Chapter 1

Convergence Bounds for Monte Carlo Markov Chains

Qian Qin

1.1 Introduction

When implementing a Markov chain Monte Carlo (MCMC) algorithm, one simulates a Markov chain $(X_t)_{t=0}^{\infty}$ such that the distribution of X_t approaches a target distribution $\varpi(\cdot)$ as t increases. Inference about the target distribution is then conducted based on a Monte Carlo sample, which is a finite portion of the chain $(X_t)_{t=\underline{n}}^{\overline{n}}$. The integer $\underline{n}$ is the amount of burn-in, while the integer $\overline{n}$ is the time at which the simulation is terminated. For the Monte Carlo sample $(X_t)_{t=\underline{n}}^{\overline{n}}$ to be representative of the target distribution, the distribution of most of the X_t's, $\underline{n} \leq t \leq \overline{n}$, should be similar to $\varpi(\cdot)$. To be certain of this, one would need to know how fast the distribution of X_t converges to $\varpi(\cdot)$ as $t \to \infty$. This chapter reviews several popular methods for convergence analysis, i.e., ascertaining the convergence properties of a Monte Carlo Markov chain using mathematics. To be specific, we investigate ways to construct bounds on the distance between the distribution of a Markov chain at a given time point and the chain's stationary distribution.

Convergence analysis plays a pivotal role in the theory and application of MCMC. Some important asymptotic results for MCMC estimators, such as the central limit theorem and the strong invariance principle, rely on conditions on the convergence rate of the Monte Carlo Markov chain; see, e.g., Jones (2004) and Kuelbs and Philipp (1980). This type of condition can be verified through appropriate convergence bounds. It is also possible to derive non-asymptotic error bounds for Monte Carlo estimators based on convergence bounds (Rudolf, 2012, Theorems 3.34 and 3.41).

A class of methods related to convergence analysis is convergence diagnostics, which aims to assess the performance of an MCMC algorithm by scrutinizing its output (Brooks and Roberts, 1998; Gelman and Rubin, 1992; Gelman and Shirley, 2011; Roy, 2020). Convergence diagnostics can detect problems with an MCMC simulation, but they cannot prove that the simulation is generating a representative sample. While convergence analysis is often mathematically challenging, it offers robust theoretical guarantees that diagnostics cannot provide.

There is an enormous body of literature devoted to the topic at hand. This chapter serves primarily as an introductory guide for those embarking on further research in this area. It is assumed that readers possess a moderate level of familiarity with the languages of measure theoretic probability and linear algebra within Hilbert spaces.

The rest of this chapter is organized as follows. In Section 1.2, we lay out the basic concepts and notations. In Section 1.3, we review the coupling method for constructing convergence bounds. In Section 1.4, we describe the L^2 framework for convergence analysis, with a focus on methods involving isoperimetric inequalities. Finally, some other methods for constructing convergence bounds are listed in Section 1.5.

1.2 Basic Setup

We begin by setting up some notations. Let X be a Polish (separable and complete) metric space with metric $\psi : \mathsf{X}^2 \to [0, \infty)$, and let $\mathcal{B}$ be its Borel σ algebra. Denote by $\mathcal{P}(\mathsf{X})$ the collection of probability measures, or distributions, on $(\mathsf{X}, \mathcal{B})$. Let $(X_t)_{t=0}^{\infty}$ be a time-homogeneous Markov chain whose state space is X. Let $K : \mathsf{X} \times \mathcal{B} \to [0, 1]$ be its Markov transition kernel (Mtk) so that $P(X_{t+1} \in A \mid X_t = x) = K(x, A)$ for $x \in \mathsf{X}$ and $A \in \mathcal{B}$. For $t \in \mathbb{N}_+ := \{1, 2, \dots\}$, we can define the t-step Mtk of the chain, which is a function $K^t : \mathsf{X} \times \mathcal{B} \to [0, 1]$ satisfying $P(X_t \in A \mid X_0 = x) = K^t(x, A)$. Indeed, one simply let $K^1(x, A) = K(x, A)$, and $K^{t+1}(x, A) = \int_{\mathsf{X}} K^t(x, \mathrm{d}y) K(y, A)$. For $\mu \in \mathcal{P}(\mathsf{X})$, let $\mu K^t(\cdot) = \int_{\mathsf{X}} \mu(\mathrm{d}x) K^t(x, \cdot)$ if $t \in \mathbb{N}_+$, and, by convention, let $\mu K^0 = \mu$. Then $\mu K^t(\cdot)$ is the distribution of X_t if $X_0 \sim \mu$.

Assume that $(X_t)_{t=0}^{\infty}$ has a stationary distribution $\varpi \in \mathcal{P}(\mathsf{X})$, so that $\varpi K^t(\cdot) = \varpi(\cdot)$ for $t \in \mathbb{N} = \{0\} \cup \mathbb{N}_+$. If this chain is associated with an MCMC algorithm targeting $\varpi(\cdot)$, then the distribution of X_t should converge to $\varpi(\cdot)$ in some sense as $t \to \infty$. When conducting a convergence analysis, we seek to understand how fast $\mu K^t(\cdot)$ approaches $\varpi(\cdot)$ as t grows for some class of initial distributions $\mu(\cdot)$.

To conduct a quantitative analysis, we need to define a distance function that quantifies the difference between two probability measures. A common way to construct such a distance is as follows (Müller, 1997; Zolotarev, 1984). Let $\mathcal{F}$ be a collection of real measurable functions on X. Let $\mathcal{F}'$ be a subset of $\mathcal{P}(\mathsf{X})$ such that $\int_{\mathsf{X}} |f(x)| \, \mu(\mathrm{d}x) < \infty$ for each $f \in \mathcal{F}$ whenever $\mu \in \mathcal{F}'$. For $\mu, \nu \in \mathcal{F}'$, we define the "integral probability metric"

$$\|\mu - \nu\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mu f - \nu f|,$$

where $\mu f = \int_{\mathsf{X}} f \, \mathrm{d}\mu$. One can check that, for $\mu, \nu, \omega \in \mathcal{F}'$,

$$\|\mu - \nu\|_{\mathcal{F}} \leq \|\mu - \omega\|_{\mathcal{F}} + \|\omega - \nu\|_{\mathcal{F}}.$$

Assume that $\mathcal{F}$ is rich enough so that $\|\mu - \nu\|_{\mathcal{F}} = 0$ implies that $\mu(A) = \nu(A)$ for $A \in \mathcal{B}$.

Then $\|\mu - \nu\|_{\mathcal{F}}$ serves as a distance between μ and ν.

Below we list some commonly used distances constructed in this manner.

(I) $\mathcal{F}$ is the set of functions f such that $\sup_{x \in \mathsf{X}} |f(x)| = 1/2$, and $\mathcal{F}' = \mathcal{P}(\mathsf{X})$. Then $\|\mu - \nu\|_{\mathcal{F}}$ is the total variation distance between μ and ν for $\mu, \nu \in \mathcal{F}'$. The same goes if $\mathcal{F}$ is the set of measurable indicator functions. In this case, we write $\|\mu - \nu\|_{\mathcal{F}}$ as $\|\mu - \nu\|_{\mathrm{TV}}$. If μ and ν are absolutely continuous with respect to some σ-finite measure λ, then

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathsf{X}} \left| \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x) - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x) \right| \lambda(\mathrm{d}x) = 1 - \int_{\mathsf{X}} \min\left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x), \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(x) \right\} \lambda(\mathrm{d}x). \tag{1.2.1}$$

(II) $\mathcal{F}$ is the set of functions f such that $\sup_{x \neq y} |f(x) - f(y)|/\psi(x,y) = 1$, i.e., the set of functions whose Lipschitz constant is 1. $\mathcal{F}'$ is the set of probability measures μ such that $\int_{\mathsf{X}} \psi(x_0, x)\, \mu(\mathrm{d}x) < \infty$ for some $x_0 \in \mathsf{X}$. Then by the Kantorovich-Rubinstein duality (see, e.g., Villani, 2008, Theorem 5.10), $\|\mu - \nu\|_{\mathcal{F}}$ is the 1-Wasserstein distance between μ and ν induced by ψ. In this case, we write $\|\mu - \nu\|_{\mathcal{F}}$ as $W_{\psi}(\mu, \nu)$.

(III) $\mathcal{F}$ is the set of functions f such that $\int_{\mathsf{X}} f(x)^2\, \varpi(\mathrm{d}x) = 1$. $\mathcal{F}'$ is the set of probability measures μ such that μ is absolutely continuous with respect to ϖ (i.e., $\mu \ll \varpi$), and that

$$\int_{\mathsf{X}} \left[\frac{\mathrm{d}\mu}{\mathrm{d}\varpi}(x) \right]^2 \varpi(\mathrm{d}x) < \infty.$$

(It can be checked that, if $\mu \ll \varpi$, then the above display is equivalent to $\int_{\mathsf{X}} |f(x)|\, \mu(\mathrm{d}x)$ being bounded as f varies in $\mathcal{F}$.) Then $\|\mu - \nu\|_{\mathcal{F}}$ is the L^2 distance between μ and ν. In this case, we write $\|\mu - \nu\|_{\mathcal{F}}$ as $\|\mu - \nu\|_2$. One can show via the Cauchy-Schwarz inequality that the L^2 distance has a dual representation

$$\|\mu - \nu\|_2 = \sqrt{\int_{\mathsf{X}} \left[\frac{\mathrm{d}\mu}{\mathrm{d}\varpi}(x) - \frac{\mathrm{d}\nu}{\mathrm{d}\varpi}(x) \right]^2 \varpi(\mathrm{d}x)}. \tag{1.2.2}$$

Throughout this chapter, assume that $\mathcal{F}'$ contains the stationary distribution ϖ. We also assume that $\mu K \in \mathcal{F}'$ whenever $\mu \in \mathcal{F}'$, so that $\mu K^t \in \mathcal{F}'$ for $t \in \mathbb{N}$ whenever $\mu \in \mathcal{F}'$. It can be shown that the two assumptions always hold in scenarios (I) and (III); see, e.g., Lemma 22.1.3 of Douc et al. (2004). In scenario (II), the second assumption holds if, say, there exist a point $x_0 \in \mathsf{X}$ and finite constants c_1 and c_2 such that $\int_{\mathsf{X}} \psi(x_0, x') \, K(x, \mathrm{d}x') \leq c_1 \psi(x_0, x) + c_2$ for $x \in \mathsf{X}$. All examples herein satisfy these two assumptions.

A central goal of convergence analysis is to construct bounds on $\|\mu K^t - \varpi\|_{\mathcal{F}}$ for at least some initial distribution $\mu \in \mathcal{F}'$ that is practically feasible. We are mainly concerned with constructing upper bounds on $\|\mu K^t - \varpi\|_{\mathcal{F}}$, although lower bounds will also be touched on. In particular, we shall focus on several methods that enable us to form convergence bounds of the form

$$\|\mu K^t - \varpi\|_{\mathcal{F}} \leq C_\mu \rho^t, \quad t \in \mathbb{N}_+,$$

where $C_\mu \in (0, \infty)$ is a function of the initial distribution μ, and ρ is a constant in $[0, 1)$. This type of bound, among others, can be used to bound the (ϵ, μ)-mixing time, which is the smallest $t \in \mathbb{N}_+$ such that $\|\mu K^t - \varpi\|_{\mathcal{F}} \leq \epsilon$, where $\epsilon \in (0, \infty)$ is some prescribed level of tolerance. Indeed, denoting the (ϵ, μ)-mixing time by $t_{\mathcal{F}}(\epsilon, \mu)$, the above bound would yield

$$t_{\mathcal{F}}(\epsilon, \mu) \leq \left\lceil \frac{\log C_\mu - \log \epsilon}{-\log \rho} \right\rceil$$

if $\rho \in (0, 1)$, where $\lceil \cdot \rceil$ is the ceiling function.

The choice of the distance function would of course affect the outcome of one's analysis, but bounds in terms of one distance can often be translated to those in terms of another. For instance, using (1.2.1), (1.2.2), and Jensen's inequality, one can show that if $\mathrm{d}\mu/\mathrm{d}\varpi$ exists and is squared integrable with respect to ϖ, then

$$2\|\mu K^t - \varpi\|_{\mathrm{TV}} \leq \|\mu K^t - \varpi\|_2,$$

so an upper bound on the right-hand-side upper bounds the left-hand-side as well. See Roberts and Rosenthal (1997), Roberts and Tweedie (2001), and Kontoyiannis and Meyn (2012) for additional details on the relationship between convergence bounds in the L^2 and total variation distances. In Section 1.3.2, we discuss how to translate a bound in terms of the 1-Wasserstein distance to one in terms of the total variation distance.

For the construction of upper bounds on $\|\mu K^t - \varpi\|_{\mathcal{F}}$, we review (a) the coupling method when $\|\cdot - \cdot\|_{\mathcal{F}}$ is the total variation distance or the 1-Wasserstein distance, and (b) the L^2 theory, especially techniques based on the conductance and isoperimetric inequalities, when $\|\cdot - \cdot\|_{\mathcal{F}}$ is the L^2 distance. We will then describe a simple method for lower bounding the "convergence rate" in the L^2 framework, which quantifies how slow a chain converges. Some other important methods are listed with references at the end of the chapter.

To end Section 1.2, we give a couple running toy examples on which we will demonstrate several techniques for convergence analysis.

Example 1.2.1. *Let* $\mathsf{X} = [0,1]$, *and let* $\mathcal{B}$ *be the Borel subsets of* $[0,1]$. *Let* $s : [0,1] \to (0,\infty)$ *be a positive continuous probability density function, and denote the corresponding distribution by* $\pi_s(\cdot)$. *Note that, due to continuity,* $M_s := \sup_{x\in[0,1]} s(x) < \infty$. *For* $x, x' \in \mathsf{X}$, *let* $a_s(x,x') = \min\{1, s(x')/s(x)\}$. *Let* $(X_t)_{t=0}^{\infty}$ *be a Markov chain such that, given* X_t, *the next state* X_{t+1} *is generated using the following procedure: Draw* X' *from the uniform distribution on* $[0,1]$; *with probability* $a_s(X_t, X')$, *set* $X_{t+1} = X'$; *with probability* $1 - a_s(X_t, X')$, *set* $X_{t+1} = X_t$. *Then* $(X_t)_{t=0}^{\infty}$ *is associated with an independent Metropolis Hastings algorithm targeting* $\pi_s(\cdot)$. *Its transition kernel is*

$$K_s(x,A) = \int_A a_s(x,x')\,\mathrm{d}x' + \left[1 - \int_0^1 a_s(x,x')\,\mathrm{d}x'\right]\mathbf{1}_{x\in A}, \quad x \in [0,1],\, A \in \mathcal{B},$$

where $\mathbf{1}_{x\in A}$ *is 1 if* $x \in A$ *and 0 otherwise. It is well-known that chains associated with Metropolis Hastings algorithms are reversible with respect to their target distributions.*

Example 1.2.2. *Let* $\mathsf{X} = \mathbb{R}^p$, *where* p *is a positive integer, and let* $\mathcal{B}$ *be the Borel sets. Let* $\alpha \in [0,1)$ *be a constant. Define the Gaussian chain as a Markov chain* $(X_t)_{t=0}^{\infty}$ *such that, given* $X_t = x \in \mathbb{R}^p$, X_{t+1} *follows the* $N_p(\alpha x,\, (1-\alpha^2)I_p)$ *distribution, where* $N_p(m,V)$ *means the* p*-variate normal distribution with mean* m *and variance* V, *and* I_p *is the* $p \times p$ *identity matrix. Its transition kernel is*

$$K_{p,\alpha}(x,A) = \int_A \frac{1}{[2\pi(1-\alpha^2)]^{p/2}} \exp\left[-\frac{1}{2(1-\alpha^2)}\|x' - \alpha x\|^2\right]\mathrm{d}x', \quad x \in \mathbb{R}^p,\, A \in \mathcal{B},$$

where $\|\cdot\|$ *is the Euclidean norm. This chain is reversible with respect to the* $N_p(0, I_p)$

measure, which will be denoted by $\varpi_p(\cdot)$*. Indeed,*

$$\int_A \varpi_p(\mathrm{d}x) K_{p,\alpha}(x, B) = \frac{1}{(2\pi)^p (1-\alpha^2)^{p/2}} \int_{A \times B} \exp\left[-\frac{\|x'\|^2 + \|x\|^2 - 2\alpha x^\top x'}{2(1-\alpha^2)} \right] \mathrm{d}x \, \mathrm{d}x'$$

is a symmetric function of $A \in \mathcal{B}$ *and* $B \in \mathcal{B}$*.*

Due of the simplicity of the chains in these examples, their convergence properties are well-understood, but for illustrative purposes we will feign ignorance in most of our analyses.

1.3 Bounds via coupling

The coupling method is a powerful tool in probability theory that enables one to compare two distributions. Numerous works have utilized the technique to obtain useful convergence bounds for a wide range of important Markov chains. See Aldous (1983); Bou-Rabee et al. (2020); Bubley and Dyer (1997); Burdzy and Kendall (2000); Durmus and Moulines (2019); Eberle and Majka (2019); Lindvall and Rogers (1986); Pillai and Smith (2017), just to name several. In this section, we describe the general idea of this approach, and illustrate it through a few simple examples. In particular, we use it to derive a convergence bound from a set of “drift and minorization conditions.”

1.3.1 Basic theory

For $\mu, \nu \in \mathcal{P}(\mathsf{X})$, a coupling of theirs is a distribution in $\mathcal{P}(\mathsf{X}^2)$, say γ, such that $\gamma(A \times \mathsf{X}) = \mu(A)$ and $\gamma(\mathsf{X} \times A) = \nu(A)$ for $A \in \mathcal{B}$. In other words, γ is a coupling of μ and ν if it is the joint distribution of some random vector (X, Y) such that, marginally, $X \sim \mu$ and $Y \sim \nu$. Denote the set of all couplings of μ and ν by $C(\mu, \nu)$. Suppose that we can find a measurable function $D : \mathsf{X}^2 \to [0, \infty]$ such that, for $f \in \mathcal{F}$,

$$|f(x) - f(y)| \leq D(x, y), \quad (x, y) \in \mathsf{X}^2. \tag{1.3.1}$$

(The function $D(\cdot, \cdot)$ is often some semi-metric.) Then, for $\mu, \nu \in \mathcal{F}'$ and $\gamma \in C(\mu, \nu)$,

$$\begin{aligned} \|\mu - \nu\|_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |\mu f - \nu f| \\ &= \sup_{f \in \mathcal{F}} \left| \int_{\mathsf{X}^2} [f(x) - f(y)] \, \gamma(\mathrm{d}(x, y)) \right| \\ &\leq \int_{\mathsf{X}^2} D(x, y) \, \gamma(\mathrm{d}(x, y)). \end{aligned} \tag{1.3.2}$$

If one can construct a random vector (X, Y) whose joint distribution is in $C(\mu, \nu)$, then $\|\mu - \nu\|_{\mathcal{F}} \leq E[D(X, Y)]$. In particular, if one can, on some probability space, define a copy of X_t along with a random element Y_t such that $X_t \sim \mu K^t$ and $Y_t \sim \varpi$, i.e., $(X_t, Y_t) \sim \gamma_t \in C(\mu K^t, \varpi)$, then $\|\mu K^t - \varpi\|_{\mathcal{F}} \leq E[D(X_t, Y_t)]$. Usually, to obtain a sharp bound, X_t and Y_t need to be correlated in some suitable manner.

This approach can be used to bound the total variation and the 1-Wasserstein distances. Indeed, if $\mathcal{F}$ is the set of functions f such that $\sup_{x \in \mathsf{X}} |f(x)| = 1/2$, then (1.3.1) holds for $f \in \mathcal{F}$ when $D(x, y) \geq \mathbf{1}_{x \neq y}$. Thus, if $(X, Y) \sim \gamma \in C(\mu, \nu)$, then taking $D(x, y) = \mathbf{1}_{x \neq y}$ yields

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \int_{\mathsf{X}^2} \mathbf{1}_{x \neq y} \, \gamma(\mathrm{d}(x, y)) = P(X \neq Y). \tag{1.3.3}$$

If $\mathcal{F}$ is the set of functions f such that $\sup_{x \neq y} |f(x) - f(y)|/\psi(x, y) = 1$, then (1.3.1) holds for $f \in \mathcal{F}$ when $D(x, y) \geq \psi(x, y)$. Thus, if $(X, Y) \sim \gamma \in C(\mu, \nu)$, then taking $D(x, y) = \psi(x, y)$ yields

$$W_\psi(\mu, \nu) \leq \int_{\mathsf{X}^2} \psi(x, y) \, \gamma(\mathrm{d}(x, y)) = E[\psi(X, Y)].$$

This is obvious if one knows the more standard definition of the 1-Wasserstein distance:

$$W_\psi(\mu, \nu) = \inf_{\gamma \in C(\mu, \nu)} \int_{\mathsf{X}^2} \psi(x, y) \, \gamma(\mathrm{d}(x, y)).$$

It is worth noting that, in the above display, there always exists a coupling γ that attains the infimum (see, e.g., Villani, 2008, Theorem 4.1).

It is common (but not always optimal) that couplings of μK^t and $\varpi = \varpi K^t$ are constructed in a Markovian manner. That is, one constructs a bivariate Markov chain $(X_t, Y_t)_{t=0}^{\infty}$

with state space X^2 such that the distribution of (X_t, Y_t) is in $C(\mu K^t, \varpi K^t)$ for $t \in \mathbb{N}$. This can be achieved if $X_0 \sim \mu$, $Y_0 \sim \varpi$, and the Mtk of the bivariate chain, denoted by $\tilde{K} : \mathsf{X}^2 \times \mathcal{B}^2 \to [0,1]$, is a coupling kernel of K in the following sense: For $x, y \in \mathsf{X}$, $\tilde{K}((x,y), \cdot)$ is in $C(\delta_x K, \delta_y K)$, where δ_x is the point mass at x (i.e., $\delta_x(A) = \mathbf{1}_{x \in A}$ for $A \in \mathcal{B}$) so that $\delta_x K(\cdot) = K(x, \cdot)$. In other words, given $(X_t, Y_t) = (x, y)$, X_{t+1} is distributed as $K(x, \cdot)$, and Y_{t+1} is distributed as $K(y, \cdot)$.

A coupling kernel always exists, since we can let

$$\int_A \tilde{K}((x,y), \mathrm{d}(x', y')) = \int_A K(x, \mathrm{d}x') K(y, \mathrm{d}y'), \quad (x,y) \in \mathsf{X}^2, \; A \in \mathcal{B}^2.$$

But this construction wouldn't be very helpful since there is no dependence between X_t and Y_t conditioning on (X_0, Y_0), which usually renders the bound $\|\mu K^t - \varpi\|_{\mathcal{F}} \leq E[D(X_t, Y_t)]$ too loose. The following is an elementary result that provides a more useful coupling kernel under a simple but restrictive condition.

Theorem 1.3.1. *Suppose that there exist $\varepsilon > 0$ and a probability measure $\nu \in \mathcal{P}(\mathsf{X})$ such that, for $x \in \mathsf{X}$ and $A \in \mathcal{B}$,*

$$K(x, A) \geq \varepsilon \nu(A).$$

(This is called Doeblin's, or a global minorization condition.) Then one may construct a coupling kernel $\tilde{K}$ of K such that

$$\int_{\mathsf{X}^2} \mathbf{1}_{x' \neq y'} \tilde{K}((x,y), \mathrm{d}(x', y')) \leq (1 - \varepsilon) \mathbf{1}_{x \neq y} \tag{1.3.4}$$

for $(x, y) \in \mathsf{X}^2$.

Remark 1.3.1. *If $(X_t, Y_t)_{t=0}^{\infty}$ is a bivariate chain whose Mtk satisfies* (1.3.4)*, then, for $t \in \mathbb{N}$, $P(X_{t+1} \neq Y_{t+1} \mid X_t, Y_t) = E(\mathbf{1}_{X_{t+1} \neq Y_{t+1}} \mid X_t, Y_t) \leq (1 - \varepsilon) \mathbf{1}_{X_t \neq Y_t}$.*

Proof of Theorem 1.3.1. Let $(X_t, Y_t)_{t=0}^{\infty}$ be a bivariate Markov chain with state space X^2 that evolves as follows. Suppose that the current state is $(X_t, Y_t) = (x, y) \in \mathsf{X}^2$. If $x = y$, let $X_{t+1} = Y_{t+1}$ be distributed as $K(x, \cdot)$. When $x \neq y$, do the following. With probability ε, let $X_{t+1} = Y_{t+1}$ be distributed as ν; with probability $1 - \varepsilon$ (if $\varepsilon < 1$), let X_{t+1} be distributed according to the probability measure

$$A \mapsto \frac{K(x, A) - \varepsilon\nu(A)}{1 - \varepsilon}, \tag{1.3.5}$$

and, independently, let Y_{t+1} be distributed according to the probability measure

$$A \mapsto \frac{K(y, A) - \varepsilon\nu(A)}{1 - \varepsilon}. \tag{1.3.6}$$

Note that the two measures are well-defined whenever $\varepsilon \in (0, 1)$ due to Doeblin's condition. Evidently, given $(X_t, Y_t) = (x, y)$, X_{t+1} is distributed as $K(x, \cdot)$, and Y_{t+1} is distributed as $K(y, \cdot)$. Thus, the Mtk of the bivariate chain, which we denote by $\tilde{K}$, is a coupling kernel of K.

By construction, for $x, y \in \mathsf{X}^2$,

$$\int_{\mathsf{X}^2} \mathbf{1}_{x'=y'} \tilde{K}((x, y), \mathrm{d}(x', y')) \geq \begin{cases} \varepsilon, & x \neq y, \\ 1, & x = y. \end{cases}$$

This establishes (1.3.4) for $(x, y) \in \mathsf{X}^2$. □

Consider the independent Metropolis Hastings chain in Example 1.2.1. Since $s(x) \leq M_s$ for $x \in \mathsf{X}$, it holds that, for $x \in \mathsf{X} = [0, 1]$ and $A \in \mathcal{B}$,

$$K_s(x, A) \geq \int_A \inf_{x \in [0,1]} a_s(x, x') \, \mathrm{d}x' = \int_A \min\left\{1, \frac{s(x')}{M_s}\right\} \mathrm{d}x' = \int_A \frac{s(x')}{M_s} \, \mathrm{d}x' = \frac{1}{M_s} \pi_s(A). \tag{1.3.7}$$

That is, Doeblin's condition holds with $\varepsilon = 1/M_s$. Hence, there exists a coupling kernel of K_s satisfying (1.3.4) for $(x, y) \in [0, 1]^2$ with $1 - \varepsilon = 1 - 1/M_s$. On the other hand, the Gaussian chain in Example 1.2.2 does not satisfy Doeblin's condition with any positive ε since, for any bounded $A \in \mathcal{B}$, $\inf_{x \in \mathsf{X}} K_{p,\alpha}(x, A) = 0$.

As implied by Remark 1.3.1, (1.3.4) is a type of "contraction condition" that indicates $\mathbf{1}_{X_t \neq Y_t}$ decreases in expectation at a geometric rate as t grows. Let us now show exactly how a coupling kernel that satisfies a contraction condition can be used to construct a convergence bound.

Theorem 1.3.2. *Let $\tilde{K}$ be a coupling kernel of K. Suppose that there exist a constant $\rho < 1$ and a measurable function $D : \mathsf{X}^2 \to [0, \infty]$ satisfying (1.3.1) for all $f \in \mathcal{F}$ such that the following contraction condition holds:*

$$\int_{\mathsf{X}^2} D(x', y')\, \tilde{K}((x, y), \mathrm{d}(x', y')) \leq \rho D(x, y) \tag{1.3.8}$$

for $(x, y) \in \mathsf{X}^2$. Then, for $\mu \in \mathcal{F}'$ and $t \in \mathbb{N}$,

$$\|\mu K^t - \varpi\|_{\mathcal{F}} \leq \int_{\mathsf{X}^2} D(x, y)\, \gamma(\mathrm{d}(x, y))\, \rho^t,$$

where γ is any coupling of μ and ϖ.

Proof. Let $\mu \in \mathcal{F}'$ be arbitrary. Let $(X_t, Y_t)_{t=0}^{\infty}$ be a bivariate chain associated with $\tilde{K}$ such that $(X_0, Y_0) \sim \gamma \in C(\mu, \varpi)$. Then the distribution of (X_t, Y_t) is in $C(\mu K^t, \varpi)$. Thus, by (1.3.2), for $t \in \mathbb{N}$,

$$\|\mu K^t - \varpi\|_{\mathcal{F}} \leq E[D(X_t, Y_t)].$$

On the other hand, by (1.3.8), for $t \in \mathbb{N}$,

$$E[D(X_{t+1}, X_{t+1}) \mid X_t, Y_t] \leq \rho D(X_t, Y_t).$$

By the two displays above and the tower property of conditional expectations,

$$\|\mu K^t - \varpi\|_{\mathcal{F}} \leq E[D(X_t, Y_t)] \leq \rho^t E[D(X_0, Y_0)] = \int_{\mathsf{X}^2} D(x, y)\, \gamma(\mathrm{d}(x, y)) \rho^t.$$

□

We may apply Theorem 1.3.2 to Example 1.2.1. It was already demonstrated through Theorem 1.3.1 that there is a coupling kernel of K_s that satisfies (1.3.8) for $(x, y) \in [0, 1]^2$ with $D(x, y) = \mathbf{1}_{x \neq y}$ and $\rho = 1 - 1/M_s$. Let $\mathcal{F}$ be the set of functions f such that $\sup_{x \neq \mathsf{X}} |f(x)| = 1/2$ so that $\| \cdot - \cdot \|_{\mathcal{F}}$ corresponds to the total variation distance, and (1.3.1) holds for $f \in \mathcal{F}$.

By Theorem 1.3.2, for $\mu \in \mathcal{P}(\mathsf{X})$,

$$\|\mu K_s^t - \pi_s\|_{\mathrm{TV}} \leq \int_{\mathsf{X}^2} \mathbf{1}_{x \neq y}\, \mu(\mathrm{d}x)\, \pi_s(\mathrm{d}y) \left(1 - \frac{1}{M_s}\right)^t \leq \left(1 - \frac{1}{M_s}\right)^t.$$

If M_s is known, then this is a fully computable convergence bound for the independent Metropolis Hastings chain.

Turning to the Gaussian chain in Example 1.2.2, we note that Theorem 1.3.1 is insufficient to provide a coupling kernel with a proper contraction condition. To effectively utilize Theorem 1.3.2 in this context, we would need some other techniques for constructing coupling kernels. This will be resolved in Section 1.3.2.

1.3.2 One-shot coupling

To obtain a sharp convergence bound using Theorem 1.3.2, one needs to find a coupling kernel $\tilde{K}$ such that ρ is small (or at the very least, strictly less than 1) in the contraction condition (1.3.8). This is not always easy, but for some functions $D(\cdot, \cdot)$ it may be easier than for others. Of course, the choice of $D(\cdot, \cdot)$ depends on the distance function $\| \cdot - \cdot \|_{\mathcal{F}}$ because of the restriction (1.3.1). When a bound in terms of the total variation distance is desired, $D(x, y)$ is often taken to be a weighted version of $\mathbf{1}_{x \neq y}$, i.e., $\mathbf{1}_{x \neq y}\, h(x, y)$ for some $h(x, y) \geq 1$; see Section 1.3.3. When a bound in terms of the 1-Wasserstein distance induced by $\psi(\cdot, \cdot)$ is desired, one may consider $D(x, y)$ of the form $\psi(x, y)^r h(x, y)^{1-r}$ for some $h(x, y) \geq 1$ and $r \in (0, 1]$ (Douc et al., 2018, Section 20.4). Sometimes, for the distance function $\| \cdot - \cdot \|_{\mathcal{F}}$ that we are interested in, it is too difficult to establish a good contraction condition for a function $D(\cdot, \cdot)$ that satisfies (1.3.1) for $f \in \mathcal{F}$. In such cases, it may be helpful to consider some other distance function $\| \cdot - \cdot \|_{\mathcal{G}}$ first, establish a good contraction condition suitable for that distance, and then transform the resulting convergence bound to one in terms of the original distance.

Starting from a convergence bound in terms of the 1-Wasserstein distance, it is often possible to obtain a bound in terms of the total variation distance through a technique called "one-shot coupling" (Madras and Sezer, 2010; Roberts and Rosenthal, 2002). In particular, one can use the following theorem.

Theorem 1.3.3. *(Madras and Sezer, 2010, Theorem 12) Let $\mathcal{F}'$ be the set of probability measures $\mu \in \mathcal{P}(\mathsf{X})$ such that $\int_{\mathsf{X}} \psi(x_0, x)\, \mu(\mathrm{d}x)$ for some $x_0 \in \mathsf{X}$. Let ν be a σ-finite measure on $(\mathsf{X}, \mathcal{B})$. Assume that there is a measurable function $k : \mathsf{X}^2 \to [0, \infty)$ such that, for $x \in \mathsf{X}$ and $A \in \mathcal{B}$, $K(x, A) = \int_A k(x, x')\, \nu(\mathrm{d}x')$. Suppose further that there exists a constant $b < \infty$ such that*

$$1 - \int_{\mathsf{X}} \min\{k(x, x'), k(y, x')\}\, \nu(\mathrm{d}x') \le b\psi(x, y) \tag{1.3.9}$$

for $x, y \in \mathsf{X}$. Then, for $t \in \mathbb{N}$ and $\mu \in \mathcal{F}'$,

$$\|\mu K^{t+1} - \varpi\|_{TV} \le bW_\psi(\mu K^t, \varpi).$$

Proof. Recall that

$$W_\psi(\mu K^t, \varpi) = \inf_{\gamma_t \in C(\mu K^t, \varpi)} \int_{\mathsf{X}^2} \psi(x, y)\, \gamma(\mathrm{d}(x, y)),$$

and one can find $\gamma_t \in C(\mu K^t, \varpi)$ that attains the infimum. Let $(X, Y) \in \gamma_t$, so that

$$W_\psi(\mu K^t, \varpi) = E[\psi(X, Y)] \tag{1.3.10}$$

For $x, y \in \mathsf{X}$, let

$$q_{x,y}(z) = \min\{k(x, z), k(y, z)\},$$

and set $a_{x,y} = \int_{\mathsf{X}} q_{x,y}(z)\, \nu(\mathrm{d}z)$, so that $1 - a_{x,y}$ is the left-hand-side of (1.3.9). Given $(X, Y) = (x, y)$, generate (X', Y') in the following manner: With probability $a_{x,y}$, let $X' = Y'$ be distributed according to the probability density function $q_{x,y}(\cdot)/a_{x,y}$; with probability $1 - a_{x,y}$ (if $a_{x,y} < 1$), let X' be distributed according to the probability measure

$$A \mapsto \frac{K(x, A) - \int_A q_{x,y}(z)\, \nu(\mathrm{d}z)}{1 - a_{x,y}}$$

and, independently, let Y' be distributed according to the probability measure

$$A \mapsto \frac{K(y, A) - \int_A q_{x,y}(z)\, \nu(\mathrm{d}z)}{1 - a_{x,y}}.$$

It is easy to see that, given $(X, Y) = (x, y)$, X' is distributed as $K(x, \cdot)$ and Y' is distributed as $K(y, \cdot)$. Thus, marginally, X' is distributed as

$$\int_{\mathsf{X}^2} K(x, \cdot)\, \gamma_t(\mathrm{d}(x, y)) = \int_{\mathsf{X}} \mu K^t(\mathrm{d}x) K(x, \cdot) = \mu K^{t+1}(\cdot),$$

while Y' is distributed as $\varpi K(\cdot) = \varpi(\cdot)$. Hence, the joint distribution of (X', Y') is in $C(\mu K^{t+1}, \varpi)$.

By (1.3.3), (1.3.9), and (1.3.10),

$$\begin{aligned}
\|\mu K^{t+1} - \varpi\|_{\mathrm{TV}} &\leq P(X' \neq Y') \\
&= E[P(X' \neq Y' \mid X, Y)] \\
&\leq E(1 - a_{X,Y}) \\
&\leq bE[\psi(X, Y)] \\
&= bW_\psi(\mu K^t, \varpi).
\end{aligned}$$

□

Application to the Gaussian chain

Let us apply Theorem 1.3.2 and Theorem 1.3.3 to the Gaussian chain in Example 1.2.2, and use it as a stage for discussing some practical issues concerning the application of coupling methods.

Suppose that our goal is to bound $\|\mu K_{p,\alpha}^t - \varpi_p\|_{\mathrm{TV}}$ from above. To apply Theorem 1.3.2, one could let $D(x, y) = \mathbf{1}_{x \neq y}$ for $x, y \in \mathbb{R}^p$. But, in this case, it is impossible to establish a nontrivial contraction condition. Indeed, by (1.3.3), for any coupling kernel $\tilde{K}$ of $K_{p,\alpha}$ and

$x, y \in \mathbb{R}^p$,

$$\int_{\mathbb{R}^p \times \mathbb{R}^p} \mathbf{1}_{x' \neq y'} \, \tilde{K}((x,y), \mathrm{d}(x',y')) \geq \|\delta_x K_{p,\alpha} - \delta_y K_{p,\alpha}\|_{\mathrm{TV}}.$$

But, by (1.2.1), whenever $\alpha > 0$, the total variation distance between $\mathrm{N}_p(\alpha x, (1-\alpha^2)I_p)$ and $\mathrm{N}_p(\alpha y, (1-\alpha^2)I_p)$ goes to 1 as $\|x - y\| \to \infty$. Thus, the contraction condition

$$\int_{\mathbb{R}^p \times \mathbb{R}^p} \mathbf{1}_{x' \neq y'} \, \tilde{K}((x,y), \mathrm{d}(x',y')) \leq \rho \mathbf{1}_{x \neq y}, \quad x, y \in \mathbb{R}^p,$$

can only hold when $\rho \geq 1$. One could get somewhere if $D(x,y)$ is taken to be, say, $\mathbf{1}_{x \neq y} \, [p^{-1}\|x\|^2 + p^{-1}\|y\|^2 + 1]^{1-r}$ for some $r \in (0,1)$; see Section 1.3.3. However, it turns out that it is much easier to construct a sharp convergence bound by first considering the 1-Wasserstein distance induced by the Euclidean distance, and then utilize one-shot coupling.

For $x, y \in \mathbb{R}^p$, let $\psi(x,y) = \|x - y\|$. Let $\mathcal{F}_p'$ be the set of probability measures μ such that $\int_{\mathbb{R}^p} \|x\| \, \mu(\mathrm{d}x) < \infty$. We will bound $W_\psi(\mu K_{p,\alpha}^t, \varpi_p)$ from above for $\mu \in \mathcal{F}_p'$ using Theorem 1.3.2.

There are many possible ways of constructing coupling kernels. Here, we use a construction sometimes referred to as the "common random number coupling." Let N be an $\mathrm{N}_p(0, I_p)$ distributed random vector. For $(x,y) \in \mathbb{R}^p$, let $\tilde{K}_{p,\alpha}((x,y),\cdot)$ be the joint distribution of $\alpha x + \sqrt{1-\alpha^2}N$ and $\alpha y + \sqrt{1-\alpha^2}N$. Since $\alpha x' + \sqrt{1-\alpha^2}N \sim K_{p,\alpha}(x', \cdot)$ for $x' \in \mathbb{R}^p$, $\tilde{K}_{p,\alpha}$ is a coupling kernel of $K_{p,\alpha}$. For $(x,y) \in \mathbb{R}^p \times \mathbb{R}^p$,

$$\begin{aligned}
\int_{\mathbb{R}^p \times \mathbb{R}^p} \|x' - y'\| \, \tilde{K}_{p,\alpha}((x,y), \mathrm{d}(x',y')) &= E[\|(\alpha x + \sqrt{1-\alpha^2}N) - (\alpha y + \sqrt{1-\alpha^2}N)\|] \\
&= \alpha \|x - y\|.
\end{aligned}$$

Thus, the contraction condition (1.3.8) holds on $\mathbb{R}^p \times \mathbb{R}^p$ for $\tilde{K} = \tilde{K}_{p,\alpha}$ with $D(\cdot,\cdot) = \psi(\cdot,\cdot)$ and $\rho = \alpha$. By Theorem 1.3.2 and the triangle inequality, for $t \in \mathbb{N}$,

$$W_\psi(\mu K_{p,\alpha}^t, \varpi_p) \leq \left(\int_{\mathbb{R}^p} \|x\| \, \mu(\mathrm{d}x) + \int_{\mathbb{R}^p} \|y\| \, \varpi_p(\mathrm{d}y) \right) \alpha^t. \tag{1.3.11}$$

This bound reveals that $\mu K_{p,\alpha}^t$ approaches ϖ_p in the 1-Wasserstein distance at a geometric rate α.

If this were a practical problem, the stationary distribution ϖ_p would likely be intractable, and one would not be able to evaluate $\int_{\mathbb{R}^p} \|y\| \, \varpi_p(\mathrm{d}y)$. It is however possible to bound the integral from above via the following result for a generic Markov chain with Mtk $K(\cdot,\cdot)$.

Theorem 1.3.4. *(Hairer, 2006, Proposition 4.24) Let $h : \mathsf{X} \to [0,\infty)$ be a measurable function. Suppose that there exist $\lambda \in [0,1)$ and $L \in [0,\infty)$ such that, for $x \in \mathsf{X}$,*

$$\int_{\mathsf{X}} h(x') K(x, \mathrm{d}x') \leq \lambda h(x) + L.$$

(This is called a drift condition.) Then

$$\int_{\mathsf{X}} h(x) \, \varpi(\mathrm{d}x) \leq \frac{L}{1-\lambda}.$$

Proof. By the drift condition, for $x \in \mathsf{X}$ and $t \in \mathbb{N}_+$,

$$\int_{\mathsf{X}} h(x') K^t(x, \mathrm{d}x') \leq \lambda^t h(x) + \frac{L(1-\lambda^t)}{1-\lambda}.$$

For $n \in \mathbb{N}_+$, let $h_n : \mathsf{X} \to [0,\infty)$ be such that $h_n(x) = \min\{h(x), n\}$. Then

$$\int_{\mathsf{X}} h_n(x') K^t(x, \mathrm{d}x') \leq \min\left\{ \lambda^t h(x) + \frac{L(1-\lambda^t)}{1-\lambda}, n \right\}.$$

Integrating with respect to ϖ yields

$$\int_{\mathsf{X}} h_n(x) \, \varpi(\mathrm{d}x) = \int_{\mathsf{X}^2} h_n(x') K^t(x, \mathrm{d}x') \, \varpi(\mathrm{d}x) \leq \int_{\mathsf{X}} \min\left\{ \lambda^t h(x) + \frac{L(1-\lambda^t)}{1-\lambda}, n \right\} \varpi(\mathrm{d}x).$$

By the dominated convergence theorem, letting $t \to \infty$ shows that

$$\int_{\mathsf{X}} h_n(x) \, \varpi(\mathrm{d}x) \leq \min\left\{ \frac{L}{1-\lambda}, n \right\}.$$

By the monotone convergence theorem, letting $n \to \infty$ yields the desired result. □

In Example 1.2.2, we can pretend that we know little of ϖ_p, and bound $\int_{\mathbb{R}^p} \|y\| \, \varpi_p(\mathrm{d}y)$

in the following manner. Note that

$$\int_{\mathbb{R}^p} \|x'\|^2 \, K_{p,\alpha}(x, \mathrm{d}x') = \alpha^2 \|x\|^2 + (1-\alpha^2)p.$$

Thus, by Theorem 1.3.4,

$$\int_{\mathbb{R}^p} \|x\| \, \varpi_p(\mathrm{d}y) \leq \sqrt{\int_{\mathbb{R}^p} \|x\|^2 \, \varpi_p(\mathrm{d}y)} \leq \sqrt{\frac{(1-\alpha^2)p}{1-\alpha^2}} = \sqrt{p}. \tag{1.3.12}$$

Let us now apply Theorem 1.3.3 to transform the Wasserstein distance bound into a total variation bound. For $x \in \mathbb{R}^p$, let $k_{p,\alpha}(x, \cdot)$ be the density function of the $\mathrm{N}_p(\alpha x, (1-\alpha^2) I_p)$ distribution, i.e., $K_{p,\alpha}(x, \cdot)$, with respect to the Lebesgue measure. A bit of calculus reveals that, for $x, y \in \mathbb{R}^p$,

$$1 - \int_{\mathbb{R}^p} \min\{k_{p,\alpha}(x, x'), k_{p,\alpha}(y, x')\} \, \mathrm{d}x' = 1 - 2F_N\left(- \frac{\alpha \|x-y\|}{2\sqrt{1-\alpha^2}} \right) \leq \frac{\alpha}{\sqrt{(2\pi)(1-\alpha^2)}} \|x-y\|, \tag{1.3.13}$$

where $F_N(\cdot)$ is the cumulative distribution function of the one-dimensional standard normal distribution. By Theorem 1.3.3 along with (1.3.11) to (1.3.13), for $\mu \in \mathcal{F}_p'$ and $t \in \mathbb{N}$,

$$\|\mu K_{p,\alpha}^{t+1} - \varpi_p\|_{\mathrm{TV}} \leq \frac{\alpha}{\sqrt{(2\pi)(1-\alpha^2)}} \left(\int_{\mathbb{R}^p} \|x\| \, \mu(\mathrm{d}x) + \sqrt{p} \right) \alpha^t. \tag{1.3.14}$$

This bound indicates that $\mu K_{p,\alpha}^t$ approaches ϖ_p in the total variation distance at a geometric rate α, which does not depend on the dimension p. By considering the L^2 convergence rate of the chain (see Section 1.4.1) and utilizing Theorem 2.1 of Roberts and Rosenthal (1997), it is possible to show that the convergence rate indicated by (1.3.14) is in fact sharp. That is, it is not possible to find a bound of the form $\|\mu K_{p,\alpha}^t - \varpi_p\|_{\mathrm{TV}} \leq C_{p,\alpha,\mu} \rho_{p,\alpha}^t$, where $C_{p,\alpha,\mu} < \infty$ and $\rho_{p,\alpha} < \alpha$, that holds for all $\mu \in \mathcal{F}_p'$ and $t \in \mathbb{N}_+$. This is an example of the convergence bound scaling well with the dimension in the sense that the convergence rate indicated by the bound does not deteriorate (go to 1) faster than the true convergence rate does as the dimension grows. Good scaling is not always easily achieved, as obtaining sharp convergence bounds often becomes more difficult in problems with higher dimensions. In

Section 1.3.3, we give a type of bound that is tremendously popular and powerful in certain settings but scales poorly with the dimension when applied to Example 1.2.2.

We may also use (1.3.14) to obtain a bound on the mixing time. For $\mu \in \mathcal{F}_p'$ and $\epsilon > 0$, let $t_{\mathrm{TV}}(\epsilon, \mu)$ be the smallest $t \in \mathbb{N}_+$ such that $\|\mu K^t - \varpi\|_{\mathrm{TV}} \leq \epsilon$. Then (1.3.14) implies that

$$t_{\mathrm{TV}}(\epsilon, \mu) \leq \left\lceil (-\log\alpha)^{-1} \left\{ \log\left[\frac{\alpha}{\sqrt{(2\pi)(1-\alpha^2)}} \left(\int_{\mathbb{R}^p} \|x\| \, \mu(\mathrm{d}x) + \sqrt{p} \right) \right] - \log\epsilon \right\} \right\rceil$$

if $\alpha \in (0,1)$. If $\int_{\mathbb{R}^p} \|x\| \, \mu(\mathrm{d}x) = O(p)$ as $p \to \infty$, then $t_{\mathrm{TV}}(\epsilon, \mu) = O(\log p)$.

1.3.3 Drift and minorization

As a final application of the coupling method, we use it to derive a convergence bound based on a set of drift and minorization conditions. Many important convergence bounds in the literature are constructed based on this type of condition (Mengersen and Tweedie, 1996; Meyn and Tweedie, 2012; Roberts and Tweedie, 1999; Rosenthal, 1995; Tierney, 1994). While these bounds are often far from sharp (Qin and Hobert, 2020), they are very powerful for establishing qualitative results like geometric ergodicity (Hobert and Geyer, 1998; Jarner and Hansen, 2000; Jones and Hobert, 2001, 2004; Khare and Hobert, 2013; Livingstone et al., 2019; Roy and Hobert, 2007).

Recall the drift condition in Theorem 1.3.4. We say that the Mtk K satisfies a drift condition with drift function $h : \mathsf{X} \to [0, \infty)$ if there exist $\lambda \in [0, 1)$ and $L \in [0, \infty)$ such that, for $x \in \mathsf{X}$,

$$Kh(x) := \int_{\mathsf{X}} h(x') K(x, \mathrm{d}x') \leq \lambda h(x) + L.$$

We say that K satisfies a minorization condition associated with h if there exist $\varepsilon > 0$, a probability measure $\nu \in \mathcal{P}(\mathsf{X})$, and $\Delta > 2L/(1-\lambda)$ such that

$$K(x, A) \geq \varepsilon \nu(A)$$

for $A \in \mathcal{B}$ whenever $h(x) \leq \Delta$.

The following convergence bound is reminiscent of a famous result from Rosenthal (1995). Its proof uses ideas from Butkovsky (2014); Douc et al. (2018); Hairer and Mattingly (2011); Hairer et al. (2011).

Theorem 1.3.5. *Suppose that the above drift and minorization conditions hold. Then, for* $\mu \in \mathcal{P}(\mathsf{X})$*,* $r \in (0,1)$*, and* $t \in \mathbb{N}$*,*

$$\|\mu K^t - \varpi\|_{TV} \leq \left(\mu h + \frac{L}{1-\lambda} + 1 \right) \rho^t.$$

where

$$\rho = \max \left\{ (1-\varepsilon)^r (2L+1)^{1-r}, \left(\lambda + \frac{2L+1-\lambda}{\Delta+1} \right)^{1-r} \right\}. \tag{1.3.15}$$

Remark 1.3.2. *Note that* $\Delta > 2L/(1-\lambda)$ *implies that* $\lambda + (2L+1-\lambda)/(\Delta+1) < 1$*. Thus, there exists* $r \in (0,1)$ *such that* ρ*, as given in* (1.3.15)*, is strictly less than 1.*

Proof of Theorem 1.3.5. We will make use of Theorem 1.3.2. Take $\mathcal{F}$ to be the set of functions f such that $\sup_{x \in \mathsf{X}} |f(x)| = 1/2$ and $\mathcal{F}' = \mathcal{P}(\mathsf{X})$, so that $\| \cdot - \cdot \|_{\mathcal{F}} = \| \cdot - \cdot \|_{\mathrm{TV}}$.

Fix $r \in (0,1)$. For $(x,y) \in \mathsf{X}^2$, let

$$D(x,y) = \mathbf{1}_{x \neq y} [h(x) + h(y) + 1]^{1-r}.$$

Then $D(x,y) \geq \mathbf{1}_{x \neq y} \geq |f(x) - f(y)|$ for $f \in \mathcal{F}$, i.e., (1.3.1) holds for $f \in \mathcal{F}$.

Let us now construct a coupling kernel that satisfies a contraction condition. The construction is somewhat similar to the one in the proof of Theorem 1.3.1. Let $S = \{x \in \mathsf{X} : h(x) \leq \Delta\}$. Let $(X_t, Y_t)_{t=0}^{\infty}$ be a bivariate Markov chain that evolves as follows. Suppose that the current state is $(X_t, Y_t) = (x, y)$. When $x = y$, simply generate $X_{t+1} = Y_{t+1}$ according to the probability measure $K(x, \cdot)$. When $x \neq y$, do the following. If $(x,y) \notin S^2$, then generate $X_{t+1} \sim K(x, \cdot)$ and $Y_{t+1} \sim K(y, \cdot)$ independently. If $(x,y) \in S^2$, then, with probability ε, let $X_{t+1} = Y_{t+1}$ be distributed as ν; with probability $1-\varepsilon$, let X_{t+1} be distributed according to the probability measure (1.3.5), and, independently, let Y_{t+1} be distributed according to the probability measure (1.3.6). It is straightforward to check that the Mtk of this chain, denoted by $\tilde{K}$, is a coupling kernel of K.

Now, for $(x,y) \in \mathsf{X}^2$, by Höder's inequality,

$$\begin{aligned}
&\int_{\mathsf{X}^2} D(x',y')\,\tilde{K}((x,y),\mathrm{d}(x',y')) \\
&= \int_{\mathsf{X}^2} \mathbf{1}^r_{x'\neq y'}\,[h(x')+h(y')+1]^{1-r}\,\tilde{K}((x,y),\mathrm{d}(x',y')) \\
&\le \left[\int_{\mathsf{X}^2} \mathbf{1}_{x'\neq y'}\,\tilde{K}((x,y),\mathrm{d}(x',y'))\right]^r \left\{\int_{\mathsf{X}^2}[h(x')+h(y')+1]\,\tilde{K}((x,y),\mathrm{d}(x',y'))\right\}^{1-r} \\
&= \left[\int_{\mathsf{X}^2} \mathbf{1}_{x'\neq y'}\,\tilde{K}((x,y),\mathrm{d}(x',y'))\right]^r [Kh(x)+Kh(y)+1]^{1-r}.
\end{aligned}$$

By construction,

$$\int_{\mathsf{X}^2} \mathbf{1}_{x'\neq y'}\,\tilde{K}((x,y),\mathrm{d}(x',y')) \le \begin{cases} (1-\varepsilon)\mathbf{1}_{x\neq y}, & (x,y)\in S^2, \\ \mathbf{1}_{x\neq y}, & \text{otherwise.} \end{cases}$$

By the drift condition,

$$\begin{aligned}
Kh(x)+Kh(y)+1 &\le \lambda h(x)+\lambda h(y)+2L+1 \\
&= \left[\lambda + \frac{2L+1-\lambda}{h(x)+h(y)+1}\right][h(x)+h(y)+1] \\
&\le \begin{cases} (2L+1)[h(x)+h(y)+1], & (x,y)\in S^2, \\ \left(\lambda+\frac{2L+1-\lambda}{\Delta+1}\right)[h(x)+h(y)+1], & \text{otherwise.} \end{cases}
\end{aligned}$$

Combining terms, we find that

$$\int_{\mathsf{X}^2} D(x',y')\,\tilde{K}((x,y),\mathrm{d}(x',y')) \le \rho D(x,y),$$

where ρ is given in (1.3.15).

Let γ be a coupling of μ and ϖ. By Theorem 1.3.2, for $\mu \in \mathcal{P}(\mathsf{X})$ and $t \in \mathbb{N}$,

$$\begin{aligned}\|\mu K^t - \varpi\|_{\mathrm{TV}} &\leq \int_{\mathsf{X}^2} D(x,y)\,\gamma(\mathrm{d}(x,y))\rho^t \\ &\leq \int_{\mathsf{X}^2} [h(x)+h(y)+1]\,\gamma(\mathrm{d}(x,y))\rho^t \\ &= (\mu h + \varpi h + 1)\rho^t.\end{aligned}$$

Applying Theorem 1.3.4 to bound ϖh gives us the desired result. □

For alternative derivations of drift and minorization-based convergence bounds that rely less on the coupling method, see Baxendale (2005); Jerison (2019); Meyn and Tweedie (1994).

The drift and minorization condition presented here is just one among several commonly used forms. For some other useful versions of drift and minorization, including those used for establishing subgeometric convergence, see Andrieu et al. (2015); Butkovsky (2014); Douc et al. (2004, 2008); Durmus et al. (2016); Jarner and Roberts (2002); Zhou et al. (2022).

Application to the Gaussian chain

We now apply Theorem 1.3.5 to the Gaussian chain in Example 1.2.2. A drift function we can use is $h(x) = p^{-1}\|x\|^2$, $x \in \mathbb{R}^p$. Then

$$K_{p,\alpha}h(x) = \alpha^2 h(x) + 1 - \alpha^2.$$

So a drift condition holds with $\lambda = \alpha^2$ and $L = 1-\alpha^2$. Let $\Delta > 2L/(1-\lambda) = 2$, and let $S = \{x \in \mathbb{R}^p : h(x) \leq \Delta\}$. Recall that $k_{p,\alpha}(x,\cdot)$ is the density of $K_{p,\alpha}(x,\cdot)$ for $x \in \mathsf{X}$. For $x' \in \mathbb{R}^p$, let

$$q(x') = \inf_{x\in S} k_{p,\alpha}(x,x') = \frac{1}{[2\pi(1-\alpha^2)]^{p/2}} \exp\left[-\frac{(\|x'\| + \alpha\sqrt{p\Delta})^2}{2(1-\alpha^2)}\right].$$

Then, for $x \in S$ and $x' \in \mathbb{R}^p$,

$$k_{p,\alpha}(x, x') \geq \varepsilon \frac{q(x')}{\int_{\mathbb{R}^p} q(x'') \, \mathrm{d}x''},$$

where

$$\varepsilon = \int_{\mathbb{R}^p} q(x'') \, \mathrm{d}x'' = \int_{\mathbb{R}^p} \frac{1}{[2\pi(1-\alpha^2)]^{p/2}} \exp\left[-\frac{(\|x''\| + \alpha\sqrt{p\Delta})^2}{2(1-\alpha^2)} \right] \mathrm{d}x''.$$

Thus, a minorization condition holds with this value of ε. Applying Theorem 1.3.5 shows that, for $\mu \in \mathcal{P}(\mathbb{R}^p)$ and $t \in \mathbb{N}$,

$$\|\mu K_{p,\alpha}^t - \varpi_p\|_{\mathrm{TV}} \leq (\mu h + 2) \, \rho^t,$$

where, for some $r \in (0,1)$,

$$\rho = \max\left\{ (1-\varepsilon)^r (3 - 2\alpha^2)^{1-r}, \left(\alpha^2 + \frac{3(1-\alpha^2)}{\Delta+1} \right)^{1-r} \right\}. \qquad (1.3.16)$$

For concreteness, take $p = 10$ and $\alpha = 1/2$. Then

$$\begin{aligned} \varepsilon &= \int_{\mathbb{R}^{10}} \left(\frac{2}{3\pi} \right)^5 \exp\left[-\frac{2(\|x''\| + \sqrt{2.5\Delta})^2}{3} \right] \mathrm{d}x'' \\ &= \int_0^\infty \frac{2^6 u^9}{3^5 \times 4!} \exp\left[-\frac{2(u + \sqrt{2.5\Delta})^2}{3} \right] \mathrm{d}u. \end{aligned}$$

For instance, if $\Delta = 4$, then $\varepsilon \approx 2.28 \times 10^{-7}$. In (1.3.16), we can optimize the value of r to find the smallest value of ρ, which yields $\rho \approx 1 - 6 \times 10^{-8}$. Other choices of $\Delta \in (2, \infty)$ would result in values of ρ that are very close to unity as well. When t is large, the resulting upper bound on $\|\mu K_{p,\alpha}^t - \varpi_p\|_{\mathrm{TV}}$, which is proportional to ρ^t, is extremely conservative. Indeed, recall that the much sharper bound from (1.3.14) is proportional to $\alpha^t = 0.5^t$. Through experiments one can find that the conservativeness of (1.3.16) is exacerbated when the dimension p is increased. This is consistent with empirical evidence and theoretical analyses in the existing literature, which suggest that bounds based on drift and minorization conditions typically scale poorly with dimensions (Qin and Hobert, 2020). Oftentimes, this type of bound is more suitable for establishing qualitative results such as geometric ergodicity.

1.4 L^2 theory

The L^2 theory for Markov chains is a framework for studying the convergence properties of a Markov chain, usually reversible, in terms of the L^2 distance by examining the linear operator associated with the chain's transition kernel. A substantial body of literature works within this theoretical framework, offering a diverse array of analytical techniques (Amit, 1996; Andrieu et al., 2022; Diaconis et al., 2000, 2008; Dwivedi et al., 2019; Hobert and Marchev, 2008; Khare and Hobert, 2011; Liu et al., 1994; Roberts and Rosenthal, 1997; Roberts and Sahu, 1997, to name some). We will first review some basic concepts. Then, we explain how isoperimetric inequalities, a type of inequality that regulates the geometric features of the target distribution, can be leveraged to analyze Markov chains within this framework. Finally, we review some simple techniques for showing how slow a chain converges.

1.4.1 Basic theory

Throughout Section 1.4, let $\mathcal{F}$ be the set of functions f such that $\int_{\mathsf{X}} f(x)^2 \, \varpi(\mathrm{d}x) = 1$, and let $\mathcal{F}'$ be the set of probability measures μ such that $\mathrm{d}\mu/\mathrm{d}\varpi$ is squared integrable with respect to ϖ. Then $\|\mu - \nu\|_{\mathcal{F}}$ is the L^2 distance $\|\mu - \nu\|_2$ for $\mu, \nu \in \mathcal{F}'$.

The L^2 theory for Markov chains begins with the examination of a linear space formed by some functions on X. Denote by $L^2(\varpi)$ the set of real measurable functions f on X such that $\int_{\mathsf{X}} f(x)^2 \, \varpi(\mathrm{d}x) < \infty$, with the understanding that two functions are equal if their difference is ϖ-almost everywhere vanishing. For $c \in \mathbb{R}$ and $f, g \in L^2(\varpi)$, let $(-f)(x) = -f(x)$, $(cf)(x) = cf(x)$, $(f+g)(x) = f(x) + g(x)$, and $(f-g)(x) = f(x) - g(x)$ for $x \in \mathsf{X}$. Then $L^2(\varpi)$ forms a real linear space. For $f, g \in L^2(\varpi)$, define their inner product as

$$\langle f, g \rangle = \int_{\mathsf{X}} f(x) g(x) \, \varpi(\mathrm{d}x),$$

and let $\|f\|_2 = \langle f, f \rangle^{1/2}$. Then $\|\cdot\|_2$ is a norm, and shall be referred to as the $L^2(\varpi)$ norm. It can be shown that $L^2(\varpi)$ is a Hilbert space (see, e.g., Bruckner et al., 2008, Theorem 13.15).

The space $L^2(\varpi)$ provides a natural stage for studying the L^2 distance between distributions. Indeed, a distribution μ is in $\mathcal{F}'$ if and only if the function $\mathrm{d}\mu/\mathrm{d}\varpi$ exists and is in $L^2(\varpi)$. Moreover, using the Cauchy-Schwarz inequality, one can derive (1.2.2), which states that, for $\mu, \nu \in \mathcal{F}'$,

$$\|\mu - \nu\|_2 = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\varpi} - \frac{\mathrm{d}\nu}{\mathrm{d}\varpi} \right\|_2 .$$

It will be convenient to work with the subspace $L_0^2(\varpi)$, which consists of functions $f \in L^2(\varpi)$ such that $\varpi f = 0$.

Remark 1.4.1. *If a random element X is distributed as ϖ, and f and g are in $L_0^2(\varpi)$, then $E[f(X)] = E[g(X)] = 0$, $\langle f, g \rangle = cov[f(X), g(X)]$, and $\|f\|_2^2 = var[f(X)]$.*

The Mtk $K(\cdot, \cdot)$, which satisfies $\varpi K(\cdot) = \varpi(\cdot)$, can be regarded as a bounded linear operator (called a Markov operator) on $L_0^2(\varpi)$ in the following way. For $f \in L_0^2(\varpi)$, let

$$Kf(x) = \int_{\mathsf{X}} f(x') K(x, \mathrm{d}x'), \quad x \in \mathsf{X}.$$

The map $f \mapsto Kf$ is clearly linear. To see that its range is in $L_0^2(\varpi)$, note that when $f \in L_0^2(\varpi)$,

$$\int_{\mathsf{X}} Kf(x)\, \varpi(\mathrm{d}x) = \int_{\mathsf{X}} \int_{\mathsf{X}} f(x') K(x, \mathrm{d}x')\, \varpi(\mathrm{d}x) = \int_{\mathsf{X}} f(x)\, \varpi(\mathrm{d}x) = 0,$$

and, by Jensen's inequality,

$$\int_{\mathsf{X}} \left[\int_{\mathsf{X}} K(x, \mathrm{d}x') f(x') \right]^2 \varpi(\mathrm{d}x) \leq \int_{\mathsf{X}} \int_{\mathsf{X}} f(x')^2 K(x, \mathrm{d}x')\, \varpi(\mathrm{d}x) = \int_{\mathsf{X}} f(x')^2\, \varpi(\mathrm{d}x') < \infty. \tag{1.4.1}$$

The operator norm of K is

$$\|K\|_2 = \sup_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\|Kf\|_2}{\|f\|_2} = \sup_{f \in L_0^2(\varpi),\, \|f\|_2 = 1} \|Kf\|_2 .$$

By (1.4.1), $\|K\|_2 \le 1$.

Remark 1.4.2. *Let $(X_t)_{t=0}^{\infty}$ be a chain associated with K. Then, for $f \in L_0^2(\varpi)$, $Kf(X_0) = E[f(X_1) \mid X_0]$. If, furthermore, $X_0 \sim \varpi$, then, for $f, g \in L_0^2(\varpi)$,*

$$\|Kf\|_2^2 = var\{E[f(X_1) \mid X_0]\}, \quad \langle f, Kg\rangle = cov(f(X_0), g(X_1)).$$

For $t \in \mathbb{N}_+$, the t-step Mtk $K^t(\cdot,\cdot)$ also defines an operator K_t on $L_0^2(\varpi)$, with $K_1 f = Kf$, and

$$K_t f(x) = \int_{\mathsf{X}} f(x')K^t(x, \mathrm{d}x') = \int_{\mathsf{X}} Kf(x')K_{t-1}(x, \mathrm{d}x')$$

when $t \ge 2$. Note that $K_t f$ is K applied to f t times. In other words, K_t is just K^t, the product (composition) of t K's. By convention, K^0 is the identity operator.

The convergence behavior of a Markov chain associated with the Mtk $K(\cdot,\cdot)$ is tied to the properties of the operator K. Indeed, the following theorem is well-known, and can be found in, e.g., Roberts and Rosenthal (1997).

Theorem 1.4.1. *For $\mu \in \mathcal{F}'$ and $t \in \mathbb{N}$,*

$$\|\mu K^t - \varpi\|_2 \le \|\mu - \varpi\|_2 \|K^t\|_2 \le \|\mu - \varpi\|_2 \|K\|_2^t. \tag{1.4.2}$$

Proof. The second inequality follows from the sub-multiplicity of operator norms. We will focus on establishing the first inequality. Note that if $f \in \mathcal{F}$ so that $\|f\|_2 = 1$, then $f - \varpi f \in L_0^2(\varpi)$, and $\|f - \varpi f\|_2^2 = \|f\|_2^2 - (\varpi f)^2 \le 1$. (For a constant c, $f - c$ means the function satisfying $(f-c)(x) = f(x) - c$.) Then

$$\begin{aligned}
\|\mu K^t - \varpi\|_2 &= \sup_{f\in\mathcal{F}} |(\mu K^t)f - (\varpi K^t)f| \\
&= \sup_{f\in\mathcal{F}} |(\mu K^t)(f - \varpi f) - (\varpi K^t)(f - \varpi f)| \\
&= \sup_{f\in\mathcal{F}} \left| \int_{\mathsf{X}^2} [\mu(\mathrm{d}x) - \varpi(\mathrm{d}x)] K^t(x, \mathrm{d}x')[f(x') - \varpi f] \right| \\
&= \sup_{f\in\mathcal{F}} \left\langle \frac{\mathrm{d}\mu}{\mathrm{d}\varpi} - 1, K^t(f - \varpi f) \right\rangle \\
&\leq \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\varpi} - 1 \right\|_2 \|K^t\|_2 \sup_{f\in\mathcal{F}} \|f - \varpi f\|_2 \\
&\leq \|\mu - \varpi\|_2 \|K^t\|_2.
\end{aligned}$$

Note that we have used the Cauchy-Schwarz inequality and the definition of the operator norm. □

Theorem 1.4.1 implies that, if $\|K\|_2 \leq \rho$ for some $\rho < 1$, then asymptotically $\|\mu K^t - \varpi\|_2$ decreases with t at a geometric rate of ρ^t or faster. In fact, if the Mtk $K(\cdot,\cdot)$ is reversible with respect to $\varpi(\cdot)$, then the converse is true as well. Note that $K(\cdot,\cdot)$ is reversible if and only if

$$\int_{\mathsf{X}^2} f(x) g(x') K(x, \mathrm{d}x') \, \varpi(\mathrm{d}x) = \int_{\mathsf{X}^2} g(x) f(x') K(x, \mathrm{d}x') \, \varpi(\mathrm{d}x)$$

for $f, g \in L_0^2(\varpi)$, i.e., $\langle f, Kg \rangle = \langle Kf, g \rangle$. In other words, the Mtk $K(\cdot,\cdot)$ is reversible if and only if the operator K is self-adjoint. Using this fact we can establish the following result.

Theorem 1.4.2. *(Roberts and Rosenthal, 1997, Theorem 2.1) Suppose that $K(\cdot,\cdot)$ is reversible. Suppose further that there exist $C : \mathcal{F}' \to [0,\infty)$ and $\rho < 1$ such that, for $\mu \in \mathcal{F}'$ and $t \in \mathbb{N}$,*

$$\|\mu K^t - \varpi\|_2 \leq C(\mu)\rho^t. \tag{1.4.3}$$

Then $\|K\|_2 \leq \rho$.

Proof. By assumption, K is a self-adjoint operator. Then K has a spectral decomposition:

for $f, g \in L_0^2(\varpi)$ and $t \in \mathbb{N}$,

$$\langle K^t f, g\rangle = \int_{-\infty}^{\infty} \lambda^t \langle E_K(\mathrm{d}\lambda) f, g\rangle,$$

where $E_K(\cdot)$ is the spectral measure of K, which is a projection-valued measure that is supported on the spectrum of K. See, e.g., Conway (1990), §IX.2; Arveson (2006), Section 2.7. An important property of $E_K(\cdot)$ is that, for a measurable subset B of $\mathbb{R}$, if a non-vanishing function $f \in L_0^2(\varpi)$ is in the range of the orthogonal projection operator $E_K(B)$, then $\|f\|_2^{-2}\langle f, E_K(\cdot) f\rangle$ is a probability measure concentrated on B.

Suppose that $\|K\|_2 > \rho$ so that $\|K\|_2 \geq \rho + \varepsilon$ for some $\varepsilon > 0$. Then there exists a non-vanishing function $f \in L_0^2(\varpi)$ in the range of the projection operator $E_K([\rho+\varepsilon, \infty) \cup (-\infty, -\rho-\varepsilon])$. Let $f_+(x) = \max\{f(x), 0\}$ and $f_-(x) = \max\{-f(x), 0\}$ for $x \in \mathsf{X}$, so that $f = f_+ - f_-$. Let

$$\mu_+(A) = \frac{2\int_A f_+(x)\, \varpi(\mathrm{d}x)}{\|f\|_1}, \quad \mu_-(A) = \frac{2\int_A f_-(x)\, \varpi(\mathrm{d}x)}{\|f\|_1}$$

for $A \in \mathcal{B}$, where

$$\|f\|_1 = \int_{\mathsf{X}} |f(x)|\, \varpi(\mathrm{d}x) = 2\int_{\mathsf{X}} f_+(x)\, \varpi(\mathrm{d}x) = 2\int_{\mathsf{X}} f_-(x)\, \varpi(\mathrm{d}x).$$

Then μ_+ and μ_- are in $\mathcal{F}'$. Moreover, $f/\|f\|_2$ is in $\mathcal{F}$. By the spectral decomposition, for $t \in \mathbb{N}$,

$$\begin{aligned}
\|\mu_+ K^{2t} - \mu_- K^{2t}\|_2 &\geq \int_{\mathsf{X}^2} [\mu_+(\mathrm{d}x) - \mu_-(\mathrm{d}x)] K^{2t}(x, \mathrm{d}x') \frac{f(x')}{\|f\|_2} \\
&= 2\left\langle \frac{f_+ - f_-}{\|f\|_1}, K^{2t}\frac{f}{\|f\|_2}\right\rangle \\
&= \frac{2}{\|f\|_1\|f\|_2}\int_{-\infty}^{\infty} \lambda^{2t}\langle f, E_K(\mathrm{d}\lambda) f\rangle \\
&\geq \frac{2\|f\|_2}{\|f\|_1}(\rho+\varepsilon)^{2t}.
\end{aligned}$$

The last line holds because the probability measure $\|f\|_2^{-2}\langle f, E_K(\cdot) f\rangle$ is concentrated on

$[\rho+\varepsilon,\infty)\cup(-\infty,-\rho-\varepsilon]$. The L^2 distance satisfies the triangle inequality, so

$$\|\mu_+K^t-\varpi\|_2+\|\mu_-K^t-\varpi\|_2\geq\|\mu_+K^t-\mu_-K^t\|_2\geq\frac{2\|f\|_2}{\|f\|_1}(\rho+\varepsilon)^t$$

for $t\in\mathbb{N}$. Then it is impossible for (1.4.3) to hold for all $\mu\in\mathcal{F}'$ and $t\in\mathbb{N}$. Thus, it must hold that $\|K\|_2\leq\rho$. □

Theorems 1.4.1 and 1.4.2 show that, if $K(\cdot,\cdot)$ is reversible, then $\|K\|_2$, which lies in $[0,1]$, can be regarded the convergence rate of the corresponding chain. The smaller $\|K\|_2$ is, the faster the chain converges.

In some simple scenarios, it is possible to calculate $\|K\|_2$ directly using functional analytic techniques. For instance, in Example 1.2.2, it can be shown, using orthogonal polynomials, that $\|K_{p,\alpha}\|_2=\alpha$ (see, e.g., Diaconis et al., 2008). One can compare this rate with that in Section 1.3 involving the total variation distance. In more complex scenarios, one hopes to derive some reasonably sharp bounds on $\|K\|_2$. Of course, to apply Theorem 1.4.1 in practice, one would also need to get a handle on $\|\mu-\varpi\|_2$. But since $\|K\|_2$ is the more important quantity in (1.4.2) when t is large, it will be our focus herein.

We end Section 1.4.1 with an elementary method for bounding $\|K\|_2$ from above which can be applied to Example 1.2.1. More sophisticated methods will be given later.

Theorem 1.4.3. *Suppose that there exists $\varepsilon>0$ such that $K(x,A)\geq\varepsilon\varpi(A)$ for $x\in\mathsf{X}$ and $A\in\mathcal{B}$. Then $\|K\|_2\leq1-\varepsilon$.*

Proof. It is clear that $\varepsilon\leq1$. If $\varepsilon=1$ then $Kf=\varpi f=0$ for $f\in L_0^2(\varpi)$, indicating that $\|K\|_2=0$. Suppose that $\varepsilon<1$. Let $R(x,A)=(1-\varepsilon)^{-1}[K(x,A)-\varepsilon\varpi(A)]$ for $x\in\mathsf{X}$ and $A\in\mathcal{B}$. Then $R(\cdot,\cdot)$ is an Mtk such that $\varpi R=\varpi$, and we may view R as a Markov operator on $L_0^2(\varpi)$. Replacing $K(\cdot,\cdot)$ with $R(\cdot,\cdot)$ in (1.4.1), we find that $\|R\|_2\leq1$. Thus, for $f\in L_0^2(\varpi)$,

$$\|Kf\|_2=\|\varepsilon\varpi f+(1-\varepsilon)Rf\|_2=(1-\varepsilon)\|Rf\|_2\leq(1-\varepsilon)\|f\|_2.$$

This implies that $\|K\|_2 \leq 1 - \varepsilon$. □

Let us apply Theorem 1.4.3 to the independent Metropolis Hastings chain in Example 1.2.1. Recall from (1.3.7) that, for $x \in \mathsf{X} = [0,1]$ and $A \in \mathcal{B}$, $K_s(x, A) \geq (1/M_s)\pi_s(A)$. Hence, by Theorem 1.4.3, $\|K_s\|_2 \leq 1 - 1/M_s$. In Section 1.4.4, it will be shown that this bound is tight.

1.4.2 The spectral gap and the conductance

Let $K(\cdot, \cdot)$ be reversible, so that the corresponding operator is self-adjoint. Consider the task of bounding $\|K\|_2$, particularly from above. Self-adjoint-ness implies that

$$\|K\|_2 = \max\left\{ \sup_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\langle f, Kf \rangle}{\|f\|_2^2}, - \inf_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\langle f, Kf \rangle}{\|f\|_2^2} \right\} \tag{1.4.4}$$

(see, e.g., Helmberg, 2014, §14, Corollary 5.1). In certain cases, it is possible to bound the second term in the maximum quite easily. For instance, if K is associated with a random-scan Gibbs algorithm or a data augmentation algorithm, then K is positive semi-definite in the sense that it is self-adjoint, and $\langle f, Kf \rangle \geq 0$ for $f \in L_0^2(\varpi)$, so the second term is at most zero. See Theorem 3 of Liu et al. (1995) and Theorem 3.2 of Liu et al. (1994). This is also the case if $K(\cdot, \cdot)$ is the two-step Mtk of some reversible chain, i.e., $K = T^2$ for some Mtk $T(\cdot, \cdot)$ that is reversible with respect to $\varpi(\cdot)$. Finally, if the chain is lazy in the sense that $K(x, \{x\}) \geq c$ for every x and some positive constant c, then the second term is at most $1 - 2c$. For various MCMC algorithms, much effort has been spent on bounding the first term from above, or equivalently, bounding the spectral gap, defined as

$$G(K) = 1 - \sup_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\langle f, Kf \rangle}{\|f\|_2^2},$$

from below.

It is worth mentioning that the spectral gap is also closely related to the asymptotic variance of Monte Carlo estimators. To be more precise, let $(X_t)_{t=0}^{\infty}$ be a chain associated with K, and let $f \in L^2(\varpi)$. Then, under regularity conditions, $n^{1/2}[n^{-1}\sum_{i=1}^n f(X_i) - \varpi f]$ is asymptotically normally distributed as $n \to \infty$, and the asymptotic variance is upper bounded by $[2 - G(K)]/G(K)$. See, e.g., Chan and Geyer (1994), equation (7).

One important approach for bounding the spectral gap is relating it to a quantity called the "conductance" (Jerrum and Sinclair, 1988). The conductance of K is defined to be

$$\Phi_K = \inf_{A \in \mathcal{B},\, 0 < \varpi(A) < 1} \phi_K(A), \text{ where } \phi_K(A) = \frac{\int_A \varpi(\mathrm{d}x) K(x, A^c)}{\varpi(A)\varpi(A^c)}.$$

Loosely speaking, $\phi_K(A)$ measures the probability flow from A to its complement A^c, after adjusting for the probability masses of A and A^c. A large conductance indicates that the chain can freely move around the state space X, and vice versa.

The conductance is related to the spectral gap through the following remarkable result, called Cheeger's inequality.

Theorem 1.4.4. *(Lawler and Sokal, 1988, Theorem 2.1) For the reversible Mtk $K(\cdot,\cdot)$, $\Phi_K^2/8 \leq G(K) \leq \Phi_K$.*

Proof. For $A \in \mathcal{B}$ and $x \in \mathsf{X}$, let $\eta_A(x) = \mathbf{1}_{x \in A} - \varpi(A)$. Then $\eta_A \in L_0^2(\varpi)$, and it is straightforward to check that

$$\phi_K(A) = 1 - \frac{\langle \eta_A, K\eta_A \rangle}{\|\eta_A\|_2^2}, \quad A \in \mathcal{B}.$$

Then $G(K) \leq \phi_K(A)$ for any A and thus, $G(K) \leq \Phi_K$.

We now establish the other inequality. Using the fact that $\varpi K = \varpi$, one can obtain

$$\begin{aligned} G(K) &= \inf_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\int_{\mathsf{X}} f(x)^2 \,\varpi(\mathrm{d}x) - \int_{\mathsf{X}^2} f(x)f(x')K(x, \mathrm{d}x')\,\varpi(\mathrm{d}x)}{\|f\|_2^2} \\ &= \frac{1}{2} \inf_{f \in L_0^2(\varpi),\, f \neq 0} \frac{\int_{\mathsf{X}^2} [f(x) - f(x')]^2 K(x, \mathrm{d}x')\,\varpi(\mathrm{d}x)}{\|f\|_2^2} \end{aligned} \tag{1.4.5}$$

For now, fix $f \in L_0^2(\varpi)$ such that $f \neq 0$ and $s \in \mathbb{R}$. Let $f_s(x) = f(x) - s$ for $x \in \mathsf{X}$. By the

Cauchy-Schwarz inequality,

$$\begin{aligned}\int_{\mathsf{X}^2}[f(x)-f(x')]^2K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x) &= \int_{\mathsf{X}^2}[f_s(x)-f_s(x')]^2K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\\ &\geq \frac{\left\{\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\right\}^2}{\int_{\mathsf{X}^2}[f_s(x)+f_s(x')]^2K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)}\\ &\geq \frac{\left\{\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\right\}^2}{2\int_{\mathsf{X}^2}[f_s(x)^2+f_s(x')^2]K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)}\\ &= \frac{\left\{\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\right\}^2}{4\|f_s\|_2^2}\end{aligned} \tag{1.4.6}$$

Let $A_{s,f}(t)=\{x:\ f_s(x)^2\geq t\}$ for $t\in[0,\infty)$. Then

$$\begin{aligned}&\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\\ =&\int_{\mathsf{X}^2}\mathbf{1}_{f_s(x')^2<f_s(x)^2}\int_{f_s(x')^2}^{f_s(x)^2}\mathrm{d}t\,K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)+\int_{\mathsf{X}^2}\mathbf{1}_{f_s(x)^2<f_s(x')^2}\int_{f_s(x)^2}^{f_s(x')^2}\mathrm{d}t\,K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\\ =&\int_0^\infty\int_{\mathsf{X}^2}\mathbf{1}_{f_s(x')^2<t\leq f_s(x)^2}\,K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\,\mathrm{d}t+\int_0^\infty\int_{\mathsf{X}^2}\mathbf{1}_{f_s(x)^2<t\leq f_s(x')^2}\,K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\,\mathrm{d}t\\ =&\int_0^\infty\int_{A_{s,f}(t)}\varpi(\mathrm{d}x)K(x,A_{s,f}(t)^c)\,\mathrm{d}t+\int_0^\infty\int_{A_{s,f}(t)^c}\varpi(\mathrm{d}x)K(x,A_{s,f}(t))\,\mathrm{d}t\\ \geq&2\Phi_K\int_0^\infty\varpi(A_{s,f}(t))[1-\varpi(A_{s,f}(t))]\,\mathrm{d}t.\end{aligned} \tag{1.4.7}$$

A similar calculation reveals that

$$\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|\,\varpi(\mathrm{d}x')\,\varpi(\mathrm{d}x)=2\int_0^\infty\varpi(A_{s,f}(t))[1-\varpi(A_{s,f}(t))]\,\mathrm{d}t. \tag{1.4.8}$$

Letting f and s vary, we have the following:

$$\begin{aligned}G(K)&\geq\inf_{f\in L_0^2(\varpi),\,f\neq0}\ \sup_{s\in\mathbb{R}}\frac{\left\{\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|K(x,\mathrm{d}x')\,\varpi(\mathrm{d}x)\right\}^2}{8\|f\|_2^2\|f_s\|_2^2}\quad\text{by (1.4.5) and (1.4.6)}\\ &\geq\frac{\Phi_K^2}{8}\inf_{f\in L_0^2(\varpi),\,f\neq0}\ \sup_{s\in\mathbb{R}}\left(\frac{\int_{\mathsf{X}^2}|f_s(x)^2-f_s(x')^2|\,\varpi(\mathrm{d}x')\,\varpi(\mathrm{d}x)}{\|f\|_2\|f_s\|_2}\right)^2\quad\text{by (1.4.7) and (1.4.8)}\\ &=\frac{\Phi_K^2}{8}\inf_{f\in L_0^2(\varpi),\,f\neq0}\ \sup_{s\in\mathbb{R}}\left(\frac{E[|(X-s)^2-(Y-s)^2|]}{\sqrt{\mathrm{var}(X)}\sqrt{E[(X-s)^2]}}\right)^2,\end{aligned}$$

where X and Y are independently and identically (iid) distributed as $\varpi \circ f^{-1}$, i.e., the distribution of $f(W)$ with $W \sim \varpi$. Note that $E(X) = \varpi f = 0$, and $\text{var}(X) = \|f\|_2^2 \in (0, \infty)$.

To conclude the proof, it suffices to show that, for two iid random variables X and Y with mean zero and some standard deviation $\sigma > 0$,

$$\sup_{s \in \mathbb{R}} \frac{E[|(X-s)^2 - (Y-s)^2|]}{\sigma\sqrt{E[(X-s)^2]}} \geq 1. \tag{1.4.9}$$

By the dominated convergence theorem, $\lim_{s\to\infty} E[s^{-2}(X-s)^2] = 1$, while

$$\begin{aligned}\lim_{s\to\infty} E[s^{-1}|(X-s)^2 - (Y-s)^2|] &= 2E[|X-Y|]\\ &= 2E[E(|X-Y| \mid X)]\\ &\geq 2E[|E(X-Y \mid X)|]\\ &= 2E(|X|) \quad \text{since } E(Y \mid X) = E(Y) = 0.\end{aligned}$$

Thus,

$$\sup_{s \in \mathbb{R}} \frac{E[|(X-s)^2 - (Y-s)^2|]}{\sqrt{E[(X-s)^2]}} \geq \lim_{s\to\infty} \frac{E[|(X-s)^2 - (Y-s)^2|]}{\sqrt{E[(X-s)^2]}} \geq 2E(|X|). \tag{1.4.10}$$

On the other hand, using the assumption that $E(X^2) = E(Y^2) = \sigma^2$ and the fact that $|u^2 - \sigma^2| \geq (|u| - \sigma)^2$ for $u \in \mathbb{R}$, we have

$$E(|X^2 - Y^2|) \geq E[|E(X^2 - Y^2 \mid X)|] = E(|X^2 - \sigma^2|) \geq E[(|X| - \sigma)^2] = 2\sigma^2 - 2\sigma E(|X|).$$

Thus,

$$\sup_{s \in \mathbb{R}} \frac{E[|(X-s)^2 - (Y-s)^2|]}{\sqrt{E[(X-s)^2]}} \geq \frac{E|(X-0)^2 - (Y-0)^2|}{\sqrt{E[(X-0)^2]}} \geq 2\sigma - 2E(|X|). \tag{1.4.11}$$

Combining (1.4.10) and (1.4.11) gives (1.4.9). □

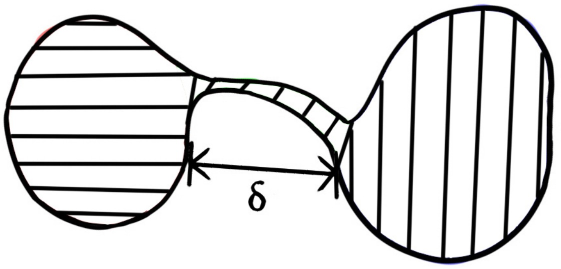

Figure 1.1: A dumbbell-shaped domain partitioned into three parts. The sets S_1 (horizontal stripes) and S_2 (vertical stripes) are separated by a narrow corridor S_3 (diagonal stripes).

1.4.3 Bounds via isoperimetric inequalities

We now describe a method for bounding Φ_K, and in turn, $G(K)$ from below. The method is particularly powerful when $\varpi(\cdot)$ admits a log-concave density function. It is based on a certain type of isoperimetric inequality. Let dist : $\mathcal{B}^2 \to [0, \infty]$ be some function that quantifies how far two sets are from each other. We say $\varpi(\cdot)$ satisfies a three-set isoperimetric inequality of Cheeger type if one can find some $\delta \in (0, \infty)$ and $\kappa \in (0, \infty)$ such that, for any partition of X consisting of three measurable sets, say, $\{S_1, S_2, S_3\}$,

$$\varpi(S_3) \geq \kappa\, \varpi(S_1)\varpi(S_2) \text{ whenever } \operatorname{dist}(S_1, S_2) \geq \delta. \tag{1.4.12}$$

Ideally, κ is not close to zero, especially if δ is large. This would indicate that, if two sets S_1 and S_2 are not too close, then $S_3 = (S_1 \cup S_2)^c$ must have a non-negligible probability mass relative to the masses of S_1 and S_2. Loosely speaking, this means that the state space X cannot exhibit two disjoint subdomains, each possessing substantial probability mass, where transitioning from one subdomain to the other necessitates covering extensive distances through a low-probability region.

Figure 1.1 shows a scenario where a good isoperimetric inequality, i.e., one with a large value of κ, would eliminate. Here, a dumbbell-shaped domain X is presented. Let $\varpi(\cdot)$ be the uniform distribution on X. We can partition X into three regions, S_1, S_2, and S_3, which are filled with, respectively, horizontal, vertical, and diagonal stripes. Imagine that $\delta = \mathrm{dist}(S_1, S_2)$, and $\varpi(S_3)$ is small relative to $\varpi(S_1)\varpi(S_2)$. Then (1.4.12) cannot hold for large values of κ. Scenarios like this could prevent the fast mixing of some Markov chains whose stationary distributions are $\varpi(\cdot)$. Indeed, it may be difficult for a chain to travel from S_1 to S_2 due to how narrow the corridor connecting the two sets is.

There is a large literature devoted to establishing isoperimetric inequalities, especially for distributions on Euclidean spaces with log-concave density functions. We will not attempt to establish an inequality ourselves since this typically requires some sophisticated analysis. Instead, we state, without proof, a well-known inequality for distributions with strongly log-concave density functions (see, e.g., Bobkov, 2003; Ledoux, 1999).

Theorem 1.4.5. *(Bobkov, 2003, Theorem 1) Let $\mathsf{X} = \mathbb{R}^p$ for some $p \in \mathbb{N}_+$. Suppose that ϖ admits a probability density function with respect to the Lebesgue measure that is proportional to $\exp[-\|x\|^2/(2\sigma^2)]g(x)$, where $\sigma > 0$, and $x \mapsto g(x)$ is log-concave, i.e., for all $x, y \in \mathbb{R}^p$ and $\lambda \in (0,1)$, the inequality $g(\lambda x + (1-\lambda)y) \geq g(x)^\lambda g(y)^{1-\lambda}$ holds. Let $\{S_1, S_2, S_3\}$ be a measurable partition of $\mathbb{R}^p$. Then, for $i = 1, 2$,*

$$\varpi(S_3) \geq F_N\left(F_N^{-1}(\varpi(S_i)) + \frac{\mathit{dist}\,(S_1, S_2)}{\sigma}\right) - \varpi(S_i),$$

where $F_N(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\mathit{dist}(S_1, S_2) = \inf_{x \in S_1, y \in S_2} \|x - y\|$.

We can get an inequality of the form (1.4.12) from Theorem 1.4.5 using some calculus. Let $r = \min\{\varpi(S_1), \varpi(S_2)\}$ so that $F_N^{-1}(r) \leq 0$, and denote by $f_N(\cdot)$ the density function of the standard normal distribution. Under the assumption of Theorem 1.4.5, if $\mathrm{dist}(S_1, S_2) \geq \delta$

for some $\delta \in (0, \infty)$, then

$$\begin{aligned}
\varpi(S_3) &\geq \int_{F_N^{-1}(r)}^{F_N^{-1}(r)+\delta/\sigma} f_N(t)\,\mathrm{d}t \\
&\geq \begin{cases} f_N\left(F_N^{-1}(r) + \delta/\sigma\right)\,\delta/\sigma & \text{if } F_N^{-1}(r) + \delta/\sigma \geq -F_N^{-1}(r) \geq 0 \\ f_N(F_N^{-1}(r))\,\delta/\sigma & \text{otherwise} \end{cases} \\
&\geq \begin{cases} f_N\left(\delta/\sigma\right)\,\delta/\sigma & \text{if } F_N^{-1}(r) + \delta/\sigma \geq -F_N^{-1}(r) \geq 0 \\ f_N(F_N^{-1}(r))\,\delta/\sigma & \text{otherwise} \end{cases} \\
&\geq \sqrt{2\pi}\, f_N(F_N^{-1}(r))\, f_N\left(\delta/\sigma\right) \frac{\delta}{\sigma}.
\end{aligned}$$

By (4) in Sampford (1953), one can show that $f_N(q)/[1 - F_N(q)] \geq 4F_N(q)/\sqrt{2\pi}$ for $q \geq 0$. Letting $q = -F_N^{-1}(r)$ yields $f_N(F_N^{-1}(r)) \geq 4r(1-r)/\sqrt{2\pi}$. Moreover, $r(1-r) \geq \varpi(S_1)\varpi(S_2)$. Thus, for an arbitrary choice of $\delta \in (0, \infty)$,

$$\varpi(S_3) \geq \frac{4\delta f_N\left(\delta/\sigma\right)}{\sigma}\varpi(S_1)\varpi(S_2) \text{ whenever } \operatorname{dist}(S_1, S_2) \geq \delta,$$

i.e., (1.4.12) holds with $\delta \in (0, \infty)$ and $\kappa = 4(\delta/\sigma) f_N(\delta/\sigma)$.

For other examples of three-set isoperimetric inequalities, see, e.g., Theorem 2.6 of Lovász and Simonovits (1993), Theorem 2.1 of Kannan and Li (1996), Theorem 1 of Lovász (1999), and Theorem 4.2 of Cousins and Vempala (2014). Three-set isoperimetric inequalities can also be obtained through more standard forms of isoperimetric inequalities that involve the perimeter and volume of an arbitrary measurable set (see, e.g., Andrieu et al., 2024; Bobkov and Houdré, 1997).

We now give a bound on $G(K)$ based on a three-set isoperimetric inequality of Cheeger type. It is largely similar to existing results from, e.g., Lovász (1999), Belloni and Chernozhukov (2009), and Dwivedi et al. (2019).

Theorem 1.4.6. *Let $\psi' : \mathsf{X}^2 \to [0, \infty)$ be a measurable function (not necessarily a metric), and for $A, B \in \mathcal{B}$, let*

$$\mathit{dist}(A, B) = \inf_{x \in A,\, y \in B} \psi'(x, y).$$

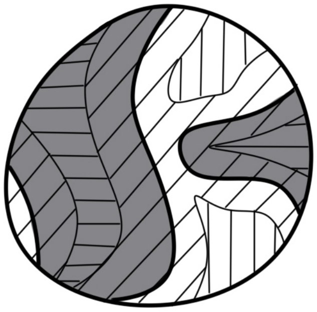

Figure 1.2: A domain partitioned into A (grey) and A^c. Depending on the amount of probability flow from each point in A (resp. A^c) to A^c (resp. A), the domain can be alternatively partitioned into S_1 (horizontal stripes), S_2 (vertical stripes), and S_3 (diagonal stripes).

Suppose that there exist $\delta \in (0,\infty)$ and $\varepsilon \in (0,1]$ such that the following "close coupling condition" holds:

$$\|\delta_x K - \delta_y K\|_{TV} \le 1-\varepsilon \quad \textit{whenever } \psi'(x,y) < \delta. \tag{1.4.13}$$

(Recall that δ_x is the point mass at x, so $\delta_x K(\cdot) = K(x,\cdot)$.) Suppose further that $\varpi(\cdot)$ satisfies a three-set isoperimetric inequality of Cheeger type with δ given above and some $\kappa \in (0,\infty)$. Then, for $A \in \mathcal{B}$ such that $\varpi(A) \in (0,1)$ and $a \in (0,1)$,

$$\phi_K(A) := \frac{\int_A \varpi(\mathrm{d}x) K(x,A^c)}{\varpi(A)\varpi(A^c)} \ge \varepsilon \min\left\{\frac{1-a}{2}, \frac{a^2\kappa}{4}\right\}. \tag{1.4.14}$$

This inequality holds even when $K(\cdot,\cdot)$ is non-reversible.

Proof. Let $A \in \mathcal{B}$ be such that $\varpi(A) \in (0,1)$. Let

$$S_1 = \{x \in A : K(x,A^c) < \varepsilon/2\}, \quad S_2 = \{x \in A^c : K(x,A) < \varepsilon/2\}, \quad S_3 = (S_1 \cup S_2)^c.$$

See Figure 1.2. Fix $a \in (0,1)$. We will establish (1.4.14) in three cases: (i) $\varpi(S_1) \le a\varpi(A)$, (ii) $\varpi(S_2) \le a\varpi(A^c)$, and (iii) $\varpi(S_1) > a\varpi(A)$ and $\varpi(S_2) > a\varpi(A^c)$. Note that these three cases exhaust all possibilities.

Case (i): By the definition of S_3,

$$\int_A \varpi(\mathrm{d}x)K(x,A^c) \geq \int_{S_3\cap A} \varpi(\mathrm{d}x)K(x,A^c) \geq \frac{\varepsilon}{2}\varpi(S_3\cap A). \tag{1.4.15}$$

In Case (i),

$$\varpi(S_3\cap A) = \varpi(A) - \varpi(S_1) \geq (1-a)\varpi(A) \geq (1-a)\varpi(A)\varpi(A^c),$$

so, by (1.4.15), we have (1.4.14).

Case (ii): Since $\varpi K = \varpi$,

$$\begin{aligned}\int_A \varpi(\mathrm{d}x)K(x,A^c) &= \int_{\mathsf{X}} \varpi(\mathrm{d}x)K(x,A^c) - \int_{A^c} \varpi(\mathrm{d}x)K(x,A^c)\\ &= \varpi(A^c) - \int_{A^c} \varpi(\mathrm{d}x)[1-K(x,A)]\\ &= \int_{A^c} \varpi(\mathrm{d}x)K(x,A).\end{aligned}$$

Then

$$\int_A \varpi(\mathrm{d}x)K(x,A^c) \geq \int_{S_3\cap A^c} \varpi(\mathrm{d}x)K(x,A) \geq \frac{\varepsilon}{2}\varpi(S_3\cap A^c). \tag{1.4.16}$$

In Case (ii),

$$\varpi(S_3\cap A^c) = \varpi(A^c) - \varpi(S_2) \geq (1-a)\varpi(A^c) \geq (1-a)\varpi(A)\varpi(A^c),$$

so, by (1.4.16), we have (1.4.14).

Case (iii): By the definition of the total variation distance (see Section 1.1), for $x \in S_1$ and $y \in S_2$,

$$\|\delta_x K - \delta_y K\|_{\mathrm{TV}} \geq K(x,A) - K(y,A) > 1-\varepsilon.$$

By the close coupling condition (1.4.13), $\psi'(x,y) \geq \delta$ for $x \in S_1$ and $y \in S_2$, so $\operatorname{dist}(S_1,S_2) \geq \delta$. By the isoperimetric inequality,

$$\varpi(S_3) \geq \kappa\varpi(S_1)\varpi(S_2).$$

Note that (1.4.15) and (1.4.16) still hold. Combining them with the display above yields

$$\int_A \varpi(\mathrm{d}x)K(x, A^c) \geq \frac{\varepsilon\varpi(S_3)}{4} \geq \frac{\varepsilon\kappa}{4}\varpi(S_1)\varpi(S_2) \tag{1.4.17}$$

In Case (iii),

$$\varpi(S_1)\varpi(S_2) \geq a^2\varpi(A)\varpi(A^c),$$

so, by (1.4.17), we have (1.4.14).

□

Application to the Gaussian chain

Recall that, in Example 1.2.2, $\mathsf{X} = \mathbb{R}^p$, $\varpi_p(\cdot)$ is the $\mathrm{N}_p(0, I_p)$ distribution, and $K_{p,\alpha}(x, \cdot)$ is the $\mathrm{N}_p(\alpha x, (1-\alpha^2)I_p)$ distribution for $x \in \mathbb{R}^p$, where $\alpha \in [0, 1)$. We can place an upper bound on $\|K_{p,\alpha}\|_2$ using Theorems 1.4.4, 1.4.5, and 1.4.6.

In light of the discussion in Section 1.4.2, we first show that $K_{p,\alpha}$, as a linear operator on $L_0^2(\varpi_p)$, is positive semi-definite. Recall first that the Mtk $K_{p,\alpha}(\cdot, \cdot)$ is reversible with respect to $\varpi_p(\cdot)$. This is equivalent to $K_{p,\alpha}$ being self-adjoint. Next, note that, for $f \in L_0^2(\varpi_p)$ and $x \in \mathbb{R}^p$,

$$\begin{aligned} K_{p,\alpha}f(x) &= \frac{1}{[2\pi(1-\alpha^2)]^{p/2}} \int_{\mathbb{R}^p} f(x') \exp\left[-\frac{1}{2(1-\alpha^2)}\|x' - \alpha x\|^2\right] \mathrm{d}x' \\ &= \frac{1}{[2\pi(1-\alpha)]^p} \int_{\mathbb{R}^p} f(x') \int_{\mathbb{R}^p} \exp\left[-\frac{\|x' - \sqrt{\alpha}x''\|^2}{2(1-\alpha)} - \frac{\|x'' - \sqrt{\alpha}x\|^2}{2(1-\alpha)}\right] \mathrm{d}x''\mathrm{d}x' \\ &= K_{p,\sqrt{\alpha}}^2 f(x). \end{aligned}$$

Since $K_{p,\sqrt{\alpha}}(\cdot, \cdot)$ is also reversible with respect to $\varpi_p(\cdot)$, the corresponding operator is self-adjoint. As a result, for $f \in L_0^2(\varpi_p)$,

$$\langle f, K_{p,\alpha}f\rangle = \langle f, K_{p,\sqrt{\alpha}}K_{p,\sqrt{\alpha}}f\rangle = \langle K_{p,\sqrt{\alpha}}f, K_{p,\sqrt{\alpha}}f\rangle = \|K_{p,\sqrt{\alpha}}f\|_2^2 \geq 0.$$

Hence, $K_{p,\alpha}$ is positive semi-definite. By (1.4.4) and Cheeger's inequality (Theorem 1.4.4),

$$\|K_{p,\alpha}\|_2 = 1 - G(K_{p,\alpha}) \leq 1 - \Phi^2_{K_{p,\alpha}}/8.$$

We now bound $\Phi_{K_{p,\alpha}}$ from below. We may pretend that the only thing we know about $\varpi_p(\cdot)$ is the following: It has a density function of the form $e^{-h_p(x)}$, where $\nabla^2 h_p(x) - I_p$ is positive semi-definite for $x \in \mathbb{R}^p$. ($\nabla^2 h_p(x)$ denotes the Hessian matrix of h_p.) Then, by Theorem 1.4.5, $\varpi_p(\cdot)$ satisfies a three-set isoperimetric inequality of Cheeger type with an arbitrary positive δ and $\kappa = 4\delta f_N(\delta)$. On the other hand, by (1.2.1) and (1.3.13), for $x, y \in \mathbb{R}^p$,

$$\|\delta_x K - \delta_y K\|_{\mathrm{TV}} = 1 - \int_{\mathbb{R}^p} \min\{k_{p,\alpha}(x, \mathrm{d}x'), k_{p,\alpha}(y, \mathrm{d}x')\}\, \mathrm{d}x' = 1 - 2F_N\left(-\frac{\alpha\|x-y\|}{2\sqrt{1-\alpha^2}}\right),$$

where $k_{p,\alpha}(x,\cdot)$ is the density function of $K_{p,\alpha}(x,\cdot)$. Applying the bound in Theorem 1.4.6 with $a = 1/2$ and $\delta = \sqrt{1-\alpha^2}/\alpha$ yields

$$\Phi_{K_{p,\alpha}} \geq \min\left\{\frac{1}{4}, \frac{\delta f_N(\delta)}{4}\right\} \times 2F_N\left(-\frac{1}{2}\right) = \frac{\sqrt{1-\alpha^2}}{2\sqrt{2\pi}\alpha} \exp\left(-\frac{1-\alpha^2}{2\alpha^2}\right) F_N\left(-\frac{1}{2}\right).$$

Thus,

$$\|K_{p,\alpha}\|_2 \leq 1 - \frac{1-\alpha^2}{64\pi\alpha^2} \exp\left(-\frac{1-\alpha^2}{\alpha^2}\right) F_N\left(-\frac{1}{2}\right)^2.$$

Recall that, in truth, $\|K_{p,\alpha}\|_2 = \alpha$. The bound correctly indicates that $\|K_{p,\alpha}\|_2$ is bounded away from unity as $p \to \infty$, and that $(1 - \|K_{p,\alpha}\|_2)/(1-\alpha)$ is bounded away from zero as $\alpha \to 1$.

1.4.4 Lower bounds on $\|K\|_2$

Let us consider the problem of bounding $\|K\|_2$ from below, which, by Theorem 1.4.2, would quantify how slowly the chain converges when $K(\cdot,\cdot)$ is reversible. The problem is not as frequently studied as that of bounding $\|K\|_2$ from above, but it has been examined in the context of some important MCMC algorithms (Andrieu et al., 2024; Chewi et al., 2021;

Johndrow et al., 2018; Wu et al., 2022). We provide a concise overview of some of the basic techniques employed in these studies. Although this problem is most meaningful when $K(\cdot,\cdot)$ is reversible, the results we present would not require the assumption of reversibility.

Let $f \in L^2(\varpi)$ be a non-constant function. Then $f - \varpi f \in L_0^2(\varpi)$, and $f - \varpi f \neq 0$. By the Cauchy-Schwarz inequality and the definition of $\|K\|_2$,

$$\langle f - \varpi f, K(f - \varpi f)\rangle \leq \|f - \varpi f\|_2 \, \|K(f - \varpi f)\|_2 \leq \|K\|_2 \|f - \varpi f\|_2^2.$$

This yields the following bound:

Theorem 1.4.7. *Let $f \in L^2(\varpi)$ be a non-constant function. Then*

$$\|K\|_2 \geq \frac{\langle f - \varpi f, K(f - \varpi f)\rangle}{\|f - \varpi f\|_2^2} = 1 - \frac{\int_{\mathsf{X}^2} [f(x) - f(x')]^2 K(x, \mathrm{d}x') \, \varpi(\mathrm{d}x)}{2\|f - \varpi f\|_2^2}.$$

Theorem 1.4.7 can be directly applied to scenarios where $\varpi(\cdot)$ has a simple form. Consider the Gaussian chain in Example 1.2.2. For $x = (x[1], \dots, x[p]) \in \mathbb{R}^p$, let $f(x) = x[1]$. Then $\varpi_p f = 0$, $\|f - \varpi_p f\|_2^2 = 1$, and

$$\langle f - \varpi_p f, K_{p,\alpha}(f - \varpi_p f)\rangle = \langle f, \alpha f\rangle = \alpha.$$

By Theorem 1.4.7, $\|K_{p,\alpha}\|_2 \geq \alpha$. Recall that α is in fact the true value of $\|K_{p,\alpha}\|_2$.

Letting f be an indicator function in Theorem 1.4.7 gives the following (after a bit of calculations), which is very similar to parts of Theorem 1.4.4.

Corollary 1.4.1. *Let $A \in \mathcal{B}$ be such that $\varpi(A) \in (0, 1)$. Then*

$$\|K\|_2 \geq 1 - \frac{\int_A \varpi(\mathrm{d}x) K(x, A^c)}{\varpi(A)\varpi(A^c)}. \tag{1.4.18}$$

Finally, from Corollary 1.4.1 we can immediately derive the result below.

Corollary 1.4.2. *Suppose that there exist $A \in \mathcal{B}$ and $\delta \leq 1$ such that $K(x, \{x\}^c) \leq \delta$ for*

$x \in A$. *Then, if* $\varpi(A) \in (0,1)$,

$$\|K\|_2 \geq 1 - \frac{\delta}{\varpi(A^c)}.$$

By Corollary 1.4.2, to obtain a large lower bound on $\|K\|_2$, we only need to find a set A such that $\varpi(A)$ is small, and that $K(x,\{x\})$ is large when $x \in A$. Oftentimes, we can take A to be an arbitrarily small neighborhood around a point x_0 where $K(x_0,\{x_0\})$ is large.

We may apply Corollary 1.4.2 to the independent Metropolis Hastings chain from Example 1.2.1. Recall that $s : [0,1] \to (0,\infty)$ is continuous, and $M_s = \sup_{x\in[0,1]} s(x) < \infty$. Then, for $\varepsilon > 0$, one can find an interval $A_\varepsilon = (a_\varepsilon, b_\varepsilon) \subset [0,1]$ such that $0 < b_\varepsilon - a_\varepsilon < \varepsilon$, and that $s(x) > M_s - \varepsilon$ for $x \in (a_\varepsilon, b_\varepsilon)$. Then, whenever $\varepsilon \in (0, M_s)$, it holds that $\pi_s(A_\varepsilon) \leq \varepsilon M_s$, and, for $x \in A_\varepsilon$,

$$K_s(x,\{x\}^c) = \int_0^1 a_s(x,x')\,\mathrm{d}x' = \int_0^1 \min\left\{1, \frac{s(x')}{s(x)}\right\} \mathrm{d}x' \leq \int_0^1 \frac{s(x')}{M_s - \varepsilon}\,\mathrm{d}x' = \frac{1}{M_s - \varepsilon}.$$

Hence, by Corollary 1.4.2, if $\varepsilon < M_s$ and $\varepsilon < 1/M_s$,

$$\|K_s\|_2 \geq 1 - \frac{1}{(M_s - \varepsilon)(1 - \varepsilon M_s)}.$$

Since ε can be arbitrarily small, we have the bound $\|K_s\|_2 \geq 1 - 1/M_s$. In Section 1.4.1, it was shown that $\|K_s\| \leq 1 - 1/M_s$. Thus, we may conclude that $\|K_s\|_2 = 1 - 1/M_s$.

Corollary 1.4.2 can be used to analyze Markov chains associated with general Metropolis Hastings algorithms. See Brown and Jones (2022) for a detailed discussion on this topic.

There are also results showing the slowness of Markov chains outside the L^2 framework. See, e.g., Roberts and Rosenthal (2011), Wang (2022), Brown and Jones (2022).

Example: a random walk Metropolis Hastings algorithm

We now illustrate Theorem 1.4.7 and Corollary 1.4.2 through a semi-toy example, which is a simplified version of a study from Andrieu et al. (2024).

Consider a random walk Metropolis Hastings (RWMH) algorithm on $\mathbb{R}^p$ targeting the p-dimensional standard normal distribution, which we denote by $\varpi_p(\cdot)$. Let σ be a positive constant. Given the current state $x \in \mathbb{R}^p$, the RWMH algorithm proceeds as follows. Draw X' from the $\mathrm{N}_p(x, \sigma^2 I_p)$ distribution, and call its realization x'. Let

$$a(x, x') = \min\left\{1, \frac{\exp(-\|x'\|^2/2)}{\exp(-\|x\|^2/2)}\right\}.$$

With probability $a(x, x')$, set the next state to x', and with probability $1 - a(x, x')$, set the next state to x. The underlying Markov chain is reversible with respect to $\varpi_p(\cdot)$. Denote the Mtk of this algorithm by $T_{p,\sigma}$. Then, for $x \in \mathbb{R}^p$ and $A \in \mathcal{B}$,

$$\begin{aligned} T_{p,\sigma}(x, A) = & \int_A \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right) a(x, x') \, \mathrm{d}x' + \\ & \int_{\mathbb{R}^p} \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right) [1 - a(x, x')] \, \mathrm{d}x' \, \mathbf{1}_{x \in A}. \end{aligned}$$

We now apply Theorem 1.4.7 to bound $\|T_{p,\sigma}\|_2$ from below. For $x = (x[1], \dots, x[p]) \in \mathbb{R}^p$, let $f(x) = x[1]$. Then $\|f - \varpi_p f\|_2^2 = 1$. Moreover, for $x \in \mathbb{R}^p$,

$$\begin{aligned} \int_{\mathbb{R}^p} [f(x) - f(x')]^2 \, T_{p,\sigma}(x, \mathrm{d}x') &= \int_{\mathbb{R}^p} \frac{[f(x) - f(x')]^2}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right) a(x, x') \, \mathrm{d}x' \\ &\leq \int_{\mathbb{R}^p} \frac{[f(x) - f(x')]^2}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right) \mathrm{d}x' = \sigma^2. \end{aligned}$$

By Theorem 1.4.7, $\|T_{p,\sigma}\|_2 \geq 1 - \sigma^2/2$.

Another lower bound on $\|T_{p,\sigma}\|_2$ can be obtained through Corollary 1.4.2. Elementary calculations show that

$$\begin{aligned} T_{p,\sigma}(0, \{0\}^c) &= \int_{\mathbb{R}^p} \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x' - 0\|^2}{2\sigma^2}\right) a(0, x') \, \mathrm{d}x' \\ &= \frac{1}{(\sigma^2 + 1)^{p/2}}. \end{aligned}$$

Moreover, one can verify that $x \mapsto T_{p,\sigma}(x, \{x\}^c)$ is a continuous function. Hence, there exists a sequence of open neighborhoods of 0, say, $A_1, A_2, \dots$, such that $\lim_{n\to\infty} \varpi_p(A_n) = 0$, and that

$$\lim_{n\to\infty} \sup_{x\in A_n} T_{p,\sigma}(x, \{x\}^c) = \frac{1}{(\sigma^2+1)^{p/2}}.$$

By Corollary 1.4.2,

$$\|T_{p,\sigma}\|_2 \geq 1 - \lim_{n\to\infty} \frac{\sup_{x\in A_n} T(x,\{x\}^c)}{1-\varpi_p(A_n)} = 1 - \frac{1}{(\sigma^2+1)^{p/2}}.$$

Combining the two bounds, we see that

$$\|T_{p,\sigma}\|_2 \geq 1 - \min\left\{\frac{\sigma^2}{2}, \frac{1}{(\sigma^2+1)^{p/2}}\right\}. \tag{1.4.19}$$

In practice, we may tune σ in the hopes of making $\|T_{p,\sigma}\|_2$ small. But how small can $\|T_{p,\sigma}\|_2$ go? We can partially answer this by minimizing the lower bound in (1.4.19). The bound is minimized when $\sigma = \sigma_p$, where σ_p satisfies

$$\frac{\sigma_p^2}{2} = \frac{1}{(\sigma_p^2+1)^{p/2}}.$$

It can be shown that, when p is sufficiently large, $1/p \leq \sigma_p^2 \leq 2(\log p)/p$, and

$$1 - \min\left\{\frac{\sigma_p^2}{2}, \frac{1}{(\sigma_p^2+1)^{p/2}}\right\} \geq 1 - \frac{\log p}{p}.$$

This means that, regardless of how σ is chosen, $\|T_{p,\sigma}\|_2$ is always lower bounded by $1 - (\log p)/p$ for large values of p.

When $\sigma^2 = 1/p$, (1.4.19) implies that $1 - \|T_{p,\sigma}\|_2 \leq 1/(2p)$. In this case, the bound gives the correct order when $p \to \infty$. Indeed, using isoperimetric inequalities, it is possible to establish a lower bound on $1 - \|T_{p,\sigma}\|_2$ that is also of the order $1/p$ when $\sigma^2 = 1/p$. See Andrieu et al. (2024), who studied RWMH algorithms targeting general distributions with strongly log-concave densities.

1.5 Other methods

We end this chapter by listing some other important methods for constructing convergence bounds.

The canonical path technique is a powerful tool for analyzing Markov chains taking values in a discrete state space (Diaconis and Stroock, 1991; Sinclair, 1992; Yang et al., 2016).

The transition law of a Markov chain can be written into a random function, and convergence bounds may be formed by studying the local contractive behavior of this function (Jarner and Tweedie, 2001; Qin and Hobert, 2022; Qu et al., 2023; Steinsaltz, 1999).

The convergence properties of a Markov chain with a complicated transition law can be studied by comparing it to a simpler Markov chain or process (Andrieu et al., 2018; Ascolani and Zanella, 2024; Dalalyan, 2017; Jones et al., 2014; Łatuszyński and Rudolf, 2024; Pillai and Smith, 2014; Rudolf and Schweizer, 2018). In particular, the optimal scaling framework provides a unique perspective for studying the properties of a high-dimensional Metropolis-Hastings algorithm by relating it to a certain diffusion process (Atchadé et al., 2011; Gelman et al., 1997; Pillai et al., 2012; Yang et al., 2020).

One can also decompose an intricate transition law into simpler components (Ge et al., 2018; Guan and Krone, 2007; Jerrum et al., 2004; Madras and Randall, 2002; Qin et al., 2023; Woodard et al., 2009). Related to this approach, techniques based on spectral independence and stochastic localization have recently received an increasing amount of attention (Anari et al., 2021; Chen and Eldan, 2022; Chen et al., 2021a,b; Feng et al., 2022; Qin and Wang, 2024).

Finally, some MCMC algorithms can be conceptualized as certain deterministic optimization algorithms over a space of distributions. These algorithms can be analyzed using the theory of gradient flows. See Cheng and Bartlett (2018), Durmus et al. (2019), and references therein.

Acknowledgments

This work was partially supported by the NSF. The author thanks Austin Brown, Galin L. Jones, and Youngwoo Kwon for their helpful comments.

Bibliography

Aldous, D. (1983). Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire de Probabilités (Strasbourg), Tome 17*, pages 243–297. Springer.

Amit, Y. (1996). Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Annals of Statistics*, 24:122–140.

Anari, N., Liu, K., and Gharan, S. O. (2021). Spectral independence in high-dimensional expanders and applications to the hardcore model. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 1319–1330. IEEE.

Andrieu, C., Fort, G., and Vihola, M. (2015). Quantitative convergence rates for subgeometric Markov chains. *Journal of Applied Probability*, 52:391–404.

Andrieu, C., Lee, A., Power, S., and Wang, A. Q. (2022). Comparison of Markov chains via weak Poincaré inequalities with application to pseudo-marginal MCMC. *Annals of Statistics*, 50:3592–3618.

Andrieu, C., Lee, A., Power, S., and Wang, A. Q. (2024+). Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. *Annals of Applied Probability,* to appear.

Andrieu, C., Lee, A., and Vihola, M. (2018). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24:842–872.

Arveson, W. (2006). *A Short Course on Spectral Theory*, volume 209. Springer Science & Business Media.

Ascolani, F. and Zanella, G. (2024+). Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models. *Annals of Statistics,* to appear.

Atchadé, Y. F., Roberts, G. O., and Rosenthal, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21:555–568.

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Annals of Applied Probability*, 15:700–738.

Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Annals of Statistics*, 37:2011–2055.

Bobkov, S. G. (2003). Localization proof of the Bakry–Ledoux isoperimetric inequality and some applications. *Theory of Probability and Its Applications*, 47:308–314.

Bobkov, S. G. and Houdré, C. (1997). Some connections between isoperimetric and Sobolev-type inequalities. *Memoirs of the American Mathematical Society*, 129:1–111.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *Annals of Applied Probability*, 30:1209–1250.

Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335.

Brown, A. and Jones, G. L. (2022). Lower bounds on the rate of convergence for accept-reject-based Markov chains. arXiv preprint.

Bruckner, A. M., Bruckner, J. B., and Thomson, B. S. (2008). *Real Analysis.* ClassicalRealAnalysis.com, 2nd edition.

Bubley, R. and Dyer, M. (1997). Path coupling: A technique for proving rapid mixing in Markov chains. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 223–231. IEEE.

Burdzy, K. and Kendall, W. S. (2000). Efficient Markovian couplings: examples and counterexamples. *Annals of Applied Probability*, 10:362–409.

Butkovsky, O. (2014). Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *Annals of Applied Probability*, 24:526–552.

Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1747–1758.

Chen, Y. and Eldan, R. (2022). Localization schemes: A framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*, pages 110–122. IEEE.

Chen, Z., Galanis, A., Štefankovič, D., and Vigoda, E. (2021a). Rapid mixing for colorings via spectral independence. In *Proceedings of the Thirty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1548–1557. SIAM.

Chen, Z., Liu, K., and Vigoda, E. (2021b). Optimal mixing of Glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550.

Cheng, X. and Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR.

Chewi, S., Lu, C., Ahn, K., Cheng, X., Le Gouic, T., and Rigollet, P. (2021). Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR.

Conway, J. B. (1990). *A Course in Functional Analysis*. Springer-Verlag, 2nd edition.

Cousins, B. and Vempala, S. (2014). A cubic algorithm for computing Gaussian volume. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1215–1228. SIAM.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79:651–676.

Diaconis, P., Holmes, S., and Neal, R. M. (2000). Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, 10:726–752.

Diaconis, P., Khare, K., and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statistical Science*, 23:151–200.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability*, 1:36–61.

Douc, R., Fort, G., Moulines, E., and Soulier, P. (2004). Practical drift conditions for subgeometric rates of convergence. *Annals of Applied Probability*, 14:1353–1377.

Douc, R., Guillin, A., and Moulines, E. (2008). Bounds on regeneration times and limit theorems for subgeometric Markov chains. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 44:239–257.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018). *Markov Chains*. Springer.

Durmus, A., Fort, G., and Moulines, É. (2016). Subgeometric rates of convergence in Wasserstein distance for Markov chains. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52:1799–1822.

Durmus, A., Majewski, S., and Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20:2666–2711.

Durmus, A. and Moulines, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25:2854–2882.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2019). Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20:1–42.

Eberle, A. and Majka, M. B. (2019). Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24:1–36.

Feng, W., Guo, H., Yin, Y., and Zhang, C. (2022). Rapid mixing from spectral independence beyond the boolean domain. *ACM Transactions on Algorithms*, 18:1–32.

Ge, R., Lee, H., and Risteski, A. (2018). Simulated tempering Langevin monte carlo II: An improved proof using soft Markov chain decomposition. arXiv preprint.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. In *Handbook of Markov chain Monte Carlo*, pages 163–174. Chapman and Hall/CRC.

Guan, Y. and Krone, S. M. (2007). Small-world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing. *Annals of Applied Probability*, 17(1):284–304.

Hairer, M. (2006). Ergodic properties of Markov processes. Lecture notes, Univ. Warwick.

Hairer, M. and Mattingly, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer.

Hairer, M., Mattingly, J. C., and Scheutzow, M. (2011). Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probability Theory and Related Fields*, 149:223–259.

Helmberg, G. (2014). *Introduction to Spectral Theory in Hilbert Space*. Elsevier.

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67:414–430.

Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics*, 36:532–554.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and Their Applications*, 85:341–361.

Jarner, S. F. and Roberts, G. O. (2002). Polynomial convergence rates of Markov chains. *Annals of Applied Probability*, 12:224–247.

Jarner, S. F. and Tweedie, R. L. (2001). Locally contracting iterated functions and stability of Markov chains. *Journal of Applied Probability*, 38:494–507.

Jerison, D. C. (2019). Quantitative convergence rates for reversible Markov chains via strong random times. arXiv preprint.

Jerrum, M. and Sinclair, A. (1988). Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 235–244.

Jerrum, M., Son, J.-B., Tetali, P., and Vigoda, E. (2004). Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains. *Annals of Applied Probability*, 14:1741–1765.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2018). MCMC for imbalanced categorical data. *Journal of the American Statistical Association.*

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

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

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

Jones, G. L., Roberts, G. O., and Rosenthal, J. S. (2014). Convergence of conditional Metropolis-Hastings samplers. *Advances in Applied Probability*, 46:422–445.

Kannan, R. and Li, G. (1996). Sampling according to the multivariate normal density. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 204–212. IEEE.

Khare, K. and Hobert, J. P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. *Annals of Statistics*, 39:2585–2606.

Khare, K. and Hobert, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7:2150–2163.

Kontoyiannis, I. and Meyn, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields*, 154:327–339.

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing B-valued random variables. *Annals of Probability*, 8:1003–1036.

Łatuszyński, K. and Rudolf, D. (2024+). Convergence of hybrid slice sampling via spectral gap. *Advances in Applied Probability,* to appear.

Lawler, G. F. and Sokal, A. D. (1988). Bounds on the l^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality. *Transactions of the American Mathematical Society*, 309:557–580.

Ledoux, M. (1999). Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités (Strasbourg), Tome 33*, pages 120–216. Springer-Verlag.

Lindvall, T. and Rogers, L. C. G. (1986). Coupling of multidimensional diffusions by reflection. *Annals of Probability*, 14:860–872.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.

Liu, J. S., Wong, W. H., and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, Series B*, 57:157–169.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. (2019). On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25:3109–3138.

Lovász, L. (1999). Hit-and-run mixes fast. *Mathematical Programming*, 86:443–461.

Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures and Algorithms*, 4:359–412.

Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, 12:581–606.

Madras, N. and Sezer, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16:882–908.

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

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Annals of Applied Probability*, 4:981–1011.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability.* Springer Science & Business Media, 2nd edition.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443.

Pillai, N. S. and Smith, A. (2014). Ergodicity of approximate MCMC chains with applications to large data sets. arXiv preprint.

Pillai, N. S. and Smith, A. (2017). Kac's walk on n-sphere mixes in $n \log n$ steps. *Annals of Applied Probability*, 27:631–650.

Pillai, N. S., Stuart, A. M., and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Annals of Applied Probability*, 22:2320–2356.

Qin, Q. and Hobert, J. P. (2020). On the limitations of single-step drift and minorization in Markov chain convergence analysis. *Annals of Applied Probability*, 31:1633–1659.

Qin, Q. and Hobert, J. P. (2022). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *Annals of Applied Probability*, 32:124–166.

Qin, Q., Ju, N., and Wang, G. (2023). Spectral gap bounds for reversible hybrid Gibbs chains. arXiv preprint.

Qin, Q. and Wang, G. (2024+). Spectral telescope: Convergence rate bounds for random-scan Gibbs samplers based on a hierarchical structure. *Annals of Applied Probability,* to appear.

Qu, Y., Blanchet, J., and Glynn, P. (2023). Computable bounds on convergence of Markov chains in Wasserstein distance. arXiv preprint.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25.

Roberts, G. O. and Rosenthal, J. S. (2002). One-shot coupling for certain stochastic recursive sequences. *Stochastic Processes and Their Applications*, 99:195–208.

Roberts, G. O. and Rosenthal, J. S. (2011). Quantitative non-geometric convergence bounds for independence samplers. *Methodology and Computing in Applied Probability*, 13:391–403.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59:291–317.

Roberts, G. O. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and Their Applications*, 80:211–229.

Roberts, G. O. and Tweedie, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, 38:37–41.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90:558–566.

Roy, V. (2020). Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7:387–412.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69:607–623.

Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae*, 485:1–93.

Rudolf, D. and Schweizer, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24:2610–2639.

Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *Annals of Mathematical Statistics*, 24:130–132.

Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, 1:351–370.

Steinsaltz, D. (1999). Locally contractive iterated function systems. *Annals of Probability*, 27:1952–1979.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1728.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Wang, G. (2022). Exact convergence analysis of the independent Metropolis-Hastings algorithms. *Bernoulli*, 28:2012–2033.

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Annals of Applied Probability*, 19:617–640.

Wu, K., Schmidler, S., and Chen, Y. (2022). Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23:12348–12410.

Yang, J., Roberts, G. O., and Rosenthal, J. S. (2020). Optimal scaling of random-walk Metropolis algorithms on general target distributions. *Stochastic Processes and their Applications*, 130:6094–6132.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Annals of Statistics*, 44:2497–2532.

Zhou, Q., Yang, J., Vats, D., Roberts, G. O., and Rosenthal, J. S. (2022). Dimension-free mixing for high-dimensional Bayesian variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84:1751–1784.

Zolotarev, V. M. (1984). Probability metrics. *Theory of Probability and Its Applications*, 28:278–302.