CPSC 661: Sampling Algorithms in Machine Learning February 8, 2020

Lecture 3

Lecturer: Andre Wibisono *Scribe: John Lazarsfeld*

1 Outline

Today's lecture is about the spectral theory (eigenvalue analysis) of reversible Markov chains. These lecture notes cover:

- the motivation for using spectral theory and a review of the $L^2(\nu)$ space and χ^2-divergence
- spectral properties of reversible Markov chains
- a mixing time result in χ^2-divergence for reversible chains.

2 Motivation and Preliminaries

We begin by briefly recalling our notation for Markov chains, reversibility, and mixing time before giving some motivation and intuition for using spectral theory to characterize convergence rate.

2.1 Markov chains, reversibility, and mixing time

Recall from earlier lectures that given a state space $\mathcal{X}$, a Markov chain P is a collection of probability measures on $\mathcal{X}$:

$$P = (P_x : x \in \mathcal{X}) .$$

Each P_x is the distribution of the next state X_1 conditioned on being currently in state $X_0 = x$, meaning

$$P_x(A) = \Pr(X_1 \in A | X_0 = x) .$$

Starting from $X_0 \sim \rho_0$, P defines the Markov chain $X_0 \to \cdots \to X_k \to X_{k+1} \to \cdots$ where the distribution $X_{k+1} \sim \rho_{k+1}$ is given by

$$\rho_{k+1}(A) = \int_{\mathcal{X}} P_x(A) \, d\rho_k(x).$$

If ν is a stationary distribution of P, then we say P is **reversible** with respect to ν if P and ν satisfy the following *detailed balance* equations:

$$\forall x, y \in \mathcal{X} : \quad \nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x).$$

If P is reversible, then ν is preserved under the operation of the Markov chain (i.e., reversibility implies the stationary condition). This follows from integrating the equation above over $x \in \mathcal{X}$.

Given a distance or divergence d between distributions, the ϵ-**mixing time** $\tau(\epsilon)$ of a Markov chain P with target distribution ν is the smallest number of steps k needed by the chain such that $d(\rho_k, \nu) \leq \epsilon$. Formally:

$$\tau(\epsilon) = \inf\{K : d(\rho_k, \nu) \leq \epsilon, \ \forall k \geq K\}.$$

2.2 Motivation for using spectral theory

To build some intuition as to why it is natural to use spectral theory to study the convergence rate of Markov chains, we can consider a simple example in the discrete setting where $\mathcal{X} = \{1, \ldots, n\}$. Here, the space of distributions over $\mathcal{X}$ is just the n-dimensional probability simplex Δ_n, and a Markov chain P is just an $n \times n$ stochastic matrix. As an example process: consider a random walk on a graph $G = (V, E)$ with adjacency matrix A and diagonal degree matrix D. Then the Markov chain P comes from the random walk matrix $P = D^{-1}A$.

In the discrete setting, if state $X_k \sim \rho_k$ and $X_{k+1}|X_k \sim P_{X_k}$, then $X_{k+1} \sim \rho_{k+1}$, where ρ_{k+1} and ρ_k are row vectors and

$$\rho_{k+1} = \rho_k \cdot P \ .$$

So the distribution at time $(k+1)$ is obtained via a right-multiplication of ρ_k by the matrix P and

$$\rho_k = \rho_0 \cdot P^k, \tag{1}$$

since we apply this multiplication k times starting from the initial distribution ρ_0. We are interested then in the asymptotic behavior of this iteration (1), and specifically we would like to characterize the rate of convergence $\rho_k \to \nu$, where ν is the stationary distribution of P. By looking at this iteration, we see this is simply the **power method** for computing the eigenvector corresponding to maximum eigenvalue of P, and this is where the connection to spectral theory comes from.

To be more concrete, say we assume P is symmetric and thus $P_x(y) = P_y(x)$ for all $x, y \in \mathcal{X}$ (in the random walk on G example, P is symmetric when G is regular). Then the spectral theorem for symmetric matrices tells us that P has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ with an orthonormal basis of eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$. Since P is a stochastic matrix, we further know $\lambda_1 = 1$ with corresponding eigenvector $v_1 = \nu$. In the case of symmetric P, the stationary distribution ν is uniform, and from here it is straightforward to use this spectral decomposition to characterize the convergence rate of the iteration (1) by the **spectral gap** $\lambda_1 - \lambda_2 = 1 - \lambda_2$.

In the general case when P is *not* symmetric, P might have complex eigenvalues and so we will not be able to use this nice decomposition. However, if we assume P is reversible with respect to ν, then we can consider the matrix $\widetilde{P}$ given by

$$\widetilde{P}(x, y) = \sqrt{\nu(x)} \cdot \frac{P(x, y)}{\sqrt{\nu(y)}}.$$

It is easy to check that the reversibility assumption implies $\widetilde{P}$ is symmetric, and this means we can apply the spectral theorem to this normalized matrix $\widetilde{P}$. For example, for a random walk on G, we have $\widetilde{P} = D^{-1/2}AD^{-1/2}$ since the stationary distribution of $P = D^{-1}A$ is proportional to the degree distribution of G. But when performing this change of basis and using the matrix $\widetilde{P}$, we are really working in the L^2 space with base measure ν.

So we will begin by reviewing the $L^2(\nu)$ space for general (not necessarily discrete) state spaces $\mathcal{X}$, and then we will develop an analogous eigenvalue analysis for reversible Markov chains P in order to characterize the convergence rate $\rho_k \to \nu$ under χ^2-divergence.

2.3 $L^2(\nu)$ space

Let ν be a probability distribution on a space $\mathcal{X}$ (which could be continuous or discrete). Then

$$L^2(\nu) = \left\{ f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} f(x)^2 d\nu(x) < \infty \right\},$$

which is the set of square-integrable functions on $\mathcal{X}$ with respect to the measure ν.[1] Note that $L^2(\nu)$ is a Hilbert space with **inner product** $\langle \cdot, \cdot \rangle_\nu$ given by

$$\langle f, g \rangle_\nu := \int_{\mathcal{X}} f(x)g(x)d\nu(x)$$

and squared norm

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

When clear, for ease of notation we will write $\langle f, g \rangle_\nu = \int fg \, d\nu$. Observe that this inner product is really an expectation with respect to ν, with

$$\langle f, g \rangle_\nu = \mathbb{E}_\nu [fg] \tag{2}$$

$$\|f\|_\nu^2 = \mathbb{E}_\nu \left[f^2 \right]. \tag{3}$$

We also make note of the following two useful properties that will be used later on in the lecture.

- Expectation of f:

$$\mathbb{E}_\nu [f] = \int_{\mathcal{X}} f(x) \cdot 1 \, d\nu(x) = \langle f, \mathbf{1} \rangle_\nu \tag{4}$$

  where $\mathbf{1} : \mathcal{X} \to \mathbb{R}$ is the constant 1 function (i.e., $\mathbf{1}(x) = 1$ for all $x \in \mathcal{X}$).

- Variance of f when $\mathbb{E}_\nu [f] = 0$:

$$\mathrm{Var}_\nu [f] = \mathbb{E}_\nu \left[(f - \mathbb{E}_\nu [f])^2 \right] = \mathbb{E}_\nu \left[f^2 \right] = \|f\|_\nu^2 . \tag{5}$$

[1] When $\mathcal{X}$ is infinite, there are some technical restrictions that have to be placed on the functions $f : \mathcal{X} \to \mathbb{R}$, and for example we should consider the functions f and g to be the same if $\int |f - g|^2 \, d\nu = 0$. We will ignore these technical issues in this lecture.

2.4 χ^2-divergence

When working in the $L^2(\nu)$ space, we want to define a a notion of "distance" between two distributions ρ and ν, and we can do this by measuring the squared norm of ρ in $L^2(\nu)$. We first write the densities (Radon-Nikodym derivatives)

$$h = \frac{d\rho}{d\nu} \quad \text{and} \quad \mathbf{1} = \frac{d\nu}{d\nu}$$

where both $h, \mathbf{1} \in L^2(\nu)$. Note also that

$$\mathbb{E}_\nu [h] = \int_{\mathcal{X}} h \, d\nu = \int_{\mathcal{X}} d\rho = \rho(\mathcal{X}) = 1.$$

We then define the **χ^2-divergence of ρ with respect to ν** to be

$$\chi^2_\nu(\rho) \triangleq \|h - \mathbf{1}\|^2_\nu. \tag{6}$$

Using (3), we can rewrite the RHS of the definition as an expectation

$$\chi^2_\nu(\rho) = \mathbb{E}_\nu \left[(h-1)^2\right] = \mathrm{Var}_\nu [h] = \mathrm{Var}_\nu \left[\tfrac{d\rho}{d\nu}\right], \tag{7}$$

where the second equality comes from our observation that $\mathbb{E}_\nu [h] = 1$. So working in the $L^2(\nu)$ space, the χ^2-divergence of ρ is the variance of its density with respect to the reference measure ν.

2.4.1 Drawbacks of χ^2-divergence

While working in the $L^2(\nu)$ space gives us some nice geometric notions, we should note two drawbacks of using χ^2-divergence to measure "distance" between distributions.

- First, χ^2-divergence is *not a metric* as it is not symmetric: in general $\chi^2_\nu(\rho) \neq \chi^2_\rho(\nu)$.
- The range of $\chi^2_\nu(\rho)$ can be large, for example growing exponentially in n when our space is $\mathcal{X} = \mathbb{R}^n$. To illustrate this, suppose $\rho = \mathcal{N}(0, \alpha I)$ and $\nu = \mathcal{N}(0, I)$ are both zero-mean Gaussian distributions with $\alpha > 0$. Then we can compute $\chi^2_\nu(\rho)$ to find

  $$\chi^2_\nu(\rho) = \begin{cases} \frac{1}{(\alpha(2-\alpha))^{n/2}} & \text{if } 0 < \alpha < 2 \\ \infty & \text{otherwise} \end{cases}.$$

  Figure 1 plots this relationship between $\chi^2_\nu(\rho)$ and α for several values of n. So even in this "nice" case where ρ and ν are both normal, for α between 0 and 2, the quantity $\chi^2_\nu(\rho)$ grows exponentially in n as α approaches 0 or 2.

We will nonetheless use χ^2-divergence as our notion of distance between distributions for our mixing time analysis in this lecture.

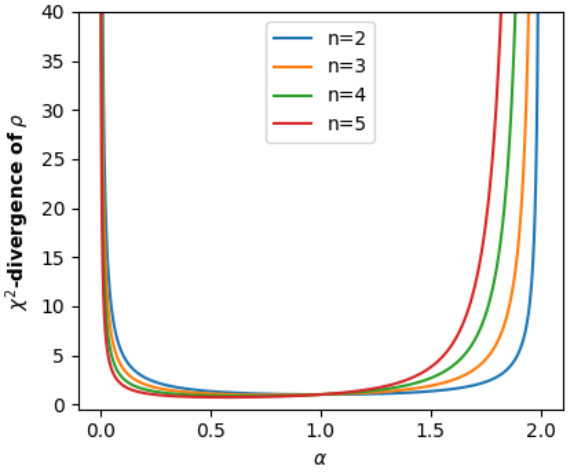


Figure 1: $\chi^2_\nu(\rho)$ vs. α when $\mathcal{X} = \mathbb{R}^n$, $\nu = \mathcal{N}(0, I)$ and $\rho = \mathcal{N}(0, \alpha I)$ where $0 < \alpha < 2$ for several values of n. χ^2-divergence grows exponentially in n as α approaches 0 and 2.

3 Spectral Properties of Reversible Markov Chains

3.1 Markov chains as linear operators

Let P be a reversible Markov chain wrt ν on space $\mathcal{X}$. Then we can think of P as defining a linear operator $P : L^2(\nu) \to L^2(\nu)$ mapping $f \mapsto Pf$, given by

$$(Pf)(x) = \int_{y \in \mathcal{X}} f(y)\, dP_x(y) \tag{8}$$

for $f \in L^2(\nu)$ and $x \in \mathcal{X}$. Since $P_x(\cdot)$ is the conditional probability distribution of state X_1 of the Markov chain given $X_0 = x$, we can see that Pf is really the conditional expectation $(Pf)(x) = \mathbb{E}[f(X_1) \mid X_0 = x]$.

Note that when $\mathcal{X}$ is discrete, the map $f \mapsto Pf$ is a matrix multiplication by P on the *left*:

$$\begin{bmatrix} \\ f \\ \\ \end{bmatrix} \mapsto \begin{bmatrix} \\ P \\ \\ \end{bmatrix} \begin{bmatrix} \\ f \\ \\ \end{bmatrix}.$$

On the other hand, recall that for a distribution ρ_k (a row vector), we obtain the distribution ρ_{k+1} via matrix multiplication by P on the *right* (i.e., the map $\rho_k \mapsto \rho_k P$):

$$\begin{bmatrix} \quad \rho \quad \end{bmatrix} \mapsto \begin{bmatrix} \quad \rho \quad \end{bmatrix} \begin{bmatrix} \\ P \\ \\ \end{bmatrix}.$$

When P is symmetric, these two maps are equivalent, but in general we are interested in P that may not be symmetric. However, when working in the $L^2(\nu)$ space and under the assumption that P is reversible, we can show that P is a self-adjoint operator (which is equivalent to having a symmetric transition matrix in the discrete setting). This in turn allows us to better characterize the spectral properties of P in the general case.

3.2 Reversibility implies symmetry

We will now state and prove the prior remark more formally:

Lemma 1. *Assume P is reversible with respect to ν. Then P is **self-adjoint**:*

$$\forall f, g \in L^2(\nu): \quad \langle f, Pg \rangle_\nu = \langle Pf, g \rangle_\nu \, . \tag{9}$$

Proof. By the definition of $\langle \cdot, \cdot \rangle_\nu$ we have for any $f, g \in L^2(\nu)$ that

$$\begin{aligned} \langle f, Pg \rangle_\nu &= \int f(x)(Pg)(x) \, d\nu(x) \\ &= \int f(x) \left(\int g(y) \, dP_x(y) \right) \, d\nu(x) \\ &= \int \int f(x) g(y) \, dP_x(y) \, d\nu(x), \end{aligned} \tag{10}$$

where the second inequality follows from the definition of Pg from (8). Now by reversibility, it follows that $dP_x(y) \, d\nu(x) = dP_y(x) \, d\nu(y)$, and substituting this back into (10) yields

$$\begin{aligned} \langle f, Pg \rangle_\nu &= \int \int f(x) g(y) \, dP_y(x) \, d\nu(y) \\ &= \int g(y)(Pf)(y) \, d\nu(y) \\ &= \langle Pf, g \rangle_\nu. \end{aligned}$$

□

3.3 Spectral Theorem for self-adjoint operators

The previous lemma is useful because a self-adjoint operator P has many nice properties — in particular, we have the following Spectral Theorem characterizing the eigenvalues of P. Note that we are assuming here our space $L^2(\nu)$ is separable (meaning it has a countable basis), which will always be true in our settings; see [Bakry et al., 2013, Appendix A.4] for the general/continuous spectrum.

Theorem 1. *A self-adjoint operator P on $L^2(\nu)$ has **real** eigenvalues $\lambda_i \in \mathbb{R}$:*

$$P\phi_i = \lambda_i \phi_i \tag{11}$$

for $i = 1, 2, \ldots, n$ $(n = \infty)$ with an orthonormal basis of eigenfunctions $\phi_i \in L^2(\nu)$:

$$\langle \phi_i, \phi_j \rangle_\nu = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

We can view Theorem 1 as the infinite-dimensional analog of the symmetric matrix decomposition when our Markov chain is given by the symmetric transition matrix $P \in \mathbb{R}^{n \times n}$. Now combining Lemma 1 and Theorem 1 tells us that if P is reversible with respect to ν, then P has real eigenvalues. Ultimately, our goal then is to bound the eigenvalues of P, and we can achieve this using the Courant-Fischer characterization.

3.4 Courant-Fischer characterization and bounding eigenvalues

The following theorem is sometimes referred to as the **min-max theorem**, the **variational theorem** or the **Courant-Fischer-Weyl min-max principle**. Note that for functions $f, g \in L^2(\nu)$, we write $f \perp g$ to denote that f and g are orthogonal (i.e., that $\langle f, g \rangle_\nu = 0$).

Theorem 2. *Assume P is self-adjoint with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ $(n = \infty)$ and corresponding eigenfunctions $\phi_1, \ldots, \phi_n \in L^2(\nu)$. Then:*

$$\lambda_1 = \max_{0 \neq f \in L^2(\nu)} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu} \tag{12}$$

$$\lambda_2 = \max_{\substack{0 \neq f \in L^2(\nu) \\ f \perp \phi_1}} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu} \tag{13}$$

$$\lambda_n = \min_{0 \neq f \in L^2(\nu)} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu} \ . \tag{14}$$

Our goal is to bound the spectrum of P by obtaining upper and lower bounds on λ_1 and λ_n respectively. Using the Courant-Fischer characterization, we can achieve this by deriving an upper and lower bound on the quantity $\langle f, Pf \rangle_\nu$ with respect to $\langle f, f \rangle_\nu$ for all f, as stated in this next lemma.

Lemma 2. *Assume P is reversible with respect to ν. Then for all $f \in L^2(\nu)$:*

(a) $\langle f, Pf \rangle_\nu \leq \langle f, f \rangle_\nu$

(b) $\langle f, Pf \rangle_\nu \geq -\langle f, f \rangle_\nu$.

Using the notation $A \preceq B$ to denote that $\langle f, Af\rangle_\nu \leq \langle f, Bf\rangle_\nu$ for all $f \in L^2(\nu)$, Lemma 2 says that $-I \preceq P \preceq I$. To bound the eigenvalues of P, observe that part (a) of the lemma combined with (12) of Theorem 2 implies that $\lambda_1 \leq 1$. Similarly, part (b) of the lemma combined with (14) implies $\lambda_n \geq -1$.

Proof. We first write

$$\begin{aligned}\langle f, f\rangle_\nu = \int_{\mathcal{X}} f(x)^2 \, d\nu(x) &= \frac{1}{2}\int_{\mathcal{X}} f(x)^2 \, d\nu(x) + \frac{1}{2}\int_{\mathcal{X}} f(y)^2 \, d\nu(y) \\ &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^2 \, dP_x(y) \, d\nu(x) + \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} f(y)^2 \, dP_y(x) \, d\nu(y) \qquad (15)\end{aligned}$$

where the last equality holds since $\int_{\mathcal{X}} dP_x(y) = 1$ for any x. By reversibility we know $dP_x(y)\, d\nu(x) = dP_y(x)\, d\nu(y)$, and so we rewrite (15) to yield

$$\langle f, f\rangle_\nu = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \left(f(x)^2 + f(y)^2\right) \, dP_x(y) \, d\nu(x). \qquad (16)$$

Note also that by the definition of Pf we have

$$\langle f, Pf\rangle_\nu = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x)f(y) \, dP_x(y) \, d\nu(x). \qquad (17)$$

To prove part (a) of the lemma, we will show $\langle f, f\rangle_\nu - \langle f, Pf\rangle_\nu \geq 0$. So using (16) and (17):

$$\begin{aligned}\langle f, f\rangle_\nu - \langle f, Pf\rangle_\nu &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \left(f(x)^2 + f(y)^2 - 2f(x)f(y)\right) \, dP_x(y) \, d\nu(x) \\ &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \left(f(x) - f(y)\right)^2 \, dP_x(y) \, d\nu(x) \\ &\geq 0\end{aligned}$$

since we have expressed the difference as a sum of squares. By a similar calculation we can show

$$\begin{aligned}\langle f, f\rangle_\nu + \langle f, Pf\rangle_\nu &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \left(f(x) + f(y)\right)^2 \, dP_x(y) \, d\nu(x) \\ &\geq 0\end{aligned}$$

which implies part (b) of the lemma and thus concludes the proof. □

So when P is reversible, the eigenvalues of P satisfy

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 1.$$

We remark that $\lambda_2 < \lambda_1 = 1$ when P is irreducible, which will always be true in our settings. Moreover it can be shown that $\lambda_n = -1$ only when the Markov chain has only two connected components (i.e., when the random walk graph is bipartite in the discrete case, and thus the chain has period 2 and never converges to ν). For our spectral mixing time analysis, we will be interested in the magnitude of the gaps between $\lambda_1 = 1$ and λ_2 and between λ_n and -1. We will now formalize these quantities as the *spectral gap* of P.

3.5 Spectral Gap

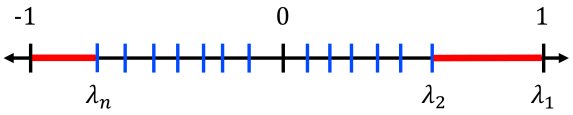
Assume P is reversible and irreducible with eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1.$$

We denote the **(absolute) spectral gap** of P by $\gamma^\star$ where

$$\gamma^\star = \min\{1 - \lambda_2, 1 + \lambda_n\}.$$

This absolute gap can be visualized in the following illustration as the smaller of the two red portions of the line.



Given our ordering on the λ_i, observe that all eigenvalues of P satisfy $|\lambda_i| \leq 1 - \gamma^\star$ for $i \geq 2$.

3.5.1 Lazifying P

Note that given reversible P, we can *lazify* P to obtain a new chain $\widehat{P}$ where all eigenvalues λ_i of $\widehat{P}$ are non-negative. We define the new chain by

$$\widehat{P} = \frac{1}{2}I + \frac{1}{2}P,$$

meaning at each step, with probability $1/2$ the chain remains in its current state, and with probability $1/2$ the chain transitions according to P. We can see that $\widehat{P} \succeq 0$ (meaning all $\lambda_i \geq 0$), by observing that for all f:

$$\begin{aligned}\langle f, \widehat{P}f\rangle_\nu = \frac{1}{2}\langle f, (I+P)f\rangle_\nu &= \frac{1}{2}\langle f, f\rangle_\nu + \frac{1}{2}\langle f, Pf\rangle_\nu \\ &\geq 0\end{aligned}$$

where the final inequality follows by part (b) of Lemma 2.

So we usually do not need to worry about the smallest eigenvalue λ_n of P, since we can always guarantee $\lambda_n \geq 0$ by lazifying P. This leads us to define the **spectral gap** γ of P to be

$$\gamma = 1 - \lambda_2,$$

meaning that $\gamma = \gamma^\star$ if P is lazy or otherwise positive semidefinite.

3.5.2 Laplacian and its spectral gap

As it will be used often in later lectures, here we define and make several remarks on the Laplacian of P and its spectral gap. Given P, we define the Laplacian $L : L^2(\nu) \to L^2(\nu)$ by

$$L = I - P.$$

If P is reversible, then by part (a) of Lemma 2, L is positive semidefinite ($L \succeq 0$) since for any $f \in L^2(\nu)$:

$$\langle f, Lf\rangle_\nu = \langle f, (I-P)f\rangle_\nu = \langle f, f\rangle_\nu - \langle f, Pf\rangle_\nu \geq 0.$$

As an alternative way of showing $L \succeq 0$, note that for any f we define the **Dirichlet form** $\mathcal{E}(f, f)$ by

$$\mathcal{E}(f, f) \triangleq \langle f, Lf\rangle_\nu,$$

and it follows from our earlier computation in the proof of Lemma 2 that

$$\mathcal{E}(f, f) = \tfrac{1}{2}\, \mathbb{E}_{(\nu,P)}\left[(f(X_1) - f(X_0))^2\right] \geq 0,$$

where the probability in the expectation is over $X_0 \sim \nu$ and $X_1|X_0 \sim P_{X_0}$.

The eigenvalues of L satisfy $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n < 2$ with corresponding eigenfunctions $\mathbf{1} = \phi_1, \phi_2, \ldots, \phi_n$, and we define the **spectral gap of L** to be

$$\gamma = \lambda_2(L) \;=\; \min_{\substack{0 \neq f \in L^2(\nu) \\ f \perp \mathbf{1}}} \frac{\langle f, Lf\rangle_\nu}{\langle f, f\rangle_\nu} \;=\; \min_{0 \neq f \in L^2(\nu)} \frac{\mathcal{E}(f, f)}{\mathrm{Var}_\nu\left[f\right]}.$$

4 Mixing time in χ^2-divergence

Using the spectral characterization of reversible chains P developed in the preceding section, we now state our desired mixing time result:

Theorem 3. *Let P be a reversible Markov chain reversible with respect to ν with spectral gap $\gamma\ (= \gamma^\star) > 0$. For any $X_0 \sim \rho_0$, along the Markov chain $X_k \sim \rho_k$:*

$$\chi^2_\nu(\rho_k) \leq (1-\gamma)^{2k} \cdot \chi^2_\nu(\rho_0)\ .$$

As $(1-\gamma) < 1$, the theorem states that the χ^2-divergence of ρ_k with respect to ν decreases exponentially fast in k, the number of steps taken by the chain. Letting $\tau(\epsilon, \rho_0)$ denote the mixing time of P with an initial distribution ρ_0, we have the following corollary:

Corollary 1. *To reach $\chi^2_\nu(\rho_k) \leq \epsilon$, it is enough to take*

$$\tau(\epsilon, \rho_0) = \frac{1}{2\gamma} \log \frac{\chi^2_\nu(\rho_0)}{\epsilon}.$$

Proof. The corollary follows by observing from Theorem 3 that

$$\chi^2_\nu(\rho_k) \;\le\; (1-\gamma)^{2k} \cdot \chi^2_\nu(\rho_0) \;\le\; e^{-2\gamma k} \cdot \chi^2_\nu(\rho_0),$$

where the second inequality holds since $1 - x \le e^{-x}$ for all x. Then to ensure $\chi^2_\nu(\rho_k) \le \epsilon$, it is sufficient to set

$$e^{-2\gamma k} \cdot \chi^2_\nu(\rho_0) \;\le\; \epsilon \quad \iff \quad k \;\ge\; \frac{1}{2\gamma} \cdot \log \frac{\chi^2_\nu(\rho_0)}{\epsilon}.$$

□

Note that this mixing time guarantee has a logarithmic dependence on the χ^2-divergence of the initial distribution ρ_0 with respect to ν. Recall from Section 2.4 that the range of χ^2_ν can be large, so this dependence on $\log \chi^2_\nu(\rho_0)$ can be non-negligible. For example, we saw that $\chi^2_\nu(\rho_0)$ is exponential in n when $\mathcal{X} = \mathbb{R}^n$ and ρ_0 and ν are Gaussian, and so $\log \chi^2_\nu(\rho_0) = O(n)$ in this case. However, we usually write this spectral mixing-time guarantee from Corollary 1 as $\tau = \tilde{O}(1/\gamma)$, where the tilde-O notation hides the logarithmic dependence on ϵ and $\chi^2_\nu(\rho_0)$.

We did not have time to prove Theorem 3 in this lecture, and so we defer this to the start of the next lecture.

References

[Bakry et al., 2013] Bakry, D., Gentil, I., and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media.