Everything about Markov Chains

Leran Cai
University of Cambridge
Cambridge, UK
leran.cai@cl.cam.ac.uk

Contents

Part I

Preliminaries

Chapter 1

Intuitions

1.1 Why intuitions, not introductions

In this first chapter, usually we should call it introduction. However I think a better word should be intuitions. To understand a topic like an expert, we need to gain intuitions so even for the newcomers they can understand complicated stuff very easily.

Together with basic concepts, we should explain them with simple intuitions so after ten years we may forget the concepts but we remember the feelings. Mathematicians like to write things formally, but we want to explain them intuitively without losing formality.

We should not waste too much time on presenting rigorous proofs most of the time (if we have time we can type them in to complete the book). Instead, we should read the original books/references and then present the intuitions/understandings here, for those are the most important things.

The main references for this book are [2], [3] and also Part 1B/3 courses at University of Cambridge.

1.2 Necessary concepts

Basic concepts explain why we are interested in such a topic.

1.2.1 Markov chains

We have a countable set of states. It is possible to stay at any of them and in each step we have a certain probability to jump from one state to another. Hence we need a set Ω and a transition matrix P to define a Markov chain.

1.2.2 Irreducibility

This concept means if we can always transit from one state to another state in the state space. If we can then the chain is irreducible, otherwise it is reducible. Intuitively if some states are not reachable from a state, then we can group those states and maybe compress them into one state. This is where the name comes from.

1.2.3 Periodicity

Remember the odd-even chain. That is why we are interested in the periodicity and also if a chain is periodic then we cannot talk about its mixing time easily.

1.2.4 Stationary/equilibrium distribution/in variant measure

Formally it is a probability measure on the state space Ω. Usually we call it π. This is the most important property of a Markov chain. It means that whenever you look at the state, the probability that it is on a certain state remains unchanged.

Reversibility This a property of π. It means $\pi(x)P(x,y) = \pi(y)P(y,x)$. Intuitively, this means that the edge cost you go from x to y is the same as the other way around.

1.2.5 Distances

To measure how far a probability distribution is from a stationary distribution, we use different measures.

Total variance/l_1 This is a state-wise distance. For each state, there is a difference between μ and ν. We sum them up.

1.3 Mixing times

Each time we do a transition it is equivalent to multiply a P to the current distribution. So in the end we are checking all rows in P^t. If the chain mixes, then all rows are the same.

It is essentially a *cutoff phenomenon*, which says that convergence to the equilibrium/stationary distribution usually happens abruptly asymptotically as $n \to \infty$. The time this convergence happens is called **mixing times**.

In other words, the distance between your distribution and the stationary distribution reduces to a very small value which is close to zero.

1.3.1 Methods to stuty mixing times

Probabilistic techniques: coupling, martingales, evolving sets. Spectral methods: eigenvalues and eigenfunctions, functional and geometric inequalities like Cheeger's inequality, Poincare and Nash inequalities. Representation theory. Statistical methods: Glauber dynamics for the Ising model.

Chapter 2

Formal Definitions and Basic Theorems

Mathematicians like formal things. So we need a chapter to present formal definitions of everything in this topic. In the previous chapter, we have seen the intuitions of them. Now we want to formally define them.

There is some "common knowledge" in every topic, which every expert should be familiar with but is not sufficient to form an entire chapter. We collect them all here to make a decent chapter.

(For now I do not have time to type all those things because I do not learn much from doing so. Let's assume one day in the future I can finish those things to make this book complete.)

Definition 2.1.

$$\mathcal{T}(x) = \{t : P^t(x, x) > 0\}$$

This is the set of return times. If the greatest common divisor of them is 1, then the chain is **aperiodic**.

Definition 2.2. *The total variation distance between μ and ν is*

$$\|\mu - \nu\|_{TV} = \sum_{A \subseteq \Omega} |\mu(A) - \nu(A)|$$

Also

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{s \in \Omega} |\mu(s) - \nu(s)|$$

Definition 2.3. *Let P be an irreducible, aperiodic transition matrix on a finite state space Ω, and let π denote its stationary distribution. Define the distance function for all $t = 0, 1, ...$ by:*

$$d(t) = \max_{x \in \Omega} \left\| P^t(x, \cdot) - \pi(\cdot) \right\|_{TV} \tag{2.1}$$

$d(t)$ is the total variation distance between the distribution of the Markov chain at time t and its equilibrium, started from the worst possible starting point x, so that if $d(t)$ is small we know that the chain is close to equilibrium no matter what was its starting point. The ergodic theorem implies that $d(t) \to 0$ as $t \to \infty$. In fact, elementary linear algebra tells us that, asymptotically as $t \to \infty$ (this can be found in [2]), the distance $d(t)$ decays exponentially fast, with a rate of decay control by the **spectral gap** of the chain.

Proposition 2.4. *$d(t)$ is non-increasing with time.*

Proof.

$$\begin{aligned}
\left\|P^{t+1}(x,\cdot)-\pi(\cdot)\right\|_{TV} &= \frac{1}{2}\sum_{i\in\Omega}\left|P^{t+1}(x,i)-\pi(i)\right| \\
&= \frac{1}{2}\sum_{i\in\Omega}\left|\sum_{j\in\Omega}\left(P^t(x,j)P(j,i)-\pi(j)P(j,i)\right)\right| \\
&= \frac{1}{2}\sum_{i\in\Omega}\left|\sum_{j\in\Omega}\left(P^t(x,j)-\pi(j)\right)P(j,i)\right| \\
&= \frac{1}{2}\left|\sum_{i\in\Omega}\sum_{j\in\Omega}\left(P^t(x,j)-\pi(j)\right)P(j,i)\right| \\
&= \frac{1}{2}\left|\sum_{j\in\Omega}\sum_{i\in\Omega}\left(P^t(x,j)-\pi(j)\right)P(j,i)\right| \\
&= \frac{1}{2}\left|\sum_{j\in\Omega}\left(P^t(x,j)-\pi(j)\right)\right| \\
&\leq \frac{1}{2}\sum_{j\in\Omega}\left|P^t(x,j)-\pi(j)\right| \\
&= \left\|P^t(x,\cdot)-\pi(\cdot)\right\|_{TV}
\end{aligned}$$

The inequality comes from $|a+b|\leq|a|+|b|$. To see this, our a or b is any $P^t(x,j)-\pi(j)$ since they might be positive or negative. □

Proposition 2.5. *Let ρ be defined by:*

$$\rho(t)=\max_{x,y\in\Omega}\left\|P^t(x,\cdot)-P^t(y,\cdot)\right\|_{TV}$$

Then

$$d(t)\leq\rho(t)\leq 2d(t)$$

Proof. Triangle inequality. □

$d(t)$ is the distance between P^t and π. $\rho(t)$ is the distance between $P^t(x,\cdot)$ and $P^t(y,\cdot)$. Essentially they are different measures of the distances.

Theorem 2.6 (The Convergence Theorem)**.** *P is irreducible and aperiodic, with stationary distribution π. There exists a constant $\alpha\in(0,1)$ and $C>0$ such that*

$$\max_{x\in\Omega}\left\|P^t(x,\cdot)-\pi\right\|_{TV}\leq C\alpha^t$$

Proof. It is irreducible and aperiodic. This guarantees that there exists r such that P^r has strictly positive entries. □

A state is said to be **recurrent** if $\mathbb{P}_x\left[\tau_x^+<\infty\right]=1$ and **transient** if it is less than 1. If $\mathbb{E}_x\left[\tau_x^+<\infty\right]$, it is said to be **positive recurrent**. If $\mathbb{E}_x\left[\tau_x^+=\infty\right]$ then it is **null recurrent**.

Theorem 2.7 (The ergodic theorem)**.** *If P is irreducible, aperiodic and positive recurrent, then for all starting distribution μ on S, then the Markov chain X started from μ converges to the* ***unique*** *stationary distribution π in the long run.*

Remark 2.8. The stationary probability can be not unique, the ergodic theorem states when it is unique.

Definition 2.9 (Expectation and Variance)**.** *Let $f : \Omega \to \mathbb{R}$ be any real-valued function. Define the expectation*

$$\mathbb{E}_\pi [f] = \sum_x \pi(x) f(x)$$

and the variance

$$\begin{aligned}
\mathrm{Var}_\pi[f] &= \sum_x \pi(x)(f(x) - \mathbb{E}_\pi [f])^2 \\
&= \sum_x \pi(x) f(x)^2 - (\mathbb{E}_\pi [f])^2 \\
&= \sum_x \pi(x) f(x)^2 \sum_y \pi(y) - \sum_x \pi(x) f(x) \sum_y \pi(y) f(y) \\
&= \frac{1}{2} \sum_{x,y} \pi(x)\pi(y)(f(x) - f(y))^2
\end{aligned}$$

and the entropy

$$\mathrm{Ent}_\pi[f] = \mathbb{E}_\pi \left[f \log \frac{f}{\mathbb{E}_\pi [f]} \right] = \mathbb{E}_\pi [f \log f - f \log \mathbb{E}_\pi [f]]$$

2.1 Examples

2.1.1 Example: Simple random walk on a graph

Consider a simple random walk on a graph $G = (V, E)$. For any vertex $y \in V$

$$\sum_{x \in V} \deg(x) P(x, y) = \sum_{x \sim y} \frac{\deg(x)}{\deg(x)} = \deg(y)$$

This satisfies the definition of the stationary distribution.

Part II

Probabilistic Methods

Chapter 3

Coupling Methods

The technique of coupling is one of the most powerful probabilistic tools to obtain quantitative estimates about mixing times.

The very intuition about coupling is that, we have an arbitrary starting distribution X and a nice distribution that we expect Y. We apply the same source of the randomness on them, and then prove that at some point t, they become equal with high probability.

Definition 3.1. *A coupling of μ and ν is the realisation of a pair of random variables (X, Y) on the same probability space such that $X \sim \mu$ and $Y \sim \nu$.*

So to construct a coupling we seek two random variables which have the correct distributions μ and ν respectively, but there is complete freedom over how they are correlated.

Coupling itself is not mysterious. X and Y can be correlated or not. As long as they have the correct distributions we are good. An important theorem here is:

Theorem 3.2. *For all couplings (X, Y) of μ and ν, we have:*

$$\|\mu - \nu\|_{TV} \leq \mathbb{P}[X \neq Y] \tag{3.1}$$

Furthermore, there always is a coupling (X, Y) which achieves equality.

Proof. We can have intuitive proof if we draw a picture:

□

What does this tell us? Whether or not $X = Y$ will occur with a high probability depends on the total variation distance. Intuitively, if the total variantion is large, it means that on some states, the probabilities of X and Y are much different. So they are unlikely to be the same. In other words, how far these two distributions are is upper bounded by the probability that they are different.

Proposition 3.3. *ρ is submultiplicative: for all $s, t \geq 0$:*

$$\rho(t + s) \leq \rho(t)\rho(s)$$

Proof. The proof uses the Markov property. Note that here it does not mean Markov's inequality. It means the memoryless property. □

3.1 Example: Random to top shuffling

This shuffle means we take the card at the position i and put it on the top of the deck.

Theorem 3.4. *The random-to-top chain exhibits cutoff at time* $t_{\text{mix}} = n \log n$.

Proof. The most important idea in this proof is that we need to prove two bounds: an upper bound which shows $d((1+\epsilon)n\log n) \to 0$ and an lower bound which shows $d((1-\epsilon)n\log n) \to 1$.

Upper bound. We should learn the idea here: we prove that with high probability X_t, which is the state of the deck after t moves, can be coupled with a uniform deck. In this way we can have the conclusion that t steps can make it converge.

We apply the coupling trick here. Each time we randomly pick a card number, then find them in the two decks, move them to the top. Note that the same card number may appear in different positions in these two decks. The intuition/trick here is to realize that once a card i has been chosen, their positions in the deck would become the same. The randomness is caused by selecting the number instead of the card, but the probability a card/number is chosen is the same.

Hence the time when the two decks are the same would be the time when all the cards have been chosen for at least once. This is a classical coupon collector problem. In the proof of the lecture notes at Cambridge, they use second moment method to prove the concentration.

Lower bound. The idea is that if we do not have enough time, then a lot of cards would remain in their original relative order.

This relies on the fact that we need all cards to be touched to make it uniform. Let A_j be the event that the j bottom cards are in their original relative order, meaning they remain untouched (if they are touched, they would have been moved to the top). For a uniform permutation, the probability of this event should be:

$$\pi(A_j) = \frac{1}{j!}$$

To see this, we know that there are $n!$ permutations. Then we fix j of them of the relative order, this gives us $\binom{n}{j}(n-j)! = n!/j!$. The quotient of them is just $1/j!$. Let

$$\tau_j = \inf\{t \geq 0 : \ j \text{ cards have been selected at least once}\}$$

We check the expected time when n^ϵ cards have not been touched. It is

$$\mathbb{E}[\tau_j] = \sum_{j=1}^{n-n^\epsilon} \frac{1}{p_j} = \frac{n}{n} + ... + \frac{n}{n^\epsilon} \sim n(\log n - \log n^\epsilon) = (1-\epsilon)n\log n$$

We use a similar proof, if we pick $t = (1-2\epsilon)n\log n$, then by concentration

$$\mathbb{P}[\tau_j \leq (1-2\epsilon)n\log n] \leq \frac{\text{Var}(\tau_j)}{\epsilon^2 \mathbb{E}[\tau_j]^2} \to 0$$

Note that we pick $j = n - n^\epsilon$. Let A be the event that n^ϵ cards have not been touched by time t. According to the above calculation, if we pick $t \leq (1-2\epsilon)n\log n$, then we know with probability $1 - o(1)$ we cannot touch all $n - n^\epsilon$ cards. Hence it means with high probability at least n^ϵ have not been touched. □

3.2 Example: random walk on a hypercube

A n-dimensional hypercube is $H_n = \{0,1\}^n$.

One special thing about this chain is that it is irreducible but not aperiodic. To solve it we consider l**azy chain or continuous chain**. We wait an exponential random variable with mean 1. The idea is we couple the walk with a uniform distribution.

Theorem 3.5. *$d_L(t)$ and $d_C(t)$ are the distances of the lazy chain and continuous time chain respectively. For any $\epsilon > 0$, $d_L((1+\epsilon)n\log n) \to 0$ and $d_C((1/2)(1+\epsilon)n\log n) \to 0$.*

Proof. For the lazy chain, we use a coupling trick. Y_0 is uniform on H_n. Couple X_t and Y_t as follows: pick $1 \le i \le n$ at random and flip a coin at every step. If $X_t(i) = Y_t(i)$, this means the ith bits are the same for them, if the coin is head we flip them simultaneously otherwise we do not flip. If they are not the same, then flip $X_t(i)$ if the coin is head and $Y_t(i)$ if it is tail. So with probability $1/2$ that we flip a bit. The trick is that they have the same source of randomness. Once a bit has been chosen, they become the same for X and Y. So it is a coupon collector problem. □

3.3 Example: couplings for random transpositions

It is a Markov chain on the symmetric group S_n. For a deck of cards, in each step, we swap any two of them. These two cards can be the same. A deck of cards can be seen as one permutation $\sigma \in S_n$.

The main reference here is chapter 9 of Diaconis [1]. Use some marking schemes to boundthe mixing time.

Part III

Advanced Methods

Chapter 4

Spectral methods

Chapter 12 of [2] is the main reference for this chapter.

Proposition 4.1. *Let P be a transition matrix.*

1. *If λ is a possibly complex eigenvalue, then $|\lambda| \leq 1$*
2. *If P is irreducible then the eigenspace with $\lambda = 1$ is one-dimensional and is generated by $\mathbf{1}$*
3. *If P is irreducible and aperiodic then -1 is not an eigenvalue*

Proof. For the third item, we need to know that in graph theory, G is a bipartite graph if and only if -1 is an eigenvalue of its adjacency matrix $A(G)$. Since $P = D^{-1}A$, this is intuitively true if we have a regular graph.

The idea is to assume -1 is an eigenvalue and we derive that $\mathcal{T}(x) \subseteq 2\mathbb{Z}$. In other words, we prove that it is not aperiodic and also it is essentially the same as proving it is bipartite.

First of all for odd t we have

$$\sum_y P^t(x,y)f(y) = -f(x) \Rightarrow \sum_y P^t(x,y)\frac{f(y)}{f(x)} = -1$$

We pick $f(x) = \|f\|_\infty$. Then we can prove that all the entries in f have the same absolute value. Without loss of generality we assume they are all 1 or -1. Furthermore, we have noticed that when $x = y$, $f(y)/f)(x)$ is 1, hence $P^t(x,x)$ must be zero otherwise we cannot get -1 on the right hand side. Hence we proved that only even number t can satisfy such things. □

Note that some times we write $\langle f, f_1\rangle_\pi = 0$ as $\mathbb{E}_\pi[f]$, which is a very commonly used expression and it makes a lot more sense when considering the space intuition.

The general logic here is that to use the spectral information, we need to decompose the original transition matrix. Then multiplying them is equivalent to multiplying their eigenvalues.

To understand this type of methods better, we should introduce in functional analysis. One needs to view the transition matrix P as an operator on functions $f : \Omega \to \mathbb{R}$ by setting

$$(Pf)(x) = \sum_y P(x,y)f(y)$$

Here f is a function on the state space. The inner product on real-valued functions on Ω is defined as

$$\langle f, g\rangle_\pi = \sum_{x\in\Omega} f(x)g(x)\pi(x)$$

Eigenfunctions means that $Pf(x) = \lambda f(x)$.

Theorem 4.2. *Assume that π is reversible with respect to P. Then:*

- *There exists a set of eigenfunctions $f_1, ..., f_n$ which are orthonormal for $\langle\cdot,\cdot\rangle_\pi$ and f_1 is the constant vector $(1, ..., 1)^T$.*
- *P^t can be decomposed as:*

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{j=1}^{n} f_j(x)f_j(y)\lambda_j^t$$

Proof. This form guarantees that the new matrix is symmetric. Then we apply the classical spectral theorem.

$$A(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}P(x,y)$$

Then A is symmetric. It can be decomposed into

$$A = \sum_i \lambda_i\phi_i\phi_i^T$$

By some basic calculations we find

$$f_i(x) = \frac{\phi_i(x)}{\sqrt{\pi(x)}}, i.e., f_i = D_\pi^{-1/2}\phi_i$$

I made a mistake here, I accidently wrote $P = \sum_i \lambda_i f_i f_i^T$ but this is wrong because P is not syymetric. We can have

$$P(x,y) = \sqrt{\frac{\pi(y)}{\pi(x)}}A(x,y) = \sqrt{\frac{\pi(y)}{\pi(x)}}\sum_i \lambda_i\phi_i(x)\phi_i(y)$$

Then we get what we have easily. □

4.1 The spectral gap and the relaxation time

Definition 4.3. *P is irreducible aperiodic. $\lambda_* = \max|\lambda| : \lambda \neq 1$. $\gamma_* = 1 - \lambda_*$ is called the absolute spectral gap, and $\gamma = 1 - \lambda_2$ is called the spectral gap of P. The relaxation time t_{rel}*

$$t_{\text{rel}} = \frac{1}{\gamma_*}$$

The important thing is to understand where these definitions come from and where they are used. The special thing is that this relaxation time is defined by the absolute spectral gap. Why this is a time? Why do we call it relaxation? Also why we define γ as $1 - \lambda$?

Definition 4.4 (l_2 distance)**.**

$$d_2(t) = \sup_{x\in\Omega} \left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - \mathbf{1} \right\|_2 = \sup_{x\in\Omega} \left(\sum_{y\in\Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right|^2 \right)^{1/2}$$

We've seen this many times. We never discuss the intuition. Note that when talking about the spectral things, we use a norm $l^p(\pi)$.

$$\|f\|_{p,\pi} = \left(\sum_x |f(x)|^p \pi(x) \right)^{1/p}$$

An important thing is that l_1 distance is **dominated** by an l_2 distance. What does dominate mean?

Lemma 4.5. *Assume P is irreducible, aperiodic. Then $d(t) \le (1/2)d_2(t)$.*

Proof.

$$\begin{aligned} 2d(t) = 2\left\|P^t(x,\cdot) - \pi\right\|_{TV} &= \sum_{y\in\Omega} |P^t(x,y) - \pi(y)| \\ &= \sum_{y\in\Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right| \\ &= \left\| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right\|_{1,\pi} \end{aligned}$$

Note that this can be also seen as an expectation. Taking the square and using Jensen's inequality: $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$ where $f(X) = X^2$

$$\begin{aligned} 4d^2(t) = 4\left\|P^t(x,\cdot) - \pi\right\|_{TV}^2 &= \mathbb{E}_\pi \left[\left| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right| \right]^2 \\ &\le \mathbb{E}_\pi \left[\left| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right|^2 \right] \\ &\le \sup_{x\in\Omega} \left(\sum_{y\in\Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - \mathbf{1} \right|^2 \right) \\ &= d_2^2(t) \end{aligned}$$

□

Why do we care about this l_2 norm and why do we have a minus 1? If we expand the definition of l_2 it does not give us any intuition. So we need to use the previous decomposition. Also the main purpose of this definition serves to show the relationship between the relaxation time and the mixing time.

Theorem 4.6. *Fix $0 < \epsilon < 1$ arbitrary. Assume that P is aperiodic, irreducible and reversible with respect to π. Then*

$$(t_{\text{rel}} - 1)\log\left(\frac{1}{2\epsilon}\right) \le t_{\text{mix}}(\epsilon) \le t_{\text{rel}} \log\left(\frac{1}{2\epsilon\sqrt{\pi_{\min}}}\right)$$

Proof. **For the upper bound**, we expand the l_2 norm:

$$\begin{aligned}
\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - \mathbf{1} \right\|_{2,\pi}^2 &= \left\| \sum_{j=1}^{n} f_j(x) f_j(\cdot) \lambda_j^t - \mathbf{1} \right\|_{2,\pi}^2 \\
&= \left\| \sum_{j=2}^{n} f_j(x) f_j(\cdot) \lambda_j^t \right\|_{2,\pi}^2 \\
&= \sum_{y\in\Omega} \pi(y) \left| \sum_{j=2}^{n} f_j(x) f_j(y) \lambda_j^t \right|^2 \\
&= \sum_{y\in\Omega} \pi(y) \left(\sum_{j=2}^{n} f_j^2(x) f_j^2(y) \lambda_j^{2t} + \sum_{j\neq k} f_j(x) f_j(y) \lambda_j^t f_k(x) f_k(y) \lambda_k^t \right)
\end{aligned}$$

Fix j, k in the second term, and $\sum_{y\in\Omega} \pi(y) f_j(y) f_k(y) = 0$ because fs are orthonormal.

$$= \sum_{y\in\Omega} \pi(y) \sum_{j=2}^{n} f_j^2(x) f_j^2(y) \lambda_j^{2t}$$

Again, since fs are normalized

$$\begin{aligned}
&= \sum_{j=2}^{n} f_j^2(x) \lambda_j^{2t} \\
&\leq \lambda_*^{2t} \sum_{j=2}^{n} f_j^2(x)
\end{aligned}$$

Note that here it maybe not λ_2. So how large is $\sum_{j=2}^{n} f_j^2(x)$? To see the result, we need to observe when we would have $f_j^2(x)$. We want a vector decomposed on fs.

$$\delta_x = \sum_{j=1}^{n} \langle \delta_x, f_j \rangle_\pi f_j = \sum_{j=1}^{n} f_j(x) \pi(x) f_j$$

Remember that this trick of decomposition onto different directions when having eigen bases is very commonly used in the literature.

Hence we find that

$$\langle \delta_x, \delta_x \rangle_\pi = \sum_{j=1}^{n} f_j^2(x) \pi^2(x) \langle f_j, f_j \rangle_\pi = \sum_{j=1}^{n} f_j^2(x) \pi^2(x) = \pi^2(x) \sum_{j=1}^{n} f_j^2(x)$$

Since we know that $\langle \delta_x, \delta_x \rangle_\pi = \pi(x)$, we derive that $\sum_{j=2}^{n} f_j^2(x) = 1/\pi(x)$. Therefore by combining the previous results we have

$$
\begin{aligned}
4d^2(t) = 4\left\|P^t(x,\cdot)-\pi\right\|_{TV}^2 &\le \sup_{x\in\Omega}\left(\sum_{y\in\Omega}\pi(y)\left|\frac{P^t(x,y)}{\pi(y)}-\mathbf{1}\right|^2\right)\\
&\le \lambda_*^{2t}\sum_{j=2}^{n} f_j^2(x)\\
&\le \lambda_*^{2t}\frac{1}{\pi(x)}\\
&\le \lambda_*^{2t}\frac{1}{\pi_{\min}(x)}\\
&\le (1-\gamma_*)^{2t}\pi_{\min}^{-1}\\
&\le e^{-2\gamma_* t}\pi_{\min}^{-1}
\end{aligned}
$$

Hence maximizing over x and taking the square root, we get

$$d(t) \le \frac{1}{2}e^{\gamma_* t}\sqrt{\pi_{\min}^{-1}}$$

Solving for the right-hand side equal to ϵ gives us $d(t)\le\epsilon$ as soon as $t \ge \frac{1}{\gamma_*}\log\left(\frac{1}{2\epsilon\sqrt{\pi_{\min}}}\right)$.

Note that at first we define $\gamma = 1-\lambda$ but then use $1-\gamma$ back to represent λ. The reason lies above: we need a cleaner expression.

Till now we've seen why we want γ_* and why it directly influences the mixing rate. Also we see how important our l_2 norm is in our field.

For the lower bound, we need it to show the bound is **tight**, this is very important. To prove a lower bound we need to find something like $d(t)\ge$ .... The l_2 norm is not going to be used. So where do we start? Maybe we start from the definition of $2d(t) = \sum_{y\in\Omega}|P^t(x,y)-\pi(y)|$. We pick an eigenfunction orthogonal to f_1, then since $\mathbb{E}_\pi[f] = \sum_{y\in\Omega}\pi(y)f(y) = 0$.

$$|\lambda^t f(x)| = |P^t f(x)| = \left|\sum_{y\in\Omega}P^t(x,y)f(y)-\pi(y)f(y)\right| \le 2\|f\|_\infty d(t)$$

Taking $f(x) = \|f\|_\infty$. We can obtain

$$|\lambda|^t \le 2d(t)$$

and take $\lambda = \lambda_*$ evaluating at $t = t_{\text{mix}}(\epsilon)$ we have

$$\lambda_*^{t_{\text{mix}}(\epsilon)} \le 2\epsilon$$

hence

$$\frac{1}{2\epsilon} \le \frac{1}{\lambda_*^{t_{\text{mix}}(\epsilon)}}$$

If we take the log we would get

$$t_{\text{mix}}(\epsilon) \ge -\log\left(\frac{1}{2\epsilon}\right)\frac{1}{\log(1-\gamma_*)}$$

Using $-(1-x)\log(1-x)\le x$ for $x\in[0,1]$ on γ_*, we have

$$t_{\text{mix}}(\epsilon) \ge \log\left(\frac{1}{2\epsilon}\right(t_{\text{rel}}-1)$$

as desired. □

4.2 Examples

4.2.1 Example: Random walk on the cycle

Since we've developed the spectral method, we see how to use it. Clearly we need eigenvalues and eigenfunctions of the transition matrix. For circle, this is not hard, we omit the proof here and give the eigenfunctions/eigenvales.

For cycle, the n eigenfunctions/values are $\phi_j(z) = z^j$ is an eigenfunction with eigenvalue $\cos(2\pi j/n)$ for $1 \le j \le n$. If n is even, then the chain is periodic and the absolute $(1 - \lambda_*)$ spectral gap is 0. If it is odd, then the absolute spectral gap is $\gamma_* = 1 - \cos(\pi/n) \approx \pi^2/2n^2$. So $t_{\text{rel}} = 2n^2/\pi^2$. However we have $\pi_{\min} = 1/n$ so in Theorem 4.6, the upper bound is $n^2 \log n$ which does not match the lower bound.

To fix the problem, we use all the information from all eigenvalues. We omit the proof because there is no more clever things. Maybe later we come back to this as a practice.

4.2.2 Example: Random walk on the hypercube

Finding the eigenvalues and eigenfunctions is a bit complicated. It also requires some familiarity of the product chains. We can go back to some technical details later. For now just memorize $\gamma_* = 1/n$.

4.3 Spectra and their matrices

Graph/Markov chain	Matrix	spectra
d-regular	A, adjacency	$\lambda_1 = d$
any finite Markov chain	P, transition matrix	$\lvert\lambda\rvert \le 1$
irreducible, aperiodic	P, transition matrix	-1 is not an eigenvalue
bipartite (periodic)	P, transition matrix	-1 is an eigenvalue
d-regular	lazy random walk	positive semi-definite
any graph	$L = D - A$, Laplacian	positive semi-definite

Note that the PSD property exists for lazy walks because we can prove something like $x^T P x = \sum c_{ij}(x_i + x_j)^2$.

Chapter 5

Comparison Methods and Geometric Methods

First why do we need comparison methods in the first place? In many cases, computing the spectral gap explicitly is hard. So the spectral gap has to be estimated. Amon all the comparison methods, geometric methods are the most important ones. Common methods for estimating the spectral gap is: canonical paths of Diaconis and Saloff-Coste, which gives a Poincaré inequality and thus an estimate of the spectral gap by a path counting argument. The second is Cheeger's inequality which relates the spectral gap to bottleneck ratios.

Definition 5.1. *Let $f, g : \Omega \to \mathbb{R}$. The **Dirichlet form** associated with a **reversible** P is defined by*

$$\begin{aligned}\mathcal{E}(f,g) &= \langle (I-P)f, g\rangle_\pi \\ &= \sum_x \pi(x)[f(x) - Pf(x)]g(x) \\ &= \sum_x \pi(x)\left[\sum_y P(x,y)(f(x)-f(y))\right] g(x) \\ &= \sum_{x,y} \pi(x)P(x,y)g(x)(f(x)-f(y))\end{aligned}$$

Remark 5.2. Compared with the variance definition in Definition 2.9, the Dirichlet form is the **local variance** by considering only adjacent pairs.

When P is reversible, $\pi(x)P(x,y) = \pi(y)P(y,x)$ in the last line and we have

$$\mathcal{E}(f,g) = \sum_{x,y} \pi(y)P(y,x)g(x)(f(x)-f(y))$$

We swap x, y which gives us

$$\mathcal{E}(f,g) = \sum_{x,y} \pi(x)P(x,y)g(y)(f(y)-f(x)) = \sum_{x,y} \pi(x)P(x,y)(-g(y))(f(x)-f(y))$$

Summing them yields

$$\mathcal{E}(f,g) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))(g(x)-g(y))\pi(x)P(x,y)$$

a much more useful expression.

When $f = g$,

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\pi(x)P(x,y)$$

Because P is reversible, the operator $f \to Pf$ is **self-adjoint** on l_2 with eigenvalues $1 = \lambda_1 > \lambda_2 \geq ... \geq \lambda_{|\Omega|-1} \geq -1$.

We call

$$Q(e) = \pi(x)P(x,y)$$

and

$$\nabla f(e) = f(y) - f(x)$$

These are more like potential functions.

$$\mathcal{E}(f,g) = \frac{1}{2}\sum_{e} Q(e)\nabla f(e)\nabla g(e) = \int_D \nabla f(e)\nabla g(e)$$

When $f = g$, this energy **measures how rough or how smooth the function is**.

The following variational characterisation/minimax characterization of the spectral gap in terms of the Dirichlet form is very useful. **It essentially links the spectral gap with the Dirichlet form!** Then using the Dirichlet form we convert P to be comparable with $\tilde{P}$.

Theorem 5.3 (Variational characterisation)**.** (P,π) *is reversible, then*

$$1-\lambda_2 = \gamma = \min_{\substack{f:\Omega\to\mathbb{R} \\ \mathbb{E}_\pi[f]=0, \|f\|_{2,\pi}=1}} \mathcal{E}(f,f) = \min_{\substack{f:\Omega\to\mathbb{R} \\ \mathbb{E}_\pi[f]=0}} \frac{\mathcal{E}(f,f)}{\|f\|_{2,\pi}^2}$$

Proof. Check [2]. $\mathbb{E}_\pi[f] = 0$ is equivalent to saying $f \perp_\pi \mathbf{1}$. It involves diagonalization and since we have $I-P$ in the Dirichlet form we get $1-\lambda_2 = \gamma$. Let $n = |\Omega|$, and $f_1, ..., f_n$ are the eigenfunctions of P corresponding to the decreasing ordered eigenvalues and are orthonormal. $f_1 = \mathbf{1}$. Hence, if $\|f\|_{2,\pi} = 1$ and $f \perp_\pi \mathbf{1}$, then $f = \sum_{j=2}^n a_j f_j$ (because it is perpendicular to f_1 so it can be only represented by these eigenvectors) and notice that after we take l_2-norm on both sides we have $\sum_{j=2}^n a_j^2 = 1$ (because the forms like $a_i a_j \langle f_i, f_j\rangle_\pi = 0$ and $\langle f_i, f_i\rangle_\pi = \|f_i\|_{2,\pi}^2 = 1$). Thus,

$$\langle (I-P)f, f\rangle_\pi = \sum_{j=2}^n a_j^2(1-\lambda_j) \geq (1-\lambda_2)$$

As for the third equality, note that we can define $\tilde{f} := f/\|f\|_{2,\pi}$ and replace it in the second equality. We know that $\left\|\tilde{f}\right\|_{2,\pi} = 1$ and we can find that $\mathcal{E}\left(\tilde{f},\tilde{f}\right) = \mathcal{E}(f,f)/\|f\|_{2,\pi}^2$. It follows from the standard variational characterization of eigenvalues (minimax theorem/The Courant-Fischer Theorem) of symmetric matrices; since P is not necessarily symmetric, but is reversible, and hence similar to a symmetric matrix, the standard formula has to be suitably weighted by the principal eigenvector π. □

Remark 5.4. Try to remember that $\mathbb{E}_\pi[f] = 0$ means it is perpendicular to the first eigenvector and also

$$\mathrm{Var}_\pi[f] = \mathbb{E}_\pi\left[(f-\mathbb{E}_\pi[f])^2\right] = \mathbb{E}_\pi\left[f^2\right] = \sum_x \pi(x)f^2(x) = \|f\|_{2,\pi}^2$$

. Based on the variational characterization theorem we know that for $f \perp_\pi \mathbf{1}$,

$$\gamma \|f\|_{2,\pi}^2 \leq \mathcal{E}(f,f)$$

hence

$$\mathrm{Var}_\pi[f] \leq \frac{1}{\gamma}\mathcal{E}(f,f)$$

5.1 Geometric Bounds

How do we use the comparison technique? For two **reversible** transition matrix $P, \tilde{P}$. P, π is the chain of interest with edge set $E = \{(x,y) : P(x,y) > 0\}$, $\tilde{P}, \tilde{\pi}$ is the chain with **known eigenvalues** with edge set $\tilde{E} = \{(x,y) : \tilde{P}(x,y) > 0\}$. For each pair $x \neq y$ with $\tilde{P}(x,y) > 0$, fix a sequence of steps $x = x_0, ..., x_k = y$ with $P(x_i, x_{i+1}) > 0$ called **E-path** denoted as Γ_{xy} with length $|\Gamma_{xy}| = k$.

$\mathcal{E}(f,f)$ is the Dirichlet form corresponding to P and $\tilde{\mathcal{E}}(f,f)$ is the Dirichlet form corresponding to $\tilde{P}$.

Definition 5.5 (congestion ratio)**.** *Supposing for each $(x,y) \in \tilde{E}, \tilde{P}(x,y) > 0$, there is an E-path from x to y, Γ_{xy}. **The congestion ratio** is defined as*

$$C = \max_{e \in E}\left(\frac{1}{Q(e)}\sum_{\substack{x,y \\ \Gamma_{xy} \ni e}} \tilde{Q}(x,y)|\Gamma_{xy}|\right)$$

Intuitively, we fix an edge $e \in E$. Then we look at all the edges $(x,y) \in \tilde{E}$ with their flows $\tilde{Q}(x,y)$ and for each of them we find one path from x to y in E which goes through e. It is important to write $\Gamma_{xy} \ni e$ because here we sum all possible paths that pass e with ending points x, y. To **maximize** C, we hope that $Q(e)$ is small and the number of possible paths $\Gamma_{xy} \ni e$ is large. In other words, e has many paths through it in E, and its ergodic flow is small.

Lemma 5.6. *Let P and $\tilde{P}$ be reversible transition matrices with stationary distributions π and $\tilde{\pi}$. If $\tilde{\mathcal{E}}(f,f) \leq \alpha\mathcal{E}(f,f)$ for all f, then*

$$\tilde{\gamma} \leq \left[\max_{x \in \Omega}\frac{\pi(x)}{\tilde{\pi}(x)}\right]\alpha\gamma$$

Proof. The proof is ommited for now. □

Theorem 5.7 (The canonical paths method)**.** *Given **a choice** of E-paths, use C as defined in Definition 5.5 for all functions $f : \Omega \to \mathbb{R}$,*

$$\tilde{\mathcal{E}}(f,f) \leq C\mathcal{E}(f,f)$$

Consequently,

$$\tilde{\gamma} \leq \left[\max_{x \in \Omega}\frac{\pi(x)}{\tilde{\pi}(x)}\right]C\gamma \tag{5.1}$$

Proof. Define $e = (z,w), \nabla f(e) = f(w) - f(z)$. One trick in this proof used the trick that $f(x) - f(y) = \sum_{e \in \Gamma_{xy}} \nabla f(e) = (f(x_k) - f(x_{k-1})) + (f(x_{k-1}) - f(x_{k-2})) + ... + (f(x_1) - f(x_0))$. Hence

$$
\begin{aligned}
2\tilde{\mathcal{E}}(f,f) &= \sum_{x,y}\tilde{Q}(x,y)(f(x)-f(y))^2 = \sum_{x,y}\tilde{Q}(x,y)\left(\sum_{e\in\Gamma_{xy}}\nabla f(e)\right)^2 \\
&\leq \sum_{x,y}\tilde{Q}(x,y)|\Gamma_{xy}|\sum_{e\in\Gamma_{xy}}(\nabla f(e))^2 = \sum_{e\in E}\left[\sum_{\Gamma_{xy}\ni e}\tilde{Q}(x,y)|\Gamma_{xy}|\right](\nabla f(e))^2 \\
&\leq \sum_{e\in E} C\cdot Q(e)(\nabla f(e))^2 = 2C\mathcal{E}(f,f)
\end{aligned}
\tag{5.2}
$$

Note the technical details here: first we start from the Dirichlet form $\tilde{\mathcal{E}}(f,f)$, we do the sum for all x,y (because of the definition of the Dirichlet form). Then we just pick **one single** arbitrary E-path Γ_{xy}. Why can we do $E-$path when calculating the Dirichlet form for $\tilde{P}$? Because this f is the same function for both P and $\tilde{P}$, hence $(f(x)-f(y))^2$ part would not be influenced. At first $(f(x)-f(y))^2$ is from the definition, then the second equality comes from its nature of being a potential. Hence this is **the key point we link $\tilde{P}$ back to P.**

Then we use Cauchy-Schwarz. Note that here the detail is:

$$
\begin{aligned}
\left(\sum_{e\in\Gamma_{xy}}\nabla f(e)\right)^2 &= \left(\sum_{e\in\Gamma_{xy}}\nabla f(e)\cdot 1\right)^2 \\
&\leq \left(\sum_{e\in\Gamma_{xy}}(\nabla f(e))^2\right)\left(\sum_{e\in\Gamma_{xy}}1^2\right) \\
&\leq |\Gamma_{xy}|\sum_{e\in\Gamma_{xy}}(\nabla f(e))^2
\end{aligned}
$$

Next is a technical trick because the sum for all x,y basically means all pairs of nodes. Since the graph is connected, all $|\Gamma_{xy}|\geq 1$. For sure all the edges in P will be covered because we exhaust all the paths, at least we can pick the two end point of each edge as our x,y. The last sum changes the perspective, it says for each edge e we count how many paths actually cover it. In the edge, we sum the same $(\nabla f(e))^2$s.

The last non-trivial detail is that for the definition of Dirichlet form, we notice that the range of sum can be changed from all x,y to $e\in E$ in general. This is because $Q(x,y)=0$ if the edge (x,y) is not in the edge set. However this might cause confusion in the above calculation if we use $e\in E$ style at the beginning even though it is a true equality. To see this, imagine we have,

$$
\sum_{(x,y)\in\tilde{E}}\tilde{Q}(x,y)|\Gamma_{xy}|\sum_{e\in\Gamma_{xy}}(\nabla f(e))^2 = \sum_{e\in E}\left[\sum_{\Gamma_{xy}\ni e}\tilde{Q}(x,y)|\Gamma_{xy}|\right](\nabla f(e))^2
$$

This looks like we miss some $e\in E$, though this does not influence the bound and the correctness of the above equality. Imagine we have $(u,v)\in E$ but u is only connected with v, and $(u,v)\notin\tilde{E}$, then we would never check any $E-$path going through (u,v) because they will not be chosen as ending points and since u is not connected to any other vertices in E, no other $E-$path would use it. So in the first sum, there is no way that $(f(v)-f(u))^2$ would appear. Remember this will not influence the equality even though in the second sum $(u,v)\in E$ is considered. Since none of our previous Γ_{xy}s contain such an edge, it is just 0 in the second sum. Anyway, I am being a bit harsh.

□

Since we mentioned that to upper bound the mixing time, we want a lower bound on the spectral gap. We can have all the information we have except γ in the equation 5.1. $\tilde{\gamma}$ is known since we know the eigenvalues of $\tilde{P}$. We have got the lower bound of γ!

Definition 5.8. *P satisfies **Poincaré inequality** with constant C if, for all functions $f : \Omega \to \mathbb{R}$,*

$$\mathrm{Var}_\pi [f] \leq C\mathcal{E}(f, f)$$

Notice that

$$\mathrm{Var}_\pi [f] \leq \frac{1}{\gamma}\mathcal{E}(f, f)$$

from the variational characterization. Hence to make sure this always holds, $C \geq 1/\gamma$, meaning $\gamma \geq 1/C$.

The Poincaré inequality is a control on the spectral gap. It is only **sharp up to logarithms** when we use the standard relation between spectral gap and mixing times (Theorem 4.6). Here sharp means something similar to tight.

Corollary 5.8.1. *Let P be reversible and irreducible with π. Suppose Γ_{xy} is a choice of E-path for each x, y and let*

$$C = \max_{e \in E} \left(\frac{1}{Q(e)} \sum_{\substack{x,y \\ e \in \Gamma_{xy}}} \pi(x)\pi(y)|\Gamma_{xy}| \right)$$

Then the spectral gap satisfies $\gamma \geq C^{-1}$. The Poincaré inequality holds with this C.

Proof. One important thing is to notice that we apply a trick here: $\tilde{P}(x, y) = \pi(y)$ and hence $\tilde{\pi} = \pi$! So the probability flow $\tilde{Q}(x, y) = \tilde{\pi}(x)\tilde{P}(x, y) = \pi(x)\pi(y)$.

$$\begin{aligned}
\tilde{\mathcal{E}}(f, f) &= \frac{1}{2} \sum_{x,y \in \Omega} (f(x) - f(y))^2 \pi(x)\pi(y) \\
&= \frac{1}{2} \sum_{x,y \in \Omega} (f^2(x) - 2f(x)f(y) + f^2(y))\pi(x)\pi(y) \\
&= \frac{1}{2} \left[\left(\sum_{x \in \Omega} f^2(x)\pi(x) \sum_{y \in \Omega} \pi(y) \right) - \left(\sum_{x \in \Omega} 2f(x)\pi(x) \sum_{y \in \Omega} f(y)\pi(y) \right) + \left(\sum_{y \in \Omega} f^2(y)\pi(y) \sum_{x \in \Omega} \pi(x) \right) \right] \\
&= \frac{1}{2} \left[2 \left(\sum_{x \in \Omega} f^2(x)\pi(x) \right) \right] \\
&= \mathrm{Var}_f [\pi] = \|f\|_{2,\pi}^2
\end{aligned}$$

By the canonical path method, we know $\mathcal{E}(f, f) \geq C^{-1} \|f\|_{2,\pi}$. Together with the variational characterization $\gamma \geq C^{-1}$. □

So let's write some intuitions about this geometric method. We can see the goal is to use $\tilde{P}$ with **known eigenvalues/vectors** to estimate the spectral information about P. The trick is to decide how to compare P and $\tilde{P}$. By **the variational characterization**, we see the spectral gap is lower bounded by the Dirichlet form. Then we check the Dirichlet form of P, converting it to something

related to the Dirichlet form of $\tilde{P}$. They differ by C, the **congestion ratio**, which describes how the probability flow in $\tilde{E}$ would get stuck through a worst edge e in E. Regarding the conversion of the Dirichlet form, we use the trick that $f(x) - f(y) = \sum_{e \in \Gamma_{xy}} \nabla f(e)$ is true for both E and $\tilde{E}$. Then we find the relationship between their Dirichlet forms and hence find the correspoinding relationship between the spectral gap of P and $\tilde{P}$ and successfully lower bound/estimate γ by Lemma 5.6. To be more specific, by Corollary 5.8.1, we lower bound γ by C^{-1}. To memorize this, it means that t_{rel} is upper bounded by C, which means the congestion ration determines the mixing time.

5.2 Examples

5.2.1 Example: random walk on a box

Note that a box is also a common term used to describe an area in a space. Formally

$$[n]^d := \{1, ..., n\}^d$$

with the restriction of the edges of $\mathbb{Z}^d$ to these vertices. Then there exist $c > 0$ such that $\gamma \geq c(dn)^{-2}$. The congestion ratio is:

$$C = \max_e \left(\frac{1}{Q(e)} \sum_{x,y: e \in \Gamma_{xy}} |\Gamma_{xy}| \pi(x) \pi(y) \right) \leq \max_e \left(\frac{d^2 O(1)}{n^{d-1}} |\{\Gamma_{xy} : e \in \Gamma_{xy}\}| \right)$$

We know that $\pi(x) \leq \kappa / n^d$, and $Q(e) \geq \pi(x) P(x, y) \geq \kappa' / n^d \cdot 1/d$. Also the length is at most dn. The number of such paths is at most n^{d+1}. Hence in the end C (which is roughly t_{rel}) is roughly n^2.

5.2.2 Example: random walk on a tree

We consider a tree with node n and max degree d and max height H. So in our formula of the congestion ratio: $|\Gamma_{xy}| \leq H$, $\pi(x) \leq \frac{d}{n}$, $1/Q(e) \leq dn$, $\sum_{x,y: e \in \Gamma_{xy}} \pi(x) \pi(y) \leq 1/4$. So

$$\gamma \geq \frac{1}{dnH}$$

5.2.3 Example: random walk on a convex set

5.2.4 Example: random walk on the n-dog

5.3 Cheeger's inequality

Cheeger's inequality basically links the conductance of the entire graph to the spectral information. Then the way of computing the conductance can be linked with the Dirichlet form. Then everything here is connected.

The canonical paths method kinda shows that if an edge is passed by many paths, then this slows down our mixing time. Cheeger handles this systematically.

The conductance is a bit different from what we have in the spectral graph theory but essentially they are the same. $Q(A, A^c) = \sum_{x \in A, y \in B} Q(x, y)$.

$$\Phi(A) = \frac{Q(A, A^c)}{\pi(A)} = \frac{\sum_{x \in A, y \in A^c} \pi(x) P(x, y)}{\pi(A)}$$

Definition 5.9. *The **bottleneck ratio** of the Markov chain is defined by*

$$\Phi_* = \min_{A:\pi(A)\leq 1/2} \Phi(A)$$

A very important thing is Cheeger's inequality. Either bound can be **sharp** in some examples. A more precise result can be proved when taking into account the whole **isoperimetric profile**.

Theorem 5.10. *Suppose P is reversible and let $\gamma = 1 - \lambda_2$ be the spectral gap, then*

$$\frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_*$$

Proof. The proof of this is very worth learning. The connection between the spectral gap and the conductance lies in their definitions.

Upper bound, easier Our goal is to prove $\gamma \leq 2\Phi_*$. How to think about this? If we expand the definition on both side.

First, by variational characterization,

$$\gamma = \min_{\substack{f:\Omega\to\mathbb{R} \\ \mathbb{E}_\pi[f]=0}} \frac{\mathcal{E}(f,f)}{\|f\|_{2,\pi}^2}$$

The trick is to define a clever f.

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\pi(x)P(x,y)$$

You see here we have $\pi(x)P(x,y)$ which also appears in the conductance. We want to introduce in A and A^c. So f should treat states in these two parts differently.

We pick $f(x) = -\pi(A^c)$ if $x \in A$, and $f(x) = \pi(A)$ if $x \in A^c$. So if x, y are in the same cut, $(f(x)-f(y))^2$ is 0, if they belong to different sets, $(f(x)-f(y))^2 = (-1)^2 = 1$. Hence

$$\mathcal{E}(f,f) = \frac{1}{2}(Q(A,A^c)+Q(A^c,A)) = Q(A,A^c)$$

On the other hand, we check $\|f\|_{2,\pi}^2$.

$$\|f\|_{2,\pi}^2 = \sum_{x\in A}\pi(x)\pi(A^c)^2 + \sum_{x\in A^c}\pi(x)\pi(A)^2 = \pi(A)\pi(A^c) \geq \pi(A)/2$$

Consequently,

$$\gamma \leq \frac{Q(A,A^c)}{\pi(A)/2}$$

Taking the minimum over all sets A gives $\gamma \leq 2\Phi_*$. This direction is eazy because an arbitrary f would do, it does not have to satisfy the expectation being 0.

Lower bound, harder Our goal is to prove $\gamma \geq \frac{\Phi_*^2}{2}$. Can we still use definition? We should, and also we should realize that the trick also hides in the designs of f.

To lower bound γ, by variational characterization, we should lower bound $\mathcal{E}(f,f)$ and upper bound $\|f\|_{2,\pi}^2 = \mathbb{E}_\pi\left[f^2\right]$.

First let's upper bound $\mathbb{E}_\pi\left[f^2\right]$.

Lemma 5.11. *Let $g : \Omega \to [0,\infty)$ be a nonnegative function such that $\pi(g > 0) \leq 1/2$. Order Ω so that g is non-increasing. Then*

$$\mathbb{E}_\pi[g] \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(g(x) - g(y))$$

Proof. Fix $\epsilon > 0$ and let $A = \{x : g(x) > \epsilon\}$, since Φ_* is the minimum bottleneck ratio.

$$\Phi_* \leq \frac{Q(A, A^c)}{\pi(A)} = \frac{\sum_{x,y:g(x)>\epsilon\geq g(y)} Q(x,y)}{\pi(g > \epsilon)}$$

Why do we define a set $g > t$? Because the expectation of the nonnegative function g can also be calculated by

$$\mathbb{E}_\pi[g] = \int_{x=0}^{\infty} \pi(g > x)dx$$

Rewrite the inequality and take the integral.

$$\mathbb{E}_\pi[g] \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(g(x) - g(y))$$

□

Since we can use this to upper bound the l_2 norm of a nonnegative function. We should design such a function to minimize the Dirichlet energy. Who are the candidates?

Let f_2 be the eigenfunction corresponding to λ_2. Why f_2? Because it is linked to the Dirichlet form and it minimizes variational characterization. We also know the expectation is 0, hence there are some positive and negative entries.

To apply the above lemma we need an non-negative function so we let $f = \max(0, f_2)$. Now $\mathbb{E}_\pi\left[f^2\right]$ is upper bounded by

$$\mathbb{E}_\pi\left[f^2\right] \leq \Phi_*^{-1} \sum_{x<y} Q(x,y)(f^2(x) - f^2(y))$$

To prepare for it, we need some link between this and

Note that here $x < y$ serves for the order we defined in Ω to make f non-increasing. By Cauchy-Schwarz

$$\begin{aligned}
\mathbb{E}_\pi\left[f^2\right]^2 &\leq \Phi_*^{-2} \left(\sum_{x<y} Q(x,y)(f^2(x) - f^2(y))\right)^2 \\
&\leq \Phi_*^{-2} \left(\sum_{x<y} Q(x,y)^{-1/2}(f(x) - f(y))Q(x,y)^{-1/2}(f(x) + f(y))\right)^2 \\
&\leq \Phi_*^{-2} \left(\sum_{x<y} Q(x,y)(f(x) - f(y))^2\right)\left(\sum_{x<y} Q(x,y)(f(x) + f(y))^2\right)
\end{aligned}$$

$a^2 + b^2 \geq 2ab$

$$\leq \Phi_*^{-2} \cdot \mathcal{E}(f,f) \left(2 \sum_{x<y} Q(x,y)(f(x)^2 + f(y)^2) \right)$$

$\sum_{x<y} Q(x,y) f(x)^2 = \sum_{x<y} Q(x,y) f(y)^2 \leq (1/2) \sum_x \pi(x) f^2(x) \leq (1/2)\mathbb{E}_\pi \left[f^2 \right]$

$$\leq \Phi_*^{-2} \cdot \mathcal{E}(f,f) \left(2\mathbb{E}_\pi \left[f^2 \right] \right)$$

Hence we have

$$\mathbb{E}_\pi \left[f^2 \right] \leq 2\mathcal{E}(f,f) \Phi_*^{-2}$$

Then

$$\gamma \geq \frac{\Phi_*^2}{2}$$

□

To be honest, the harder half is not that hard. Essentially we just need to play with the Dirichlet form and use Cauchy-Schwarz and some inequalities. Of course it is hard to come up with this proof, so we need to read this proof over and over again.

5.4 Expander graphs

If we look at Cheeger, we find that it bounds the spectral gap. If the conductance is large, which means the graph does not have bottleneck, then the spectral gap is large and the mixing time is small, which is consistent with the intuition.

The **best** graphs from this point of view are those for which the Cheeger constant is bounded below. We call such graphs **expanders**.

Definition 5.12. *A family of graphs $\{G_n\}$ is called an **expander family** if the Cheeger constant satisfies $\Phi_* \geq \alpha$ for some $\alpha > 0$*

Theorem 5.13. *Let G_n be a graph uniformly chosen among all $d-$regular graphs on n vertices, then there exists $\alpha > 0$ sufficiently small that with probability tending to 1 as $n \to \infty$, G_n is an α-expander.*

Bibliography

[1] Persi Diaconis. Group representations in probability and statistics. *Lecture notes-monograph series*, 11:i–192, 1988.

[2] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[3] Ravi Montenegro, Prasad Tetali, et al. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006.