# *[CS3958: Lecture 13] Variational Characterization, Relaxation Time, Graph Expansion*

*Instructor: Chihao Zhang, Scribed by Yulin Wang*

*January 2, 2023*

## 1 Variational Characterization of Eigenvalues

In this section, we will always assume that $P$ is a reversible chain with respect to $\pi$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ are its eigenvalues with corresponding orthonormal eigenvectors $v_1, \dots, v_n$. Hence $P = \sum_{i=1}^{n} \lambda_i v_i v_i^\top \Pi$. Assume $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n$, then the following proposition is immediate.

**Proposition 1** $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$.

*Proof.* Since $P$ is a stochastic matrix, we have $|\lambda_i| \leq 1$ for all $i \in [n]$. Moreover, 1 is an eigenvalue of $P$ with $v_1 = \mathbf{1}$. □

Define the *Rayleigh quotient*

$$R_P(x) = \frac{\langle \mathbf{x}, P\mathbf{x} \rangle_\Pi}{\langle \mathbf{x}, \mathbf{x} \rangle_\Pi}.$$

We can write $\lambda_1$ as the optimum of the following optimization problem:

$$\lambda_1 = \max_{\mathbf{x} \neq 0} R_P(\mathbf{x}).$$

This can be easily verified using the spectral decomposition of $P$. Supposing $\mathbf{x} = \sum_{i=1}^{n} a_i v_i$, we have

$$R_P(\mathbf{x}) = \frac{\sum_{i=1}^{n} \lambda_i a_i^2}{\sum_{i=1}^{n} a_i^2} = \sum_{i=1}^{n} \frac{a_i^2}{\sum_{j=1}^{n} a_j^2} \cdot \lambda_i.$$

Therefore, the Rayleigh quotient of any each nonzero $\mathbf{x}$ can be viewed as a weighted sum of all $P$'s eigenvalues. Of course the maximum is achieved at $\mathbf{x} = v_1$ — putting all the weight on the maximum eigenvalue. We also have

$$\lambda_n = \min_{\mathbf{x} \neq 0} R_P(\mathbf{x}).$$

Similar arguments can be used to prove that $\lambda_2 = \max_{\substack{\mathbf{x} \neq 0 \\ \mathbf{x} \perp v_1}} R_P(\mathbf{x})$, or more generally

$$\lambda_k = \max_{\substack{\mathbf{x} \neq 0 \\ \mathbf{x} \perp \mathrm{span}(v_1, \dots, v_{k-1})}} R_P(\mathbf{x}),$$

where $\mathbf{x} \perp \mathbf{y}$ means $\langle \mathbf{x}, \mathbf{y} \rangle_\Pi = 0$.

We can use another two-stage optimization problem to characterize $\lambda_k$:

$$\lambda_k = \max_{\substack{k-\text{dimenional subspace} \\ V \subseteq \mathbb{R}^n}} \min_{\mathbf{x} \in V \setminus \{0\}} R_P(\mathbf{x}).$$

To justify this, imagining the following game between two players: a *max* player and a *min* player.

- The goal of the max player is to maximize $R_P(\mathbf{x})$, and what he can do is to provide some $k$-dimensional subspace $V \subseteq \mathbb{R}^n$;
- The goal of the min player is to minimize $R_P(\mathbf{x})$ and he can only choose those nonzero vectors from the space $V$ provided by the max player.

Recall that for each nonzero vector $\mathbf{x}$, the value of $R_P(\mathbf{x})$ can be viewed as a weighted sum of $P$'s eigenvalues. So the min players strategy must be choosing the vector whose mass is concentrated on small eigenvalues. The max player should choose the collection of vectors so that the min player's strategy does not perform well, so his strategy must be choosing $V = span(v_1, \dots, v_k)$. This yields the min player to choose $\mathbf{x} = c \cdot v_k$, and therefore $R_P(\mathbf{x}) = \lambda_k$.

## 2 FTMC for Reversible Chains

Recall the *Fundamental Theorem of Markov Chains*:

**Theorem 2 (Fundamental theorem of Markov chains)** . *If a* finite *Markov chain $P \in \mathbb{R}^{n \times n}$ is* irreducible *and* aperiodic, *then it has a unique stationary distribution $\pi \in \mathbb{R}^n$. Moreover, for any distribution $\mu \in \mathbb{R}^n$,*

$$\lim_{t \to \infty} \mu^{\top} P^t = \pi^{\top}.$$

Today we will give another proof of the theorem for *reversible chains* using spectral decomposition. The proof is elegant, insightful and can be generalized to studying the rate of convergence.

Let us collect what we know about $P$. First, we have the spectral decomposition

$$P = \sum_{i=1}^{n} \lambda_i v_i v_i^{\top} \Pi = V \Lambda V^{\top} \Pi,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$ are eigenvalues of $P$ with corresponding orthonormal (w.r.t. the inner product $\langle \cdot, \cdot \rangle_{\Pi}$) eigenvectors $v_1, \dots, v_n$. Moreover, we know $\lambda_1 = 1$ and $v_1 = \mathbf{1}$.

With this decomposition, it is easy to compute

$$P^t = \sum_{i=1}^{n} \lambda_i^t v_i v_i^{\top} \Pi = \mathbf{1}\pi^{\top} + \sum_{i=2}^{n} \lambda_i^t v_i v_i^{\top} \Pi.$$

Note that $\mathbf{1}\pi^{\top} = \begin{bmatrix} \pi^{\top} \\ \pi^{\top} \\ \vdots \\ \pi^{\top} \end{bmatrix}$ and therefore for any distribution $\mu$, it holds $\mu^{\top}\mathbf{1}\pi^{\top} = \pi^{\top}$. This implies that

$$\lim_{t \to \infty} \mu^{\top} P^t = \pi^{\top} + \lim_{t \to \infty} \mu^{\top} \left( \sum_{i=2}^{n} \lambda_i^t v_i v_i^{\top} \Pi \right).$$

Therefore we only need to argue when

$$\lim_{t \to \infty} \mu^{\top} \left( \sum_{i=2}^{n} \lambda_i^t v_i v_i^{\top} \Pi \right) = 0.$$

Since $P$ is stochastic, we know that $|\lambda_i| \leq 1$ for all eigenvalues $\lambda_i$ of $P$. Therefore, there are two ways to prohibit $\lim_{t \to \infty} \mu^{\top} \left( \sum_{i=2}^{n} \lambda_i^t v_i v_i^{\top} \Pi \right) = 0$, $\lambda_2 = 1$ or $\lambda_n = -1$. We will now show that the two cases correspond to reducibility and periodicity of $P$ respectively.

Since we assume $P$ is reversible, all edges in $P$ can be viewed as being undirected, namely $(u, v) \in E \iff (v, u) \in E$. As a result, reducibility is equivalent to that the transition graph is disconnected.

We now prove

**Proposition 3** $\lambda_2 = 1$ *if and only if the transition graph of $P$ is disconnected.*

*Proof.* The main tool to prove the proposition is the variational characterization of eigenvalues. Recall that

$$\begin{aligned}
\lambda_2 &= \max_{\substack{\mathbf{x}\neq 0\\ \mathbf{x}\perp\mathbf{1}}} R_P(\mathbf{x}) \\
&= \max_{\substack{\mathbf{x}\neq 0\\ \mathbf{x}\perp\mathbf{1}}} \frac{\langle \mathbf{x}, P\mathbf{x}\rangle_\Pi}{\langle \mathbf{x}, \mathbf{x}\rangle_\Pi} \\
&= \max_{\substack{\mathbf{x}\neq 0\\ \mathbf{x}\perp\mathbf{1}}} \frac{\sum_{(i,j)\in V^2} \pi(i)P(i,j)x(i)x(j)}{\sum_{i\in V}\pi(i)x(i)^2} + 1 - 1 \\
&= \max_{\substack{\mathbf{x}\neq 0\\ \mathbf{x}\perp\mathbf{1}}} 1 - \frac{\sum_{\{i,j\}\in E} \pi(i)P(i,j)(x(i)-x(j))^2}{\sum_{i\in V}\pi(i)x(i)^2}
\end{aligned}$$

Therefore, $\lambda_2 = 1$ if and only if we can find some nonzero $\mathbf{x} \perp \mathbf{1}$ such that $\frac{\sum_{\{i,j\}\in E} \pi(i)P(i,j)(x(i)-x(j))^2}{\sum_{i\in V}\pi(i)x(i)^2} = 0$. Clearly this is equivalent to that $P$ is disconnected. □

In fact, we can prove that $\lambda_k = 1$ iff $P$ contains at least $k$ connected components.

Since $P$ is reversible, if $P$ is connected and contains more than one vertex, then each vertex is on a cycle of length two. Therefore, $P$ is periodic iff it does not contain odd cycles, or equivalently, it is bipartite.

**Proposition 4** $\lambda_n = -1$ *if and only if the transition graph of $P$ is bipartite.*

*Proof.* Again by the variational characterization of $\lambda_n$, we have

$$\begin{aligned}
\lambda_n &= \min_{\mathbf{x}\neq 0} R_P(\mathbf{x}) \\
&= \min_{\mathbf{x}\neq 0} \frac{\sum_{(i,j)\in V^2} \pi(i)P(i,j)x(i)x(j)}{\sum_{i=1}^n \pi(i)x(i)^2} + 1 - 1 \\
&= \min_{\mathbf{x}\neq 0} \frac{\sum_{(i,j)\in V^2} \pi(i)P(i,j)x(i)x(j) + \sum_{(i,j)\in V^2} \pi(i)P(i,j)x(i)^2}{\sum_{i=1}^n \pi(i)x(i)^2} - 1 \\
&= \min_{\mathbf{x}\neq 0} \frac{2\sum_{(i,j)\in V^2} \pi(i)P(i,j)x(i)x(j) + \sum_{(i,j)\in V^2} \pi(i)P(i,j)(x(i)^2 + x(j)^2)}{2\sum_{i=1}^n \pi(i)x(i)^2} - 1 \\
&= \min_{\mathbf{x}\neq 0} \frac{\sum_{(i,j)\in V^2} \pi(i)P(i,j)(x(i)+x(j))^2}{2\sum_{i=1}^n \pi(i)x(i)^2} - 1
\end{aligned}$$

Clearly the numerator is zero for some nonzero $\mathbf{x}$ if and only if $P$ is bipartite. □

### 2.1 Relaxation Time

Therefore by discussion above, if $\lambda_2 < 1$ and $\lambda_n > -1$, or equivalently $P$ is aperiodic and irreducible, then

$$\lim_{t\to\infty} \mu^\top \left(\sum_{i=2}^n \lambda_i^t v_i v_i^\top \Pi\right) = 0.$$

The gap between 1 and the absolute value of these two eigenvalues also determines how fast $P^t$ converges to $\mathbf{1}\pi^\top$.

Let $\lambda^* \triangleq \max\{|\lambda_2|, |\lambda_n|\}$, then the *relaxation time* of $P$ is defined to be

$$\tau_{\text{rel}} \triangleq \frac{1}{1-\lambda^*}.$$

A lazy chain, for example, $P' = \frac{1}{2}(P+I) \geq 0$, which implies that $\lambda_n(P') \geq 0$. So we consider $\lambda^* = |\lambda_2|$ most of the time in applications.

It measures the rate of convergence and is related to the *mixing time* as follows:

**Theorem 5** *Let $P$ be a reversible chain with stationary distribution $\pi$, then*

$$(\tau_{\text{rel}} - 1)\log\frac{1}{2\varepsilon} \leq \tau_{\text{mix}}(\varepsilon) \leq \tau_{\text{rel}}\log\frac{1}{\varepsilon\pi_{\min}},$$

*where $\pi_{\min} = \min_{x\in\Omega}\pi(x)$.*

## 3 From Coupling to Spectral Gap

**Theorem 6 (Mu-Fa Chen, 1998)** *If there is a coupling $\{\omega_t\}$ such that*

$$\mathbf{E}_{(X_{t+1},Y_{t+1})\sim\omega_t}\left[d(X_{t+1},Y_{t+1}) \mid (X_t,Y_t)\right] \leq (1-\alpha)d(X_t,Y_t),$$

*then $|\lambda^*| \leq 1-\alpha$.*

*Proof.* Define $\mathtt{Lip}(f) \triangleq \max_{x,y\in\Omega}\frac{|f(x)-f(y)|}{d(x,y)}$. For any $f : \Omega \to \mathbb{R}$, we claim that

$$\mathtt{Lip}(Pf) \leq (1-\alpha)\mathtt{Lip}(f). \tag{1}$$

And then we have for any eigenvector $f$ of $\lambda_2$,

$$|\lambda_2|\mathtt{Lip}(f) = \mathtt{Lip}(\lambda f) = \mathtt{Lip}(Pf) \leq (1-\alpha)\mathtt{Lip}(f).$$

Therefore, $|\lambda_2| \leq 1-\alpha$. Now we only need to prove eq. (1). For any $x, y$,

$$\begin{aligned}\frac{|Pf(x)-Pf(y)|}{d(x,y)} =& \frac{\mathbf{E}\left[f(X_1)\right] - \mathbf{E}\left[f(Y_1)\right]}{d(x,y)} \\ \leq& \mathbf{E}_{(X_1,Y_1)\sim\omega}\left[|f(X_1)-f(Y_1)|\right]d(x,y) \\ \leq& \frac{\mathbf{E}\left[\mathtt{Lip}(f)d(X_1,Y_1)\right]}{d(x,y)} = (1-\alpha)\mathtt{Lip}(f),\end{aligned}$$

where $(X_1, Y_1)$ is generated from $(x, y)$ according to the coupling. □

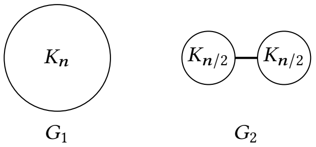
Here we consider a Markov chain as a operator $P$ such that for any function $f : \Omega \to \mathbb{R}$,

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

Intuitively, $Pf(x)$ is the expectation of $f(y)$ that one walks from $x$ to $y$ by one step.

## 4 Graph Expansion

**Example 1** *Consider a simple random walk on a graph that one walks to a neighbor uniformly at random for each step. Therefore, $P(i,j) = \frac{1}{\deg(i)}$ and $\pi(i) \sim \deg(i)$. It is obviously that the simple random walk on $G_1 = K_n$ mixes much faster than that on $G_2$.*



$K_n$ denotes complete graph of $n$ vertices.

Let $P$ be a reversible chain on $\Omega$. For any $i, j \in \Omega$, define the probability flow from $i$ to $j$ as $Q(i, j) \triangleq \pi(i)P(i, j)$. Similarly, for any $S \subset \Omega$, the flow from $S$ to $\bar{S}$, denoted by $Q(S, \bar{S})$, is $\sum_{i\in S, j\in\Omega\setminus S} Q(i, j)$, and we define the *expansion* of $S$ as

$$\Phi(S) = \frac{Q(S, \bar{S})}{\pi(S)},$$

where $\pi(S) = \sum_{i\in S} \pi(i)$.

Suppose $X_t \sim \pi$, then a direct calculation can prove that

$$\Phi(S) = \mathbf{Pr}\left[X_{t+1} \notin S \mid X_t \in S\right].$$

The expansion of $P$ is

$$\Phi(P) = \min_{S\subseteq\Omega:\pi(S)\le\frac{1}{2}} \Phi(S).$$

The following theorem justify our intuition that small expansion implies slow mixing.

**Theorem 7** *Let $P$ be a reversible chain.*

$$\tau_{mix}(\varepsilon) \ge \frac{1-2\varepsilon}{2}\frac{1}{\Phi(P)}.$$

The relation between mixing time and graph expansion obtained from Cheeger's inequality and Theorem 5 is much weaker than this direct one.

*Proof.* Let $X_0 \sim \pi$, and we use $P$ to generate $X_1, X_2, \ldots$

Let $S = \arg\min_{\pi(S)\le\frac{1}{2}} \Phi(S)$.

$$\begin{aligned}
\mathbf{Pr}\left[X_t \in \bar{S} \,\middle|\, X_0 \in S\right] &= \frac{\mathbf{Pr}\left[X_t \in \bar{S} \wedge X_0 \in S\right]}{\mathbf{Pr}\left[X_0 \in S\right]} \\
&\le \frac{\sum_{i=0}^{t-1}\mathbf{Pr}\left[X_i \in S \wedge X_{i+1} \in \bar{S}\right]}{\mathbf{Pr}\left[X_0 \in S\right]} \\
&= \frac{t\cdot\mathbf{Pr}\left[X_1 \in \bar{S} \wedge X_0 \in S\right]}{\mathbf{Pr}\left[X_0 \in S\right]} \\
&= t\cdot\mathbf{Pr}\left[X_1 \in \bar{S} \,\middle|\, X_0 \in S\right] \\
&= t\cdot\Phi(S).
\end{aligned}$$

So there exists $x_0$ such that

$$\mathbf{Pr}\left[X_t \in S \mid X_0 = x_0\right] \ge 1 - t\cdot\Phi(S).$$

Therefore,

$$d_{TV}(P^t(x_0,\cdot),\pi) \ge 1 - t\cdot\Phi(S) - \pi(S) \ge \frac{1}{2} - t\cdot\Phi(S) \ge \varepsilon,$$

as long as $t \le \frac{1-2\varepsilon}{2}\frac{1}{\Phi(P)}$. □