MA431 Spectral Graph Theory: Lecture 8

Ahmad Abdi Neil Olver

16 Intermezzo: matrix functions

Given a function $f : \mathbb{R} \to \mathbb{R}$, we can extend this to a function on symmetric matrices in the following way. First, given a diagonal matrix D, we define $f(D)$ by applying f to each diagonal entry of D: So $f(D)$ is diagonal, and $f(D)_{ii} = f(D_{ii})$. Then given a symmetric matrix M with diagonalization $M = P^{-1}DP$, define $f(M) = P^{-1}f(D)P$.

Put differently, $f(M)$ applies f to each eigenvalue, while keeping the eigenspace associated with each eigenvalue unchanged.

We will particularly use this for $f(x) = \exp(x)$. For any analytic function, this definition of f does correspond to the definition one gets from taking the Taylor series of f and directly applying this to matrices (exercise).

Finally, we remark that the same definition can be applied to matrices that are merely diagonalizable, and not necessarily symmetric.

17 Markov chains

Consider a connected graph $G = (V, E)$. An infinite random sequence $(X_0, X_1, X_2, \dots)$ is called a *simple random walk* on G if, for each $i \geq 0$, the condition distribution of X_{i+1} given X_i is the uniform distribution over the neighbours of X_i, and the choice of neighbour is independent of the history of the process before step i (that is, the process is Markovian).

For any $i \geq 1$, let $p^{(i)} \in \mathbb{R}^V_+$ denote the probability distribution of X_i; assume the initial distribution $p^{(0)}$ is given. Let A be the adjacency matrix of G, and let D be the diagonal matrix where D_{vv} is the degree of node v. Now let $W = AD^{-1}$.

Lemma 17.1. *$p^{(i+1)} = Wp^{(i)}$ for any i. That is, W is the transition matrix for simple random walk on G.*

Proof. Exercise. □

Observe that W is *not* symmetric, which is awkward. Fortunately, it is similar to a symmetric matrix, meaning we can still apply what we know about the spectrum of symmetric matrices. Let

$$\mathcal{A} = D^{-1/2}WD^{1/2} = D^{-1/2}AD^{-1/2}.$$

$\mathcal{A}$ is called the *normalized adjacency matrix of* G. You can think of it as the adjacency matrix of the weighted graph (G, w) where the weight of an edge $e = \{u, v\}$ is $w_e = 1/\sqrt{\deg(u) \cdot \deg(v)}$. By the Perron-Frobenius theorem, the eigenspace of the largest eigenvalue of $\mathcal{A}$ has dimension 1; the same must then hold for W. Note the vector d defined by $d_v = \deg(v)$ for all $v \in V$ is an eigenvector of W of eigenvalue 1, since

$$(Wd)_v = \sum_{w:\{v,w\}\in E} \frac{d_w}{\deg(w)} = \deg(v).$$

Further, no eigenvalue of W can be larger than 1; if x is an eigenvector of W, and v is an index for which $x_v/\deg(v)$ is maximal, then

$$(Wx)_v = \sum_{w:\{v,w\}\in E} \frac{x_w}{\deg(w)} \leq \sum_{w:\{v,w\}\in E} \frac{x_v}{\deg(v)} = x_v.$$

So d spans the eigenspace of W with eigenvalue 1. Define $\bar{p}$ to be the unique multiple of d with $\sum_v \bar{p}_v = 1$. This is a probability distribution, and we have determined that it is the unique probability distribution for which $W\bar{p} = \bar{p}$. This is the *stationary distribution*, the unique distribution that is invariant to the action of W.

18 Mixing times and the spectral gap

While the simple random walk is the most "obvious" random walk to define on G, it's not always entirely convenient. One pathology is that if G is bipartite, memory of the initial condition will remain encoded in the current time step i. If $X_0 = v$, then X_i will be on the same side of the bipartition for i even, and on the opposite side for i odd. In particular, $p^{(i)}$ does not converge to $\bar{p}$ (or to anything) as $i \to \infty$.

There are a number of possible alternative walks that avoid this problem. The first possibility is to consider the *lazy random walk*: in each step, we stay put with probability $1/2$, and otherwise we take a step along a uniformly chosen adjacent edge of the graph. The problem with simple random walk relates to a potential -1 eigenvalue, but lazy random walk has all eigenvalues in $[0, 1]$.

We will take a different route, and consider a *continuous time* random walk. It's just a bit nicer mathematically (admittedly, there is a bit more formal measure-theoretic machinery needed in the background, which we will mostly gloss over). There are two particularly natural continuous time walks:

- The *node-based* continuous time random walk (CTRW). Imagine that every node of the graph has an associated independent Poisson process of unit intensity. If the walk is at some vertex v at time t, it will remain there until the node "rings"—that is, until the smallest $s > t$ for which s is an increment of the point process for v—and then move to a uniformly random neighbour of v. So in other words, we are embedding the simple random walk in continuous time, such that after each step, we wait a random amount of time given by an exponential of unit rate.

  The stationary distribution is the same as simple random walk.

- Even nicer in many ways is the *edge-based* CTRW. Here, every *edge* of the graph has an associated independent Poisson process of unit intensity. If the walk is at some vertex v at time t, it will remain there until the first moment after t when one of the edges adjacent to v rings, at which point the walk moves to the other endpoint of this edge. Equivalently, we embed simple random walk in continuous time, but the amount of time spent at a node v has distribution given by an exponential of rate $\deg(v)$.

  Notice that the expected time of an individual visit for a node v is now $1/\deg(v)$, and so the stationary distribution becomes simply the uniform distribution.

We will continue by discussing the edge-based random walk, since things are cleanest, and then remark at the end how results transfer over to the node-based CRTW or lazy walk.

Let L denote the Laplacian of G, and $\lambda_1 = 0 < \lambda_2 < \cdots$ its spectrum. Let $p(t)$ for $t \in \mathbb{R}_+$ be the probability distribution for an edge-based CTRW on G, with $p(0)$ given.

Lemma 18.1. *For any $t \geq 0$,*

$$\frac{dp(t)}{dt} = -Lp(t).$$

Proof. (Informal.) Condition on $X(t) = v$, and consider some small interval $[t, t+\epsilon)$. To first order in ϵ, the probability that any adjacent edge $\{v, w\}$ rings is ϵ, and these are all independent events for different edges. So for any w adjacent to v, $X(t+\epsilon) = w$ with probability about ϵ, and with the remaining $1 - \deg(v)\epsilon$ probability, $X(t+\epsilon) = v$. So of the probability mass $p_v(t)$ at v at time t, approximately $\epsilon \deg(v) p_v(t)$ will leave in the time interval, and $\epsilon p_w(t)$ will arrive from each neighbour w. Thus (again to first order in ϵ)

$$p_v(t+\epsilon) \approx p_v(t)(1 - \epsilon \deg(v)) + \sum_{w:\{v,w\}\in E} \epsilon p_w(t) = -\epsilon(Lp(t))_v.$$

Thus $p(t+\epsilon) = p(t) - \epsilon Lp(t) + O(\epsilon^2)$, which is exactly the desired differential equation in the limit $\epsilon \to 0$. □

Remark 18.2. If one takes G to be a large d-dimensional grid graph, then this is a discrete approximation to the *heat equation*. If you think of the probability masses as representing some kind of temperature, we start with an initial temperature distribution $p(0)$, and through diffusion, this gradually equilibriates until eventually everything is at the same temperature.

It is from this differential equation that L can be seen as a "smoothing" operator.

Recalling perhaps some long-ago differential equations class, we can write down the solution to this differential equation in Lemma 18.1 immediately. It's the same as the case of a single-variable differential equation ($f'(x) = -\alpha f(x)$ has solution $f(x) = e^{-\alpha} f(0)$), but with matrices. A proof is included for completeness.

Lemma 18.3. *For any $t \geq 0$,*

$$p(t) = e^{-tL} p(0).$$

Proof. Let $Q^\top DQ$ be the orthogonal diagonalization of L, and make the change of variables $y = Qp$. The differential equation becomes

$$\frac{dy}{dt} = -Dy,$$

which decouples completely into independent single-variable differential equations. This yields $y_i(t) = y_i(0)e^{-tD_{ii}}$, that is, $y(t) = e^{-tD}y(0)$; changing back to the original basis yields the lemma. □

The stationary distribution for the edge-based continuous time walk is the uniform distribution, which we denote here by $\bar{q}$ (since if $p(t) = \bar{q}$, then $\frac{dp(t)}{dt} = -L\bar{q} = 0$). Fix some initial distribution $p(0)$. Observe that $p(0) - \bar{q}$ is orthogonal to $\bar{q}$, since $p(0)$ and $\bar{q}$ are both probability distributions. Now L is positive semidefinite, and $\bar{q}$ spans the eigenspace corresponding to the 0 eigenvalue. Thus $p(0) - \bar{q}$ lies in the span of the positive eigenspaces of L. This already guarantees convergence, at a rate controlled by the smallest positive eigenvalue of L. The proof just says the above again, slightly more carefully.

Lemma 18.4. *Let λ_2 be the second smallest eigenvalue of L. Then*

$$\|p(t) - \bar{q}\|_2 \le e^{-\lambda_2 t}\|p(0) - \bar{q}\|_2.$$

Proof. Recall that the matrix exponential preserves the eigenspaces but exponentiates the eigenvalues. So e^{-2tL} has spectrum $1 = e^{-2t\lambda_1} > e^{-2t\lambda_2} \ge e^{-2t\lambda_3} \ge \cdots$, and $\mathrm{span}(\{\bar{q}\})$ is the eigenspace associated with eigenvalue 1. Thus $e^{-2tL}\bar{q} = \bar{q}$, and for any q orthogonal to $\bar{q}$, $\frac{q^T e^{-2tL} q}{q^T q} \le e^{-t\lambda 2}$, by the Courant-Fischer Theorem. But

$$\langle p(0) - \bar{q}, \bar{q}\rangle = \tfrac{1}{n}\sum_{i\in V}(p_i(0) - \bar{q}_i) = 0,$$

since $p(0)$ and $\bar{q}$ are both probability distributions. So

$$\|p(t) - \bar{q}\|_2^2 = \|e^{-tL}p(0) - e^{-tL}\bar{q}\|_2^2 = (p(0) - \bar{q})^\top e^{-2\lambda_2 t}(p(0) - \bar{q}) \le e^{-2\lambda_2 t}\|p(0) - \bar{q}\|_2^2,$$

as required. □

A more standard measure of “distance” between two probability distributions is the *total variation distance*: given two probability distributions p, q on V,

$$d_{\mathrm{TV}}(p, q) := \tfrac{1}{2}\|p - q\|_1.$$

A reason that this is a standard measure is that it turns out that

$$d_{\mathrm{TV}}(p, q) = \max_{R\subseteq V}\left(\sum_{i\in R} p_i - \sum_{i\in R} q_i\right);$$

the probability of any event differs by at most $d_{\mathrm{TV}}(p, q)$. In any case, the difference between different measures won’t matter too much for our purposes. We certainly have

$$\|p - q\|_2 \le 2d_{\mathrm{TV}}(p, q) \le \sqrt{n}\|p - q\|_2.$$

As such,

Lemma 18.5. *For any initial distribution $p(0)$ and any $\epsilon > 0$, $d_{\mathrm{TV}}(p(t), \bar{q}) \leq \epsilon$ as long as*

$$t \geq \frac{1}{2\lambda_2} \ln(n/\epsilon).$$

Proof. Clearly $\|p(0) - \bar{q}\|_2 \leq \|p(0)\|_2 + \|\bar{q}\|_2 \leq 2$. So $d_{\mathrm{TV}}(p(t), \bar{q}) \leq \sqrt{n} e^{-\lambda_2 t}$, and the claim follows. □

The time required until the total variation distance stays below ϵ (for some given parameter ϵ) is called the *mixing time* of the random walk. Modulo the logarithmic term that we won't focus on, this lemma tells us that the mixing time is controlled by the second eigenvalue of L. (It's also clear from the proof that the mixing time cannot be much less than $1/\lambda_2$; the multiplying logarithmic terms will differ though, as they depend on how large the component of $p(0)$ in the λ_2-eigenspace can be, under the constraint that it must be a probability distribution. We won't go into these details.)

18.1 Node-based CTRW and lazy random walk

We'll quickly sketch what changes with node-based CTRW; results for lazy random walk are also similar.

The differential equation changes to

$$\frac{dp(t)}{dt} = -LD^{-1}p(t),$$

and hence

$$p(t) = e^{-tLD^{-1}}p(0).$$

Since LD^{-1} is not symmetric, this introduces some awkwardness. But we can define the *normalized Laplacian* $\mathcal{L}$:

$$\mathcal{L} = I - \mathcal{A} = D^{-1/2}LD^{-1/2}.$$

This is symmetric, and similar to LD^{-1}, so has the same spectrum. (It is not generally the Laplacian of some weighted graph, even though $\mathcal{A}$ can be viewed that way.) The second eigenvalue of $\mathcal{L}$ (which we will usually denote by ν_2 in these notes) thus controls the mixing time of the node-based CTRW: this logarithmic factor changes slightly, but the mixing time can be bounded by $1/\nu_2 \ln\bigl(1/(\epsilon \min_i \bar{p}_i)\bigr)$. The mixing time for lazy random walk is also the same, up to constant factors.

The value ν_2 is often called the *spectral gap* (the name comes from looking at the eigenvalues of $\mathcal{A}$ instead; ν_2 is the difference between the largest and second-largest eigenvalue of $\mathcal{A}$).

19 Conductance and Cheeger's inequality

We'll work with the node-based CTRW or lazy walk now, so that the mixing time is controlled by the second eigenvalue of $\mathcal{L}$. What is a good "combinatorial reason" that random walk might not mix on a given graph?

Suppose that there is a set $S \subseteq V$ of nodes for which:

- there are relatively few edges crossing S, and

- S and $V \setminus S$ are both not too small, and both contain many edges.

Then if we start our random walk at some node inside S, it may take a long time before it traverses any of the few edges crossing S. So the probability mass on $V \setminus S$ may stay quite small for a long time. Since $V \setminus S$ is not too small, this will make a significant contribution to $d_{\mathrm{TV}}(p(t), \bar{p})$.

To make this precise, define the *conductance* of a set $S \subseteq V$, written $\phi(S)$, by

$$\phi(S) := \frac{|\delta(S)|}{\min\{\operatorname{vol}(S), \operatorname{vol}(V \setminus S)\}},$$

where

$$\operatorname{vol}(S) := \sum_{v \in S} \deg(v).$$

Notice that the volume of S is, up to scaling, the same as $\sum_{v \in S} \bar{p}_v$. The conductance $\phi(G)$ of graph G is simply the minimum conductance over all choices of vertex sets S (excluding the empty set and the entire vertex set). Observe that the conductance of a cut, and hence of any graph, is bounded by 1.

Cheeger's inequality shows that the conductance of G *does* control the spectral gap, and hence the mixing time. The relationship is not exact: there is a quadratic gap between upper and lower bounds.

Theorem 19.1 (Cheeger's inequality)**.** *With ν_2 denoting the second eigenvalue of $\mathcal{L}$, we have*

$$\tfrac{1}{2}\nu_2 \leq \phi(G) \leq \sqrt{2\nu_2}.$$

The first inequality is called the "easy direction": it shows that if we have a cut of small conductance, then mixing is slow. The second inequality is called the "hard direction": it shows the less intuitively clear fact that if mixing is slow, then there must be a cut of reasonably low conductance that certifies this.

We will prove the two inequalities under the assumption that G is d-regular, meaning that $\mathcal{L} = \frac{1}{d}L$. The proof in the general case follows similar lines, though some additional technical awkwardnesses do need to be overcome.

19.1 The easy direction

By the CFT, and using $\mathcal{L} = \frac{1}{d}L$, we have

$$\nu_2 = \min_{x \neq 0, x \perp \mathbf{1}} \frac{x^\top L x}{d x^\top x} =: R(x).$$

Observe that $R(\mathbf{1}_S) = |\delta(S)| / \operatorname{vol}(S)$, and so $\phi(S) = \min\{R(\mathbf{1}_S), R(\mathbf{1}_{V \setminus S})\}$. However, $\mathbf{1}_S$ and $\mathbf{1}_{V \setminus S}$ are not generally orthogonal to $\mathbf{1}$, so we cannot deduce that $\nu_2 \leq \phi(S)$ from this. Instead, we choose an appropriate convex combination of $\mathbf{1}_S$ and $\mathbf{1}_{V \setminus S}$ that is orthogonal to $\mathbf{1}$ (we took a different approach in the lecture).

So let $z = \frac{1}{|S|}\mathbf{1}_S - \frac{1}{|V\setminus S|}\mathbf{1}_{V\setminus S}$, so that $z \perp \mathbf{1}$. Then $\nu_2 \leq R(z)$. We then just compute:

$$\begin{aligned}
R(z) &= \frac{\sum_{vw\in E}(z_v - z_w)^2}{d\sum_{v\in V} z_v^2} \\
&= \frac{|\delta(S)| \cdot \left(\frac{1}{|S|} + \frac{1}{|V\setminus S|}\right)^2}{d\left(\frac{|S|}{|S|^2} + \frac{|V\setminus S|}{|V\setminus S|^2}\right)} \\
&= \frac{|\delta(S)|}{d} \cdot \left(\frac{1}{|S|} + \frac{1}{|V\setminus S|}\right) \\
&\leq 2\frac{|\delta(S)|}{d} \cdot \max\{1/|S|, 1/|V \setminus S|\} \\
&= 2\phi(S).
\end{aligned}$$

Thus $\nu_2 \leq 2\phi(G)$ as required.

19.2 The hard direction

We will see the proof of the hard direction, also for a d-regular graph G, next time. The proof of the hard direction will be *constructive*: it will provide us an efficient way of finding a cut of small conductance (if there is one). This will be of independent interest, since it provides a way of partitioning the graph.

Acknowledgements

See, e.g., [1] for much more detail and rigour on Markov chains and mixing.

References

[1] D. Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, 2009.

Exercises

1. Show that the spectral approach to extending a function $f : \mathbb{R} \to \mathbb{R}$ to symmetric matrices matches the Taylor expansion approach, whenever f is analytic.

2. Give an example to demonstrate that the hitting time is not a symmetric quantity.

3. Is the commute time monotone, in the sense that adding edges to a network can only decrease the commute time between two nodes?

4. The *cover time* of a connected graph G (which we will denote by $\mathrm{Cover}(G)$ is the maximum over all choices of starting node s, of the expected time needed for a simple random walk starting from s to visit each node of G.

   Let R be the maximum over $u, v \in V$ of the effective resistance between u and v in G; also let m denote the number of edges of G. Show that

$$\mathrm{Cover}(G) = \Omega(mR) \text{ and } \mathrm{Cover}(G) = O(\log|V| \cdot mR).$$

5. Show that the commute time between any two nodes of a connected graph with n nodes is at most n^3. Find an example where the commute time between some pair of vertices is $\Omega(n^3)$.

6. Give a bound of the form $O\Big(1/\nu_2 \log(K/\epsilon)\Big)$ for the mixing time of the lazy random walk. (One option would be to show a bit more, and argue that the mixing time for lazy simple random walk is within constant factor of mixing time of node-based CTRW).

7. Show that the spectrum of the normalized Laplacian of a graph lies in $[0, 2]$.

8. Consider C_n, the cycle of length n. Show that one direction of Cheeger's inequality is tight for C_n up to constant factors, for all n.

9. Consider Q_n, the n-dimensional binary hypercube; identify its vertex set with $\{-1, 1\}^n$.

   Determine the second eigenvalue of the Laplacian of Q_n. *(Hint: for $a \in \{-1, 1\}^n$, consider vectors $x \in \mathbb{R}^{V(Q_n)}$ of the form $x_v = (-1)^{a \cdot v}$ for all $v \in \{-1, 1\}^n$.)*

   Hence show that one direction of Cheeger's inequality is tight (up to constant factors) for the hypercube. Also give bounds on the mixing time, both directly from the second eigenvalue, and from Cheeger's inequality.