Mathematics for Data Sciences **December 2nd, 2014**

Lecture 11. Lumpable Markov Chain & Multiple Spectral Clustering

Instructor: Yuan Yao, Peking University *Scribe: Your Name*

1 Lumpability of Markov Chain

Let P be the transition matrix of a Markov chain on graph $G = (V, E)$ with $V = \{1, 2, \cdots, n\}$, i.e. $P_{ij} = \text{Prob}\{x_t = j : x_{t-1} = i\}$. Assume that V admits a partition Ω:

$$V = \cup_{i=1}^{k} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \ i \neq j.$$

$$\Omega = \{\Omega_s : s = 1, \cdots, k\}.$$

Observe a sequence$\{x_0, x_1, \cdots, x_t\}$ sampled from the Markov chain with initial distribution π_0.

Definition (Lumpability, Kemeny-Snell 1976)**.** P is lumpable with respect to partition Ω if the sequence $\{y_t\}$ is Markovian. In other words, the transition probabilities do not depend on the choice of initial distribution π_0 and history, *i.e.*

$$\text{Prob}_{\pi_0}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}, \cdots, x_0 \in \Omega_{k_0}\} = \text{Prob}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}\}. \tag{1}$$

Relabel $x_t \mapsto y_t \in \{1, \cdots, k\}$ by

$$y_t = \sum_{s=1}^{k} s\mathcal{X}_{\Omega_s}(x_t).$$

Thus we obtain a sequence (y_t) which is a coarse-grained representation of original sequence. The lumpability condition above can be rewritten as

$$\text{Prob}_{\pi_0}\{y_t = k_t : y_{t-1} = k_{t-1}, \cdots, y_0 = k_0\} = \text{Prob}\{y_t = k_t : y_{t-1} = k_{t-1}\}. \tag{2}$$

Theorem 1.1. **I.** (Kemeny-Snell 1976) P is lumpable with respect to partition $\Omega \Leftrightarrow \forall \Omega_s, \Omega_t \in \Omega$, $\forall i, j \in \Omega_s$, $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$, where $\hat{P}_{i\Omega_t} = \sum_{j \in \Omega_t} P_{ij}$.

II. (Meila-Shi 2001) P is lumpable with respect to partition Ω and $\hat{P}$ ($\hat{p}_{st} = \sum_{i \in \Omega_s, j \in \Omega_t} p_{ij}$) is nonsingular $\Leftrightarrow P$ has k independent piecewise constant right eigenvectors in span$\{\chi_{\Omega_s} : s = 1, \cdots, k\}$, χ is the characteristic function.

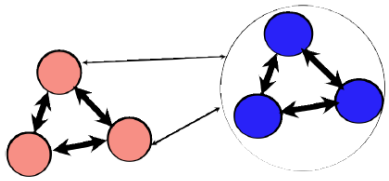

图 1: Lumpability condition $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$

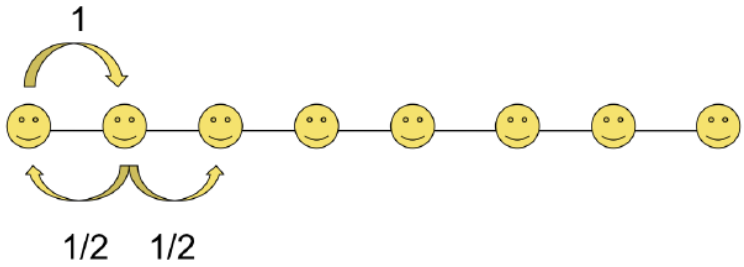

图 2: A linear chain of $2n$ nodes with a random walk.

Example *. Consider a linear chain with $2n$ nodes (Figure 2) whose adjacency matrix and degree matrix are given by

$$A = \begin{bmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 0 & 1 & \\ & & & 1 & 0 & \end{bmatrix}, \quad D = \mathrm{diag}\{1, 2, \cdots, 2, 1\}$$

So the transition matrix is $P = D^{-1}A$ which is illustrated in Figure 2. The spectrum of P includes two eigenvalues of magnitude 1, *i.e.* $\lambda_0 = 1$ and $\lambda_{n-1} = -1$. Although P is not a *primitive* matrix here, it is *lumpable*. Let $\Omega_1 = \{\text{odd nodes}\}$, $\Omega_2 = \{\text{even nodes}\}$. We can check that I and II are satisfied.

To see I, note that for any two even nodes, say $i = 2$ and $j = 4$, $\hat{P}_{i\Omega_2} = \hat{P}_{j\Omega_2} = 1$ as their neighbors are all odd nodes, whence I is satisfied. To see II, note that ϕ_0 (associated with $\lambda_0 = 1$) is a constant vector while ϕ_1 (associated with $\lambda_{n-1} = -1$) is constant on even nodes and odd nodes respectively. Figure 3 shows the lumpable states when $n = 4$ in the left.

Note that lumpable states might not be optimal bi-partitions in $NCUT = Cut(S)/\min(vol(S), vol(\bar{S}))$. In this example, the optimal bi-partition by Ncut is given by $S = \{1, \ldots, n\}$, shown in the right of Figure 3. In fact the second largest eigenvalue $\lambda_1 = 0.9010$ with eigenvector

$$v_1 = [0.4714, 0.4247, 0.2939, 0.1049, -0.1049, -0.2939, -0.4247, -0.4714],$$

give the optimal bi-partition.

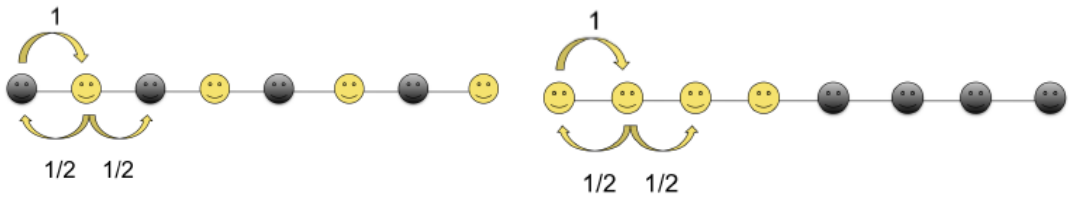

图 3: Left: two lumpable states; Right: optimal-bipartition of Ncut.

Example. Uncoupled Markov chains are lumpable, e.g.

$$P_0 = \begin{bmatrix} \Omega_1 & & \\ & \Omega_2 & \\ & & \Omega_3 \end{bmatrix}, \ \hat{P}_{it} = \hat{P}_{jt} = 0.$$

A markov chain $\tilde{P} = P_0 + O(\epsilon)$ is called nearly uncoupled Markov chain. Such Markov chains can be approximately represented as uncoupled Markov chains with *metastable states*, $\{\Omega_s\}$, where within metastable state transitions are fast while cross metastable states transitions are slow. Such a separation of scale in dynamics often appears in many phenomena in real lives, such as protein folding, or your life transitions *primary schools* $\mapsto$ *middle schools* $\mapsto$ *high schools* $\mapsto$ *college/university* $\mapsto$ *work unit*, which can be visualized in figure 4. In this figure, there are a variety of links between the specified states such as the dorm or the dinning hall. However, in the view of lumpable Markov Chain, the states can be divided into several subsets of the state space. Specifically speaking, the chain can be seen as the life transitions in a person's youth, ranging from high school to the place he or she works. Thus, it gives a vivid depiction of the application of lumpable Markov Chains.

Before the proof of the theorem, we note that condition I is in fact equivalent to

$$VUPV = PV, \tag{3}$$

where U is a k-by-n matrix where each row is a uniform probability that

$$U_{is}^{k\times n} = \frac{1}{|\Omega_s|}\chi_{\Omega_s}(i), \quad i \in V, \ s \in \Omega,$$

and V is a n-by-k matrix where each column is a characteristic function on Ω_s,

$$V_{sj}^{n\times k} = \chi_{\Omega_s}(j).$$

With this we have $\hat{P} = UPV$ and $UV = I$. Such a matrix representation will be useful in the derivation of condition II. Now we give the proof of the main theorem.

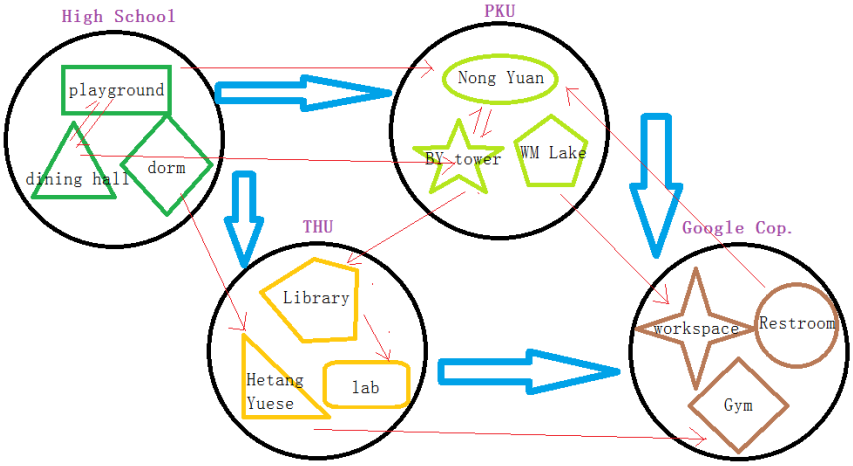

图 4: life transition path

Proof *. I. "⇒" To see the necessity, P is lumpable w.r.t. partition Ω, then it is necessary that

$$\text{Prob}_{\pi_0}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \text{Prob}_{\pi_0}\{y_1 = t : y_0 = s\} = \hat{p}_{st}$$

which does not depend on π_0. Now assume there are two different initial distribution such that $\pi_0^{(1)}(i) = 1$ and $\pi_0^{(2)}(j) = 1$ for $\forall i, j \in \Omega_s$. Thus

$$\hat{p}_{i\Omega_t} = \text{Prob}_{\pi_0^{(1)}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \hat{p}_{st} = \text{Prob}_{\pi_0^{(2)}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \hat{p}_{j\Omega_t}.$$

$\forall i \in \Omega_s$, let π_0 have a 1 in its i-th compoment. Then

$$\hat{p}_{i\Omega_t} = \text{Prob}\{x_1 \in \Omega_t : x_0 = i\} = \text{Prob}_{\pi_0}\{y_1 = t : y_0 = s\} = \hat{p}_{st}.$$

"⇐" To show the sufficiency, we are going to show that if the condition is satisfied, then the probability

$$\text{Prob}_{\pi_0}\{y_t = t : y_{t-1} = s, \cdots, y_0 = k_0\}$$

depends only on $\Omega_s, \Omega_t \in \Omega$. Probability above can be written as $\text{Prob}_{\pi_{t-1}}(y_t = t)$ where π_{t-1} is a distribution with support only on Ω_s which depends on π_0 and history up to $t-1$. But since $\text{Prob}_i(y_t = t) = \hat{p}_{i\Omega_t} \equiv \hat{p}_{st}$ for all $i \in \Omega_s$, then $\text{Prob}_{\pi_{t-1}}(y_t = t) = \sum_{i\in\Omega_s} \pi_{t-1}\hat{p}_{i\Omega_t} = \hat{p}_{st}$ which only depends on Ω_s and Ω_t.

II.

"⇒"

Since $\hat{P}$ is nonsingular, let $\{\psi_i, i = 1, \cdots, k\}$ are independent right eigenvectors of $\hat{P}$, i.e., $\hat{P}\psi_i = \lambda_i\psi_i$.

Define $\phi_i = V\psi_i$, then ϕ_i are independent piecewise constant vectors in span$\{\chi_{\Omega_i}, i = 1, \cdots, k\}$. We have

$$P\phi_i = PV\psi_i = VUPV\psi_i = V\hat{P}\psi_i = \lambda_i V\psi_i = \lambda_i\phi_i,$$

i.e. ϕ_i are right eigenvectors of P.

"$\Leftarrow$"

Let $\{\phi_i, i = 1, \cdots, k\}$ be k independent piecewise constant right eigenvectors of P in span$\{\mathcal{X}_{\Omega_i}, i = 1, \cdots, k\}$. There must be k independent vectors $\psi_i \in \mathbb{R}^k$ that satisfied $\phi_i = V\psi_i$. Then

$$P\phi_i = \lambda_i\phi_i \Rightarrow PV\psi_i = \lambda_i V\psi_i,$$

Multiplying VU to the left on both sides of the equation, we have

$$VUPV\psi_i = \lambda_i VUV\psi_i = \lambda_i V\psi_i = PV\psi_i, \ (UV = I),$$

which implies

$$(VUPV - PV)\Psi = 0, \quad \Psi = [\psi_1, \ldots, \psi_k].$$

Since Ψ is nonsingular due to independence of ψ_i, whence we must have $VUPV = PV$.

1.1 Normalized Laplacian

The definitions of matrices A, D are the same as in the previous sections. Here we have:

$$\mathcal{L} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = \Theta^{\frac{1}{2}}z[D_B^{-\frac{1}{2}}BD_B^{-\frac{1}{2}}]z^T\Theta^{\frac{1}{2}} = \Theta^{\frac{1}{2}}zB_nz^T\Theta^{\frac{1}{2}}$$

Our goal is to find the eigenvalue decomposition.

$$B_n = U\Lambda U^T$$

Notice that B_n is a symmetric matrix but not semi-definite positive. Here

$$V = \Theta^{\frac{1}{2}}ZU$$

is the EVD (eigenvalue decomposition) of $\mathcal{L}$:

$$\mathcal{L} = V\Lambda V^T = \Theta^{\frac{1}{2}}zU\Lambda U^Tz^T\Theta^{\frac{1}{2}}$$

Then let

$$\Phi = D^{-\frac{1}{2}}V = zD^{-\frac{1}{2}}U, \;\; \Psi = D^{\frac{1}{2}}V = \Theta z D_B^{\frac{1}{2}}U$$

Therefore,

$$\Phi\Lambda\Psi^T = zD_B^{-\frac{1}{2}}U\Lambda U^T D^{\frac{1}{2}}z^T\Theta = zP_B z^T\Theta = P$$

Hence we acquire two embeddings:

- Φ is piecewise constants on S_k
- V is piecewise conic lines (polar lines) on S_k.

1.2 Optimal Lumpable Approximation

For a Markov Chain that has stationary distribution, let P be its transition probability matrix and μ be its stationary distribution. By definition we have

$$\mu^T P = \mu^T$$

Let $\mu_{S_k} = \sum_{i\in S_k}\mu_i$, our aim here is to find a lumpable Markov Chain whose transition probability matrix $\hat{P}$ is lumpable with regard to state subsets $S_1, \cdots, S_k$ that minimizes the following value

$$\min_{\{S_k\}}\|P - \hat{P}\|_*$$

Where the '*' under the norm can be chosen as the HS-norm. With the stationary distribution and previous analysis, we can write $\hat{P}$ in the form of

$$\hat{P} = Z\hat{P}_B Z^T\theta$$

where $\theta_i = \frac{\mu_i}{\mu_{Z_i}}, \mu_{Z_i} = \sum_{i\in Z_i}\mu_i$.

In this way the minimization problem can be transformed as

$$\min_{\{S_k\}}\|P - \hat{P}\|_{HS} := \min_{\{S_k\}}\|\mathcal{L}_p - \mathcal{L}_{\hat{P}}\|_F = \min_{\{S_k\}}\|D^{-\frac{1}{2}}AD^{\frac{1}{2}} - \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{\frac{1}{2}}\|_F$$

This is a NP-Hard problem. Since the lumpable matrix Laplacian is a rank-k matrix, we can transform the problem into a easier one. We have

$$\|P - \hat{P}(\{S_k\})\|_{HS} \geq \|P - \hat{P}_k\|_{HS}$$

Thus, a lower bound can be found by rank-k approximation of the transition probability matrix P. The rank-k approximating matrix is $\hat{P}_k$ in the above inequality. Furthermore, we can find the optimal coefficient k

after minimizing the approximation error for every fixed k.

Example. (DC-SBM) Degree-Corrected Stochastic Block Model. Assume $\tilde{G}$ is a random graph, $\tilde{A}_{ij}$ are sampled independently from Bernoulli model such that

$$\mathbb{E}[\tilde{A}_{ij}] = A_{ij},$$

where

$$A = \Theta Z B Z \Theta, \quad B = B^T,$$

$Z^{N\times k}$ with row vectors $\in \{0,1\}^k$ is the block membership,

$$z_{ik} = \begin{cases} 1 & z_i \in S_k, \\ 0 & otherwise. \end{cases}$$

and $\Theta = diag(\theta_i)$ with each $\theta_i = const \geq 0$ is the degree expected,

$$\sum_{i\in z_i} \theta_i = 1.$$

Then we may define a Markov chain with transition probability matrix

$$P = D^{-1}A = ZP_BZ^T\Theta,$$

where

$$P_B = D_B^{-1}B, \quad D_B = diag(\sum_t B_{st}),$$

since

$$d_i = \sum_j \theta_i B_{z_i z_j} \theta_j = \theta\left(\sum_{z_j} B_{z_i z_j}\right), \quad \sum_{j\in z_j} \theta_j = 1.$$

2 Transition Path Theory

The transition path theory was originally introduced in the context of continuous-time Markov process on continuous state space and discrete state space. The following material is adapted to the setting of discrete time Markov chain with transition probability matrix P. We assume reversibility in the following presentation, which can be extended to non-reversible Markov chains.

Assume that an irreducible Markov Chain on graph $G = (V, E)$ admits the following decomposition $P = D^{-1}W = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$. Here $V_l = V_0 \cup V_1$ denotes the labeled vertices with source set V_0 (e.g. reaction state in

chemistry) and sink set V_1 (e.g. product state in chemistry), and V_u is the unlabeled vertex set (intermediate states). That is,

- $V_0 = \{i \in V_l : f_i = f(x_i) = 0\}$
- $V_1 = \{i \in V_l : f_i = f(x_i) = 1\}$
- $V = V_0 \cup V_1 \cup V_u$ where $V_l = V_0 \cup V_1$

Given two sets V_0 and V_1 in the state space V, the transition path theory tells how these transitions between the two sets happen (mechanism, rates, etc.). If we view V_0 as a reactant state and V_1 as a product state, then one transition from V_0 to V_1 is a reaction event. The reactve trajectories are those part of the equilibrium trajectory that the system is going from V_0 to V_1.

Let the hitting time of V_l be

$$\tau_i^k = \inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, \quad k = 0, 1.$$

To make the notion more precise, define the ordered family of times $\{n_j^A, n_j^B\}$ such that

$$\begin{aligned} &X_{n_j^A} \in A, \quad , X_{n_j^B} \in B, \\ &X_n \in V \backslash (A \cup B), \quad \forall n, n_j^A < n < n_j^B. \end{aligned}$$

Hence, a reaction happens from time n_j^A to time n_j^B.

Definition. Given any equilibrium trajectory $\{X_n\}$, we call each portion of the trajectory of between n_j^A and n_j^B a *AB-reactive trajectory*. We call the time during which the reaction occurs the *reactive times*

$$R = \bigcup_{j \in \mathbb{Z}} (n_j^A, n_j^B). \tag{4}$$

The central object in transition path theory is the committor function. Its value at $i \in V_u$ gives the probability that a trajectory starting from i will hit the set V_1 first than V_0, *i.e.*, the success rate of the transition at i.

Proposition 2.1. For $\forall i \in V_u$, define the *committor function*

$$q_i := Prob(\tau_i^1 < \tau_i^0) = Prob(\text{trajectory starting from } x_i \text{ hit } V_1 \text{ before } V_0)$$

which satisfies the following Laplacian equation with Dirichlet boundary conditions

$$(Lq)(i) = [(I - P)q](i) = 0, \quad i \in V_u$$

$$q_{i\in V_0} = 0, q_{i\in V_1} = 1.$$

The solution is

$$q_u = (D_u - W_{uu})^{-1} W_{ul} q_l.$$

Proof. By definition,

$$q_i = Prob(\tau_i^1 < \tau_i^0) = \begin{cases} 1 & x_i \in V_1 \\ 0 & x_i \in V_0 \\ \sum_{j\in V} P_{ij} q_j & i \in V_u \end{cases}$$

This is because $\forall i \in V_u$,

$$\begin{aligned} q_i &= Pr(\tau_{iV_1} < \tau_{iV_0}) \\ &= \sum_j P_{ij} q_j \\ &= \sum_{j\in V_1} P_{ij} q_j + \sum_{j\in V_0} P_{ij} q_j + \sum_{j\in V_u} P_{ij} q_j \\ &= \sum_{j\in V_1} P_{ij} + \sum_{j\in V_u} P_{ij} q_j \end{aligned}$$

$$\therefore \quad q_u = P_{ul} q_l + P_{uu} q_u = D_u^{-1} W_{ul} q_l + D_u^{-1} W_{uu} q_u$$

multiply D_u to both side and reorganize

$$(D_u - W_{uu}) q_u = W_{ul} q_l$$

If $D_u - W_{uu}$ is reversible, we get

$$q_u = (D_u - W_{uu})^{-1} W_{ul} q_l.$$

□

Given two sets A and B in the state space, q satisfies the equation

$$\begin{cases} \sum_{y\in V} p_{xy} q(y) - q(x) = 0, & x \notin A \cup B; \\ q(x) = 0, & x \in A; \\ q(x) = 1, & x \in B, \end{cases} \tag{5}$$

The committor function provides natural decomposition of the graph. If $q(x)$ is less than 0.5, i is more likely to reach V_0 first than V_1; so that $\{i \mid q(x) < 0.5\}$ gives the set of points that are more attached to set V_0.

Once the committor function is given, the statistical properties of the reaction trajectories between V_0 and V_1 can be quantified. We state several propositions characterizing transition mechanism from V_0 to V_1. The proof of them is an easy adaptation of and will be omitted.

Proposition 2.2 (Probability distribution of reactive trajectories)**.** The probability distribution of reactive trajectories

$$\pi_R(x) = \mathbb{P}(X_n = x, n \in R) = \lim_{T\to\infty} \frac{1}{2T} \sum_{n\in R} \delta(X_n = i), \tag{6}$$

is given by

$$\pi_R(x) = \pi(x)q(x)(1 - q(x)). \tag{7}$$

The distribution π_R gives the equilibrium probability that a reactive trajectory visits x. It provides information about the proportion of time the reactive trajectories spend in state x along the way from V_0 to V_1.

Proposition 2.3 (Reactive current from V_0 to V_1)**.** The reactive current from A to B, defined by

$$J(xy) = \mathbb{P}(X_n = x, X_{n+1} = y, \{n, n + 1\} \subset R), \tag{8}$$

is given by

$$J(xy) = \begin{cases} \pi(x)(1 - q(x))P_{xy}q(y), & x \neq y; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The reactive current $J(xy)$ gives the average rate the reactive trajectories jump from state x to y. From the reactive current, we may define the effective reactive current on an edge and transition current through a node which characterizes the importance of an edge and a node in the transition from A to B, respectively.

Definition. The *effective current* of an edge xy is defined as

$$J^+(xy) = \max(J(xy) - J(yx), 0). \tag{10}$$

The *transition current* through a node $x \in V$ is defined as

$$T(x) = \begin{cases} \sum_{y\in V} J^+(xy), & x \in A \\ \sum_{y\in V} J^+(yx), & x \in B \\ \sum_{y\in V} J^+(xy) = \sum_{y\in V} J^+(yx), & x \notin A \cup B \end{cases} \tag{11}$$

In applications one often examines partial transition current through a node connecting two communities $V^- = \{x : q(x) < 0.5\}$ and $V^+ = \{x : q(x) \geq 0.5\}$, *e.g.* $\sum_{y\in V^+} J^+(xy)$ for $x \in V^-$, which shows relative importance of the node in bridging communities.

The reaction rate ν, defined as the number of transitions from V_0 to V_1 happened in a unit time interval, can be obtained from adding up the probability current flowing out of the reactant state. This is stated by the next proposition.

Proposition 2.4 (Reaction rate)**.** The reaction rate is given by

$$\nu = \sum_{x\in A, y\in V} J(xy) = \sum_{x\in V, y\in B} J(xy). \tag{12}$$

Finally, the committor functions also give information about the time proportion that an equilibrium trajectory comes from A (the trajectory hits A last rather than B).

Proposition 2.5. The proportion of time that the trajectory comes from A (resp. from B) is given by

$$\rho^A = \sum_{x\in V} \pi(x)q(x), \quad \rho^B = \sum_{x\in V} \pi(x)(1-q(x)). \tag{13}$$

3 Semi-supervised Learning

3.1 Introduction

Problem: $x_1, x_2, ..., x_l \in V_l$ are labled data, that is data with the value $f(x_i), f \in V \to \mathbb{R}$ observed. $x_{l+1}, x_{l+2}, ..., x_{l+u} \in V_u$ are unlabled. Our concern is how to fully exploiting the information (like geometric structure in disbution) provided in the labeled and unlabeled data to find the unobserved labels.

This kind of problem may occur in many situations, like ZIP Code recognition. We may only have a part of digits labeled and our task is to label the unlabeled ones.

3.2 Harmonic Extension of Functions on Graph

Suppose the whole graph is $G = (V, E, W)$, where $V = V_l \cup V_u$ and weight matrix is partitioned into blocks $W = \begin{pmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{pmatrix}$. As before, we define $D = \text{diag}(d_1, d_2, ..., d_n) = \text{diag}(D_l, D_u),\ d_i = \sum_{j=1}^n W_{ij},\ L = D - W$. The goal is to find $f_u = (f_{l+1}, ..., f_{l+u})^T$ such that

$$\begin{aligned} \min \quad & f^T L f \\ s.t. \quad & f(V_l) = f_l \end{aligned}$$

where $f = \begin{pmatrix} f_l \\ f_u \end{pmatrix}$. Note that

$$f^T L f = (f_l^T, f_u^T) L \begin{pmatrix} f_l \\ f_u \end{pmatrix} = f_u^T L_{uu} f_u + f_l^T L_{ll} f_l + 2 f_u^T L_{ul} f_l$$

So we have:

$$\frac{\partial f^T L f}{\partial f_u} = 0 \Rightarrow 2L_{uu}f_u + 2L_{lu}f_u = 0 \Rightarrow f_u = -L_{uu}^{-1}L_{ul}f_l = (D_u - W_{uu})^{-1}W_{ul}f_l$$

3.3 Explanation from Transition Path Theory

We can also view the problem as a random walk on graph. Constructing a graph model with transition matrix $P = D^{-1}W = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$. Assume that the labeled data are binary (classification). That is, for $x_i \in V_l$, $f(x_i) = 0 \, or \, 1$. Denote

- $V_0 = \{i \in V_l : f_i = f(x_i) = 0\}$
- $V_1 = \{i \in V_l : f_i = f(x_i) = 1\}$
- $V = V_0 \cup V_1 \cup V_u$ where $V_l = V_0 \cup V_1$

With this random walk on graph P, f_u can be interpreted as hitting time or first passage time of V_1.

Proposition 3.1. Define hitting time

$$\tau_i^k = \inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, \quad k = 0, 1$$

Then for $\forall i \in V_u$,

$$f_i = Prob(\tau_i^1 < \tau_i^0)$$

i.e.

$$f_i = Prob(\text{trajectory starting from } x_i \text{ hit } V_1 \text{ before } V_0)$$

Note that the probability above also called committor function in Transition Path Theory of Markov Chains.

Proof. Define the committor function,

$$q_i^+ = Prob(\tau_i^1 < \tau_i^0) = \begin{cases} 1 & x_i \in V_1 \\ 0 & x_i \in V_0 \\ \sum_{j \in V} P_{ij} q_j^+ & i \in V_u \end{cases}$$

This is because $\forall i \in V_u$,

$$\begin{aligned} q_i^+ &= Pr(\tau_{iV_1} < \tau_{iV_0}) \\ &= \sum_j P_{ij} q_j^+ \\ &= \sum_{j\in V_1} P_{ij} q_j^+ + \sum_{j\in V_0} P_{ij} q_j^+ + \sum_{j\in V_u} P_{ij} q_j^+ \\ &= \sum_{j\in V_1} P_{ij} + \sum_{j\in V_u} P_{ij} q_j^+ \end{aligned}$$

$$\therefore \quad q_u^+ = P_{ul} f_l + P_{uu} q_u^+ = D_u^{-1} W_{ul} f_l + D_u^{-1} W_{uu} q_u^+$$

multiply D_u to both side and reorganize:

$$(D_u - W_{uu}) q_u^+ = W_{ul} f_l$$

If $D_u - W_{uu}$ is reversible, we get:

$$q_u^+ = (D_u - W_{uu})^{-1} W_{ul} f_l = f_u$$

i.e. f_u is the committor function on V_u. □

The result coincides with we obtained through the view of gaussian markov random field.

3.4 Explanation from Gaussian Markov Random Field

If we consider $f : V \to \mathbb{R}$ are Gaussian random variables on graph nodes whose inverse covariance matrix (precision matrix) is given by unnormalized graph Laplacian L (sparse but singular), i.e. $f \sim \mathcal{N}(0, \Sigma)$ where $\Sigma^{-1} = L$ (interpreted as a pseudo inverse). Then the conditional expectation of f_u given f_l is:

$$f_u = \Sigma_{ul} \Sigma_{ll}^{-1} f_l$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{ll} & \Sigma_{lu} \\ \Sigma_{ul} & \Sigma_{uu} \end{bmatrix}$$

Block matrix inversion formula tells us that when A and D are invertible,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = I \Rightarrow \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} S_D^{-1} & -A^{-1} B S_A^{-1} \\ -D^{-1} C S_D^{-1} & S_A^{-1} \end{bmatrix}$$

$$\begin{bmatrix} X & Y \\ Z & W \end{bmatrix} \cdot \begin{bmatrix} A & B \\ C & D \end{bmatrix} = I \Rightarrow \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} S_D^{-1} & -S_D^{-1} B D^{-1} \\ -S_A^{-1} C A^{-1} & S_A^{-1} \end{bmatrix}$$

where $S_A = D - CA^{-1}B$ and $S_D = A - BD^{-1}C$ are called Schur complements of A and D, respectively. The matrix expressions for inverse are equivalent when the matrix is invertible.

The graph Laplacian

$$L = \begin{bmatrix} D_l - W_{ll} & W_{lu} \\ W_{ul} & D_u - W_{uu} \end{bmatrix}$$

is not invertible. $D_l - W_{ll}$ and $D_u - W_{uu}$ are both strictly diagonally dominant, i.e. $D_l(i,i) > \sum_j |W_{ll}(i,j)|$, whence they are invertible by Gershgorin Circle Theorem. However their Schur complements $S_{D_u - W_{uu}}$ and $S_{D_l - W_{ll}}$ are still not invertible and the block matrix inversion formula above can not be applied directly. To avoid this issue, we define a regularized version of graph Laplacian

$$L_\lambda = L + \lambda I, \quad \lambda > 0$$

and study its inverse $\Sigma_\lambda = L_\lambda^{-1}$.

By the block matrix inversion formula, we can set Σ as its right inverse above,

$$\Sigma_\lambda = \begin{bmatrix} S^{-1}_{\lambda + D_u - W_{uu}} & -(\lambda + D_l - W_{ll})^{-1} W_{lu} S^{-1}_{\lambda + D_l - W_{ll}} \\ -(\lambda + D_u - W_{uu})^{-1} W_{ul} S^{-1}_{\lambda + D_u - W_{uu}} & S^{-1}_{\lambda + D_l - W_{ll}} \end{bmatrix}$$

Therefore,

$$f_{u,\lambda} = \Sigma_{ul,\lambda} \Sigma^{-1}_{ll,\lambda} f_l = (\lambda + D_u - W_{uu})^{-1} W_{ul} f_l,$$

whose limit however exits $\lim_{\lambda \to 0} f_{u,\lambda} = (D_u - W_{uu})^{-1} W_{ul} f_l = f_u$. This implies that f_u can be regarded as the conditional mean given f_l.

Note: the reasoning above is not quite right as Schur complements $S_{D_u - W_{uu}}$ and $S_{D_l - W_{ll}}$ are not invertible!

3.5 Well-posedness

One natural problem is: if we only have a fixed amount of labeled data, can we recover labels of an infinite amount of unobserved data? This is called well-posedness. [Nadler-Srebro 2009] gives the following result:

- If $x_i \in \mathbb{R}^1$, the problem is well-posed.
- If $x_i \in \mathbb{R}^d (d \geq 3)$, the problem is ill-posed in which case $D_u - W_{uu}$ becomes singular and f becomes a bump function (f_u is almost always zeros or ones except on some singular points).

Here we can give a brief explanation: From multivariable calculus,

This means in high dimensional case, to obtain a smooth generalization, we have to add constraints more than the norm of the first order derivatives. We also have a theorem to illustrate what kind of constraint is enough

for a good generalization:

Theorem 3.2 (Sobolev embedding Theorem)**.** $f \in \mathbf{W}^{s,p}(\mathbb{R}^d) \iff f$ has s'th order weak derivative $f^{(s)} \in \mathbf{L}_p$,

$$s > \frac{d}{2} \Rightarrow \mathbf{W}^{s,2} \hookrightarrow \mathbf{C}(\mathbb{R}^d).$$

So in $\mathbb{R}^d$, to obtain a continuous function, one needs smoothness regularization $\int \|\nabla^s f\|$ with degree $s > d/2$. To implement this in discrete Laplacian setting, one may consider iterative Laplacian L^s which might converge to high order smoothness regularization.