CS 466 / 666 Algorithm Design and Analysis . Spring 2020

Lecture 14 : Mixing time

We study convergence rate to the stationary distribution in undirected graphs using spectral analysis.

Random walks in undirected graphs

We will prove the fundamental theorem of Markov chains in the special case of undirected graphs using an algebraic approach.

Furthermore, this approach can be used to analyze the mixing time of the random walks.

Matrix formulation

Let $G=(V,E)$ be an undirected graph.

Let $p_0 : V \to \mathbb{R}$ be the initial distribution, i.e. $p_0(v) \geq 0 \ \forall v \in V$ and $\sum_{v\in V} p_t(v) = 1$.

A common setting is we start a random walk at a specific vertex i, in which case p_0 is the i-th unit vector.

In each step of the random walk, we move to a uniformly random neighbor of the current vertex.

Let $p_t : V \to \mathbb{R}$ be the probability distribution on the vertices after t steps of the random walk.

Then, $p_t(u) = \sum_{v\in N(u)} p_{t-1}(v) \cdot \frac{1}{d(v)}$ for all $u \in V$ and for all $t \geq 1$, where $d(v)$ denotes the degree of vertex v.

Let A be the adjacency matrix of G, and D be the diagonal degree matrix where $D_{i,i} = d(i)$.

Then, the above linear equations can be written compactly as one matrix equation $p_{t+1} = AD^{-1}p_t$.

Thus, by a simple induction, we have $p_t = (AD^{-1})^t p_0$.

Stationary distribution

Recall that a probability distribution $\pi : V \to \mathbb{R}$ is a stationary distribution of the random walk if $\pi = (AD^{-1})\pi$

Observe that π is an eigenvector of AD^{-1} with eigenvalue 1.

In undirected graphs, there is a natural stationary distribution based on the degree of vertices.

Let $\vec{d} : V \to \mathbb{R}$ be the vector with the i-th entry being $d(i)$, and $m = |E|$.

Claim $\pi = \frac{\vec{d}}{2m}$ is a stationary distribution of the random walk.

proof It is easy to check that $\sum_{v\in V} \pi(v) = 1$.

Also, we can check that $\pi = (AD^{-1})\pi$, as $(AD^{-1})\pi = (AD^{-1})\frac{\vec{d}}{2m} = \frac{1}{2m}A\vec{1} = \frac{\vec{d}}{2m} = \pi$. □

Fundamental theorem for undirected graphs

Does $p_t \to \frac{\vec{d}}{2m}$ when $t \to \infty$ no matter what is the initial distribution p_0?

Not necessarily.

If the graph is disconnected, then the distribution p_t depends on the initial distribution, e.g. which component does the starting vertex belongs to.

Note that an undirected graph is connected iff the corresponding Markov chain is irreducible.

Even if the graph is connected, p_t may not be converging to a distribution.

Consider a connected bipartite graph. If we start from a vertex on the left side of the bipartite graph, then the current vertex will be on the left side in even time steps and on the right side in odd time steps. So, p_t is not converging.

Note that a connected bipartite graph is non-bipartite iff the corresponding Markov chain is aperiodic.

It turns out that these are the only obstacles, and we will prove the following special case of the fundamental theorem.

<u>Theorem</u> For any finite, connected, non-bipartite graph, $p_t = (AD^{-1})^t p_0 \to \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0.

Lazy random walks

We can remove the assumption of non-bipartiteness if we consider a slightly modified process called the lazy random walk, in which we stay at the current vertex with probability $1/2$ and move to a uniformly random neighbor with probability $1/2$.

We can just think of lazy random walks as adding a self loop on each vertex to remove periodicity.

More precisely, $p_t(u) = \frac{1}{2} p_{t-1}(u) + \frac{1}{2} \sum_{v \in N(u)} p_{t-1}(v) \cdot \frac{1}{d(v)}$ for all $u \in V$ and all $t \geq 1$.

More compactly, $p_t = (\frac{1}{2} I + \frac{1}{2} AD^{-1}) p_{t-1}$ and inductively $p_t = (\frac{1}{2} I + \frac{1}{2} AD^{-1})^t p_0$.

<u>Theorem</u> For any finite and connected graph, $p_t = (\frac{1}{2} I + \frac{1}{2} AD^{-1})^t p_0 \to \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0.

Spectral analysis

Let $W = AD^{-1}$ be the random walk matrix, and $Z = \frac{1}{2} I + \frac{1}{2} AD^{-1}$ be the lazy random walk matrix.

To study the behavior of the repeated applications of a linear transformation such as $p_t = W^t p_0$, it is very useful to understand the spectrum of W.

The matrix W is not symmetric (but is symmetric if the graph is d-regular), so a priori we do not know whether W has real eigenvalues and an orthonormal basis of eigenvectors.

Fortunately, W is <u>similar</u> to a symmetric matrix, i.e. $W = B^{-1} A B$ for some symmetric matrix A, and thus W has real eigenvalues.

Specifically, $D^{-\frac{1}{2}} W D^{\frac{1}{2}} = D^{-\frac{1}{2}} (AD^{-1}) D^{\frac{1}{2}} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = \mathcal{A}$, the normalized adjacency matrix of G.

and so $W = D^{\frac{1}{2}} \mathcal{A} D^{-\frac{1}{2}}$.

Lemma W and $\mathcal{A}$ have the same spectrum $1 = \alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n \geq -1$.

proof Let $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$ be the eigenvalues of $\mathcal{A}$, with corresponding orthonormal eigenvectors $v_1, v_2, \dots, v_n$.

We know from L13 that $1 \geq \alpha_1$ and $\alpha_n \geq -1$.

Note that $u_i = D^{\frac{1}{2}} v_i$ is an eigenvector of W with eigenvalue α_i, because

$$W u_i = (AD^{-1})(D^{\frac{1}{2}} v_i) = AD^{-\frac{1}{2}} v_i = D^{\frac{1}{2}}(D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) v_i = D^{\frac{1}{2}} \mathcal{A} v_i = \alpha_i D^{\frac{1}{2}} v_i = \alpha_i u_i.$$

Since $v_1, \dots, v_n$ are linearly independent and $D^{\frac{1}{2}}$ is of full rank when the graph is connected, $u_1, \dots, u_n$ are also linearly independent, and thus the spectrum of W is $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$. □

Idea

We illustrate the idea of the spectral analysis in the simpler setting when W has real eigenvalues $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$ with orthonormal eigenvectors $v_1, v_2, \dots, v_n$ (note that it happens when G is regular).

We write $p_0 = c_1 v_1 + \dots + c_n v_n$, a linear combination of the eigenvectors.

Then $W^t p_0 = c_1 \alpha_1^t v_1 + c_2 \alpha_2^t v_2 + \dots + c_n \alpha_n^t v_n$, where $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$ are eigenvalues of $\mathcal{A}$.

We know that $\alpha_1 = 1$.

Also, we know that $\alpha_2 < 1$ if and only if G is connected.

Furthermore, we know that $\alpha_n > -1$ if and only if G is non-bipartite (HW4).

Using these connections between eigenvalues and combinatorial properties, we can conclude that when G is connected and non-bipartite, then $W^t p_0 \to c_1 v_1$ as $t \to \infty$ because $\alpha_i^t \to 0$ for $2 \leq i \leq n$ as $t \to \infty$.

This proves that there is a unique limiting distribution when G is connected and non-bipartite.

To analyze the convergence rate, let $\lambda = \min\{1 - \alpha_2, \alpha_n + 1\}$ be the "spectral gap", we can then bound the mixing time in terms of λ in a pretty natural way (i.e. how fast $\alpha_i^t \to 0$).

For lazy random walks, the spectrum of Z is $\frac{1}{2}(1+\alpha_1) \geq \dots \geq \frac{1}{2}(1+\alpha_n) \geq 0$, i.e. no negative eigenvalues.

Then, the "spectral gap" is just $\lambda = \frac{1-\alpha_2}{2}$, which is related to the graph conductance through Cheeger's inequality, and so we can conclude that the lazy random walk of an expander graph converges quickly.

In general, W does not have an orthonormal basis of eigenvectors, but we will again use the fact that W is similar to $\mathcal{A}$ and $\mathcal{A}$ has an orthonormal basis of eigenvectors to carry out the analysis.

Proof of the fundamental theorem for undirected graphs

We are ready to prove that for any finite, connected, non-bipartite graph, $p^t = W^t p_0 \to \frac{\vec{d}}{2m}$ as $t \to \infty$ $\forall p_0$.

Let $1 = \alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n \geq -1$ be the eigenvalues of $\mathcal{A}$ and $v_1, v_2, \dots, v_n$ the corresponding orthonormal eigenvectors.

Write $W^t p_0 = \left(D^{\frac{1}{2}} \mathcal{A} D^{-\frac{1}{2}}\right)^t p_0 = D^{\frac{1}{2}} \mathcal{A}^t D^{-\frac{1}{2}} p_0$.

To take advantage of the orthonormal basis of eigenvectors of $\mathcal{A}$, we write $D^{-\frac{1}{2}} p_0 = c_1 v_1 + \dots + c_n v_n$.

Then, $W^t p_0 = D^{\frac{1}{2}} \mathcal{A}^t D^{-\frac{1}{2}} p_0 = D^{\frac{1}{2}} \mathcal{A}^t \left(\sum_{i=1}^n c_i v_i \right) = D^{\frac{1}{2}} \left(\sum_{i=1}^n c_i \alpha_i^t v_i \right) = c_1 D^{\frac{1}{2}} \alpha_1^t v_1 + \sum_{i=2}^n c_i D^{\frac{1}{2}} \alpha_i^t v_i$.

Recall that $\alpha_1 = 1$, $\alpha_2 < 1$ when G is connected, and $\alpha_n > -1$ when G is non-bipartite.

This implies that $|\alpha_i| < 1$ for $2 \leq i \leq n$, and so $\alpha_i^t \to 0$ as $t \to \infty$.

Hence, $W^t p_0 \to c_1 D^{\frac{1}{2}} v_1$ as $t \to \infty$. This shows that the random walk has a unique limiting distribution.

It remains to check that $c_1 D^{\frac{1}{2}} v_1 = \frac{\vec{d}}{2m}$.

Recall that $D^{\frac{1}{2}} \vec{1}$ is an eigenvector of $\mathcal{A}$ with eigenvalue 1 (check this).

So, $v_1 = D^{\frac{1}{2}} \vec{1} / \| D^{\frac{1}{2}} \vec{1} \|$ as $\|v_1\| = 1$.

Since $\| D^{\frac{1}{2}} \vec{1} \|^2 = \sum_{i=1}^n d(i) = 2m$, we have $v_1 = D^{\frac{1}{2}} \vec{1} / \sqrt{2m}$.

Using $v_1, \dots, v_n$ are orthonormal and $D^{-\frac{1}{2}} p_0 = c_1 v_1 + \dots + c_n v_n$, we have

$$c_1 = \langle D^{-\frac{1}{2}} p_0, v_1 \rangle = \langle D^{-\frac{1}{2}} p_0, D^{\frac{1}{2}} \vec{1} / \sqrt{2m} \rangle = \frac{1}{\sqrt{2m}} \langle p_0, \vec{1} \rangle = \frac{1}{\sqrt{2m}}$$ as p_0 is a probability distribution.

Therefore, $c_1 D^{\frac{1}{2}} v_1 = \frac{1}{\sqrt{2m}} D^{\frac{1}{2}} \left(D^{\frac{1}{2}} \vec{1} / \sqrt{2m} \right) = \frac{1}{2m} D \vec{1} = \frac{\vec{d}}{2m}$, as desired.

This proves the fundamental theorem for undirected graph.

The proof for the lazy random walk version is essentially the same, with the spectrum of Z being $\frac{1}{2}(1+\alpha_1) \geq \dots \geq \frac{1}{2}(1+\alpha_n) \geq 0$, and so we only need to use the fact that $\alpha_2 < 1$ when G is connected (but don't need to assume anything about α_n). We leave the proof as an exercise.

Mixing time

We would like to understand how quickly $W^t p_0$ converges to $\pi = \frac{\vec{d}}{2m}$.

A standard measure of the closeness is $d_{TV}(\pi, p_t) = \frac{1}{2} \sum_{i=1}^n |\pi(i) - p_t(i)| = \frac{1}{2} \| \pi - p_t \|_1$.

Definition The ε-mixing time of the random walk is defined as the smallest t such that

$\| \pi - p_t \|_1 \leq \varepsilon$ regardless of p_0.

We will bound the mixing time of the random walk based on the "spectral gap".

Define the spectral gap $\lambda = \min\{1 - \alpha_2, 1 - |\alpha_n|\}$, so that $|\alpha_i| \leq 1 - \lambda$ for $2 \leq i \leq n$.

Theorem The ε-mixing time of random walk is upper bounded by $\frac{1}{\lambda} \log\left(\frac{n}{\varepsilon}\right)$.

Proof We start from $p_t = W^t p_0 = \pi + \sum_{i=2}^n c_i \alpha_i^t D^{\frac{1}{2}} v_i$ in the proof of the fundamental theorem above.

So, $\| p_t - \pi \|_2^2 = \| D^{\frac{1}{2}} \sum_{i=2}^n c_i \alpha_i^t v_i \|_2^2 \leq \| D^{\frac{1}{2}} \|_{op}^2 \| \sum_{i=2}^n c_i \alpha_i^t v_i \|_2^2$ where $\|A\|_{op} := \max \underline{\|Ax\|_2}$ is the operator norm

Proof We start from $p_t = W^t p_0 = \pi + \sum_{i=2} c_i \alpha_i^t D^{\frac{1}{2}} v_i$ in the proof of the fundamental theorem above.

So, $\|p_t - \pi\|_2^2 = \|D^{\frac{1}{2}} \sum_{i=2}^n c_i \alpha_i^t v_i\|_2^2 \leq \|D^{\frac{1}{2}}\|_{op}^2 \|\sum_{i=2}^n c_i \alpha_i^t v_i\|_2^2$ where $\|A\|_{op} := \max \frac{\|Ax\|_2}{\|x\|_2}$ is the operator norm

$= d_{max} \|\sum_{i=2}^n c_i \alpha_i^t v_i\|_2^2$ where d_{max} is the maximum degree in G

$= d_{max} \cdot \sum_{i=2}^n c_i^2 \alpha_i^{2t}$ as $v_1, \ldots, v_n$ are orthonormal

$\leq d_{max} (1-\lambda)^{2t} \sum_{i=2}^n c_i^2$ as $|\alpha_i| \leq 1-\lambda$ for $2 \leq i \leq n$

$\leq d_{max} (1-\lambda)^{2t} \|D^{-\frac{1}{2}} p_0\|_2^2$ as $D^{-\frac{1}{2}} p_0 = \sum_{i=1}^n c_i v_i$ and $v_1, \ldots, v_n$ are orthonormal

$\leq \frac{d_{max}}{d_{min}} (1-\lambda)^{2t}$ as $\|D^{-\frac{1}{2}} p_0\|_2 \leq \|D^{-\frac{1}{2}}\|_{op} \|p_0\|_2 = \frac{\|p_0\|_2}{d_{min}} \leq \frac{\|p_0\|_1}{d_{min}} = \frac{1}{d_{min}}$

It follows that $\|p_t - \pi\|_2 \leq \sqrt{\frac{d_{max}}{d_{min}}} (1-\lambda)^t \leq \sqrt{n} (1-\lambda)^t \leq \sqrt{n} e^{-\lambda t}$.

By Cauchy Schwarz, $\|p_t - \pi\|_1 \leq \sqrt{n} \|p_t - \pi\|_2 \leq n e^{-\lambda t}$.

Therefore, when $t \geq \frac{1}{\lambda} \log\left(\frac{n}{\varepsilon}\right)$, we have $\|p_t - \pi\|_1 \leq \varepsilon$. □

This implies that when $\lambda = \Omega(1)$, then the mixing time is only $O(\log \frac{n}{\varepsilon})$, logarithmic in the graph size.

For example, when the graph is regular (so that $\pi = \frac{\vec{1}}{n}$) and $\lambda = \Omega(1)$, then we only need to simulate the random walk for $O(\log n)$ steps to sample an almost uniform vertex from the graph.

Lazy random walk

What graphs have a constant spectral gap?

It is easier to answer this question by considering lazy random walks.

The same mixing time analysis works for lazy random walks.

The spectrum for the lazy random walk matrix is $1 = \frac{1}{2}(1+\alpha_1) \geq \frac{1}{2}(1+\alpha_2) \geq \ldots \geq \frac{1}{2}(1+\alpha_n)$,

and so the spectral gap for lazy random walk is simply $\frac{1}{2}(1-\alpha_2) = \frac{1}{2}\lambda_2$,

where λ_2 is the second smallest eigenvalue of the normalized Laplacian matrix.

By Cheeger's inequality, we know that $\lambda_2 \geq \phi(G)^2/2$ from the hard direction.

Corollary The ε-mixing time of lazy random walk is upper bounded by $\frac{2}{\phi(G)^2} \log\left(\frac{n}{\varepsilon}\right)$.

So, whenever a graph has constant conductance, then we know that the lazy random walk mixes in $O(\log n)$ time, a sublinear time algorithm to generate an almost uniform sample.

This is a very important result for the analysis of Markov chains and the design of fast random sampling algorithm.

Random Sampling

One of the most important applications of random walks is to design fast algorithms for random sampling.

We just discuss the main ideas here without going into the details.

Card shuffling

We have a deck of 52 cards. Our goal is to obtain a random permutation of the cards using simple operations.

Let's say in every step we pick a uniform random card and put it on the top of the deck.

① Will we get an almost uniformly random permutation if we repeat this step for sufficiently many steps?

② If so, how many steps are enough to guarantee an almost random permutation?

These correspond to the first two basic questions of random walks in this specific setting.

We can view these as questions about random walks on a bigger "state" graph, where each vertex corresponds to a permutation of the card, and permutation A has a directed edge to permutation B if we can obtain permutation B by moving one card in permutation A to the top.

So, we have a state graph with $52!$ vertices and each vertex is of indegree 52 and outdegree 52.

The random shuffling by the simple operation corresponds exactly to a random walk on this state graph.

It should be clear that ① is just asking what is the limiting distribution of this random walk, and indeed the unique limiting distribution is the uniform distribution.

And ② is just asking the mixing time of this random walk, and this is the key question to analyze.

With this idea, one can then study the performance of different shuffling rules.

For example, a famous result is that seven steps of "riffle" shuffling are enough to "mix" things up.

Random perfect matching / random spanning tree

You may wonder a random permutation is not difficult to generate anyway, but the real power of this method is to do random sampling of some complicated objects.

For example, can we generate a random perfect matching in a graph efficiently?

To do so, we can try to define some simple random walk among the matchings of the graph (e.g. add an edge, delete an edge, find an augmenting path of length three, etc).

Then, show that there is a unique limiting distribution, and all perfect matchings are equally likely.

This step is usually easy. The difficult step is to prove that the random walk converges quickly.

There are different methods to bound the convergence rate, and this is a vast topic with own textbooks.

Some popular methods include "coupling", second eigenvalue, and graph conductance.

Cheeger's inequality is useful here to connect second eigenvalue to graph conductance.

A major result in this area is a polynomial time algorithm to sample a perfect matching in a bipartite graph. and this is used to design a good approximation algorithm for estimating the permanent of a non-negative matrix. where deterministic polynomial time algorithms get nowhere close!

Also, a new result (2020) shows that one can generate a random spanning tree in near-linear time using this random walk approach.

It is interesting to note how Cheeger's inequality is used.

In situations where we want to bound the conductance (say in explicit constructions of constant degree expander graphs). we don't know how to bound the conductance directly and instead we bound the second eigenvalue (since the constructions are usually algebraic) and use Cheeger's to bound the conductance.

In situations where we want to bound the mixing time, we want to bound the second eigenvalue, but somehow we don't know how to bound it directly and instead we bound the conductance (since the problems are combinatorial) and use Cheeger's inequality to establish the mixing time.

But this is exactly the power of Cheeger's inequality. relating an algebraic property to a combinatorial property. so that we can use a totally different perspective to look at a problem.

References

- Course notes of Dan Spielman on "spectral graph theory", and CS 860 Spring 2019.
- Course notes by Shayan Oveis Gharan on "counting and sampling".
- A book "Markov chains and mixing time" by Levin, Peres, Wilmer.