf{E}[\bar{Z}^2] \cdot \frac{k!}{2}\left(\frac{1}{3}\right)^{k-2}.$$

Applying Theorem 6.3 to $\bar{Z}_1, \ldots, \bar{Z}_N$ with $V = \mathbf{E}[\bar{Z}^2] = \operatorname{Var}(Z)$ and $c = 1/3$ proves Theorem 6.2.

The empirical Bernstein inequality of Maurer and Pontil [MP09] replaces the unknown quantity $\mathrm{Var}(Z)$ with the sample variance

$$\mathbf{V}_N = \frac{1}{N(N-1)} \sum_{1\le i<j\le N} (Z_i - Z_j)^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Z_i - \frac{1}{N}\sum_{j=1}^{N} Z_j \right)^2 . \tag{6.8}$$

They prove an empirical version of Theorem 6.2:

Theorem 6.4. *Let $Z_1, \dots, Z_N$ be iid random variables distributed as Z, taking values in the interval $[0,1]$ with probability 1. For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}[Z] \le \sqrt{\frac{2\mathbf{V}_N \log(2/\delta)}{N}} + \frac{7\log(2/\delta)}{3(N-1)}.$$

As before, one can replace Z_i with $1 - Z_i$ to get the same bound on the other tail. In addition, if Z is almost surely contained in an interval of length L (rather than $[0,1]$), scaling by a factor of L and translating leads to the bound

$$\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}[Z] \le \sqrt{\frac{2\mathbf{V}_N \log(2/\delta)}{N}} + \frac{7L\log(2/\delta)}{3(N-1)} \tag{6.9}$$

with probability at least $1-\delta$. Note that the first term on the right side is unchanged, because the left side scales by L and $\mathbf{V}_N$ scales by L^2, but the second term on the right side gains a factor of L.

It seems plausible that there should also be an empirical version of Theorem 6.3. This is exactly what is needed to provide tighter bounds on quantile estimation, since both random variables Ξ, τ in (6.6) are subexponential. The following theorem, which is the main result in this chapter, is a bound of this type.

Theorem 6.5. *Let $Z_1, \dots, Z_N$ be iid random variables distributed as Z, where $\mathbf{P}(|Z| \ge z) \le Ae^{-az}$ for fixed constants A, a and all $z \ge 0$. For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}[Z] \le \sqrt{\frac{2\mathbf{V}_N \log(4/\delta)}{N}} + \frac{(14/a)\log(2AN/\delta)\log(4/\delta)}{3(N-1)} + \frac{\delta}{2aN}\left[1 + \log(2AN/\delta)\right].$$

Replacing each Z_i with $-Z_i$ gives the same bound on the other tail.

This result is probably not optimal. The second and third terms on the right side could likely be improved, especially in the dependence on δ. The classical Bernstein inequality interpolates between two regimes: the one where δ is fixed as $N \to \infty$, and the one where the right side of (6.7) is fixed and δ decays as ce^{-dN} for some constants c, d. In the regime where δ is fixed, the dominant term in the right side of Theorem 6.5 is the first one, which differs from the first term in Theorem 6.4 only by a factor of 2 in the coefficient of $1/\delta$. In the other regime, the right side of Theorem 6.5 is fixed when δ decays as $ce^{-d\sqrt{N}}$. This is the part that could be improved.

Proof of Theorem 6.5. The proof is by a truncation argument. Fix

$$b = \frac{1}{a} \log\left(\frac{2AN}{\delta}\right).$$

Then

$$\mathbf{P}(|Z| \geq b) \leq Ae^{-ab} = \frac{\delta}{2N},$$

which implies that $\mathbf{P}(\max\{|Z_1|, \ldots, |Z_N|\} \geq b) \leq \delta/2$. Let $Z = Z' + Z''$, where $Z' = Z\mathbf{1}\{|Z| \leq b\}$ and $Z'' = Z\mathbf{1}\{|Z| > b\}$; likewise decompose each $Z_i = Z_i' + Z_i''$. The Z_i' are iid and distributed as Z', which is contained in $[-b, b]$ almost surely. By (6.9), with probability at least $1 - \delta/2$,

$$\frac{1}{N}\sum_{i=1}^{N} Z_i' - \mathbf{E}[Z'] \leq \sqrt{\frac{2\mathbf{V}_N' \log(4/\delta)}{N}} + \frac{14b\log(4/\delta)}{3(N-1)}, \tag{6.10}$$

where $\mathbf{V}_N'$ is the sample variance of $(Z_1', \ldots, Z_N')$ as in (6.8). Also,

$$\begin{aligned}\mathbf{E}[|Z''|] &= \int_0^\infty \mathbf{P}(|Z''| > z)\, dz \leq b\,\mathbf{P}(|Z''| > b) + \int_b^\infty Ae^{-az}\, dz \leq b \cdot \frac{\delta}{2N} + \frac{Ae^{-ab}}{a} \\ &= \frac{\delta}{2aN}\left[1 + \log\left(\frac{2AN}{\delta}\right)\right].\end{aligned}$$

With probability at least $1 - \delta/2$, each $Z_i = Z_i'$, meaning that $\mathbf{V}_N = \mathbf{V}_N'$ and

$$\frac{1}{N}\sum_{i=1}^{N} Z_i'' - \mathbf{E}[Z''] = -\,\mathbf{E}[Z''] \leq \frac{\delta}{2aN}\left[1 + \log(2AN/\delta)\right]. \tag{6.11}$$

Adding (6.10) to (6.11) (and using a union bound on probabilities) gives the desired result. □

6.3 Worked example

This section shows how to use Theorem 6.5 to get nonasymptotic bounds on quantile estimates. The running example is the HMO premium Markov chain from Section 5.2. For comparison, an "asymptotically valid" analysis as proposed in [Dos+14] will also be performed.

Johnson and Jones [JJ10] provide a model that describes the influence of hospital admission costs on HMO premiums. They get a joint posterior distribution for three parameters:

β_0 : baseline value for an individual monthly premium

β_1 : effect on premium of a \$1000 increase in statewide average hospital admission costs

β_2 : additional cost for HMOs in New England

For an HMO plan in a state where the average hospital admission costs $\$M$ more than the national average, the predicted individual monthly premium is $\beta_0 + \beta_1(M/1000)$ if the state is not in New England and $\beta_0 + \beta_1(M/1000) + \beta_2$ if the state is in New England.

Consider the Markov chain $X_t = \begin{bmatrix} (\beta_0)_t \\ (\beta_1)_t \\ (\beta_2)_t \end{bmatrix}$ on $\mathcal{X} = \mathbf{R}^3$ started from $X_0 = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$, the least-squares estimate from Section 5.2. (In that section it was convenient to work with a slightly different chain (E_t), which was a projection of (X_t) onto $\mathbf{R}$.) Let π be the stationary distribution of (X_t), which by construction is the joint posterior distribution for the β_j. In [JJ10] the authors estimate $\pi(\beta_0) \approx 162.6$, $\pi(\beta_1) \approx 4.0$, and $\pi(\beta_2) \approx 26.3$. A long-run average of the Markov chain over 10^7 steps gives similar estimates: $\pi(\beta_0) \approx 162.3$, $\pi(\beta_1) \approx 4.3$, and $\pi(\beta_2) \approx 26.2$.

In addition to precise estimates for the means $\pi(\beta_j)$, it is also desirable to understand more about how the β_j are distributed. For example, how tightly concentrated is the law of β_1 around its mean 4.3? If $0 < \alpha < 1/2$, the $100(1-2\alpha)\%$ *credible interval* for β_1 is the closed interval $[(\beta_1)_\alpha, (\beta_1)_{1-\alpha}]$, where $(\beta_1)_\alpha, (\beta_1)_{1-\alpha}$ are quantiles as defined in Section 6.1. In turn, these quantiles $(\beta_1)_q$ can be estimated by $(\hat{\beta}_1)_{n,q}$ (see (6.1)) using the Markov chain (X_t). Carrying out this procedure with $\alpha = .025$ and $n = 10^7$ gives an estimate for the 95% credible interval for β_1:

$$[(\beta_1)_{.025}, (\beta_1)_{.975}] \approx [(\hat{\beta}_1)_{10^7,.025}, (\hat{\beta}_1)_{10^7,.975}] = [1.3, 7.2].$$

Using Theorem 6.5, one can find an interval $[c, d]$ that contains $[(\beta_1)_{.025}, (\beta_1)_{.975}]$ with high confidence (in the frequentist sense). The first step is to find an explicit one-step drift and minorization condition satisfied by the (X_t) chain. This was carried out in Section 5.2 for the (E_t) chain; the same computation works for the (X_t) chain. Let ν be the regeneration measure and T the ν-regeneration time. When the data from Section 5.2 are fed into Theorem 4.9, one obtains

$$\mathbf{P}_\nu(T \geq t) \leq Ae^{-at}, \quad \text{where} \quad A = 2.5771, \quad a = 0.0963. \tag{6.12}$$

The successive times between regenerations $\tau_1, \tau_2, \ldots$ as in (6.5) are iid as τ, where $\mathbf{P}_\nu(\tau \geq t)$ satisfies the bound (6.12). For purposes of illustration, fix $\delta = 1/300$. For any $f : \mathcal{X} \to [0, 1]$, define $\Xi_1, \Xi_2, \ldots$ as in (6.5), so that the Ξ_i are iid as Ξ. Since the range of f is $[0, 1]$, $\mathbf{P}_\nu(\Xi \geq t)$ also satisfies the bound (6.12). Theorem 6.5 says that for any $N \geq 2$, with probability at least $1-1/300$,

$$\frac{1}{N}\sum_{i=1}^{N}\Xi_i - \mathbf{E}[\Xi] \leq 3.77\sqrt{\frac{\mathbf{V}_N}{N}} + \frac{2525}{N-1} + \frac{344\log N}{N-1}, \tag{6.13}$$

and also with probability at least $1-1/300$,

$$\mathbf{E}[\Xi] - \frac{1}{N}\sum_{i=1}^{N}\Xi_i \leq 3.77\sqrt{\frac{\mathbf{V}_N}{N}} + \frac{2525}{N-1} + \frac{344\log N}{N-1}, \tag{6.14}$$

where

$$\mathbf{V}_N = \frac{1}{N(N-1)}\sum_{1\leq i<j\leq N}(\Xi_i-\Xi_j)^2.$$

Suppose $c, d \in \mathbf{R}$ are fixed. (They will eventually be chosen so that $c \leq (\beta_1)_{.025}$ and $(\beta_1)_{.975} \leq d$ with high confidence.) Define functions

$$f^{(c)}(\beta_0,\beta_1,\beta_2) = \mathbf{1}\{\beta_1 \leq c\}, \quad f^{(d)}(\beta_0,\beta_1,\beta_2) = \mathbf{1}\{\beta_1 \geq d\}, \quad f^{(1)}(\beta_0,\beta_1,\beta_2) = 1.$$

Likewise define $\Xi_i^{(c)}$, $\Xi_i^{(d)}$, and $\Xi_i^{(1)} = \tau_i$. The idea is to apply (6.13) to the variables $\Xi_i^{(1)}$ and (6.14) to the variables $\Xi_i^{(c)}$ and $\Xi_i^{(d)}$. Since

$$\pi(\beta_1 \leq c) = \pi(f^{(c)}) = \frac{\mathbf{E}[\Xi^{(c)}]}{\mathbf{E}[\Xi^{(1)}]}, \qquad \pi(\beta_1 \geq d) = \pi(f^{(d)}) = \frac{\mathbf{E}[\Xi^{(d)}]}{\mathbf{E}[\Xi^{(1)}]},$$

it follows that with probability at least $1-1/100$, both

$$\pi(\beta_1 \leq c) \leq \frac{\frac{1}{N}\sum_{i=1}^{N}\Xi_i^{(c)} + 3.77\sqrt{\frac{\mathbf{V}_N^{(c)}}{N}} + \frac{2525}{N-1} + \frac{344\log N}{N-1}}{\frac{1}{N}\sum_{i=1}^{N}\Xi_i^{(1)} - 3.77\sqrt{\frac{\mathbf{V}_N^{(1)}}{N}} - \frac{2525}{N-1} - \frac{344\log N}{N-1}} \tag{6.15}$$

and

$$\pi(\beta_1 \geq d) \leq \frac{\frac{1}{N}\sum_{i=1}^{N}\Xi_i^{(d)} + 3.77\sqrt{\frac{\mathbf{V}_N^{(d)}}{N}} + \frac{2525}{N-1} + \frac{344\log N}{N-1}}{\frac{1}{N}\sum_{i=1}^{N}\Xi_i^{(1)} - 3.77\sqrt{\frac{\mathbf{V}_N^{(1)}}{N}} - \frac{2525}{N-1} - \frac{344\log N}{N-1}}, \tag{6.16}$$

assuming N is large enough that the denominator in (6.15) and (6.16) is positive. Let $q(c,N)$ be the right side of (6.15) and $r(d,N)$ be the right side of (6.16). Then $q(c,N)$ and $r(d,N)$ are explicit functions of the sample path of the Markov chain over the first $N+1$ regenerations, and with probability 1 as $N \to \infty$, $q(c,N) \to \pi(\beta_1 \leq c)$ and $r(d,N) \to \pi(\beta_1 \geq d)$.

If indeed $\pi(\beta_1 \leq c) \leq q(c,N)$ and $\pi(\beta_1 \geq d) \leq r(d,N)$, it follows that the quantiles $(\beta_1)_{q(c,N)} \geq c$ and $(\beta_1)_{1-r(d,N)} \leq d$. Therefore, with probability at least $1-1/100$, $[(\beta_1)_{q(c,N)}, (\beta_1)_{1-r(d,N)}] \subseteq [c,d]$.

The question now is how to choose c, d, N so that $q(c, N) \leq .025$ and $r(d, N) \leq .025$. This can never be guaranteed in advance, but if it happens to be true, one concludes that $[c, d]$ contains $[(\beta_1)_{.025}, (\beta_1)_{.975}]$ with confidence at least 0.99. (It cannot be said that $[c, d] \supseteq [(\beta_1)_{.025}, (\beta_1)_{.975}]$ with *probability* at least 0.99, since c, d are not random quantities, so the probability is either 0 or 1. The word "confidence" is used in the frequentist sense. For comparison, suppose a different theorem did provide a random interval $[C, D]$ that contained $[(\beta_1)_{.025}, (\beta_1)_{.975}]$ with probability at least 0.99. If one were to draw values $c = C(\omega)$ and $d = D(\omega)$ using a sample path $\omega = (X_0, X_1, \ldots)$ of the Markov chain, it would also be true that $[c, d] \supseteq [(\beta_1)_{.025}, (\beta_1)_{.975}]$ with confidence at least 0.99.)

Here is one way of choosing c, d, N so that $q(c, N)$ and $r(d, N)$ are likely to be at most .025. The key insight is that in practice, the Markov chain probably converges much faster than can be proved rigorously. Run the chain for a relatively small number of regenerations $N_0 + 1$; say this takes n_0 steps. Pick values $0 < q_0, r_0 < .025$, for example $q_0 = r_0 = .01$, and set

$$c = (\hat{\beta}_1)_{n_0, q_0}, \qquad d = (\hat{\beta}_1)_{n_0, 1-r_0}$$

to be the estimators for the quantiles $(\beta_1)_{q_0}$ and $(\beta_1)_{1-r_0}$ as in (6.1). Using that same sample path, compute the random variables $\Xi_i^{(c)}, \Xi_i^{(d)}, \Xi_i^{(1)}$ for $1 \leq i \leq N_0$. Next, find the sample means $\mathbf{E}_{N_0}^{(c)} = \frac{1}{N_0} \sum_{i=1}^{N_0} \Xi_i^{(c)}$ (likewise $\mathbf{E}_{N_0}^{(d)}, \mathbf{E}_{N_0}^{(1)}$) and the sample variances $\mathbf{V}_{N_0}^{(c)}, \mathbf{V}_{N_0}^{(d)}, \mathbf{V}_{N_0}^{(1)}$. Consider now the quantities

$$q_{N_0}(c, N) = \frac{\mathbf{E}_{N_0}^{(c)} + 3.77\sqrt{\frac{\mathbf{V}_{N_0}^{(c)}}{N}} + \frac{2525}{N-1} + \frac{344 \log N}{N-1}}{\mathbf{E}_{N_0}^{(1)} - 3.77\sqrt{\frac{\mathbf{V}_{N_0}^{(1)}}{N}} - \frac{2525}{N-1} - \frac{344 \log N}{N-1}}, \quad r_{N_0}(d, N) = \frac{\mathbf{E}_{N_0}^{(d)} + 3.77\sqrt{\frac{\mathbf{V}_{N_0}^{(d)}}{N}} + \frac{2525}{N-1} + \frac{344 \log N}{N-1}}{\mathbf{E}_{N_0}^{(1)} - 3.77\sqrt{\frac{\mathbf{V}_{N_0}^{(1)}}{N}} - \frac{2525}{N-1} - \frac{344 \log N}{N-1}}$$

(compare with (6.15), (6.16)), which are relatively simple functions of N. If $N_0 + 1$ regenerations are enough for the sample means and variances $\mathbf{E}_{N_0}, \mathbf{V}_{N_0}$ to be near their limiting values $\mathbf{E}[\Xi], \mathrm{Var}(\Xi)$, then in an independent run of the chain for $N + 1$ regenerations, $q(c, N) \approx q_{N_0}(c, N)$ and $r(d, N) \approx r_{N_0}(d, N)$. Therefore if N is chosen large enough that $q_{N_0}(c, N), r_{N_0}(d, N) \leq \alpha < .025$ for some $\alpha < .025$, say $\alpha = .02$, it is likely that $q(c, N)$ and $r(d, N)$ will be at most .025.

This procedure was carried out with $N_0 = 2000$; it took $n_0 = 3190$ steps until the 2001st regeneration. With $q_0 = r_0 = .01$, the estimates for $(\beta_1)_{.01}$ and $(\beta_1)_{.99}$ were $c = (\hat{\beta}_1)_{n_0, .01} = 0.87$ and $d = (\hat{\beta}_1)_{n_0, .99} = 7.70$. The least value of N such that $q_{N_0}(c, N), r_{N_0}(d, N) \leq .02$ was found to be $N = 458153$. An independent run of the chain for 458154 regenerations (which took 732435 steps) found values $q(c, N) = .023$, $r(d, N) = .022$. Since both these values are less than .025, the conclusion holds that $[(\beta_1)_{.025}, (\beta_1)_{.975}] \subseteq [0.87, 7.70]$ with confidence at least 0.99. This interval is only slightly wider than the best estimate of $[1.3, 7.2]$ for $[(\beta_1)_{.025}, (\beta_1)_{.975}]$.

An immediate goal is to improve the method so that fewer steps of the Markov chain are needed to get a result of the same strength. Section 6.4 describes a few ideas in this direction. The remainder of this section considers an alternative procedure proposed by Doss, Flegal, Jones, and Neath [Dos+14].

Their setting is general: given a function $\theta : \mathcal{X} \to \mathbf{R}$ and a value $0 < q < 1$, one wants to estimate the quantile θ_q using the Markov chain estimator $\hat{\theta}_{n,q}$ as in (6.1). They provide a central limit theorem,

$$\sqrt{n}(\hat{\theta}_{n,q} - \theta_q) \to N(0, \sigma_q^2) \quad \text{in distribution},$$

and two different estimators for σ_q^2, one using batch means and the other using subsampling. This analysis considers only the subsampling estimator, denoted here by $\hat{\sigma}_{n,q}^2$ (but referred to as $\hat{\gamma}_S^2$ in Section 3.2.2 of [Dos+14]). For any $0 < r < 1$, let z_r be the standard normal quantile: $\mathbf{P}(Y \leq z_r) = r$ for $Y \sim N(0,1)$. The central limit theorem implies that

$$\lim_{n \to \infty} \mathbf{P}_\mu \left(\theta_q \leq \hat{\theta}_{n,q} + z_r \frac{\hat{\sigma}_{n,q}}{\sqrt{n}} \right) = r.$$

This is an asymptotic result; there are no guarantees for fixed n.

In the HMO chain, the true value of the quantile $(\beta_1)_{.025}$ is estimated to be 1.3275 (using a long-run average over 10^7 steps). The following procedure was performed:

1. Let $\theta = \beta_1$ and $q = .025$. Fix $n = 3000$ and $R = \{.005, .01, .05, .1, .9, .95, .99, .995\}$.
2. Run the Markov chain (X_t) for n steps started at $\hat{\beta}$, the least-squares estimator for β.
3. For each $r \in R$, determine whether $1.3275 \leq \hat{\theta}_{n,q} + z_r \frac{\hat{\sigma}_{n,q}}{\sqrt{n}}$.
4. Repeat steps 2–3 10000 times. For each $r \in R$, find the fraction of the time that the inequality in step 3 is true. Denote the fraction by $p(r)$.

The results are given in Table 6.1. This table can be interpreted as follows. Fix the quantile $\theta_q = (\beta_1)_{.025}$ and suppose $0 < r < 1$ is given. The method of [Dos+14] produces for each finite sample path $(X_0, \ldots, X_n)$ a quantity $\hat{\theta}_{n,q} + z_r \frac{\hat{\sigma}_{n,q}}{\sqrt{n}}$ whose probability of being at least θ_q is approximately equal to r. Fix $n = 3000$ and denote the actual probability of this event by

$$\phi(r) = \mathbf{P}_{\hat{\beta}} \left(\theta_q \leq \hat{\theta}_{n,q} + z_r \frac{\hat{\sigma}_{n,q}}{\sqrt{n}} \right).$$

Given 10000 independent trials, each with success probability $\phi(r)$, the fraction of successes has distribution

$$p(r) \sim \frac{1}{10000} \text{Binomial}(10000, \phi(r)).$$

If the approximation of [Dos+14] were perfectly accurate, then $\phi(r)$ would equal r and the displayed values of $p(r)$ in Table 6.1 would be very close to the r values (with standard deviation

r	$p(r)$
0.005	0.012
0.01	0.021
0.05	0.070
0.1	0.121
0.9	0.881
0.95	0.934
0.99	0.985
0.995	0.991

Table 6.1: Comparison of nominal versus empirical coverage rates for the subsampling estimator of quantiles applied to the HMO chain. The first column is the predicted fraction of the time that the true value of $(\beta_1)_{.025}$ should fall below the value given by the estimator. The second column is the actual fraction of the time this occurred for 10000 independent sample paths of length 3000 each.

$\sqrt{r(1-r)/10000}$). Instead, observe that $p(r)$ is systematically greater than r when r is near 0 and systematically less than r when r is near 1. This means that the true value of θ_q is less likely to lie within an interval $[\hat{\theta}_{n,q} + z_r \frac{\hat{\sigma}_{n,q}}{\sqrt{n}},\ \hat{\theta}_{n,q} + z_{1-r} \frac{\hat{\sigma}_{n,q}}{\sqrt{n}}]$ than the "nominal coverage rate" $1-2r$ would predict. For $r = .01$ the nominal coverage rate is 0.98, but the observed frequency (or "empirical coverage rate") is $0.985 - 0.021 = 0.964$. Still, the empirical coverage rates are reasonably close to the nominal values.

It should be noted that Flegal [Fle12] uses the subsampling method described above to estimate quantiles for a Markov chain nearly identical to the one considered in this section. (The only change is that different values are chosen for the hyperparameters b and B in Section 5.2.) He provides quantile estimates with Monte Carlo standard errors but does not compute empirical coverage rates. Such rates are computed for other Markov chains in [Fle12] and [Dos+14].

To conclude, suppose one wants to use a finite sample of the Markov chain to find an interval $[c, d]$ that contains the $100(1-2\alpha)\%$ credible interval $[\theta_\alpha, \theta_{1-\alpha}]$ with confidence at least $1-\delta$. The nonasymptotic estimates developed in this chapter accomplish this goal, though it takes many steps of the Markov chain to compute them when δ is small. The asymptotic estimates of [Dos+14] require much less computational power. They provide an interval $[c', d']$ that contains $[\theta_\alpha, \theta_{1-\alpha}]$ with *approximate* confidence $1-\delta$, but the true confidence level is unknown and likely to be somewhat less than $1-\delta$ if the behavior displayed in Table 6.1 is typical.

6.4 Further directions

Here are some ideas on how to improve this chapter's results.

1. The truncation argument used to prove the main Theorem 6.5 is rather crude. With modifications, it is likely that the dependence on δ in the error terms could be improved. Possibly one could also remove the factor of $\log N$. Any improvement of this type would reduce the number of steps needed in the Markov chain to prove statements like "$[\theta_q, \theta_{1-q}] \subseteq [c, d]$ with confidence at least 0.99."

2. The ratio estimator coming from the regeneration times of the Markov chain is attractive because both the numerator and denominator are sums of iid random variables. However, estimating the error of the numerator and denominator separately may be inefficient compared with results like Theorem 6.1. An empirical version of Theorem 6.1 (say, with the subsampling estimator $\hat{\sigma}_n$ from [Dos+14] replacing σ in (6.3)) might perform better than Theorem 6.5 even with the square-root dependence on δ. In addition, the hypothesis of Theorem 6.1 is that the function $\theta : \mathcal{X} \to \mathbf{R}$ has finite second moment. A version of Theorem 6.1 for bounded functions would likely have improved δ dependence, making it better suited for quantile estimation.

3. The procedure outlined in Section 6.3, where one uses a short run of the Markov chain to identify the interval endpoints c, d and the required length N of an independent long run, is cumbersome. It would be cleaner to prove a direct result of the form

$$\mathbf{P}_\mu(\theta_q \in [c_{n,q,\delta}, d_{n,q,\delta}]) \geq 1 - \delta$$

   for a random interval $[c_{n,q,\delta}, d_{n,q,\delta}]$ that is an explicit function of the sample path $(X_0, \dots, X_n)$. One way to obtain this type of statement might be to use concentration inequalities for empirical processes, discussed immediately below.

4. The results in this chapter are concentration inequalities for single functions of the form $f_c(x) = \mathbf{1}\{\theta(x) \leq c\}$. Empirical process theory might provide inequalities that hold uniformly over the family $\{f_c(x) : c \in \mathbf{R}\}$. This would lead to uniform bounds on $|\hat{\theta}_{n,q} - \theta_q|$ for all $0 < q < 1$. The use of the word "empirical" in "empirical process" is of course different from the use in "empirical Bernstein inequality." The desired result would be an "empirical concentration inequality for empirical processes." One such theorem was proved by Koltchinskii ([Kol06], see also [BN09]) using Rademacher processes. Extending these inequalities to apply to Markov chains is a promising direction of future research.

Chapter 7

Finite chains

This chapter discusses two applications of the method of drift and minorization to finite Markov chains. The first application is to birth and death chains, and the second is to the simple random walk on the hypercube.

7.1 Birth and death chains

A finite *birth and death chain* is a Markov chain on the state space $\{0, 1, \dots, n\}$ whose transition matrix P has the property that $P(i, j) = 0$ if $|i - j| \geq 2$. Throughout this section, the transition probabilities will be denoted by

$$p_i = P(i, i+1), \qquad q_i = P(i, i-1), \qquad r_i = 1 - p_i - q_i = P(i, i),$$

where $p_n = q_0 = 0$. All birth and death chains are reversible. A birth and death chain is irreducible if and only if $p_i > 0$ for all $0 \leq i \leq n-1$ and $q_i > 0$ for all $1 \leq i \leq n$. In that case, it has a unique stationary distribution given by

$$\pi(i) = \pi(0) \frac{p_0 p_1 \cdots p_{i-1}}{q_1 q_2 \cdots q_i},$$

where $\pi(0)$ is chosen so that $\sum_{i=0}^{n} \pi(i) = 1$. This section will consider only irreducible birth and death chains that are lazy, meaning that $r_i \geq 1/2$ for all i.

A *median state* for a birth and death chain with stationary distribution π is a state $0 \leq m \leq n$ for which $\pi(\{0, \dots, m\}) \geq 1/2$ and $\pi(\{m, \dots, n\}) \geq 1/2$. A generic birth and death chain will have a unique median state, but it is also possible for there to be two consecutive median states.

Let (X_t) be a lazy irreducible birth and death chain on $\{0, \dots, n\}$ with transition matrix P and stationary distribution π, and let m be a median state for (X_t). Define

$$t_{\text{hit}} = \max\{\mathbf{E}_0[\tau_m], \mathbf{E}_n[\tau_m]\} \tag{7.1}$$

to be the maximum expected hitting time for m. (Because of the laziness, a monotonicity argument shows that $\mathbf{E}_j[\tau_m] \le t_{\text{hit}}$ for all $0 \le j \le n$.) It turns out, as first seen by [DLP10], that the total variation mixing time $t_{\text{mix}} = t_{\text{mix}}(1/4)$ is a constant multiple of t_{hit}.

Theorem 7.1. *Let (X_t) be a lazy irreducible birth and death chain on $\{0, \dots, n\}$, and define t_{hit} by (7.1). Then*

$$\frac{t_{\text{hit}}}{25} \le t_{\text{mix}} \le 288 t_{\text{hit}}.$$

The constants $1/25$ and 288 could certainly be improved. The upper bound is from Theorem 3.1 in [CSC13b], and the lower bound is a slight extension of Theorem 3.9 in [CSC13b]. For completeness, the proof of the lower bound is given at the end of this section.

For $t_{\text{mix}}(\varepsilon)$ when $\varepsilon < 1/4$, one has the following standard result (see e.g. equation (2.3) in [DLP10]):

$$t_{\text{mix}}(\varepsilon) \le \left\lceil \frac{1}{2} \log_2(1/\varepsilon) \right\rceil t_{\text{mix}}(1/4), \qquad 0 < \varepsilon < 1/4. \tag{7.2}$$

In practice, the difference $t_{\text{mix}}(\varepsilon) - t_{\text{mix}}(1/4)$ is frequently very small compared with $t_{\text{mix}}(1/4)$. One says that a family of Markov chains indexed by n exhibits *cutoff* if for every $0 < \varepsilon, \varepsilon' < 1$,

$$t_{\text{mix}}^{(n)}(\varepsilon) \sim t_{\text{mix}}^{(n)}(\varepsilon'),$$

where $t_{\text{mix}}^{(n)}$ is the mixing time for the nth chain, and the notation $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$. Suppose sequences (c_n) and (w_n) are given. The cutoff is said to occur "at c_n" if $t_{\text{mix}}^{(n)} \sim c_n$, and it has a *window* of size w_n if $w_n = o(c_n)$ and for every $0 < \varepsilon < 1$,

$$|t_{\text{mix}}^{(n)}(\varepsilon) - c_n| \le F(\varepsilon) w_n$$

for some function $F(\varepsilon)$ not depending on n.

For families of Markov chains with cutoff, the multiplication in (7.2) does not reflect the actual behavior of $t_{\text{mix}}(\varepsilon)$ as ε decreases. Rather, one has

$$t_{\text{mix}}^{(n)}(\varepsilon) \le t_{\text{mix}}^{(n)}(1/4) + (1 + o(1)) F(\varepsilon) w_n. \tag{7.3}$$

Cutoff was first shown to exist by Diaconis and Shahshahani [DS81] for the random transposition walk on the symmetric group. Aldous and Diaconis [AD86] recognized the general phenomenon and coined the term "cutoff." Since that time there has been a large amount of research in the area.

Good introductory references are Chapter 18 of [LPW09], the survey article [Dia96], and Section 3.3 of [SC04].

A conjecture of Peres [Per04] led to the consideration of cutoff for birth and death chains. The conjecture was that a certain condition, which will be described below, is necessary and sufficient for a family of chains to exhibit cutoff. Though counterexamples were immediately found, the conjecture was (and is) still thought to hold for many natural families. The class of birth and death chains served as a test case. Diaconis and Saloff-Coste [DSC06] proved a separation-distance variant of the conjecture for birth and death chains. The original conjecture, expressed using total variation distance, was proved for birth and death chains by Ding, Lubetzky, and Peres [DLP10].

In order to motivate the conjecture of Peres, consider the question of replacing the bound (7.2) with something of the form (7.3). A well-known alternative to (7.2) is written in terms of the *relaxation time* $t_{\text{rel}} = 1/\gamma$, where γ is the spectral gap of the transition matrix P, that is, the difference between 1 and the next-largest eigenvalue.

Proposition 7.2. *Let P be the transition matrix for a reversible Markov chain on a finite state space $\mathcal{X}$ with unique stationary distribution π. Assume all the eigenvalues of P are nonnegative, and assume that $\pi_{\min} = \min\{\pi(x) : x \in \mathcal{X}\}$ is strictly positive. Then for every $0 < \varepsilon < 1$,*

$$\log\left(\frac{1}{2\varepsilon}\right)(t_{\text{rel}} - 1) \le t_{\text{mix}}(\varepsilon) \le \left\lceil \frac{1}{2}\log\left(\frac{1}{\pi_{\min}}\right)t_{\text{rel}} + \log\left(\frac{1}{2\varepsilon}\right)t_{\text{rel}} \right\rceil.$$

Proposition 7.2 was first proved in continuous time by Aldous [Ald82]. The lower bound given here is Theorem 12.4 in [LPW09], and the upper bound follows from Proposition 3 in [DS91].

The form of the upper bound in Proposition 7.2 is closer to (7.3): one has

$$t_{\text{mix}}(\varepsilon) \le (\text{term not depending on } \varepsilon) + \log\left(\frac{1}{2\varepsilon}\right)t_{\text{rel}}.$$

Suppose a family of chains indexed by n satisfies $t^{(n)}_{\text{mix}} \to \infty$. If $t^{(n)}_{\text{rel}}$ and $t^{(n)}_{\text{mix}}$ have the same order, the lower bound in Proposition 7.2 means that cutoff cannot occur. On the other hand, if $t^{(n)}_{\text{rel}} = o(t^{(n)}_{\text{mix}})$, the upper bound in Proposition 7.2 suggests that the differences $t^{(n)}_{\text{mix}}(\varepsilon) - t^{(n)}_{\text{mix}}(\varepsilon')$ might have order $t^{(n)}_{\text{rel}}$, which is small compared with $t^{(n)}_{\text{mix}}$. In that case, the family would exhibit cutoff with a window of $t^{(n)}_{\text{rel}}$.

Peres [Per04] conjectured that a family of chains with $t^{(n)}_{\text{mix}} \to \infty$ has cutoff if and only if $t^{(n)}_{\text{rel}} = o(t^{(n)}_{\text{mix}})$. The L^p version of this conjecture, for $p > 1$, was proved by Chen and Saloff-Coste [CSC08]. They also showed that the cutoff window is necessarily at most $t^{(n)}_{\text{rel}}$. When total variation distance is used, as in the original conjecture, the situation is more complicated. To begin with, there are counterexamples: see Example 18.7 in [LPW09] and Section 6 of [CSC08]. Even when the conjecture holds, as is thought to happen "most of the time," the cutoff window may be larger than $t_{\text{rel}}^{(n)}$.

A well-known example of the latter phenomenon is the lazy biased reflecting random walk on a line segment. This is a birth and death chain on $\{0, \dots, n\}$ defined as follows. Fix $1/2 < p < 1$. Let the birth probabilities be $p_i = p/2$ for $0 \le i \le n-1$ and the death probabilities be $q_i = (1-p)/2$ for $1 \le i \le n$. The holding probabilities are therefore $r_i = 1/2$ for $1 \le i \le n-1$ and $r_0 = 1 - p/2$, $r_n = 1-(1-p)/2$. If $X_t = i$ for $1 \le i \le n-1$, the expectation of X_{t+1} is $\frac{1-p}{2}(i-1) + \frac{1}{2}i + \frac{p}{2}(i+1) = i+(p-1/2)$. For this reason, the chain is said to have a bias of $\beta = p - 1/2$. This family of chains has a cutoff at $c_n = n/\beta$ with a window of size $\sqrt{n}$ (see e.g. section 18.2 of [LPW09]), but the relaxation time is constant (as follows from the explicit computation of the eigenvalues of P in Chapter XVI of [Fel68]).

This example does not contradict the upper bound in Proposition 7.2 because the term $\frac{1}{2}\log(\frac{1}{\pi_{\min}})t_{\text{rel}}$ is larger than n/β. Indeed, one can compute that

$$\frac{1}{2}\log\left(\frac{1}{\pi_{\min}}\right)t_{\text{rel}} \sim \frac{1}{4\beta^2}\left(1+\sqrt{1-4\beta^2}\right)\log\left(\frac{1+2\beta}{1-2\beta}\right)n \ge \frac{2n}{\beta}.$$

Therefore, Proposition 7.2 alone does not imply the existence of cutoff for this example, nor does it provide any information on the size of the cutoff window.

In [DLP10] it was shown that the conjecture of Peres holds for any family of lazy irreducible birth and death chains. The authors prove that if $t_{\text{rel}}^{(n)} = o(t_{\text{mix}}^{(n)})$, the family has cutoff at $t_{\text{hit}}^{(n)}$ with a window of size at most $\sqrt{t_{\text{rel}}^{(n)} t_{\text{mix}}^{(n)}}$. This is exactly the size of the window for the lazy biased reflecting random walk on $\{0, \dots, n\}$, and the authors provide a rich family of other examples for which the window size is sharp.

The confirmation of the Peres conjecture for birth and death chains in separation distance by [DSC06] and in total variation by [DLP10] complemented earlier results that found relationships between mixing parameters such as t_{rel}, t_{mix} and the birth, death, and holding probabilities p_i, q_i, r_i. Zeifman's method, developed in [Zeĭ91; Zeĭ95a; Zeĭ95b] and clearly explained in the survey [DZP10], uses Kolmogorov's forward equation to provide an exact variational formula for the spectral gap of a continuous time birth and death chain in terms of the transition rates. (Results for continuous time chains translate easily to results for lazy discrete time chains; see Chapter 20 of [LPW09] for an overview and [CSC13a] for a detailed treatment.) Zeifman's method also provides upper and lower bounds on mixing time (see e.g. Theorem 9 of [DZP10]) but these are often not precise enough to determine whether or not cutoff occurs.

As an example, consider the continuous time version of the lazy biased reflecting random walk on $\{0, \dots, n\}$, which jumps to the right with Poisson rate $p/2$ and to the left with Poisson rate $(1-p)/2$. As in discrete time, this family of chains has a bias of $\beta = p - 1/2$, and if $1/2 < p < 1$ the family exhibits cutoff at n/β with a window of size $\sqrt{n}$. Based on computations in [DZP10], Zeifman's method gives the upper bound

$$t_{\mathrm{mix}} \leq \frac{\log(p/(1-p))}{1-2\sqrt{p(1-p)}} \cdot n$$

for sufficiently large n. Numerical evaluation suggests that

$$\frac{\log(p/(1-p))}{1-2\sqrt{p(1-p)}} > \frac{1}{p-1/2}$$

when $p > 1/2$ is fixed, meaning that this bound is not sharp enough to show cutoff.

Miclo [Mic99], also working in continuous time, used discrete Hardy's inequalities to derive another formula for the spectral gap in terms of the birth and death rates. Miclo's formula, which incorporates the median state m, is explicit and non-variational but is correct only up to a universal constant factor of 8. He likewise obtained a non-variational formula for the log-Sobolev constant of the chain that is accurate to a universal constant factor. A version of Miclo's spectral gap result for discrete time birth and death chains is presented by Chen and Saloff-Coste [CSC13b].

The proof by [DLP10] that a sequence of lazy irreducible birth and death chains exhibits cutoff if and only if $t_{\mathrm{rel}}^{(n)} = o(t_{\mathrm{mix}}^{(n)})$ (or equivalently, $t_{\mathrm{rel}}^{(n)} = o(t_{\mathrm{hit}}^{(n)})$) left open the question of devising a criterion for cutoff in terms of the transition probabilities p_i, q_i, r_i. Such a criterion was developed by [CSC13b], using a formula for t_{hit} in terms of p_i, q_i, r_i due to [BBF09] and using Miclo's result to characterize t_{rel} up to a constant factor. In follow-up work [CSC14; CSC15], the authors provide methods for computing the spectral gap, cutoff time, and cutoff window size.

The proofs by Peres and Sousi [PS15] and independently Oliveira [Oli12; Gri+14] that "mixing times are hitting times of large sets" provide another perspective on the result of [DLP10]. A recent paper of Basu, Hermon, and Peres [BHP15] develops this point of view and extends the work of [DLP10] by showing that the Peres condition $t_{\mathrm{rel}}^{(n)} = o(t_{\mathrm{mix}}^{(n)})$ is equivalent to cutoff for lazy irreducible random walks on trees. When applied to birth and death chains, the results of [BHP15] are slightly more precise than those of [DLP10]: both papers prove that

$$|t_{\mathrm{mix}}(\varepsilon) - t_{\mathrm{hit}}| \leq F(\varepsilon)\sqrt{t_{\mathrm{hit}} t_{\mathrm{rel}}}, \tag{7.4}$$

but the dependence on ε of $F(\varepsilon)$ is suboptimal in [DLP10] and nearly optimal (up to constants) in [BHP15].

The purpose of this section is to use the methods of Chapter 4 to get sharp upper bounds for the mixing time of lazy birth and death chains. Corollary 7.4 below recovers a one-sided version of (7.4), namely

$$t_{\mathrm{mix}}(\varepsilon) - t_{\mathrm{hit}} \leq F(\varepsilon)\sqrt{t_{\mathrm{hit}} t_{\mathrm{rel}}}, \tag{7.5}$$

where $F(\varepsilon)$ has the same optimal dependence on ε as in [BHP15] (and with slightly better constants). The main result is the following.

Theorem 7.3. *Let P be the transition matrix for a lazy irreducible birth and death chain on $\mathcal{X} = \{0, \ldots, n\}$ with stationary distribution π. Let m be a median element for the chain. Fix $\delta > 0$. Then there are a function $V : \mathcal{X} \to [1, \infty)$ and a constant $\lambda < 1$, depending on δ, such that $PV(x) \leq \lambda V(x)$ for $x \neq m$, and the resulting bound*

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq 2V(x)\lambda^{t+1}$$

from Theorem 4.5 implies directly that

$$t_{\mathrm{mix}}(\varepsilon) \leq (1 + \delta)t_{\mathrm{hit}} + \left(1 + \frac{1}{\delta}\right) 2t_{\mathrm{rel}} \log(2/\varepsilon) \tag{7.6}$$

for all $0 < \varepsilon < 1$.

The proof of Theorem 7.3 will be given at the end of the section. For now, consider its consequences.

One can view (7.6) as a tightening of the upper bound in Proposition 7.2. It has the same form

$$t_{\mathrm{mix}}(\varepsilon) \leq (\text{term not depending on } \varepsilon) + (\text{constant})t_{\mathrm{rel}} \log(1/\varepsilon).$$

By sending $\delta \to 0$, the term not depending on ε can be brought arbitrarily close to t_{hit}, at the cost of increasing the coefficient on $t_{\mathrm{rel}} \log(1/\varepsilon)$.

For fixed ε, choosing δ to minimize the right side of (7.6) results in the following inequality.

Corollary 7.4. *For any lazy irreducible birth and death chain, and any $0 < \varepsilon < 1$,*

$$t_{\mathrm{mix}}(\varepsilon) \leq t_{\mathrm{hit}} + 2\sqrt{2t_{\mathrm{hit}}t_{\mathrm{rel}} \log(2/\varepsilon)} + 2t_{\mathrm{rel}} \log(2/\varepsilon). \tag{7.7}$$

Since t_{hit} and t_{mix} have the same order, while $t_{\mathrm{rel}} = O(t_{\mathrm{mix}})$, this recovers the upper bound (7.5). The proof shows how the $\sqrt{t_{\mathrm{hit}}t_{\mathrm{rel}}}$ term in (7.7) is a direct consequence of (7.6).

Proof of Corollary 7.4. The right side of (7.6) is

$$t_{\mathrm{hit}} + 2t_{\mathrm{rel}} \log(2/\varepsilon) + \delta t_{\mathrm{hit}} + \frac{1}{\delta} 2t_{\mathrm{rel}} \log(2/\varepsilon).$$

The product of the last two terms is $2t_{\mathrm{hit}}t_{\mathrm{rel}} \log(2/\varepsilon)$. By the arithmetic mean–geometric mean inequality, their sum is minimized by choosing δ so that both terms equal $\sqrt{2t_{\mathrm{hit}}t_{\mathrm{rel}} \log(2/\varepsilon)}$. This choice of δ yields (7.7). □

Overall, the drift function approach does not yield a completely independent proof of the Peres conjecture for birth and death chains. First, it provides only the upper bound and not the lower bound of the cutoff window. Second, certain elements of the proof of Theorem 7.3 are shared with the argument in [DLP10].

On the other hand, the derivation of the $\sqrt{t_{\mathrm{hit}} t_{\mathrm{rel}}}$ term appears to be different than in [DLP10] and [BHP15]. The inequality (7.7) can be compared with the corresponding statements in those two papers. Equation (2.31) of [DLP10] reads: if $0 < \varepsilon < 1/16$ and $t_{\mathrm{rel}} \leq \varepsilon^5 t_{\mathrm{mix}}$, then

$$t_{\mathrm{mix}}(4\varepsilon) \leq t_{\mathrm{hit}} + \frac{3}{\varepsilon}\sqrt{t_{\mathrm{hit}} t_{\mathrm{rel}}}.$$

This has worse dependence on ε than (7.7). In [BHP15], the combination of Lemma 5.1 with equation (1.4) implies: if $0 < \varepsilon \leq 1/4$ and

$$\frac{16}{5} t_{\mathrm{rel}} \log(2/\varepsilon) \leq t_{\mathrm{hit}}, \tag{7.8}$$

then

$$t_{\mathrm{mix}}(\varepsilon) \leq t_{\mathrm{hit}} + \sqrt{20 t_{\mathrm{hit}} t_{\mathrm{rel}} \log(2/\varepsilon)} + \lceil t_{\mathrm{rel}} \log(4/\varepsilon) \rceil.$$

This has the same dependence on ε as (7.7), but it has slightly worse constants and is only proved in the regime where t_{rel} is small compared to t_{hit} (as quantified by (7.8)).

For questions related to cutoff, one fixes ε and analyzes $t^{(n)}_{\mathrm{mix}}(\varepsilon)$ as a function of n. In this situation, (7.7) gives sharp control on the mixing time. One could also fix n and analyze $t_{\mathrm{mix}}(\varepsilon)$ as $\varepsilon \to 0$ for that particular chain. In that case, as long as t_{rel} is bounded away from 1, the upper bound in (7.6) is guaranteed to be within a universal constant factor of the actual value of $t_{\mathrm{mix}}(\varepsilon)$. For instance, one has the following result.

Proposition 7.5. *For any lazy irreducible birth and death chain with $t_{\mathrm{rel}} \geq 2$, and any $0 < \varepsilon \leq 1/4$,*

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{3}{2} t_{\mathrm{hit}} + 6 t_{\mathrm{rel}} \log(2/\varepsilon) \leq 74 t_{\mathrm{mix}}(\varepsilon).$$

The first inequality is (7.6) with $\delta = 1/2$, so the new statement is the second inequality.

Proof. Since $t_{\mathrm{rel}} \geq 2$ and $\varepsilon \leq 1/4$, $t_{\mathrm{rel}} \leq 2(t_{\mathrm{rel}} - 1)$ and $\log(2/\varepsilon) \leq 3\log(1/2\varepsilon)$. Thus,

$$\frac{3}{2} t_{\mathrm{hit}} + 6 t_{\mathrm{rel}} \log(2/\varepsilon) \leq \frac{3}{2} t_{\mathrm{hit}} + 36 (t_{\mathrm{rel}} - 1) \log(1/2\varepsilon).$$

By Theorem 7.1, $t_{\mathrm{hit}} \leq 25 t_{\mathrm{mix}}(1/4) \leq 25 t_{\mathrm{mix}}(\varepsilon)$. By Proposition 7.2, $(t_{\mathrm{rel}} - 1)\log(1/2\varepsilon) \leq t_{\mathrm{mix}}(\varepsilon)$. Hence

$$\frac{3}{2} t_{\mathrm{hit}} + 36 (t_{\mathrm{rel}} - 1) \log(1/2\varepsilon) \leq \left(\frac{3}{2} \times 25 + 36\right) t_{\mathrm{mix}}(\varepsilon) \leq 74 t_{\mathrm{mix}}(\varepsilon).$$ □

The constant 74 could definitely be improved. Proposition 7.5 can be interpreted as a reaffirmation that the bounds (7.6) and (7.7) have the right dependence on ε.

Proof of Theorem 7.3. The drift function will be $V(x) = \mathbf{E}_x[\lambda^{-\tau_m}]$, where $\lambda < 1$ is to be chosen later. As discussed in Chapter 2, $PV(x) = \lambda V(x)$ for $x \neq m$, so the drift condition is satisfied.

Note that for $0 \leq x \leq m$, $V(x) \leq V(0)$. This is proved by a monotone coupling argument. Let (X_t) and (X'_t) be two copies of the chain started at states 0 and x, respectively. Because the chain is lazy, there is a monotone coupling under which $X_t \leq X'_t$ for all t. If τ_m and τ'_m are the hitting times to state m for (X_t) and (X'_t), then $\tau'_m \leq \tau_m$. It follows that $V(x) \leq V(0)$. Similarly, if $m \leq x \leq n$, $V(x) \leq V(n)$.

The next ingredient is the key to this proof as well as to the arguments in [DSC06] and [DLP10]. It is an exact characterization of τ_m when the chain is started at 0. Define a transition matrix Q on $\{0, \ldots, m\}$ as follows: for $0 \leq i \leq m-1$, $Q(i,j) = P(i,j)$. Let $Q(m,m) = 1$ and $Q(m,j) = 0$ for $j \leq m-1$. So, Q is the transition matrix for the modification of P in which m is turned into an absorbing state, and the state space is restricted to $\{0, \ldots, m\}$. Denote the eigenvalues of Q by $1 = \theta_0 \geq \theta_1 \geq \ldots \geq \theta_m \geq 0$.

The following result is due to Karlin and McGregor [KM59] and sometimes attributed to Keilson [Kei79]. When the chain (X_t) with transition matrix P is started at 0, the hitting time τ_m has the same distribution as the sum of m independent geometric random variables $T_1, \ldots, T_m$, where each T_j has mean $1/(1-\theta_j)$. That is,

$$\mathbf{P}(T_j = k) = \theta_j^{k-1}(1-\theta_j), \qquad k \geq 1.$$

This means in particular that $\theta_1 < 1$.

The proofs of [KM59] and [Kei79] were analytic. More recently, Fill [Fil09b] has given a probabilistic proof; see also the independent work of Diaconis and Miclo [DM09] and the subsequent papers [Fil09a; Mic10]. An equivalent formulation of the result is that τ_m has moment generating function

$$\mathbf{E}_0[\alpha^{\tau_m}] = \prod_{j=1}^{m} \frac{\alpha(1-\theta_j)}{1-\alpha\theta_j},$$

which is finite for $\alpha < 1/\theta_1$. This is an exact formula for $V(0)$. When $\alpha\theta < 1$, one has

$$\frac{\alpha(1-\theta)}{1-\alpha\theta} = 1 + \frac{\alpha - 1}{1-\alpha\theta} \leq [1 + (\alpha-1)]^{1/(1-\alpha\theta)} = \alpha^{1/(1-\alpha\theta)},$$

where the inequality is $1 + cx \leq (1+x)^c$ for $c \geq 1$ and $x \geq -1$. (The function $x \mapsto (1+x)^c$ is convex, and $y = 1 + cx$ is the tangent line at $(0,1)$.) Therefore, if $\alpha < 1/\theta_1$, $\mathbf{E}_0[\alpha^{\tau_m}] \leq \alpha^{M(\alpha)}$ where

$$M(\alpha) = \sum_{j=1}^{m} \frac{1}{1-\alpha\theta_j}.$$

If $\alpha = 1$, then

$$M(1) = \sum_{j=1}^{m} \frac{1}{1-\theta_j} = \mathbf{E}_0[\tau_m].$$

Therefore, it is possible to choose α slightly greater than 1 for which $M(\alpha) \leq (1+\delta)\,\mathbf{E}_0[\tau_m]$. (Recall that $\delta > 0$ is fixed.) It turns out that

$$\alpha \leq \frac{1+\delta/\theta_1}{1+\delta} = \frac{1}{\theta_1}\left[1 - \frac{1-\theta_1}{1+\delta}\right] \tag{7.9}$$

is a sufficient condition. To see why, note first that $1 + \delta/\theta_1 > 1 + \delta$, so the right side of (7.9) is greater than 1. Also, because $(1-\theta_1)/(1+\delta) > 0$, the right side of (7.9) is less than $1/\theta_1$. In addition, if (7.9) is satisfied, then

$$\alpha \leq \frac{1+\delta/\theta_j}{1+\delta}$$

for all $j \geq 1$. Finally,

$$\alpha \leq \frac{1}{\theta}\left[1 - \frac{1-\theta}{1+\delta}\right] \Longleftrightarrow \frac{1}{1-\alpha\theta} \leq \frac{1+\delta}{1-\theta}.$$

Thus, if (7.9) holds, it follows that

$$\frac{1}{1-\alpha\theta_j} \leq \frac{1+\delta}{1-\theta_j}$$

for all j, and therefore $M(\alpha) \leq (1+\delta)\,\mathbf{E}_0[\tau_m]$.

The end goal of Theorem 7.3 is an inequality involving the relaxation time of the chain (X_t), which is determined by the spectral gap γ of P. So far all the bounds have been given in terms of the eigenvalues θ_j of Q. The condition (7.9) depends only on the spectral gap $\gamma' = 1 - \theta_1$ of Q. Relating the spectral gaps of P and Q was also an important step in the argument of [DLP10]. Their Lemma 2.7 says that $\gamma' \geq \gamma/2$, meaning that $\theta_1 \leq 1 - \gamma/2$.

Define

$$\lambda = \frac{1+\delta}{1+\delta/(1-\gamma/2)} < 1.$$

By the lemma, $\alpha = \lambda^{-1}$ satisfies (7.9), so $M(\alpha) \leq (1+\delta)\,\mathbf{E}_0[\tau_m]$, and

$$2V(0)\lambda^{t+1} = 2\,\mathbf{E}_0[\alpha^{\tau_m}]\alpha^{-(t+1)} \leq 2\alpha^{M(\alpha)-t-1} \leq 2\alpha^{(1+\delta)\,\mathbf{E}_0[\tau_m]-t-1}.$$

The same argument can be applied to the restriction of the chain (X_t) to the subset $\{m, \ldots, n\}$, where m is made an absorbing state. (In particular, the spectral gap of that chain is also at least

$\gamma/2$.) One obtains

$$2V(n)\lambda^{t+1} \le 2\alpha^{(1+\delta)\,\mathbf{E}_n[\tau_m]-t-1}.$$

Since $V(x) \le \max\{V(0), V(n)\}$ for all $0 \le x \le n$, the total variation bound from Theorem 4.5 yields

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \le 2V(x)\lambda^{t+1} \le 2\alpha^{(1+\delta)t_{\mathrm{hit}}-t-1}. \tag{7.10}$$

The right side of (7.10) is less than ε exactly when

$$t > (1+\delta)t_{\mathrm{hit}} + \frac{1}{\log\alpha}\log(2/\varepsilon) - 1. \tag{7.11}$$

Using that $\log\lambda \le \lambda - 1$, it follows that $1/\log\alpha \le 1/(1-\lambda)$. Therefore,

$$t_0 = \left\lceil (1+\delta)t_{\mathrm{hit}} + \frac{1}{1-\lambda}\log(2/\varepsilon) - 1 \right\rceil$$

satisfies (7.11), and

$$t_{\mathrm{mix}}(\varepsilon) \le t_0 \le (1+\delta)t_{\mathrm{hit}} + \frac{1}{1-\lambda}\log(2/\varepsilon).$$

By the definition of λ,

$$\frac{1}{1-\lambda} = \frac{2+2\delta-\gamma}{\delta\gamma} \le \frac{2+2\delta}{\delta\gamma} = 2\left(1+\frac{1}{\delta}\right)t_{\mathrm{rel}}.$$

Hence

$$t_{\mathrm{mix}}(\varepsilon) \le (1+\delta)t_{\mathrm{hit}} + 2\left(1+\frac{1}{\delta}\right)t_{\mathrm{rel}}\log(2/\varepsilon),$$

finishing the proof. □

Proof of lower bound in Theorem 7.1. This is a minor variation of the proof of Theorem 3.9 in [CSC13b]. Let m be the median state used to define t_{hit}, and set $\mu = \mathbf{E}_0[\tau_m]$. It will be shown that $t_{\mathrm{mix}}(1/4) \ge \mu/25$. By symmetry, one also has $t_{\mathrm{mix}}(1/4) \ge \mathbf{E}_n[\tau_m]/25$, which finishes the proof.

To show that $t_{\mathrm{mix}}(1/4) \ge \mu/25$, first note that when $m = 0$ the right side is zero, so the inequality is trivially true. When $m \ge 1$, it will be shown that

$$\mathbf{P}_0(\tau_m > \mu/25) > 3/4. \tag{7.12}$$

Given (7.12), for all $t \le \mu/25$, $P^t(0, \{m,\dots,n\}) \le \mathbf{P}_0(\tau_m \le t) < 1/4$, while $\pi(\{m,\dots,n\}) \ge 1/2$. It follows that $\|P^t(0,\cdot) - \pi\|_{\mathrm{TV}} > 1/4$, so $t_{\mathrm{mix}}(1/4) \ge \mu/25$. Hence it will suffice to prove (7.12).

If $m = 1$, $\tau_m = \tau_1$ is a geometric random variable when the chain is started at 0. Its distribution is

$$\mathbf{P}_0(\tau_1 = k) = \left(1 - \frac{1}{\mu}\right)^{k-1} \frac{1}{\mu}, \qquad k \geq 1.$$

For any real $t \geq 0$,

$$\mathbf{P}_0(\tau_1 > t) \geq \left(1 - \frac{1}{\mu}\right)^{t},$$

with equality when t is an integer. Note that $\mu \geq 2$ because of the laziness. It follows that

$$\mathbf{P}_0(\tau_1 > \mu/25) \geq \left(1 - \frac{1}{\mu}\right)^{\mu/25} \geq \left(1 - \frac{1}{2}\right)^{2/25} > 0.94,$$

where the second inequality used that the function $x \mapsto (1 - 1/x)^x$ is increasing for $x > 0$.

If $m \geq 2$, then $\mu \geq 4$. (See Lemma 3.3 in [CSC13b].) It is shown as part of the proof of Theorem 3.9 in [CSC13b] that if $1/\mu < a < 1$ and $t < \mu$,

$$\mathbf{P}_0(\tau_m > t) \geq \min\left\{\left(1 - \frac{1}{a\mu}\right)^{t}, \frac{(\mu - t)^2}{a\mu^2 + (\mu - t)^2}\right\}.$$

Set $a = 0.3$ and $t = \mu/25$. Then

$$\left(1 - \frac{1}{0.3\mu}\right)^{\mu/25} \geq \left(1 - \frac{1}{0.3 \times 4}\right)^{4/25} > 0.75,$$

again using that $x \mapsto (1 - 1/x)^x$ is increasing. As well,

$$\frac{(\mu - \mu/25)^2}{0.3\mu^2 + (\mu - \mu/25)^2} = \frac{(24/25)^2}{0.3 + (24/25)^2} > 0.75.$$

This proves (7.12), and the lower bound of Theorem 7.1 follows. □

7.2 Random walk on the hypercube

Let $\mathcal{X} = (\mathbf{Z}/2\mathbf{Z})^n = \{(x_1, \ldots, x_n) : \text{each } x_i \text{ is 0 or 1}\}$ be the n-dimensional hypercube. The *Hamming weight* $\|x\|$ of an element $x = (x_1, \ldots, x_n) \in \mathcal{X}$ is the number of entries x_i that equal 1. The lazy simple random walk on $\mathcal{X}$ is the Markov chain whose transition matrix P is given by

$$P(x,y) = \begin{cases} \frac{1}{2} & \text{if } x = y, \\ \frac{1}{2n} & \text{if } \|x-y\| = 1, \\ 0 & \text{if } \|x-y\| \geq 2. \end{cases}$$

This section uses the method of drift and minorization to find a lower bound on the spectral gap of P, which gives an upper bound on mixing time.

Since the hypercube walk is already well understood using other methods, some explanation is in order. The philosophy of drift and minorization is that one bounds the convergence to stationarity of a Markov chain using the hitting time of a small set C. The application to birth and death chains in the previous section is a good example: the chain converges roughly when it reaches the median element. For the hypercube, the hitting time of any particular state is exponential in n, while the mixing time has order $n \log n$. It might seem that the method of drift and minorization is not well-suited to this type of example: if the set C is very small, the hitting time is far greater than the mixing time of the chain; but if the set C is too large, the minorization is problematic.

This situation is likely to occur for any Markov chain that makes local moves on a high-dimensional space. Can the drift-minorization approach yield useful bounds in such cases? This section demonstrates that the answer is yes for the toy example of the hypercube. One caveat: the argument strongly relies on special properties of the hypercube, so it may not generalize well.

See [LPW09] for standard analyses of the hypercube walk using strong stationary times and eigenfunction decomposition. In Example 12.15 of that book it is shown that the spectral gap of the transition matrix P is $1/n$. The following theorem is the main result of this section.

Theorem 7.6. *Let P be the transition matrix for the lazy simple random walk on the n-dimensional hypercube. The spectral gap of P is at least $1/2n$.*

This theorem differs from the correct answer by a factor of 2. It is useful not for the result itself but as a demonstration of how far one can get using drift and minorization.

Note that Theorem 7.6 only provides a bound on the spectral gap, while the results of Chapter 4 also give bounds on total variation convergence. It turns out that the total variation bound from Theorem 4.4 is not nearly as sharp as the combination of Theorem 7.6 with the standard bound

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \frac{1}{2\sqrt{\pi(x)}}(1-\gamma)^t, \tag{7.13}$$

where γ is the spectral gap of P ([DS91], Proposition 3). Combining Theorem 7.6 with (7.13) shows that the lazy hypercube walk has mixing time of order at most n^2. In fact, the walk exhibits cutoff at time $\frac{1}{2}n \log n$; see Theorem 18.3 of [LPW09].

The rest of this section is devoted to the proof of Theorem 7.6. Here is an outline. The small set

$$C = \{(x_1, \dots, x_n) \in \mathcal{X} \,:\, x_n = 0\} \tag{7.14}$$

is defined to be half the space. For any $x \in \mathcal{X} \setminus C$, one has $P(x, C) = 1/2n$. Therefore the hitting time of C is a geometric random variable with parameter $1/2n$.

Since C is so large, the difficulty lies in the minorization. Certainly there is no 1-step minorization. It turns out that the m-step minorization as $m \to \infty$ is related to the mixing of the $(n-1)$-dimensional hypercube walk. This suggests a proof by induction. Theorem 7.6 in $n-1$ dimensions provides a lower bound on the minorization constant

$$\varepsilon(m) = \sum_{x \in \mathcal{X}} \min_{c \in C} P^m(c, x).$$

Indeed, it will be proved that

$$\varepsilon(m) \geq 1 - 2^{n-1}\left(1 - \frac{1}{2n}\right)^m.$$

This sets up the use of Theorem 4.6, which finishes the induction. As an aside, the general strategy of inducting on the dimension is reminiscent of a method of H.-T. Yau [Yau96] for proving log-Sobolev inequalities via multiscale analysis.

Proof of Theorem 7.6. The proof is by induction on n. The base case $n = 1$ is trivial. Suppose the theorem is true in dimension $n-1$. Define C as in (7.14). Since $P(x, C) = 1/2n$ for every $x \in \mathcal{X} \setminus C$, if one fixes λ in the range $1 - \frac{1}{2n} < \lambda < 1$, the function

$$V(x) = \begin{cases} 1 & \text{if } x \in C, \\ \frac{1/2n}{\lambda - (1 - 1/2n)} & \text{if } x \notin C, \end{cases}$$

satisfies $PV(x) = \lambda V(x)$ for $x \notin C$. In addition, $V(x) \leq M = \frac{1/2n}{\lambda - (1-1/2n)}$ for all $x \in \mathcal{X}$.

The projection onto the first $n-1$ dimensions of the lazy hypercube walk is itself a Markov chain on the state space $\mathcal{Y} = (\mathbf{Z}/2\mathbf{Z})^{n-1}$. Its transition matrix Q is defined by

$$Q(x, y) = \begin{cases} \frac{1}{2} + \frac{1}{2n} & \text{if } x = y, \\ \frac{1}{2n} & \text{if } \|x - y\| = 1, \\ 0 & \text{if } \|x - y\| \geq 2. \end{cases}$$

If P' is the transition matrix for the 1/2-lazy hypercube walk in $n-1$ dimensions, then $Q =$

$(1-\frac{1}{n})P' + \frac{1}{n}I$. By the inductive hypothesis, the spectral gap of P' is at least $1/2(n-1)$. It follows that the spectral gap of Q is at least $1/2n$.

For $m \geq 1$, the largest possible m-step minorization constant associated with the small set C is

$$\varepsilon(m) = \sum_{x \in \mathcal{X}} \min_{c \in C} P^m(c,x).$$

The next goal is to prove that

$$\sum_{x \in \mathcal{X}} \min_{c \in C} P^m(c,x) = \sum_{y \in \mathcal{Y}} \min_{d \in \mathcal{Y}} Q^m(d,y). \tag{7.15}$$

Fix $y = (y_1, \ldots, y_{n-1}) \in \mathcal{Y}$. Let $x_0 = (y_1, \ldots, y_{n-1}, 0)$ and $x_1 = (y_1, \ldots, y_{n-1}, 1)$. If

$$\min_{c \in C} P^m(c,x_0) + \min_{c \in C} P^m(c,x_1) = \min_{d \in \mathcal{Y}} Q^m(d,y), \tag{7.16}$$

then summing over all choices of $(y_1, \ldots, y_{n-1})$ will yield (7.15). Therefore it suffices to show (7.16).

For each $1 \leq j \leq n-1$, define $y_j' = 1 - y_j$. By the monotonicity of the hypercube walk, $P^m(c,x_0)$ and $P^m(c,x_1)$ are minimized at $c_0 = (y_1', \ldots, y_{n-1}', 0)$, and $Q^m(d,y)$ is minimized at $d_0 = (y_1', \ldots, y_{n-1}')$. So it suffices to show that $P^m(c_0,x_0) + P^m(c_0,x_1) = Q^m(d_0,y)$. This is true from the definition of Q as a projection of P. If (X_t) has transition matrix P and (Y_t) has transition matrix Q,

$$Q^m(d_0,y) = \mathbf{P}_{d_0}(Y_m = y) = \mathbf{P}_{c_0}(X_m \in \{x_0, x_1\}) = P^m(c_0,x_0) + P^m(c_0,x_1).$$

This proves (7.16), and (7.15) follows.

The lower bound on the spectral gap of Q gives a lower bound on the right side of (7.15). Equation (12.11) of [LPW09] implies that if R is the transition matrix for a lazy reversible Markov chain with spectral gap γ and uniform stationary distribution on a state space of size N, then

$$\left| R^t(x,y) - \frac{1}{N} \right| \leq (1-\gamma)^t$$

for any two states x, y. Applying this result to Q yields

$$\varepsilon(m) = \sum_{y \in \mathcal{Y}} \min_{d \in \mathcal{Y}} Q^m(d,y) \geq \sum_{y \in \mathcal{Y}} \left[2^{-(n-1)} - \left(1 - \frac{1}{2n}\right)^m \right] = 1 - 2^{n-1}\left(1 - \frac{1}{2n}\right)^m.$$

This lower bound on the minorization constant sets up the use of Theorem 4.6. For each $m \geq 1$, P satisfies a uniform drift and m-step minorization condition with constants λ and $\varepsilon(m)$. Theorem 4.6 implies that the spectral gap of P is at least $1 - \max\{1 - \frac{1}{2n}, \lambda\} = 1 - \lambda$. Since this result holds for any $\lambda > 1 - \frac{1}{2n}$, the spectral gap of P must be at least $\frac{1}{2n}$. This completes the proof. □

Appendix A

Proofs from Chapter 3

Proof of Proposition 3.6. First, introduce an auxiliary sequence $Y_0, Y_1, \ldots$ of Uniform$[0,1]$ random variables, independent of each other and of (X_t). Technically this is done by extending the sample space $(\Omega, \mathcal{F})$ to $(\bar{\Omega}, \bar{\mathcal{F}}) = (\Omega \times [0,1]^\infty, \mathcal{F} \otimes \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $[0,1]^\infty$ with the product topology. Denote an element of $[0,1]^\infty$ by $Y = (Y_0, Y_1, \ldots)$, and an element of $\bar{\Omega}$ by $\bar{\omega} = (\omega, Y)$. Extend each measure $\mathbf{P}_\mu \in \mathcal{P}(\Omega)$ to $\bar{\Omega}$ by letting $\mathbf{P}_\mu(E \times E') = \mathbf{P}_\mu(E)\lambda^\infty(E')$ for all events $E \in \mathcal{F}$ and $E' \in \mathcal{B}$, where λ^∞ is the usual product measure on $([0,1]^\infty, \mathcal{B})$. It is easily seen that the Y_t are Uniform$[0,1]$ random variables and that the sequence $(\omega, Y_0, Y_1, \ldots)$ is independent.

To prove that the compatibility condition (1.5) holds for the measures $\mathbf{P}_\mu$ on $\bar{\Omega}$, suppose $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ are given with μ' absolutely continuous with respect to μ. It will suffice to show for all events $E \in \mathcal{F}$ and $E' \in \mathcal{B}$ that

$$\mathbf{P}_{\mu'}(E \times E') = \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\bar{\omega} \in E \times E'\}\right]. \tag{A.1}$$

The left side of (A.1) is

$$\mathbf{P}_{\mu'}(E)\lambda^\infty(E') = \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega \in E\}\right]\lambda^\infty(E').$$

The right side of (A.1) is

$$\begin{aligned}\mathbf{E}_\mu\left[\mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega \in E\}\mathbf{1}\{Y \in E'\}\,\middle|\,\omega\right]\right] &= \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega \in E\}\,\mathbf{P}_\mu(Y \in E' \mid \omega)\right] \\ &= \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega \in E\}\right]\lambda^\infty(E').\end{aligned}$$

Thus the compatibility condition is satisfied.

The sequence (T_j) can now be defined on $\bar{\Omega}$. Let $T_1 = T$; technically this means $T_1(\omega, Y) = T(\omega)$. For $j \geq 1$, let

$$T_{j+1} - T_j = \min\{\ell \geq 0 \,:\, Y_{T_j+\ell} < r_\ell(X_{T_j}, X_{T_j+1}, \ldots, X_{T_j+\ell})\}. \tag{A.2}$$

To show that each T_j is a randomized stopping time for (X_t), fix the initial distribution μ. The proof is by induction on j. The base case is true because $T_1 = T$. For the inductive step, suppose that $k < n$. Because T_j is a randomized stopping time, the event $\{T_j = k\}$ is conditionally independent of $(X_{n+1}, X_{n+2}, \ldots)$ under $\mathbf{P}_\mu$ given $X_0, \ldots, X_n$. By (A.2), the event $\{T_{j+1} = n\}$ is conditionally independent of $(X_{n+1}, X_{n+2}, \ldots)$ given that $T_j = k$ and given $X_0, \ldots, X_n$. Hence the event $\{T_{j+1} = n, T_j = k\}$ is conditionally independent of $(X_{n+1}, X_{n+2}, \ldots)$ given $X_0, \ldots, X_n$. Summing over k yields that T_{j+1} is a randomized stopping time.

The next requirement is that $T_{j+1} - T_j$ should be conditionally independent of $(X_0, \ldots, X_{k-1})$ under $\mathbf{P}_\mu$ given that $T_j = k$ and given $X_k, X_{k+1}, \ldots$. This is clear from the definition (A.2).

Finally, it must be checked that

$$\mathbf{P}_\mu(T_{j+1} - T_j = \ell \mid T_j = k, X_k, X_{k+1}, \ldots) = f_\ell(X_k, X_{k+1}, \ldots, X_{k+\ell}).$$

The left side equals

$$\begin{aligned}\mathbf{P}_\mu &\left(Y_{k+i} \geq r_i(X_k, \ldots, X_{k+i}) \text{ for all } 0 \leq i \leq \ell - 1, \text{ and } Y_{k+\ell} < r_\ell(X_k, \ldots, X_{k+\ell})\right) \\ &= \left(\prod_{i=0}^{\ell-1} \Big[1 - r_i(X_k, \ldots, X_{k+i})\Big]\right) r_\ell(X_k, \ldots, X_{k+\ell}) = f_\ell(X_k, \ldots, X_{k+\ell}),\end{aligned}$$

where the first equality comes from the conditional independence of the Y_t given $(X_0, X_1, \ldots)$, and the second equality is (3.4). This completes the proof. □

Proof of Proposition 3.7. Fix an initial distribution μ. The following statements will be proved for all $j \geq 1$.

(1) If T and T_j are weak ν times, then for all k with $\mathbf{P}_\mu(T_j = k) > 0$,

$$\mathbf{P}_\mu(X_{T_{j+1}} \in A \mid T_j = k, X_0, \ldots, X_k) = \nu(A) \quad \text{for all } A \in \mathcal{E}.$$

(2) If T and T_j are strong ν times, then for all k, ℓ with $\mathbf{P}_\mu(T_j = k, T_{j+1} - T_j = \ell) > 0$,

$$\mathbf{P}_\mu(X_{T_{j+1}} \in A \mid T_j = k, T_{j+1} - T_j = \ell, X_0, \ldots, X_k) = \nu(A) \quad \text{for all } A \in \mathcal{E}.$$

(3) If T and T_j are ν-regeneration times, then for all k, ℓ with $\mathbf{P}_\mu(T_j = k, T_{j+1} - T_j = \ell) > 0$,

$$\mathbf{P}_\mu(X_{T_{j+1}} \in A \mid T_j = k, T_{j+1} - T_j = \ell, X_0, \ldots, X_{k+\ell-1}) = \nu(A) \quad \text{for all } A \in \mathcal{E}.$$

Statement (1) implies that T_{j+1} is a weak ν time for (X_t); statement (2) implies that T_{j+1} is a strong ν time; and statement (3) implies that T_{j+1} is a ν-regeneration time. This sets up an induction on j. Since $T_1 = T$, the base case for (1) (respectively (2), (3)) holds if T is a weak ν time (respectively strong ν time, ν-regeneration time). Using the Markov property, (1) proves (i); (2) proves (ii); and (3) proves (iii). Therefore it will suffice to show (1), (2), and (3).

Suppose $\mathbf{P}_\mu(T_j = k) > 0$. Define the measure $\eta(\cdot) = \mathbf{P}_\mu(X_k \in \cdot \mid T_j = k)$. For (2) and (3), $\eta = \nu$. For (1), η is absolutely continuous with respect to ν, since

$$\eta(A) = \frac{\mathbf{P}_\mu(X_{T_j} \in A, T_j = k)}{\mathbf{P}_\mu(T_j = k)} \leq \frac{\mathbf{P}_\mu(X_{T_j} \in A)}{\mathbf{P}_\mu(T_j = k)} = \frac{\nu(A)}{\mathbf{P}_\mu(T_j = k)} \quad \text{for all } A \in \mathcal{E}.$$

For any event $E \in \mathcal{E}^\infty$, it will now be shown that

$$\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k) = \mathbf{P}_{x_k}((X_0, X_1, \ldots) \in E). \tag{A.3}$$

(Note that the right side is indeed a measurable function of x_k, because $E \in \mathcal{E}^\infty$.) Since T_j is a randomized stopping time, the dependence on $\{T_j = k\}$ can be removed from the left side. Then, by the Markov property, the dependence on $X_0, \ldots, X_{k-1}$ can also be removed from the left side. To see that

$$\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E \mid X_k = x_k) = \mathbf{P}_{x_k}((X_0, X_1, \ldots) \in E),$$

it suffices to check the integrated form, namely that for all $A \in \mathcal{E}$,

$$\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E, X_k \in A) = \int_A \mathbf{P}_{x_k}((X_0, X_1, \ldots) \in E)\, \mathbf{P}_\mu(X_k \in dx_k).$$

Both the left and right side are equal to the integral of $\mathbf{1}\{(x_k, x_{k+1}, \ldots) \in E, x_k \in A\}$ with respect to the measure $P^k(\mu, dx_k)P(x_k, dx_{k+1})P(x_{k+1}, dx_{k+2}) \cdots$. This proves (A.3). As a consequence, for any bounded measurable function $g : \mathcal{X}^\infty \to \mathbf{R}$,

$$\mathbf{E}_\mu[g(X_k, X_{k+1}, \ldots) \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k] = \mathbf{E}_{x_k}[g(X_0, X_1, \ldots)]. \tag{A.4}$$

The next step is to show that

$$\begin{aligned}\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E, T_{j+1} - T_j = \ell \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k)& \\ = \mathbf{P}_\eta((X_0, X_1, \ldots) \in E, T = \ell \mid X_0 = x_k).&\end{aligned} \tag{A.5}$$

Let $X = (X_0, X_1, \ldots)$. The left side of (A.5) equals

$$\begin{aligned}
&\mathbf{E}_\mu[\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E, T_{j+1} - T_j = \ell \mid T_j = k, X) \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k] \\
&= \mathbf{E}_\mu[\mathbf{1}\{(X_k, X_{k+1}, \ldots) \in E\}\, \mathbf{P}_\mu(T_{j+1} - T_j = \ell \mid T_j = k, X) \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k] \\
&= \mathbf{E}_\mu[\mathbf{1}\{(X_k, X_{k+1}, \ldots) \in E\} f_\ell(X_k, \ldots, X_{k+\ell}) \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k] \\
&= \mathbf{E}_{x_k}[\mathbf{1}\{(X_0, X_1, \ldots) \in E\} f_\ell(X_0, \ldots, X_\ell)],
\end{aligned}$$

using (A.4) in the last equality. The right side of (A.5) equals

$$\begin{aligned}
&\mathbf{E}_\eta[\mathbf{P}_\eta((X_0, X_1, \ldots) \in E, T = \ell \mid X) \mid X_0 = x_k] \\
&= \mathbf{E}_\eta[\mathbf{1}\{(X_0, X_1, \ldots) \in E\}\, \mathbf{P}_\eta(T = \ell \mid X) \mid X_0 = x_k].
\end{aligned}$$

Since η is absolutely continuous with respect to ν, using (1.5),

$$\mathbf{P}_\eta(T = \ell \mid X) = \mathbf{E}_\nu\left[\frac{d\eta}{d\nu}(X_0)\mathbf{1}\{T = \ell\} \,\middle|\, X\right] = \frac{d\eta}{d\nu}(X_0)\, \mathbf{P}_\nu(T = \ell \mid X) = \frac{d\eta}{d\nu}(X_0) f_\ell(X_0, \ldots, X_\ell).$$

Therefore, the right side of (A.5) equals

$$\begin{aligned}
&\mathbf{E}_\eta\left[\mathbf{1}\{(X_0, X_1, \ldots) \in E\}\frac{d\eta}{d\nu}(X_0) f_\ell(X_0, \ldots, X_\ell) \,\middle|\, X_0 = x_k\right] \\
&= \mathbf{E}_\nu[\mathbf{1}\{(X_0, X_1, \ldots) \in E\} f_\ell(X_0, \ldots, X_\ell) \mid X_0 = x_k] \\
&= \mathbf{E}_{x_k}[\mathbf{1}\{(X_0, X_1, \ldots) \in E\} f_\ell(X_0, \ldots, X_\ell)].
\end{aligned}$$

(Note that the last equality holds ν-almost everywhere and hence also η-almost everywhere.) This proves (A.5).

With this preparation, it is now possible to prove (1)-(3). For (1),

$$\begin{aligned}
&\mathbf{P}_\mu(X_{T_{j+1}} \in A \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k) \\
&= \sum_{\ell=0}^{\infty} \mathbf{P}_\mu(X_{k+\ell} \in A, T_{j+1} - T_j = \ell \mid T_j = k, X_0 = x_0, \ldots, X_k = x_k) \\
&= \sum_{\ell=0}^{\infty} \mathbf{P}_\eta(X_\ell \in A, T = \ell \mid X_0 = x_k) = \mathbf{P}_\eta(X_T \in A \mid X_0 = x_k).
\end{aligned}$$

The final step for (1) is to show that

$$\mathbf{P}_\eta(X_T \in A \mid X_0 = x_k) = \nu(A). \tag{A.6}$$

It may not work to say that

$$\mathbf{P}_\eta(X_T \in A \mid X_0 = x_k) = \mathbf{P}_{x_k}(X_T \in A) = \nu(A),$$

because the function $x \mapsto \mathbf{P}_x(X_T \in A)$ may not be measurable, so (1.4) may not hold. Fortunately, (A.6) can be proved using the weaker compatibility condition (1.5). It is enough to check the integrated form of (A.6): for all $B \in \mathcal{E}$,

$$\mathbf{P}_\eta(X_T \in A, X_0 \in B) = \nu(A)\eta(B). \tag{A.7}$$

If $\eta(B) = 0$, both sides of (A.7) are zero. If $\eta(B) > 0$, define the probability measure η_B by $\eta_B(U) = \eta(U \cap B)/\eta(B)$. Since $\eta_B(U) \le \eta(U)/\eta(B)$, the measure η_B is absolutely continuous with respect to η. In fact,

$$\frac{d\eta_B}{d\eta}(x) = \frac{\mathbf{1}\{x \in B\}}{\eta(B)}.$$

Therefore,

$$\mathbf{P}_\eta(X_T \in A, X_0 \in B) = \mathbf{E}_\eta\left[\mathbf{1}\{X_T \in A\}\eta(B)\frac{d\eta_B}{d\eta}(X_0)\right] = \mathbf{E}_{\eta_B}[\mathbf{1}\{X_T \in A\}]\eta(B) = \nu(A)\eta(B),$$

using that T is a weak ν time in the last equality. This proves (A.6), so the proof of (1) is finished.

For (2), recall that $\eta = \nu$. Using a conditioned form of (A.5),

$$\mathbf{P}_\mu(X_{k+\ell} \in A \mid T_j = k, T_{j+1} - T_j = \ell, X_0 = x_0, \dots, X_k = x_k) = \mathbf{P}_\nu(X_\ell \in A \mid T = \ell, X_0 = x_k).$$

To prove that

$$\mathbf{P}_\nu(X_\ell \in A \mid T = \ell, X_0 = x_k) = \nu(A),$$

it is enough to check the integrated form: for all $B \in \mathcal{E}$,

$$\mathbf{P}_\nu(X_\ell \in A, X_0 \in B \mid T = \ell) = \nu(A)\,\mathbf{P}_\nu(X_0 \in B \mid T = \ell). \tag{A.8}$$

When proving (A.8) it is legal to assume that $\mathbf{P}_\nu(T = \ell) > 0$. If $\mathbf{P}_\nu(X_0 \in B \mid T = \ell) = 0$, both sides of (A.8) are zero. If $\mathbf{P}_\nu(X_0 \in B \mid T = \ell) > 0$,

$$\mathbf{P}_\nu(X_\ell \in A, X_0 \in B \mid T = \ell) = \frac{\mathbf{P}_\nu(X_\ell \in A, X_0 \in B, T = \ell)}{\mathbf{P}_\nu(X_0 \in B, T = \ell)}\,\mathbf{P}_\nu(X_0 \in B \mid T = \ell),$$

so it will suffice to show that

$$\frac{\mathbf{P}_\nu(X_\ell \in A, X_0 \in B, T = \ell)}{\mathbf{P}_\nu(X_0 \in B, T = \ell)} = \nu(A). \tag{A.9}$$

As in the proof of (1), define the probability measure ν_B by $\nu_B(U) = \nu(U \cap B)/\nu(B)$, so that

$$\frac{d\nu_B}{d\nu}(x) = \frac{\mathbf{1}\{x \in B\}}{\nu(B)}.$$

Then

$$\mathbf{P}_\nu(X_\ell \in A, X_0 \in B, T = \ell) = \mathbf{E}_\nu \left[\mathbf{1}\{X_\ell \in A, T = \ell\} \nu(B) \frac{d\nu_B}{d\nu}(X_0) \right] = \nu(B)\, \mathbf{P}_{\nu_B}(X_\ell \in A, T = \ell).$$

This can be plugged directly into the numerator of (A.9). For the denominator of (A.9), use the same computation but with A replaced by $\mathcal{X}$. The result is

$$\frac{\mathbf{P}_\nu(X_\ell \in A, X_0 \in B, T = \ell)}{\mathbf{P}_\nu(X_0 \in B, T = \ell)} = \frac{\nu(B)\, \mathbf{P}_{\nu_B}(X_\ell \in A, T = \ell)}{\nu(B)\, \mathbf{P}_{\nu_B}(T = \ell)} = \mathbf{P}_{\nu_B}(X_\ell \in A \mid T = \ell) = \nu(A),$$

using at the end that T is a strong ν time. This proves (2).

For (3), a conditioned form of (A.5) gives the result immediately:

$$\begin{aligned} &\mathbf{P}_\mu(X_{k+\ell} \in A \mid T_j = k, T_{j+1} - T_j = \ell, X_0 = x_0, \ldots, X_{k+\ell-1} = x_{k+\ell-1}) \\ &\qquad = \mathbf{P}_\nu(X_\ell \in A \mid T = \ell, X_0 = x_k, X_1 = x_{k+1}, \ldots, X_{\ell-1} = x_{k+\ell-1}) = \nu(A), \end{aligned}$$

since T is a ν-regeneration time. □

Proof of Proposition 3.9. To prove that π given by (3.5) is stationary, for any $A \in \mathcal{E}$,

$$\begin{aligned} P(\pi, A) &= \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(X_{n+1} \in A, T > n) \\ &= \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \Big[\mathbf{P}_\nu(X_{n+1} \in A, T = n+1) + \mathbf{P}_\nu(X_{n+1} \in A, T > n+1) \Big] \\ &= \frac{1}{\mathbf{E}_\nu[T]} \mathbf{P}_\nu(X_T \in A) + \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=1}^{\infty} \mathbf{P}_\nu(X_n \in A, T > n) \\ &= \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(X_n \in A, T > n) = \pi(A). \end{aligned}$$

In addition,

$$\pi(\mathcal{X}) = \frac{1}{\mathbf{E}_\nu[T]} \sum_{n=0}^{\infty} \mathbf{P}_\nu(T > n) = 1.$$

To prove uniqueness, suppose for contradiction that π_1 and π_2 are two different stationary distributions for (X_t). Then, using the Hahn decomposition for the signed measure $\pi_1 - \pi_2$, $\mathcal{X}$ can be decomposed into disjoint subsets $\mathcal{X} = \mathcal{X}_+ \cup \mathcal{X}_-$ such that $\pi_1 - \pi_2$ is a positive measure on $\mathcal{X}_+$ and a negative measure on $\mathcal{X}_-$. (This strategy is due to Nummelin and Arjas [NA76].) Since π_1 and π_2 are different, $(\pi_1 - \pi_2)(\mathcal{X}_+) = (\pi_2 - \pi_1)(\mathcal{X}_-) > 0$. Define the probability measures μ_1, μ_2 on $\mathcal{X}_+$ and $\mathcal{X}_-$ respectively by

$$\begin{aligned} \mu_1(A) &= \frac{(\pi_1 - \pi_2)(A)}{(\pi_1 - \pi_2)(\mathcal{X}_+)}, \qquad && A \in \mathcal{E}, A \subseteq \mathcal{X}_+, \\ \mu_2(B) &= \frac{(\pi_2 - \pi_1)(B)}{(\pi_2 - \pi_1)(\mathcal{X}_-)}, \qquad && B \in \mathcal{E}, B \subseteq \mathcal{X}_-. \end{aligned}$$

Then μ_1 and μ_2 are also stationary distributions for (X_t), by the following argument. Since $\pi_1 - \pi_2$ is an invariant measure for (X_t),

$$\begin{aligned} (\pi_1 - \pi_2)(\mathcal{X}_+) &= \int_{\mathcal{X}_+} P(x, \mathcal{X}_+)(\pi_1 - \pi_2)(dx) - \int_{\mathcal{X}_-} P(x, \mathcal{X}_+)(\pi_2 - \pi_1)(dx) \\ &\le \int_{\mathcal{X}_+} (\pi_1 - \pi_2)(dx) - 0 = (\pi_1 - \pi_2)(\mathcal{X}_+). \end{aligned}$$

Hence the inequality in the middle is actually equality, and in particular,

$$\int_{\mathcal{X}_-} P(x, \mathcal{X}_+)(\pi_2 - \pi_1)(dx) = 0.$$

Suppose $A \subseteq \mathcal{X}_+$. Then also

$$\int_{\mathcal{X}_-} P(x, A)(\pi_2 - \pi_1)(dx) = 0.$$

Therefore,

$$\begin{aligned} P(\mu_1, A) &= \frac{1}{(\pi_1 - \pi_2)(\mathcal{X}_+)} \int_{\mathcal{X}_+} P(x, A)(\pi_1 - \pi_2)(dx) \\ &= \frac{1}{(\pi_1 - \pi_2)(\mathcal{X}_+)} \int_{\mathcal{X}} P(x, A)(\pi_1 - \pi_2)(dx) \\ &= \frac{1}{(\pi_1 - \pi_2)(\mathcal{X}_+)} (\pi_1 - \pi_2)(A) = \mu_1(A). \end{aligned}$$

This proves that μ_1 is stationary, and the argument for μ_2 is the same.

Since T is almost surely finite started from any $x \in \mathcal{X}$,

$$\sum_{n=1}^{\infty} \mathbf{P}_{\mu_1}(X_n \in \mathcal{X}_-, T = n) = \mathbf{P}_{\mu_1}(X_T \in \mathcal{X}_-) = \nu(\mathcal{X}_-).$$

However,

$$\sum_{n=1}^{\infty} \mathbf{P}_{\mu_1}(X_n \in \mathcal{X}_-, T = n) \le \sum_{n=1}^{\infty} \mathbf{P}_{\mu_1}(X_n \in \mathcal{X}_-) = \sum_{n=1}^{\infty} \mu_1(\mathcal{X}_-) = 0.$$

Thus $\nu(\mathcal{X}_-) = 0$, and by parallel reasoning $\nu(\mathcal{X}_+) = 0$ as well. This is impossible since $\nu(\mathcal{X}) = 1$. Therefore, the stationary distribution π given by (3.5) is unique. □

Proof of Proposition 3.10. The proof is broken into several sections. The main step comes at the very beginning, with a definition of the sequence (Y_t) that will satisfy all the necessary properties. The bulk of the proof is devoted to checking each property.

Other treatments of this subject (e.g. [Num84; MT93]) include the standing assumption that the σ-algebra $\mathcal{E}$ of measurable subsets of $\mathcal{X}$ is countably generated. This assumption is necessary for results like Theorem 3.3 (existence of a small set) and Theorem 2.7 (existence of a drift-minorization condition for any geometrically ergodic chain). The main results in this work say that any chain equipped with a drift-minorization condition must have certain convergence properties. These results do not require the assumption of countable generation.

The only real difficulty in removing that assumption comes in the proof of this proposition. One natural definition of the sequence (Y_t) uses a Radon-Nikodym derivative which may fail to be measurable when $\mathcal{E}$ is not countably generated. The definition used in the proof below avoids this pitfall at the cost of extra complication.

This is also the point where the compatibility condition (1.5) is crucial. If Definition 1.4 required the map $x \mapsto \mathbf{P}_x(E)$ to be measurable, Proposition 3.10 might not be true in full generality. The construction below does satisfy condition (1.5), which is enough for the subsequent results to work.

Definitions

Extend the sample space $(\Omega, \mathcal{F})$ of (X_t) to $(\bar{\Omega}, \bar{F}) = (\Omega \times \{0,1\}^\infty, \mathcal{F} \otimes \mathcal{A})$, where $\mathcal{A}$ is the product σ-algebra on $\{0,1\}^\infty$. Represent elements of $\bar{\Omega}$ by $\bar{\omega} = (\omega, Y_0, Y_1, \ldots)$. The measures $\mathbf{P}_\mu$ will be extended to $\bar{\Omega}$ according to the following recipe. For each $\mu \in \mathcal{P}(\mathcal{X})$, $\omega \in \Omega$, and $n \geq 0$, there will be a function $f_\mu^{(n)}(\omega)$ such that $\mathbf{P}_\mu(Y_n = 1 \mid \omega) = f_\mu^{(n)}(\omega)$. The sequence (Y_n) will be conditionally independent under $\mathbf{P}_\mu$ given ω, so that for example

$$\mathbf{P}_\mu(Y_{n_1} = 0, Y_{n_2} = 1, Y_{n_3} = 0 \mid \omega) = [1 - f_\mu^{(n_1)}(\omega)] f_\mu^{(n_2)}(\omega) [1 - f_\mu^{(n_3)}(\omega)].$$

Then for $E \in \mathcal{F}$ and $E' \in \mathcal{A}$, one can define

$$\mathbf{P}_\mu(E \times E') = \int_E \mathbf{P}_\mu(E' \mid \omega)\, \mathbf{P}_\mu(d\omega).$$

In order for this to be valid, the functions $f_\mu^{(n)}(\omega)$ must be measurable, and they must be between 0 and 1. The definition of $f_\mu^{(n)}(\omega)$ will involve a Radon-Nikodym derivative of two measures defined on $(\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{E})$. Let

$$\begin{aligned}\eta_\mu^{(n)}(dx \otimes dx') &= \mathbf{P}_\mu(X_{n-m} \in dx, X_n \in dx') = P^{n-m}(\mu, dx)P^m(x, dx'),\\ \psi_\mu^{(n)}(dx \otimes dx') &= P^{n-m}(\mu, dx)\varepsilon\nu(dx').\end{aligned}$$

Since $P^m(x, \cdot) \geq \varepsilon\nu(\cdot)$ for all $x \in C$, when the measures $\eta_\mu^{(n)}$ and $\psi_\mu^{(n)}$ are restricted to $C \times \mathcal{X}$ one has $\psi_\mu^{(n)} \leq \eta_\mu^{(n)}$. Therefore, on $C \times \mathcal{X}$, the Radon-Nikodym derivative $d\psi_\mu^{(n)}/d\eta_\mu^{(n)}$ exists, and one can take a version satisfying $d\psi_\mu^{(n)}/d\eta_\mu^{(n)} \leq 1$.

Recall that for each $\omega \in \Omega$ one has a set $\mathbf{S}$ of coin-flip times. If $s \in \mathbf{S}$, then $X_s \in C$. The definition of $f_\mu^{(n)}(\omega)$ is

$$f_\mu^{(n)}(\omega) = \begin{cases} 0 & \text{if } n-m \notin \mathbf{S}, \\ \frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(X_{n-m}, X_n) & \text{if } n-m \in \mathbf{S}. \end{cases}$$

It is clear that $f_\mu^{(n)}(\omega)$ is measurable and that $0 \leq f_\mu^{(n)}(\omega) \leq 1$. Thus the extension of $\mathbf{P}_\mu$ to $\bar{\Omega}$ is defined.

Compatibility

At this stage it is necessary to check the compatibility condition. Let $\mu, \mu' \in \mathcal{P}(\mathcal{X})$, with μ' absolutely continuous with respect to μ. It will suffice to prove for events $E \in \mathcal{F}$ and $E' \in \mathcal{A}$ that

$$\mathbf{P}_{\mu'}(E \times E') = \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\bar{\omega} \in E \times E'\}\right]. \tag{A.10}$$

Fix an integer $n \geq m$. Since μ' is absolutely continuous with respect to μ, $P^{n-m}(\mu', \cdot)$ is absolutely continuous with respect to $P^{n-m}(\mu, \cdot)$. Let $g(x)$ be the Radon-Nikodym derivative of $P^{n-m}(\mu', \cdot)$ with respect to $P^{n-m}(\mu, \cdot)$. Then $g(x)$ is also a Radon-Nikodym derivative of $\eta_{\mu'}^{(n)}(dx \otimes dx') = P^{n-m}(\mu', dx)P^m(x, dx')$ with respect to $\eta_\mu^{(n)}(dx \otimes dx') = P^{n-m}(\mu, dx)P^m(x, dx')$. As well, $g(x)$ is a Radon-Nikodym derivative of $\psi_{\mu'}^{(n)}(dx \otimes dx') = P^{n-m}(\mu', dx)\varepsilon\nu(dx')$ with respect to $\psi_\mu^{(n)}(dx \otimes dx') = P^{n-m}(\mu, dx)\varepsilon\nu(dx')$.

With this information, it can be shown that

$$\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(x, x') = \frac{d\psi_{\mu'}^{(n)}}{d\eta_{\mu'}^{(n)}}(x, x') \qquad \text{on } C \times \mathcal{X}, \tag{A.11}$$

almost everywhere with respect to

$$\mathbf{P}_{\mu'}(X_{n-m} \in dx, X_n \in dx') = \eta_{\mu'}^{(n)}(dx \otimes dx').$$

To prove (A.11), it is enough to check the integrated form: for any bounded measurable function f on $C \times \mathcal{X}$,

$$\int_{C\times\mathcal{X}} f(x,x')\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(x,x')\eta_{\mu'}^{(n)}(dx\otimes dx') = \int_{C\times\mathcal{X}} f(x,x')\psi_{\mu'}^{(n)}(dx\otimes dx').$$

This is true because

$$\begin{aligned}\int_{C\times\mathcal{X}} f(x,x')\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(x,x')\eta_{\mu'}^{(n)}(dx\otimes dx') &= \int_{C\times\mathcal{X}} f(x,x')\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(x,x')g(x)\eta_{\mu}^{(n)}(dx\otimes dx')\\ &= \int_{C\times\mathcal{X}} f(x,x')g(x)\psi_{\mu}^{(n)}(dx\otimes dx')\\ &= \int_{C\times\mathcal{X}} f(x,x')\psi_{\mu'}^{(n)}(dx\otimes dx').\end{aligned}$$

Therefore, (A.11) holds. As a consequence,

$$\mathbf{P}_\mu(Y_n = 1 \mid \omega) = \mathbf{P}_{\mu'}(Y_n = 1 \mid \omega) \tag{A.12}$$

almost everywhere with respect to $\mathbf{P}_{\mu'}$. To see why, suppose ω is given. If $n-m \notin \mathbf{S}$, both sides of (A.12) are zero. If $n-m \in \mathbf{S}$, so that in particular $X_{n-m} \in C$, (A.12) reduces to (A.11).

Because the Y_n are conditionally independent given ω, it follows from (A.12) that for any event $E' \in \mathcal{A}$,

$$\mathbf{P}_\mu(E' \mid \omega) = \mathbf{P}_{\mu'}(E' \mid \omega) \tag{A.13}$$

almost everywhere with respect to $\mathbf{P}_{\mu'}$.

Now (A.10) can be proved. For any events $E \in \mathcal{F}$ and $E' \in \mathcal{A}$,

$$\mathbf{P}_{\mu'}(E\times E') = \mathbf{E}_{\mu'}[\mathbf{P}_{\mu'}(E\times E' \mid \omega)] = \mathbf{E}_{\mu'}[\mathbf{1}\{\omega\in E\}\,\mathbf{P}_{\mu'}(E' \mid \omega)] = \mathbf{E}_{\mu'}[\mathbf{1}\{\omega\in E\}\,\mathbf{P}_{\mu}(E' \mid \omega)],$$

using (A.13) in the last equality. Since (1.5) holds for Ω,

$$\begin{aligned}\mathbf{E}_{\mu'}[\mathbf{1}\{\omega\in E\}\,\mathbf{P}_{\mu}(E' \mid \omega)] &= \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\omega\in E\}\,\mathbf{P}_\mu(E' \mid \omega)\right]\\ &= \mathbf{E}_\mu\left[\mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\bar\omega\in E\times E'\}\,\middle|\,\omega\right]\right]\\ &= \mathbf{E}_\mu\left[\frac{d\mu'}{d\mu}(X_0)\mathbf{1}\{\bar\omega\in E\times E'\}\right].\end{aligned}$$

This proves (A.10), so the compatibility condition is verified.

Conditional independence

To begin the rest of the proof, a few things follow directly from the definition of $f_\mu^{(n)}(\omega)$. If $n < m$ then certainly $n - m \notin \mathbf{S}$, so $\mathbf{P}_\mu(T \geq m) = 1$. Since the functions $f_\mu^{(n)}(\omega)$ depend only on $X = (X_0, X_1, \ldots)$, the sequence (Y_n) is conditionally independent under $\mathbf{P}_\mu$ given X. A finer statement is that for every $n \geq 0$, the random variables $Y_0, \ldots, Y_n$ and the tail $(X_{n+1}, X_{n+2}, \ldots)$ are mutually conditionally independent under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_n)$. Here is the proof. Suppose $0 \leq k_0 < k_1 < \cdots < k_\ell \leq n$. Then

$$\begin{aligned}
&\mathbf{P}_\mu(\text{each } Y_{k_i} = 1, (X_{n+1}, X_{n+2}, \ldots) \in E \mid X_0, \ldots, X_n) \\
&= \mathbf{E}_\mu[\mathbf{P}_\mu(\text{each } Y_{k_i} = 1, (X_{n+1}, X_{n+2}, \ldots) \in E \mid \omega) \mid X_0, \ldots, X_n] \\
&= \mathbf{E}_\mu[\mathbf{1}\{(X_{n+1}, X_{n+2}, \ldots) \in E\}\, \mathbf{P}_\mu(\text{each } Y_{k_i} = 1 \mid \omega) \mid X_0, \ldots, X_n] \\
&= \mathbf{E}_\mu\left[\mathbf{1}\{(X_{n+1}, X_{n+2}, \ldots) \in E\} \prod_{i=1}^{\ell} f_\mu^{(k_i)}(\omega) \middle| X_0, \ldots, X_n\right] \\
&= \left(\prod_{i=1}^{\ell} \mathbf{1}\{k_i - m \in \mathbf{S}\} \frac{d\psi_\mu^{(k_i)}}{d\eta_\mu^{(k_i)}}(X_{k_i - m}, X_{k_i})\right) \mathbf{P}_\mu((X_{n+1}, X_{n+2}, \ldots) \in E \mid X_0, \ldots, X_n) \\
&= \left(\prod_{i=1}^{\ell} \mathbf{P}_\mu(Y_{k_i} = 1 \mid X_0, \ldots, X_n)\right) \mathbf{P}_\mu((X_{n+1}, X_{n+2}, \ldots) \in E \mid X_0, \ldots, X_n).
\end{aligned}$$

This proves the mutual conditional independence. For every $j \geq 1$ and $n \geq 0$, the event $\{T_j = n\}$ depends only on the values of $Y_0, \ldots, Y_n$, so it is conditionally independent of $(X_{n+1}, X_{n+2}, \ldots)$ under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_n)$. Hence each T_j is a randomized stopping time for (X_t).

Property 1

To prove property 1, the main step is to prove the following formula for all $A \in \mathcal{E}$:

$$\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid X_0, \ldots, X_{n-m}) = \varepsilon\nu(A)\mathbf{1}\{n - m \in \mathbf{S}\}. \tag{A.14}$$

Since $Y_n = 0$ if $n - m \notin \mathbf{S}$, the left side of (A.14) equals

$$\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid n - m \in \mathbf{S}, X_0, \ldots, X_{n-m})\, \mathbf{P}_\mu(n - m \in \mathbf{S} \mid X_0, \ldots, X_{n-m}),$$

and because $\mathbf{P}_\mu(n - m \in \mathbf{S} \mid X_0, \ldots, X_{n-m}) = \mathbf{1}\{n - m \in \mathbf{S}\}$, it will suffice to show that

$$\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid n - m \in \mathbf{S}, X_0, \ldots, X_{n-m}) = \varepsilon\nu(A). \tag{A.15}$$

The left side of (A.15) equals

$$\begin{aligned}\mathbf{E}_\mu[\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid n-m \in \mathbf{S}, X_0, \ldots, X_{n-m}, X_n) \mid n-m \in \mathbf{S}, X_0, \ldots, X_{n-m}]&\\ = \mathbf{E}_\mu\left[\mathbf{1}\{X_n \in A\}\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(X_{n-m}, X_n)\,\middle|\, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m}\right].&\end{aligned}$$

To see that this conditional expectation equals $\varepsilon\nu(A)$, it suffices to show that it integrates properly. Since the event $\{n-m \in \mathbf{S}\}$ depends only on the values of $X_0, \ldots, X_{n-m}$, one can view $\{n-m \in \mathbf{S}\}$ as a subset of $\mathcal{X}^{n-m+1}$. It will be enough to check that for all measurable $E \subseteq \{n-m \in \mathbf{S}\} \subseteq \mathcal{X}^{n-m+1}$,

$$\begin{aligned}\int_\Omega \mathbf{1}\{(X_0, \ldots, X_{n-m}) \in E\}\mathbf{1}\{X_n \in A\}\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(X_{n-m}, X_n)\,\mathbf{P}_\mu(d\omega)&\\ = \varepsilon\nu(A)\,\mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E).&\end{aligned} \quad (A.16)$$

Using that $\eta_\mu^{(n)}(dx \otimes dx') = \mathbf{P}_\mu(X_{n-m} \in dx, X_n \in dx')$, the left side of (A.16) equals

$$\begin{aligned}&\int_{(x,x')\in\mathcal{X}\times\mathcal{X}} \mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E \mid X_{n-m} = x, X_n = x')\mathbf{1}\{x' \in A\}\frac{d\psi_\mu^{(n)}}{d\eta_\mu^{(n)}}(x, x')\eta_\mu^{(n)}(dx \otimes dx')\\ &= \int_{x\in\mathcal{X}}\int_{x'\in\mathcal{X}} \mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E \mid X_{n-m} = x, X_n = x')\mathbf{1}\{x' \in A\}P^{n-m}(\mu, dx)\varepsilon\nu(dx').\end{aligned}$$

By the Markov property,

$$\mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E \mid X_{n-m} = x, X_n = x') = \mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E \mid X_{n-m} = x).$$

Therefore, the double integral can be written as

$$\begin{aligned}\int_{x'\in\mathcal{X}} \mathbf{1}\{x' \in A\}\varepsilon\nu(dx')\int_{x\in\mathcal{X}} \mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E \mid X_{n-m} = x)P^{n-m}(\mu, dx)&\\ = \varepsilon\nu(A)\,\mathbf{P}_\mu((X_0, \ldots, X_{n-m}) \in E),&\end{aligned}$$

which is the right side of (A.16). This proves (A.15) and thereby (A.14).

The next result directly implies that T is a ν-regeneration time when $m = 1$ and a strong ν time when $m > 1$:

$$\mathbf{P}_\mu(X_n \in A \mid T = n, X_0, \ldots, X_{n-m}) = \nu(A) \quad \text{for all } A \in \mathcal{E}. \quad (A.17)$$

The proof of (A.17) will actually use (A.15) (rather than (A.14)) at the right moment. To begin,

$$\begin{aligned}\mathbf{P}_\mu(X_n \in A \mid T = n, X_0, \ldots, X_{n-m}) &= \mathbf{P}_\mu(X_n \in A \mid T = n, T > n-m, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m}) \\ &= \frac{\mathbf{P}_\mu(X_n \in A, T = n \mid T > n-m, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})}{\mathbf{P}_\mu(T = n \mid T > n-m, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})}.\end{aligned}$$

The event $\{T > n-m\}$ is the same as $\{Y_0 = \cdots = Y_{n-m} = 0\}$. Given that $T > n-m$ and $n-m \in \mathbf{S}$, there are no elements of $\mathbf{S}$ strictly between $n-2m$ and $n-m$, so $Y_{n-m+1} = \cdots = Y_{n-1} = 0$. Hence $T = n$ if and only if $Y_n = 1$.

Suppose it can be shown that (X_n, Y_n) is conditionally independent of $(Y_0, \ldots, Y_{n-m})$ under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_{n-m})$. Then the conditioning on $T > n-m$ can be removed in both the numerator and the denominator:

$$\begin{aligned}&\mathbf{P}_\mu(X_n \in A \mid T = n, X_0, \ldots, X_{n-m}) \\ &= \frac{\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid Y_0 = \cdots = Y_{n=m} = 0, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})}{\mathbf{P}_\mu(Y_n = 1 \mid Y_0 = \cdots = Y_{n-m} = 0, n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})} \\ &= \frac{\mathbf{P}_\mu(X_n \in A, Y_n = 1 \mid n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})}{\mathbf{P}_\mu(Y_n = 1 \mid n-m \in \mathbf{S}, X_0, \ldots, X_{n-m})} \\ &= \frac{\varepsilon\nu(A)}{\varepsilon\nu(\mathcal{X})} = \nu(A),\end{aligned}$$

where the next-to-last equality used (A.15) in both the numerator and the denominator. Therefore, if it can be shown that (X_n, Y_n) is conditionally independent of $(Y_0, \ldots, Y_{n-m})$ under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_{n-m})$, (A.17) will follow.

Here is the proof of the conditional independence. To start, $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of $(X_{n-m+1}, \ldots, X_n)$ given $(X_0, \ldots, X_{n-m})$, which implies that $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of $(X_{n-m+1}, \ldots, X_{n-1})$ given $(X_0, \ldots, X_{n-m})$ and X_n. In addition, $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of Y_n given $(X_0, \ldots, X_n)$. It follows that $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of $((X_{n-m+1}, \ldots, X_{n-1}), Y_n)$ given $(X_0, \ldots, X_{n-m})$ and X_n, so in particular, $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of Y_n given $(X_0, \ldots, X_{n-m})$ and X_n. One also has that $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of X_n given $(X_0, \ldots, X_{n-m})$. Therefore, $(Y_0, \ldots, Y_{n-m})$ is conditionally independent of (X_n, Y_n) given $(X_0, \ldots, X_{n-m})$. This finishes the proof of (A.17).

To finish proving property 1, it must be shown that $\mathbf{P}_\mu(T < \infty) = 1$ for all $\mu \in \mathcal{P}(\mathcal{X})$. This will follow from the arguments used to prove property 3, so it is set aside for the moment.

Property 2

For property 2, one must check Requirements 3.5. It has already been shown that each T_j is a randomized stopping time for (X_t), and certainly $T_1 = T$. The next requirement is that if $\mathbf{P}_\mu(T_j = k) > 0$, $T_{j+1} - T_j$ is conditionally independent of $(X_0, \ldots, X_{k-1})$ under $\mathbf{P}_\mu$ given that $T_j = k$ and given $(X_k, X_{k+1}, \ldots)$. It will suffice to show that the random variables $Y_{k+1}, Y_{k+2}, \ldots$ and the path $(X_0, \ldots, X_{k-1})$ are mutually conditionally independent given that $T_j = k$ and given $(X_k, X_{k+1}, \ldots)$. Since $T_j = k$, it must be true that $k - m \in \mathbf{S}$, so no integer strictly between $k - m$ and k is in $\mathbf{S}$. Hence $Y_{k+1} = \cdots = Y_{k+m-1} = 0$, and it is enough to check the mutual conditional independence of $Y_{k+m}, Y_{k+m+1}, \ldots$ and $(X_0, \ldots, X_{k-1})$.

The strategy is the same as for the earlier mutual conditional independence statement (under the heading "Conditional independence"), but with a slight wrinkle. Suppose $k + m \le k_1 < k_2 < \cdots < k_\ell$. One has

$$\begin{aligned}
&\mathbf{P}_\mu(\text{each } Y_{k_i} = 1, (X_0, \ldots, X_{k-1}) \in E \mid T_j = k, X_k, X_{k+1}, \ldots) \\
&= \mathbf{E}_\mu[\mathbf{P}_\mu(\text{each } Y_{k_i} = 1, (X_0, \ldots, X_{k-1}) \in E \mid Y_0, \ldots, Y_k, \omega) \mid T_j = k, X_k, X_{k+1}, \ldots] \\
&= \mathbf{E}_\mu[\mathbf{1}\{(X_0, \ldots, X_{k-1}) \in E\}\, \mathbf{P}_\mu(\text{each } Y_{k_i} = 1 \mid Y_0, \ldots, Y_k, \omega) \mid T_j = k, X_k, X_{k+1}, \ldots].
\end{aligned}$$

Since the sequence (Y_t) is conditionally independent given ω, the dependence on $Y_0, \ldots, Y_k$ in the inner probability can be removed. Then, as in the previous argument,

$$\mathbf{P}_\mu(\text{each } Y_{k_i} = 1 \mid \omega) = \prod_{i=1}^{\ell} \mathbf{1}\{k_i - m \in \mathbf{S}\} \frac{d\psi_\mu^{(k_i)}}{d\eta_\mu^{(k_i)}}(X_{k_i - m}, X_{k_i}).$$

Given that $T_j = k$, so that $k - m \in \mathbf{S}$, the event $\{k_i - m \in \mathbf{S}\}$ depends only on $(X_k, X_{k+1}, \ldots)$. Indeed,

$$\mathbf{1}\{k_i - m \in \mathbf{S}\} \frac{d\psi_\mu^{(k_i)}}{d\eta_\mu^{(k_i)}}(X_{k_i - m}, X_{k_i}) = \mathbf{P}_\mu(Y_{k_i} = 1 \mid T_j = k, X_k, X_{k+1}, \ldots).$$

These terms can be taken out of the conditional expectation, yielding

$$\begin{aligned}
&\mathbf{P}_\mu(\text{each } Y_{k_i} = 1, (X_0, \ldots, X_{k-1}) \in E \mid T_j = k, X_k, X_{k+1}, \ldots) \\
&= \left(\prod_{i=1}^{\ell} \mathbf{P}_\mu(Y_{k_i} = 1 \mid T_j = k, X_k, X_{k+1}, \ldots) \right) \mathbf{P}_\mu((X_0, \ldots, X_{k-1}) \in E \mid T_j = k, X_k, X_{k+1}, \ldots),
\end{aligned}$$

which is the desired mutual conditional independence.

The final requirement for property 2 is that if $\mathbf{P}_\mu(T_j = k) > 0$,

$$\mathbf{P}_\mu(T_{j+1} - T_j = \ell \mid T_j = k, X_k, X_{k+1}, \ldots) = f_\ell(X_k, X_{k+1}, \ldots, X_{k+\ell}), \tag{A.18}$$

where f_ℓ is defined by

$$\mathbf{P}_\nu(T = \ell \mid X_0, \dots, X_\ell) = f_\ell(X_0, \dots, X_\ell).$$

To prove this, consider $\mathbf{P}_\mu$ and $\mathbf{P}_\nu$ as measures on $(\mathcal{X}^\infty \times \{0,1\}^\infty, \mathcal{E}^\infty \otimes \mathcal{A})$. (The sample space $\Omega \times \{0,1\}^\infty$ projects to $\mathcal{X}^\infty \times \{0,1\}^\infty$ using the map $X : \Omega \to \mathcal{X}^\infty$.) For $k \geq 0$, define the shift operator θ_k on $\mathcal{X}^\infty \times \{0,1\}^\infty$ by

$$\theta_k((X_0, X_1, \dots), (Y_0, Y_1, \dots)) = ((X_k, X_{k+1}, \dots), (Y_k, Y_{k+1}, \dots)).$$

It will be proved that the push-forward of the measure $\mathbf{P}_\mu(\,\cdot \mid T_j = k)$ by θ_k is almost equal to $\mathbf{P}_\nu$. One has $\mathbf{P}_\mu(Y_k = 1 \mid T_j = k) = 1$ and $\mathbf{P}_\nu(Y_0 = 1) = 0$, but aside from that, the following equation will hold for all $E \in \mathcal{E}^\infty$ and $E' \in \mathcal{A}$:

$$\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E, (Y_{k+1}, Y_{k+2}, \dots) \in E' \mid T_j = k) = \mathbf{P}_\nu((X_0, X_1, \dots) \in E, (Y_1, Y_2, \dots) \in E'). \tag{A.19}$$

This directly implies (A.18). The proof of (A.19) is by induction on j. Specifically, it will be shown that if T_j is a strong ν time for the chain (X_t) started from μ (the "inductive hypothesis"), then (A.19) holds for j and T_{j+1} is a strong ν time for (X_t) started from μ. Since $T_1 = T$, the inductive hypothesis is satisfied for $j = 1$. Assuming that T_j is a strong ν time, for any event $E \in \mathcal{E}^\infty$,

$$\begin{aligned}&\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k)\\ &\qquad = \int_{\mathcal{X}} \mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k, X_k = x_k)\, \mathbf{P}_\mu(X_k \in dx_k \mid T_j = k).\end{aligned}$$

The dependence on $T_j = k$ in the first term on the right side can be removed, by the following reasoning. First write

$$\begin{aligned}&\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k, X_k = x_k)\\ &\qquad = \mathbf{E}_\mu[\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k, Y_0, \dots, Y_k, X_0, \dots, X_k) \mid T_j = k, X_k = x_k].\end{aligned}$$

Since the event $\{T_j = k\}$ depends only on $Y_0, \dots, Y_k$, and $(Y_0, \dots, Y_k)$ is conditionally independent of $(X_k, X_{k+1}, \dots)$ given $(X_0, \dots, X_k)$, one has

$$\begin{aligned}\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k, Y_0, \dots, Y_k, X_0, \dots, X_k) &= \mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid X_0, \dots, X_k)\\ &= \mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid X_k)\end{aligned}$$

which can be taken out from the outer expectation, so that

$$\mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid T_j = k, X_k = x_k) = \mathbf{P}_\mu((X_k, X_{k+1}, \dots) \in E \mid X_k = x_k).$$

Meanwhile, the inductive hypothesis implies that $\mathbf{P}_\mu(X_k \in dx_k \mid T_j = k) = \nu(dx_k)$. Therefore,

$$\begin{aligned}\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E \mid T_j = k) &= \int_{\mathcal{X}} \mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in E \mid X_k = x_k)\nu(dx_k) \\ &= \int_{\mathcal{X}} \mathbf{P}_{x_k}((X_0, X_1, \ldots) \in E)\nu(dx_k) \\ &= \mathbf{P}_\nu((X_0, X_1, \ldots) \in E).\end{aligned}$$

The last equalities rely on (1.4), but this is not a problem since the event E is a subset of $\mathcal{X}^\infty$.

Next, note that the subset $\mathbf{S}$ behaves appropriately under θ_k, assuming that $T_j = k$. That is, suppose $(X_0, X_1, \ldots)$ is a given sample path, and $\mathbf{S}$ is the associated set of indices. Since $T_j = k$, one has $k - m \in \mathbf{S}$. Let $\mathbf{S}'$ be the set of indices associated with the sample path $(X_k, X_{k+1}, \ldots)$. Then the claim is that

$$\{s \in \mathbf{S} : s \geq k\} - k = \mathbf{S}'. \tag{A.20}$$

The first element of $\mathbf{S}'$ is $\min\{\ell \geq k : X_\ell \in C\} - k$. Because $k - m \in \mathbf{S}$, the first element of $\{s \in \mathbf{S} : s \geq k\}$ is $\min\{\ell \geq k : X_\ell \in C\}$. Therefore the first elements on the left and right side of (A.20) are equal. From there, the set equality in (A.20) follows from the inductive definition of $\mathbf{S}$.

To finish proving (A.19), it is enough to show that

$$\begin{aligned}&\mathbf{P}_\mu((Y_{k+1}, Y_{k+2}, \ldots) \in E' \mid T_j = k, X_0 = x_0, X_1 = x_1, \ldots) \\ &= \mathbf{P}_\mu((Y_{k+1}, Y_{k+2}, \ldots) \in E' \mid T_j = k, X_k = x_k, X_{k+1} = x_{k+1}, \ldots) \\ &= \mathbf{P}_\nu((Y_1, Y_2, \ldots) \in E' \mid X_0 = x_k, X_1 = x_{k+1}, \ldots),\end{aligned} \tag{A.21}$$

where the last equality holds almost everywhere with respect to the measure

$$\mathbf{P}_\mu((X_k, X_{k+1}, \ldots) \in \cdot \mid T_j = k) = \mathbf{P}_\nu((X_0, X_1, \ldots) \in \cdot).$$

If (A.21) holds, then the left side of (A.19) can be written as

$$\begin{aligned}&\int_{\mathcal{X}^\infty} \mathbf{P}_\mu(X_0 \in dx_0, X_1 \in dx_1, \ldots \mid T_j = k)\Big[\mathbf{1}\{(x_k, x_{k+1}, \ldots) \in E\}(\text{left side of (A.21)})\Big] \\ &= \int_{\mathcal{X}^\infty} \mathbf{P}_\mu(X_k \in dx_k, X_{k+1} \in dx_{k+1}, \ldots \mid T_j = k)\Big[\mathbf{1}\{(x_k, x_{k+1}, \ldots) \in E\}(\text{middle term of (A.21)})\Big] \\ &= \int_{\mathcal{X}^\infty} \mathbf{P}_\nu(X_0 \in dx_k, X_1 \in dx_{k+1}, \ldots)\Big[\mathbf{1}\{(x_k, x_{k+1}, \ldots) \in E\}(\text{right side of (A.21)})\Big]\end{aligned}$$

which equals the right side of (A.19). Finally, the induction hypothesis for $j + 1$ must be checked. This follows from (A.19):

$$\mathbf{P}_\mu(X_\ell \in A \mid T_j = k, T_{j+1} = \ell) = \mathbf{P}_\nu(X_{\ell - k} \in A \mid T = \ell - k) = \nu(A).$$

It remains to prove (A.21). Because of the conditional independence of (Y_t) given X, one need only check that

$$\begin{aligned}&\mathbf{P}_\mu(Y_\ell = 1 \mid T_j = k, X_0 = x_0, X_1 = x_1, \ldots)\\ &= \mathbf{P}_\mu(Y_\ell = 1 \mid T_j = k, X_k = x_k, X_{k+1} = x_{k+1}, \ldots)\\ &= \mathbf{P}_\nu(Y_{\ell-k} = 1 \mid X_0 = x_k, X_1 = x_{k+1}, \ldots)\end{aligned} \tag{A.22}$$

for all $\ell \geq k+1$. If $k+1 \leq \ell \leq k+m-1$, all three terms are zero. For the left and middle terms, this is because $T_j = k$, so $k - m \in \mathbf{S}$ and $\mathbf{S}$ contains no indices strictly between $k-m$ and k. For the right term, it is because $Y_i = 0$ with probability 1 for $i < m$.

For $\ell \geq k+m$, the left term depends only on $x_k, x_{k+1}, \ldots$, so the first equality is proved. To check the second equality, note that both sides are zero unless the condition involving $\mathbf{S}$ is met. As was shown earlier, the set $\mathbf{S}$ built out of the sequence $(x_k, x_{k+1}, \ldots)$ works for both the right term (by definition) and the middle term (by the argument above). Hence it can be assumed that $\ell - m \in \mathbf{S}$ for the middle term and $\ell - k - m \in \mathbf{S}$ for the right term. Equation (A.22) reduces to the statement

$$\frac{d\psi_\mu^{(\ell)}}{d\eta_\mu^{(\ell)}}(x_{\ell-m}, x_\ell) = \frac{d\psi_\nu^{(\ell-k)}}{d\eta_\nu^{(\ell-k)}}(x_{\ell-m}, x_\ell) \qquad \text{on } C \times \mathcal{X},$$

almost everywhere with respect to

$$\mathbf{P}_\mu(X_{\ell-m} \in dx_{\ell-m}, X_\ell \in dx_\ell \mid T_j = k) = \mathbf{P}_\nu(X_{\ell-k-m} \in dx_{\ell-m}, X_{\ell-k} \in dx_\ell) = \eta_\nu^{(\ell-k)}(dx_{\ell-m} \otimes dx_\ell).$$

This is the same kind of statement that was needed to obtain the compatibility condition, and the strategy of proof is the same. It will suffice to show for bounded measurable functions f on $C \times \mathcal{X}$ that

$$\int_{C\times\mathcal{X}} f(x,x') \frac{d\psi_\mu^{(\ell)}}{d\eta_\mu^{(\ell)}}(x,x') \eta_\nu^{(\ell-k)}(dx \otimes dx') = \int_{C\times\mathcal{X}} f(x,x') \psi_\nu^{(\ell-k)}(dx \otimes dx'). \tag{A.23}$$

To prove (A.23), note first that

$$\begin{aligned}\mathbf{P}_\mu(X_{\ell-m} \in A) &\geq \mathbf{P}_\mu(X_{\ell-m} \in A, T_j = k)\\ &= \mathbf{P}_\mu(T_j = k)\,\mathbf{P}_\mu(X_{\ell-m} \in A \mid T_j = k)\\ &= \mathbf{P}_\mu(T_j = k)\,\mathbf{P}_\nu(X_{\ell-k-m} \in A).\end{aligned}$$

Since $\mathbf{P}_\mu(T_j = k) > 0$, the measure $P^{\ell-k-m}(\nu, \cdot)$ is absolutely continuous with respect to $P^{\ell-m}(\mu, \cdot)$. Let $g(x)$ be the Radon-Nikodym derivative of $P^{\ell-k-m}(\nu, \cdot)$ with respect to $P^{\ell-m}(\mu, \cdot)$. Then $g(x)$ is also a Radon-Nikodym derivative of $\eta_\nu^{(\ell-k)}(dx \otimes dx') = P^{\ell-k-m}(\nu, dx)P^m(x, dx')$ with respect to

$\eta_\mu^{(\ell)}(dx \otimes dx') = P^{\ell-m}(\mu, dx)P^m(x, dx')$. As well, $g(x)$ is a Radon-Nikodym derivative of $\psi_\nu^{(\ell-k)}(dx \otimes dx') = P^{\ell-k-m}(\nu, dx)\varepsilon\nu(dx')$ with respect to $\psi_\mu^{(\ell)}(dx \otimes dx') = P^{\ell-m}(\mu, dx)\varepsilon\nu(dx')$. The left side of (A.23) can be written as

$$\begin{aligned}\int_{C\times\mathcal{X}} f(x,x')\frac{d\psi_\mu^{(\ell)}}{d\eta_\mu^{(\ell)}}(x,x')g(x)\eta_\mu^{(\ell)}(dx\otimes dx') &= \int_{C\times\mathcal{X}} f(x,x')g(x)\psi_\mu^{(\ell)}(dx\otimes dx')\\ &= \int_{C\times\mathcal{X}} f(x,x')\psi_\nu^{(\ell-k)}(dx\otimes dx'),\end{aligned}$$

which is the right side of (A.23). Since (A.23) is true, so are (A.22) and (A.21), finishing the proof of (A.19) and property 2.

Property 3

For property 3, suppose the initial distribution μ is fixed and $\mathbf{P}_\mu(\tau_C = s) > 0$. Equation (A.14) immediately shows that $\mathbf{P}_\mu(T < s+m \mid \tau_C = s) = 0$ and $\mathbf{P}_\mu(T = s+m \mid \tau_C = s) = \varepsilon$. It remains to show that when $\varepsilon < 1$,

$$\mathbf{P}_\mu(T > t \mid \tau_C = s) = (1-\varepsilon)\,\mathbf{P}_{\mu_s}(T > t-s-m)$$

for all $t \geq s+m$. Since

$$\mathbf{P}_\mu(T > t \mid \tau_C = s) = \mathbf{P}_\mu(Y_{s+m} = 0 \mid \tau_C = s)\,\mathbf{P}_\mu(T > t \mid \tau_C = s, Y_{s+m} = 0)$$

and $\mathbf{P}_\mu(Y_{s+m} = 0 \mid \tau_C = s) = 1-\varepsilon$, it is enough to show that

$$\mathbf{P}_\mu(T > t \mid \tau_C = s, Y_{s+m} = 0) = \mathbf{P}_{\mu_s}(T > t-s-m). \tag{A.24}$$

As in the proof of property 2, consider $\mathbf{P}_\mu$ and $\mathbf{P}_{\mu_s}$ as measures on $(\mathcal{X}^\infty \times \{0,1\}^\infty, \mathcal{E}^\infty \otimes \mathcal{A})$. Equation (A.24) will be proved by showing that the push-forward of $\mathbf{P}_\mu(\cdot \mid \tau_C = s, Y_{s+m} = 0)$ under the shift operator θ_{s+m} is equal to $\mathbf{P}_{\mu_s}$. In other words, for all $E \in \mathcal{E}^\infty$ and $E' \in \mathcal{A}$,

$$\begin{aligned}&\mathbf{P}_\mu((X_{s+m}, X_{s+m+1}, \ldots) \in E, (Y_{s+m}, Y_{s+m+1}, \ldots) \in E' \mid \tau_C = s, Y_{s+m} = 0)\\ &\qquad = \mathbf{P}_{\mu_s}((X_0, X_1, \ldots) \in E, (Y_0, Y_1, \ldots) \in E'). \end{aligned}\tag{A.25}$$

The argument is essentially the same as the proof of (A.19). Note first that by (A.14),

$$\mathbf{P}_\mu(X_{s+m} \in A, Y_{s+m} = 1 \mid \tau_C = s) = \varepsilon\nu(A).$$

Therefore,

$$\begin{aligned}
&\mathbf{P}_\mu(X_{s+m} \in A \mid \tau_C = s, Y_{s+m} = 0) \\
&= \frac{1}{1-\varepsilon}\left[\mathbf{P}_\mu(X_{s+m} \in A \mid \tau_C = s) - \mathbf{P}_\mu(X_{s+m} \in A, Y_{s+m} = 1 \mid \tau_C = s)\right] \\
&= \mu_s(A).
\end{aligned}$$

It can now be shown that

$$\mathbf{P}_\mu((X_{s+m}, X_{s+m+1}, \ldots) \in E \mid \tau_C = s, Y_{s+m} = 0) = \mathbf{P}_{\mu_s}((X_0, X_1, \ldots) \in E),$$

by the following argument. The left side equals

$$\int_{\mathcal{X}} \mathbf{P}_\mu((X_{s+m}, X_{s+m+1}, \ldots) \in E \mid \tau_C = s, Y_{s+m} = 0, X_{s+m} = x)\mu_s(dx).$$

The event $\{\tau_C = s\}$ depends only on $X_0, \ldots, X_s$. Since Y_{s+m} is conditionally independent of $(X_{s+m}, X_{s+m+1}, \ldots)$ under $\mathbf{P}_\mu$ given $(X_0, \ldots, X_{s+m})$, the dependence on $Y_{s+m} = 0$ in the integral can be removed. Then by the Markov property, the dependence on $\tau_C = s$ can also be removed. One obtains

$$\begin{aligned}
&\mathbf{P}_\mu((X_{s+m}, X_{s+m+1}, \ldots) \in E \mid \tau_C = s, Y_{s+m} = 0) \\
&= \int_{\mathcal{X}} \mathbf{P}_\mu((X_{s+m}, X_{s+m+1}, \ldots) \in E \mid X_{s+m} = x)\mu_s(dx) \\
&= \int_{\mathcal{X}} \mathbf{P}_x((X_0, X_1, \ldots) \in E)\mu_s(dx) \\
&= \mathbf{P}_{\mu_s}((X_0, X_1, \ldots) \in E),
\end{aligned}$$

using (1.4) at the end.

Following the proof of (A.19), the next step is to argue that the subset $\mathbf{S}$ behaves appropriately under θ_{s+m}, assuming that $\tau_C = s$. Let $(X_0, X_1, \ldots)$ be a given sample path, and let $\mathbf{S}$ be the associated set of indices. Let $\mathbf{S}'$ be the set of indices associated with the sample path $(X_{s+m}, X_{s+m+1}, \ldots)$. The first element of $\mathbf{S}'$ is $\min\{\ell \geq s+m : X_\ell \in C\} - (s+m)$. Since $s \in \mathbf{S}$, the first element of $\{t \in \mathbf{S} : t \geq s+m\}$ is $\min\{\ell \geq s+m : X_\ell \in C\}$. By the inductive definition of $\mathbf{S}$, it follows that

$$\{t \in \mathbf{S} : t \geq s+m\} - (s+m) = \mathbf{S}',$$

just as in (A.20).

To finish the proof of (A.25), it suffices to obtain the analogous version of (A.22): for all $\ell \geq s+m$,

$$\begin{aligned} &\mathbf{P}_\mu(Y_\ell = 1 \mid \tau_C = s, Y_{s+m} = 0, X_0 = x_0, X_1 = x_1, \ldots) \\ &= \mathbf{P}_\mu(Y_\ell = 1 \mid \tau_C = s, Y_{s+m} = 0, X_{s+m} = x_{s+m}, X_{s+m+1} = x_{s+m+1}, \ldots) \\ &= \mathbf{P}_{\mu_s}(Y_{\ell-s-m} = 1 \mid X_0 = x_{s+m}, X_1 = x_{s+m+1}, \ldots). \end{aligned} \tag{A.26}$$

If $\ell = s+m$, all three terms are zero. If $s+m < \ell < s+2m$, the first two terms are zero because no element of $\mathbf{S}$ is strictly between s and $s+m$, and the third term is zero because $Y_i = 0$ for $i < m$. If $\ell \geq s+2m$, the first equality holds. For the second equality, because of the good behavior of $\mathbf{S}$, one may assume that $\ell - m \in \mathbf{S}$ in the middle term and $\ell - s - 2m \in \mathbf{S}$ in the third term. It is enough to show that

$$\frac{d\psi_\mu^{(\ell)}}{d\eta_\mu^{(\ell)}}(x_{\ell-m}, x_\ell) = \frac{d\psi_{\mu_s}^{(\ell-s-m)}}{d\eta_{\mu_s}^{(\ell-s-m)}}(x_{\ell-m}, x_\ell) \qquad \text{on } C \times \mathcal{X},$$

almost everywhere with respect to

$$\begin{aligned} \mathbf{P}_\mu(X_{\ell-m} \in dx_{\ell-m}, X_\ell \in dx_\ell \mid \tau_C = s, Y_{s+m} = 0) &= \mathbf{P}_{\mu_s}(X_{\ell-s-2m} \in dx_{\ell-m}, X_{\ell-s-m} \in dx_\ell) \\ &= \eta_{\mu_s}^{(\ell-s-m)}(dx_{\ell-m} \otimes dx_\ell). \end{aligned}$$

This proceeds exactly as in the proof of (A.23). The only necessary ingredient is that the measure $P^{\ell-s-2m}(\mu_s, \cdot)$ must be absolutely continuous with respect to $P^{\ell-m}(\mu, \cdot)$. This is true because $\mathbf{P}_\mu(\tau_C = s) > 0$ and $\varepsilon < 1$, and

$$\begin{aligned} \mathbf{P}_\mu(X_{\ell-m} \in A) &\geq \mathbf{P}_\mu(X_{\ell-m} \in A, \tau_C = s, Y_{s+m} = 0) \\ &= \mathbf{P}_\mu(\tau_C = s)\,\mathbf{P}_\mu(Y_{s+m} = 0 \mid \tau_C = s)\,\mathbf{P}_\mu(X_{\ell-m} \in A \mid \tau_C = s, Y_{s+m} = 0) \\ &= \mathbf{P}_\mu(\tau_C = s)(1-\varepsilon)\,\mathbf{P}_{\mu_s}(X_{\ell-s-2m} \in A). \end{aligned}$$

Thus (A.25) is proved, and property 3 follows.

Finiteness

The only thing left to prove is that $\mathbf{P}_\mu(T < \infty) = 1$ for all $\mu \in \mathcal{P}(\mathcal{X})$. This is clear when $\varepsilon = 1$, so assume $\varepsilon < 1$. Fix the starting distribution μ. If the elements of $\mathbf{S}$ are listed in increasing order as $S_1 < S_2 < \cdots$, then since $\mathbf{P}_{\mu'}(\tau_C < \infty)$ for all $\mu' \in \mathcal{P}(\mathcal{X})$, each S_k is almost surely finite under $\mathbf{P}_\mu$. It will be proved that $\mathbf{P}_\mu(T > S_k + m) = (1-\varepsilon)^k$, which implies that $\mathbf{P}_\mu(T < \infty) = 1$. In fact, the statement to be proved is

$$\mathbf{P}_\mu(T > S_k + m \mid S_1 = s) = (1-\varepsilon)^k \qquad \text{for all } \mu \in \mathcal{P}(\mathcal{X}), s \geq 0. \tag{A.27}$$

The proof is by induction on k. For $k = 1$ this was observed during the proof of property 3 to follow directly from (A.14). (Note that $S_1 = \tau_C$.) If (A.27) is known for $k-1$,

$$\begin{aligned}\mathbf{P}_\mu(T > S_k + m \mid S_1 = s) &= \mathbf{P}_\mu(T > S_k + m, Y_{s+m} = 0 \mid \tau_C = s) \\ &= (1-\varepsilon)\,\mathbf{P}_\mu(T > S_k + m \mid \tau_C = s, Y_{s+m} = 0).\end{aligned}$$

Given that $\tau_C = s$, it was shown during the proof of property 3 that $\mathbf{S}$ behaves appropriately under the shift operator θ_{s+m}. If $\mathbf{S}$ is the set of indices associated with $(X_0, X_1, \ldots)$ and $\mathbf{S}'$ is the set of indices associated with $(X_{s+m}, X_{s+m+1}, \ldots)$, then assuming that $\tau_C = s$ for $(X_0, X_1, \ldots)$, one has

$$\mathbf{S} = \{s, S_2, S_3, \ldots\} \implies \mathbf{S}' = \{S_2, S_3, \ldots\} - (s+m).$$

Using (A.25),

$$\begin{aligned}\mathbf{P}_\mu(T > S_k + m \mid \tau_C = s, Y_{s+m} = 0) &= \sum_{\ell=s+m}^{\infty} \mathbf{P}_\mu(T > \ell + m, S_k = \ell \mid \tau_C = s, Y_{s+m} = 0) \\ &= \sum_{\ell=s+m}^{\infty} \mathbf{P}_{\mu_s}(T > \ell - s, S_{k-1} = \ell - s - m) \\ &= \mathbf{P}_{\mu_s}(T > S_{k-1} + m).\end{aligned}$$

By the inductive hypothesis, $\mathbf{P}_{\mu_s}(T > S_{k-1} + m) = (1-\varepsilon)^{k-1}$. Hence

$$\mathbf{P}_\mu(T > S_k + m \mid S_1 = s) = (1-\varepsilon)(1-\varepsilon)^{k-1} = (1-\varepsilon)^k,$$

as desired. This completes the induction, so $\mathbf{P}_\mu(T < \infty) = 1$ for all $\mu \in \mathcal{P}(\mathcal{X})$. □

Bibliography

[AB15] Radosław Adamczak and Witold Bednorz. "Exponential concentration inequalities for additive functionals of Markov chains". *ESAIM Probab. Stat.* 19 (2015), pp. 440–481. URL: http://dx.doi.org/10.1051/ps/2014032.

[AD86] David Aldous and Persi Diaconis. "Shuffling cards and stopping times". *Amer. Math. Monthly* 93.5 (1986), pp. 333–348. URL: http://dx.doi.org/10.2307/2323590.

[AD87] David Aldous and Persi Diaconis. "Strong uniform times and finite random walks". *Adv. in Appl. Math.* 8.1 (1987), pp. 69–97. URL: http://dx.doi.org/10.1016/0196-8858(87)90006-6.

[Ald82] David J. Aldous. "Some inequalities for reversible Markov chains". *J. London Math. Soc. (2)* 25.3 (1982), pp. 564–576. URL: http://dx.doi.org/10.1112/jlms/s2-25.3.564.

[ALV15] C. Andrieu, A. Lee, and M. Vihola. "Uniform Ergodicity of the Iterated Conditional SMC and Geometric Ergodicity of Particle Gibbs samplers". *ArXiv e-prints* (Apr. 2015). arXiv: 1312.6432 [math.PR].

[ALW97] David Aldous, László Lovász, and Peter Winkler. "Mixing times for uniformly ergodic Markov chains". *Stochastic Process. Appl.* 71.2 (1997), pp. 165–185. URL: http://dx.doi.org/10.1016/S0304-4149(97)00037-9.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits". *Theoret. Comput. Sci.* 410.19 (2009), pp. 1876–1902. URL: http://dx.doi.org/10.1016/j.tcs.2009.01.016.

[AN78] K. B. Athreya and P. Ney. "A new approach to the limit theory of recurrent Markov chains". *Trans. Amer. Math. Soc.* 245 (1978), pp. 493–501. URL: http://dx.doi.org/10.2307/1998882.

[Bax05] Peter H. Baxendale. "Renewal theory and computable convergence rates for geometrically ergodic Markov chains". *Ann. Appl. Probab.* 15.1B (2005), pp. 700–738. URL: http://dx.doi.org/10.1214/105051604000000710.

[BBF09] J. Barrera, O. Bertoncini, and R. Fernández. “Abrupt convergence and escape behavior for birth and death chains”. *J. Stat. Phys.* 137.4 (2009), pp. 595–623. URL: http://dx.doi.org/10.1007/s10955-009-9861-7.

[BDF10] Alexei Borodin, Persi Diaconis, and Jason Fulman. “On adding a list of numbers (and other one-dependent determinantal processes)”. *Bull. Amer. Math. Soc. (N.S.)* 47.4 (2010), pp. 639–670. URL: http://dx.doi.org/10.1090/S0273-0979-2010-01306-9.

[BE53] N. G. de Bruijn and P. Erdős. “On a recursion formula and on some Tauberian theorems”. *J. Research Nat. Bur. Standards* 50 (1953), pp. 161–164.

[Bed13] Witold Bednorz. “The Kendall theorem and its application to the geometric ergodicity of Markov chains”. *Appl. Math. (Warsaw)* 40.2 (2013), pp. 129–165. URL: http://dx.doi.org/10.4064/am40-2-1.

[Ben62] George Bennett. “Probability inequalities for the sum of independent random variables”. *Journal of the American Statistical Association* 57.297 (1962), pp. 33–45.

[BHP15] Riddhipratim Basu, Jonathan Hermon, and Yuval Peres. “Characterization of Cutoff for Reversible Markov Chains”. *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms.* SODA ’15. San Diego, California: SIAM, 2015, pp. 1774–1791. URL: http://dl.acm.org/citation.cfm?id=2722129.2722248.

[BK00] Krzysztof Burdzy and Wilfrid S. Kendall. “Efficient Markovian couplings: examples and counterexamples”. *Ann. Appl. Probab.* 10.2 (2000), pp. 362–409. URL: http://dx.doi.org/10.1214/aoap/1019487348.

[BL08] Vlad Stefan Barbu and Nikolaos Limnios. *Semi-Markov chains and hidden semi-Markov models toward applications.* Vol. 191. Lecture Notes in Statistics. Their use in reliability and DNA analysis. Springer, New York, 2008, pp. xiv+224.

[BN09] Nathanaël Berestycki and Richard Nickl. *Concentration of Measure.* 2009. URL: http://www.statslab.cam.ac.uk/~beresty/teach/cm10.pdf.

[BŁL08] Witold Bednorz, Krzysztof Łatuszyński, and Rafał Latała. “A regeneration proof of the central limit theorem for uniformly ergodic Markov chains”. *Electron. Commun. Probab.* 13 (2008), pp. 85–98. URL: http://dx.doi.org/10.1214/ECP.v13-1354.

[CG92] George Casella and Edward I. George. “Explaining the Gibbs sampler”. *Amer. Statist.* 46.3 (1992), pp. 167–174. URL: http://dx.doi.org/10.2307/2685208.

[Cha89] K. S. Chan. “A note on the geometric ergodicity of a Markov chain”. *Adv. in Appl. Probab.* 21.3 (1989), pp. 702–704. URL: http://dx.doi.org/10.2307/1427643.

[CSC08] Guan-Yu Chen and Laurent Saloff-Coste. “The cutoff phenomenon for ergodic Markov processes”. *Electron. J. Probab.* 13 (2008), no. 3, 26–78. URL: http://dx.doi.org/10.1214/EJP.v13-474.

[CSC13a] Guan-Yu Chen and Laurent Saloff-Coste. "Comparison of cutoffs between lazy walks and Markovian semigroups". *J. Appl. Probab.* 50.4 (2013), pp. 943–959. URL: http://dx.doi.org/10.1239/jap/1389370092.

[CSC13b] Guan-Yu Chen and Laurent Saloff-Coste. "On the mixing time and spectral gap for birth and death chains". *ALEA Lat. Am. J. Probab. Math. Stat.* 10.1 (2013), pp. 293–321.

[CSC14] Guan-Yu Chen and Laurent Saloff-Coste. "Spectral computations for birth and death chains". *Stochastic Process. Appl.* 124.1 (2014), pp. 848–882. URL: http://dx.doi.org/10.1016/j.spa.2013.10.002.

[CSC15] Guan-Yu Chen and Laurent Saloff-Coste. "Computing cutoff times of birth and death chains". *Electron. J. Probab.* 20 (2015), no. 76, 47.

[DF90] Persi Diaconis and James Allen Fill. "Strong stationary times via a new form of duality". *Ann. Probab.* 18.4 (1990), pp. 1483–1522. URL: http://www.jstor.org/stable/2244330.

[Dia96] Persi Diaconis. "The cutoff phenomenon in finite Markov chains". *Proc. Nat. Acad. Sci. U.S.A.* 93.4 (1996), pp. 1659–1664. URL: http://dx.doi.org/10.1073/pnas.93.4.1659.

[DKSC08] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. "Gibbs sampling, exponential families and orthogonal polynomials". *Statist. Sci.* 23.2 (2008). With comments and a rejoinder by the authors, pp. 151–178. URL: http://dx.doi.org/10.1214/07-STS252.

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". *Ann. Math. Statist.* 27 (1956), pp. 642–669.

[DLP10] Jian Ding, Eyal Lubetzky, and Yuval Peres. "Total variation cutoff in birth-and-death chains". *Probab. Theory Related Fields* 146.1-2 (2010), pp. 61–85. URL: http://dx.doi.org/10.1007/s00440-008-0185-3.

[DM09] Persi Diaconis and Laurent Miclo. "On times to quasi-stationarity for birth and death processes". *J. Theoret. Probab.* 22.3 (2009), pp. 558–586. URL: http://dx.doi.org/10.1007/s10959-009-0234-6.

[Dos+14] Charles R. Doss et al. "Markov chain Monte Carlo estimation of quantiles". *Electron. J. Stat.* 8.2 (2014), pp. 2448–2478. URL: http://dx.doi.org/10.1214/14-EJS957.

[DS81] Persi Diaconis and Mehrdad Shahshahani. "Generating a random permutation with random transpositions". *Z. Wahrsch. Verw. Gebiete* 57.2 (1981), pp. 159–179. URL: http://dx.doi.org/10.1007/BF00535487.

[DS91] Persi Diaconis and Daniel Stroock. "Geometric bounds for eigenvalues of Markov chains". *Ann. Appl. Probab.* 1.1 (1991), pp. 36–61. URL: http://www.jstor.org/stable/2959624.

[DSC06] Persi Diaconis and Laurent Saloff-Coste. "Separation cut-offs for birth and death chains". *Ann. Appl. Probab.* 16.4 (2006), pp. 2098–2122. URL: http://dx.doi.org/10.1214/105051606000000501.

[Dur10] Rick Durrett. *Probability: theory and examples.* Fourth edition. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010, pp. x+428. URL: http://dx.doi.org/10.1017/CBO9780511779398.

[DZP10] E. A. van Doorn, A. I. Zeifman, and T. L. Panfilova. "Bounds and Asymptotics for the Rate of Convergence of Birth-Death Processes". *Theory of Probability & Its Applications* 54.1 (2010), pp. 97–113. URL: http://dx.doi.org/10.1137/S0040585X97984097.

[Fel68] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney, 1968, pp. xviii+509.

[Fil09a] James Allen Fill. "On hitting times and fastest strong stationary times for skip-free and more general chains". *J. Theoret. Probab.* 22.3 (2009), pp. 587–600. URL: http://dx.doi.org/10.1007/s10959-009-0233-7.

[Fil09b] James Allen Fill. "The passage time distribution for a birth-and-death chain: strong stationary duality gives a first stochastic proof". *J. Theoret. Probab.* 22.3 (2009), pp. 543–557. URL: http://dx.doi.org/10.1007/s10959-009-0235-5.

[Fit14] Matthew Fitzpatrick. "Geometric ergodicity of the Gibbs sampler for the Poisson change-point model". *Statist. Probab. Lett.* 91 (2014), pp. 55–61. URL: http://dx.doi.org/10.1016/j.spl.2014.04.008.

[Fle12] James M. Flegal. "Applicability of subsampling bootstrap methods in Markov chain Monte Carlo". *Monte Carlo and quasi-Monte Carlo methods 2010.* Vol. 23. Springer Proc. Math. Stat. Springer, Heidelberg, 2012, pp. 363–372. URL: http://dx.doi.org/10.1007/978-3-642-27440-4_18.

[GG84] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1984), pp. 721–741.

[GO87] Donald P. Gaver and I. G. O'Muircheartaigh. "Robust empirical Bayes analyses of event rates". *Technometrics* 29.1 (1987), pp. 1–15. URL: http://dx.doi.org/10.2307/1269878.

[Gri+14] Simon Griffiths et al. “Tight inequalities among set hitting times in Markov chains”. *Proc. Amer. Math. Soc.* 142.9 (2014), pp. 3285–3298. URL: http://dx.doi.org/10.1090/S0002-9939-2014-12045-4.

[GS90] Alan E. Gelfand and Adrian F. M. Smith. “Sampling-based approaches to calculating marginal densities”. *J. Amer. Statist. Assoc.* 85.410 (1990), pp. 398–409. URL: http://www.jstor.org/stable/2289776.

[Hod98] James S. Hodges. “Some algebra and geometry for hierarchical models, applied to diagnostics”. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60.3 (1998). With discussion and a reply by the author, pp. 497–536. URL: http://dx.doi.org/10.1111/1467-9868.00137.

[Hol85] Richard Holley. “Possible rates of convergence in finite range, attractive spin systems”. *Particle systems, random media and large deviations (Brunswick, Maine, 1984)*. Vol. 41. Contemp. Math. Amer. Math. Soc., Providence, RI, 1985, pp. 215–234. URL: http://dx.doi.org/10.1090/conm/041/814713.

[HS92] Arie Hordijk and Flora Spieksma. “On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network”. *Adv. in Appl. Probab.* 24.2 (1992), pp. 343–376. URL: http://dx.doi.org/10.2307/1427696.

[Hub16] Mark L. Huber. *Perfect simulation.* Vol. 148. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2016, pp. xxii+228.

[JB15] Alicia A. Johnson and Owen Burbank. “Geometric ergodicity and scanning strategies for two-component Gibbs samplers”. *Comm. Statist. Theory Methods* 44.15 (2015), pp. 3125–3145. URL: http://dx.doi.org/10.1080/03610926.2013.823209.

[JH01] Galin L. Jones and James P. Hobert. “Honest exploration of intractable probability distributions via Markov chain Monte Carlo”. *Statist. Sci.* 16.4 (2001), pp. 312–334. URL: http://dx.doi.org/10.1214/ss/1015346317.

[JH04] Galin L. Jones and James P. Hobert. “Sufficient burn-in for Gibbs samplers for a hierarchical random effects model”. *Ann. Statist.* 32.2 (2004), pp. 784–817. URL: http://dx.doi.org/10.1214/009053604000000184.

[JJ10] Alicia A. Johnson and Galin L. Jones. “Gibbs sampling for a Bayesian hierarchical general linear model”. *Electron. J. Stat.* 4 (2010), pp. 313–333. URL: http://dx.doi.org/10.1214/09-EJS515.

[JJ15] Alicia A. Johnson and Galin L. Jones. “Geometric ergodicity of random scan Gibbs samplers for hierarchical one-way random effects models”. *J. Multivariate Anal.* 140 (2015), pp. 325–342. URL: http://dx.doi.org/10.1016/j.jmva.2015.06.002.

[JJ67] Naresh Jain and Benton Jamison. “Contributions to Doeblin’s theory of Markov processes”. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 8 (1967), pp. 19–40.

[JJN13] Alicia A. Johnson, Galin L. Jones, and Ronald C. Neath. "Component-wise Markov chain Monte Carlo: uniform and geometric ergodicity under mixing and composition". *Statist. Sci.* 28.3 (2013), pp. 360–375. URL: http://dx.doi.org/10.1214/13-STS423.

[JO10] Aldéric Joulin and Yann Ollivier. "Curvature, concentration and error estimates for Markov chain Monte Carlo". *Ann. Probab.* 38.6 (2010), pp. 2418–2442. URL: http://dx.doi.org/10.1214/10-AOP541.

[Joh09] Alicia A. Johnson. *Geometric ergodicity of Gibbs samplers.* Thesis (Ph.D.)–University of Minnesota. ProQuest LLC, Ann Arbor, MI, 2009, p. 190. URL: http://search.proquest.com/docview/304932410.

[Jon04] Galin L. Jones. "On the Markov chain central limit theorem". *Probab. Surv.* 1 (2004), pp. 299–320. URL: http://dx.doi.org/10.1214/154957804100000051.

[Kac47] M. Kac. "On the notion of recurrence in discrete stochastic processes". *Bull. Amer. Math. Soc.* 53 (1947), pp. 1002–1010.

[Kei79] Julian Keilson. *Markov chain models—rarity and exponentiality.* Vol. 28. Applied Mathematical Sciences. Springer-Verlag, New York-Berlin, 1979, pp. xiii+184.

[Ken59] David G. Kendall. "Unitary dilations of Markov transition operators, and the corresponding integral representations for transition-probability matrices". *Probability and statistics: The Harald Cramér volume (edited by Ulf Grenander).* Almqvist & Wiksell, Stockholm; John Wiley & Sons, New York, 1959, pp. 139–161.

[KH13] Kshitij Khare and James P. Hobert. "Geometric ergodicity of the Bayesian lasso". *Electron. J. Stat.* 7 (2013), pp. 2150–2163. URL: http://dx.doi.org/10.1214/13-EJS841.

[KM03] I. Kontoyiannis and S. P. Meyn. "Spectral theory and limit theorems for geometrically ergodic Markov processes". *Ann. Appl. Probab.* 13.1 (2003), pp. 304–362. URL: http://dx.doi.org/10.1214/aoap/1042765670.

[KM12] I. Kontoyiannis and S. P. Meyn. "Geometric ergodicity and the spectral gap of non-reversible Markov chains". *Probab. Theory Related Fields* 154.1-2 (2012), pp. 327–339. URL: http://dx.doi.org/10.1007/s00440-011-0373-4.

[KM59] Samuel Karlin and James McGregor. "Coincidence properties of birth and death processes". *Pacific J. Math.* 9 (1959), pp. 1109–1140.

[Kol06] Vladimir Koltchinskii. "Local Rademacher complexities and oracle inequalities in risk minimization". *Ann. Statist.* 34.6 (2006), pp. 2593–2656. URL: http://dx.doi.org/10.1214/009053606000001019.

[LDM15] Fredrik Lindsten, Randal Douc, and Eric Moulines. "Uniform ergodicity of the particle Gibbs sampler". *Scand. J. Stat.* 42.3 (2015), pp. 775–797. URL: http://dx.doi.org/10.1111/sjos.12136.

[Lez98] Pascal Lezaud. “Chernoff-type bound for finite Markov chains”. *Ann. Appl. Probab.* 8.3 (1998), pp. 849–867. URL: http://dx.doi.org/10.1214/aoap/1028903453.

[LMN13] Krzysztof Łatuszyński, Błażej Miasojedow, and Wojciech Niemiro. “Nonasymptotic bounds on the estimation error of MCMC algorithms”. *Bernoulli* 19.5A (2013), pp. 2033–2066. URL: http://dx.doi.org/10.3150/12-BEJ442.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times.* With a chapter by James G. Propp and David B. Wilson. American Mathematical Society, Providence, RI, 2009, pp. xviii+371. URL: http://pages.uoregon.edu/dlevin/MARKOV/.

[LT96] Robert B. Lund and Richard L. Tweedie. “Geometric convergence rates for stochastically ordered Markov chains”. *Math. Oper. Res.* 21.1 (1996), pp. 182–194. URL: http://dx.doi.org/10.1287/moor.21.1.182.

[LW98] László Lovász and Peter Winkler. “Mixing times”. *Microsurveys in discrete probability (Princeton, NJ, 1997).* Vol. 41. DIMACS Ser. Discrete Math. Theoret. Comput. Sci. Amer. Math. Soc., Providence, RI, 1998, pp. 85–133.

[LWK95] Jun S. Liu, Wing H. Wong, and Augustine Kong. “Covariance structure and convergence rate of the Gibbs sampler with various scans”. *J. Roy. Statist. Soc. Ser. B* 57.1 (1995), pp. 157–169. URL: http://www.jstor.org/stable/2346091.

[LZK06] Robert Lund, Ying Zhao, and Peter C. Kiessler. “A monotonicity in reversible Markov chains”. *J. Appl. Probab.* 43.2 (2006), pp. 486–499. URL: http://dx.doi.org/10.1239/jap/1152413736.

[Mas07] Pascal Massart. *Concentration inequalities and model selection.* Vol. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Springer, Berlin, 2007, pp. xiv+337.

[Mas90] P. Massart. “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality”. *Ann. Probab.* 18.3 (1990), pp. 1269–1283. URL: http://www.jstor.org/stable/2244426.

[Met+53] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[MG98] D. J. Murdoch and P. J. Green. “Exact sampling from a continuous state space”. *Scand. J. Statist.* 25.3 (1998), pp. 483–502. URL: http://dx.doi.org/10.1111/1467-9469.00116.

[Mic10] Laurent Miclo. “On absorption times and Dirichlet eigenvalues”. *ESAIM Probab. Stat.* 14 (2010), pp. 117–150. URL: http://dx.doi.org/10.1051/ps:2008037.

[Mic99] L. Miclo. “An example of application of discrete Hardy’s inequalities”. *Markov Process. Related Fields* 5.3 (1999), pp. 319–330.

[MP09] A. Maurer and M. Pontil. “Empirical Bernstein Bounds and Sample Variance Penalization”. *ArXiv e-prints* (July 2009). arXiv: `0907.3740 [stat.ML]`.

[MSA08] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. “Empirical Bernstein stopping”. *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 672–679.

[MT92] Sean P. Meyn and R. L. Tweedie. “Stability of Markovian processes. I. Criteria for discrete-time chains”. *Adv. in Appl. Probab.* 24.3 (1992), pp. 542–574. URL: `http://dx.doi.org/10.2307/1427479`.

[MT93] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993, pp. xvi+548. URL: `http://dx.doi.org/10.1007/978-1-4471-3267-7`.

[MT94] Sean P. Meyn and R. L. Tweedie. “Computable bounds for geometric convergence rates of Markov chains”. *Ann. Appl. Probab.* 4.4 (1994), pp. 981–1011. URL: `http://www.jstor.org/stable/2245077`.

[MTY95] Per Mykland, Luke Tierney, and Bin Yu. “Regeneration in Markov chain samplers”. *J. Amer. Statist. Assoc.* 90.429 (1995), pp. 233–241. URL: `http://www.jstor.org/stable/2291148`.

[NA76] Esa Nummelin and Elja Arjas. “A direct construction of the R-invariant measure for a Markov chain on a general state space”. *Ann. Probability* 4.4 (1976), pp. 674–679.

[NT78] E. Nummelin and R. L. Tweedie. “Geometric ergodicity and R-positivity for general Markov chains”. *Ann. Probability* 6.3 (1978), pp. 404–420.

[NT82] Esa Nummelin and Pekka Tuominen. “Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory”. *Stochastic Process. Appl.* 12.2 (1982), pp. 187–202. URL: `http://dx.doi.org/10.1016/0304-4149(82)90041-2`.

[Num78] E. Nummelin. “A splitting technique for Harris recurrent Markov chains”. *Z. Wahrsch. Verw. Gebiete* 43.4 (1978), pp. 309–318.

[Num84] Esa Nummelin. *General irreducible Markov chains and nonnegative operators*. Vol. 83. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1984, pp. xi+156. URL: `http://dx.doi.org/10.1017/CBO9780511526237`.

[Oli12] Roberto Imbuzeiro Oliveira. “Mixing and hitting times for finite Markov chains”. *Electron. J. Probab.* 17 (2012), no. 70, 12. URL: `http://dx.doi.org/10.1214/EJP.v17-2274`.

[Pak97] Igor Pak. *Random walks on groups: Strong uniform time approach.* Thesis (Ph.D.)–Harvard University. ProQuest LLC, Ann Arbor, MI, 1997, p. 114. URL: http://search.proquest.com/docview/304377165.

[PAR10] Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. "Empirical Bernstein inequalities for U-statistics". *Advances in Neural Information Processing Systems.* 2010, pp. 1903–1911.

[Per04] Yuval Peres. "Sharp thresholds for mixing times". *American Institute of Mathematics (AIM) Research Workshop, Palo Alto.* 2004. URL: www.aimath.org/WWN/mixingtimes/mixingtimes.ps.

[PK14] Subhadip Pal and Kshitij Khare. "Geometric ergodicity for Bayesian shrinkage models". *Electron. J. Stat.* 8.1 (2014), pp. 604–645. URL: http://dx.doi.org/10.1214/14-EJS896.

[Pop77] N. Popov. "Conditions for geometric ergodicity of countable Markov chains". *Soviet Math. Dokl.* 18 (1977), pp. 676–679.

[PS15] Yuval Peres and Perla Sousi. "Mixing times are hitting times of large sets". *J. Theoret. Probab.* 28.2 (2015), pp. 488–519. URL: http://dx.doi.org/10.1007/s10959-013-0497-9.

[Rev84] D. Revuz. *Markov chains.* Second edition. Vol. 11. North-Holland Mathematical Library. North-Holland Publishing Co., Amsterdam, 1984, pp. xi+374.

[RH15] Jorge Carlos Román and James P. Hobert. "Geometric ergodicity of Gibbs samplers for Bayesian general linear mixed models with proper priors". *Linear Algebra Appl.* 473 (2015), pp. 54–77. URL: http://dx.doi.org/10.1016/j.laa.2013.12.013.

[Ros02] Jeffrey S. Rosenthal. "Quantitative convergence rates of Markov chains: a simple account". *Electron. Comm. Probab.* 7 (2002), 123–128 (electronic). URL: http://dx.doi.org/10.1214/ECP.v7-1054.

[Ros95a] Jeffrey S. Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". *J. Amer. Statist. Assoc.* 90.430 (1995), pp. 558–566. URL: http://www.jstor.org/stable/2291067.

[Ros95b] Jeffrey S. Rosenthal. "Rates of convergence for Gibbs sampling for variance component models". *Ann. Statist.* 23.3 (1995), pp. 740–761. URL: http://dx.doi.org/10.1214/aos/1176324619.

[RR04] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". *Probab. Surv.* 1 (2004), pp. 20–71. URL: http://dx.doi.org/10.1214/154957804100000024.

[RR97] Gareth O. Roberts and Jeffrey S. Rosenthal. “Geometric ergodicity and hybrid Markov chains”. *Electron. Comm. Probab.* 2 (1997), no. 2, 13–25 (electronic). URL: http://dx.doi.org/10.1214/ECP.v2-981.

[RT00] G. O. Roberts and R. L. Tweedie. “Rates of convergence of stochastically monotone and continuous time Markov models”. *J. Appl. Probab.* 37.2 (2000), pp. 359–373.

[RT01a] G. O. Roberts and R. L. Tweedie. “Corrigendum to: “Bounds on regeneration times and convergence rates for Markov chains””. *Stochastic Process. Appl.* 91.2 (2001), pp. 337–338. URL: http://dx.doi.org/10.1016/S0304-4149(00)00074-0.

[RT01b] Gareth O. Roberts and Richard L. Tweedie. “Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains”. *J. Appl. Probab.* 38A (2001). Probability, statistics and seismology, pp. 37–41. URL: http://dx.doi.org/10.1239/jap/1085496589.

[RT99] G. O. Roberts and R. L. Tweedie. “Bounds on regeneration times and convergence rates for Markov chains”. *Stochastic Process. Appl.* 80.2 (1999), pp. 211–229. URL: http://dx.doi.org/10.1016/S0304-4149(98)00085-4.

[SC04] Laurent Saloff-Coste. “Random walks on finite groups”. *Probability on discrete structures*. Vol. 110. Encyclopaedia Math. Sci. Springer, Berlin, 2004, pp. 263–346. URL: http://dx.doi.org/10.1007/978-3-662-09444-0_5.

[Spi90] Flora Margaretha Spieksma. “Geometrically ergodic Markov chains and the optimal control of queues”. PhD thesis. Rijksuniversiteit te Leiden, 1990.

[Tie94] Luke Tierney. “Markov chains for exploring posterior distributions”. *Ann. Statist.* 22.4 (1994). With discussion and a rejoinder by the author, pp. 1701–1762. URL: http://dx.doi.org/10.1214/aos/1176325750.

[Val94] V. de Valk. *One-dependent processes: two-block factors and non-two-block factors*. Vol. 85. CWI Tract. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1994, pp. ii+178.

[VJ62] D. Vere-Jones. “Geometric ergodicity in denumerable Markov chains”. *Quart. J. Math. Oxford Ser. (2)* 13 (1962), pp. 7–28.

[Yau96] Horng-Tzer Yau. “Logarithmic Sobolev inequality for lattice gases with mixing conditions”. *Comm. Math. Phys.* 181.2 (1996), pp. 367–408. URL: http://projecteuclid.org/euclid.cmp/1104287767.

[Zeĭ91] A. I. Zeĭfman. “Some estimates of the rate of convergence for birth and death processes”. *J. Appl. Probab.* 28.2 (1991), pp. 268–277.

[Zeĭ95a] A. I. Zeĭfman. “On the estimation of probabilities for birth and death processes”. *J. Appl. Probab.* 32.3 (1995), pp. 623–634.

[Zeĭ95b] A. I. Zeĭfman. “Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes”. *Stochastic Process. Appl.* 59.1 (1995), pp. 157–173. URL: http://dx.doi.org/10.1016/0304-4149(95)00028-6.