CS 270   Combinatorial Algorithms and Data Structures, Spring 2015

Lecture 23: Mixing time

We study the convergence rate to the stationary distribution in undirected graphs using spectral analysis and discuss some algorithmic applications.

Matrix Formulation

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

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

The initial distribution is p_0, e.g. if we start at vertex i, then p_0 is the i-th standard unit vector.

Consider the random walk where in each step it moves to a uniformly random neighbor of the present node.

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

Let A be the adjacency matrix of G and D be the diagonal matrix where the i-th diagonal entry is $d(i)$.

Then the above equations can be written compactly as $p_{t+1} = AD^{-1} p_t$, and thus $p_t = (AD^{-1})^t p_0$.

Stationary Distribution

A probability distribution π is a <u>stationary distribution</u> if $\pi = (AD^{-1})\pi$. So, once we reach a stationary distribution, it stays as a stationary distribution forever. (Note that it is common to write the probability distribution as a row vector, e.g. $\pi = \pi(D^{-1}A)$, but we write it as a column.)

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

Let $\pi(v) = \frac{d(v)}{\sum_{u \in V} d(u)} = \frac{d(v)}{2m}$. Let $\vec{d}$ be the vector with the i-th entry being $d(i)$. Then $\pi = \frac{\vec{d}}{2m}$.

It is easy to check that $\pi = (AD^{-1})\pi$, as $(AD^{-1})\pi = (AD^{-1})\frac{\vec{d}}{2m} = \frac{1}{2m}(A\vec{1}) = \frac{\vec{d}}{2m} = \pi$.

Is it the unique stationary distribution?

Not necessarily. For example if the graph is disconnected, then the distribution after many steps depends on the initial distribution (e.g. which component does the starting vertex belongs to). Actually, even if the graph is connected we may not reach a stationary distribution. Consider a connected bipartite graph. If the starting vertex is on the left side, then the walk will be on the left

side on even time steps, while it is opposite if the starting vertex is on the right side.

It turns out that these are the only obstacles, that every connected non-bipartite graph has a unique
   stationary distribution, and it will be reached regardless of the initial distribution.

More generally, every connected graph (even bipartite graphs) has a unique stationary distribution
   if we consider the <u>lazy</u> random walk where we stay at the same vertex with probability $1/2$.

That is, $p_t(w) = \frac{1}{2} \cdot p_{t-1}(v) + \frac{1}{2} \sum_{u \in N(w)} p_{t-1}(u) \cdot \frac{1}{d(u)}$.

In matrix form, we have $p_t = (\frac{1}{2}I + \frac{1}{2}AD^{-1}) p_{t-1}$.

We denote $W = \frac{1}{2}I + \frac{1}{2}AD^{-1}$ as the <u>walk matrix</u>, and so $p_{t+1} = Wp_t$ and $p_t = W^t p_0$.

Spectrum

To study the behavior of the repeated applications of a linear transformation like $W^t p_0$, it is usually very
   useful to know the spectrum of the matrix.

What is the spectrum of W?

Notice that $W = \frac{1}{2}I + \frac{1}{2}AD^{-1}$ is not a symmetric matrix, so we don't know whether its eigenvalues are
   real or there is an orthonormal basis of eigenvectors.

Fortunately, W is similar to a symmetric matrix, because $D^{-\frac{1}{2}}WD^{\frac{1}{2}} = D^{-\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}AD^{-1})D^{\frac{1}{2}} = \frac{1}{2}I + \frac{1}{2}D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

Recall that $D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = \mathcal{A}$ is just the normalized adjacency matrix. So, $W = D^{\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}\mathcal{A})D^{-\frac{1}{2}}$.

Let $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$ be the eigenvalues of $\mathcal{A}$, and $v_1, v_2, \ldots, v_n$ be the corresponding eigenvectors.

Then $D^{\frac{1}{2}}v_i$ is an eigenvector of W with eigenvalue $\frac{1}{2}(1+\alpha_i)$, because $WD^{\frac{1}{2}}v_i = D^{\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}\mathcal{A})D^{-\frac{1}{2}}D^{\frac{1}{2}}v_i$

$$= D^{\frac{1}{2}}(\frac{1}{2}v_i + \frac{1}{2}\alpha_i v_i) = \frac{1}{2}(1+\alpha_i)D^{\frac{1}{2}}v_i.$$

Since $v_1, \ldots, v_n$ are linearly independent and $D^{\frac{1}{2}}$ is of full rank, we have $D^{\frac{1}{2}}v_1, \ldots, D^{\frac{1}{2}}v_n$ are linearly independent,
   and thus the spectrum of W is $\frac{1}{2}(1+\alpha_1) \geq \frac{1}{2}(1+\alpha_2) \geq \ldots \geq \frac{1}{2}(1+\alpha_n)$.

Recall from last time that $\alpha_1 = 1$, $\alpha_2 < 1$ iff G is connected, $\alpha_n > -1$ iff G is non-bipartite (HW3).

It turns out that $\alpha_2 < 1$ and $\alpha_n > -1$ is a necessary and sufficient condition for unique limiting distribution,
   and so we could prove that a connected non-bipartite graph has a unique limiting distribution.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of W, we have $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \geq 0$ if G is connected.

Furthermore, since $W\pi = (D^{\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}\mathcal{A})D^{-\frac{1}{2}})\pi = (\frac{1}{2}I + \frac{1}{2}AD^{-1})\pi = \pi$, any eigenvector of W with eigenvalue

one is a constant multiple of π.

Outline

Suppose a matrix M has an orthonormal basis of eigenvectors $v_1, \ldots, v_n$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$.

We can then write any vector $x = c_1 v_1 + \ldots + c_n v_n$, and $M^t x = M^t(c_1 v_1 + \ldots + c_n v_n) = c_1 \lambda_1^t v_1 + \ldots + c_n \lambda_n^t v_n$.

So, if $\lambda_1 = 1$ and $1 > \lambda_i \geq 0$ for $i > 1$, then $M^t x$ will converge to $c_1 v_1$, and thus M has a stationary distribution and the convergence rate depends on the gap between λ_1 and λ_2.

This is how we are going to prove that there is a unique stationary distribution for W, and the convergence rate depends on the value of λ_2.

Except that W is not a symmetric matrix, and in general does not have an orthonormal basis of eigenvectors.

But once again we use the fact that W is similar to the normalized adjacency matrix $\mathcal{A}$, and $\mathcal{A}$ has an orthonormal basis of eigenvectors.

Uniqueness

Consider $W^t p_0 = D^{\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}\mathcal{A})D^{-\frac{1}{2}} p_0$

To take advantage of the orthonormal basis of eigenvectors of $\mathcal{A}$, let's write $D^{-\frac{1}{2}} p_0 = \sum_i c_i v_i$ where $v_1, v_2, \ldots, v_n$ are the eigenvectors of $\mathcal{A}$, and $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$ are the corresponding eigenvalues of $\mathcal{A}$.

Then $W^t p_0 = D^{\frac{1}{2}}(\frac{1}{2}I + \frac{1}{2}\mathcal{A})^t (\sum_i c_i v_i) = D^{\frac{1}{2}} \sum_i (\frac{1}{2}(1+\alpha_i))^t c_i v_i = D^{\frac{1}{2}} \sum_i c_i \lambda_i^t v_i = c_1 D^{\frac{1}{2}} v_1 + D^{\frac{1}{2}} \sum_{i \geq 2} c_i \lambda_i^t v_i$,

where we used the above notation that $\lambda_i = \frac{1}{2}(1+\alpha_i)$ and the fact that $\lambda_1 = 1$.

Recall that if G is connected then $\lambda_2 < 1$, and $\lambda_i \geq 0$ for all i.

So, $W^t p_0$ will converge to $c_1 D^{\frac{1}{2}} v_1$. Let's see whether it is what we expect to be, i.e. $c_1 D^{\frac{1}{2}} v_1 = \pi = \frac{\vec{d}}{2m}$.

Recall that $\vec{d}^{\frac{1}{2}}$ is an eigenvector of $\mathcal{A}$ with eigenvalue one, and so $v_1 = \vec{d}^{1/2} / \|\vec{d}^{1/2}\| = \vec{d}^{1/2} / \sqrt{2m}$.

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

Therefore, $c_1 D^{\frac{1}{2}} v_1 = \frac{1}{\sqrt{2m}} D^{\frac{1}{2}} \left(\frac{\vec{d}^{1/2}}{\sqrt{2m}}\right) = \frac{\vec{d}}{2m} = \pi$, and thus $W^t p_0 \Rightarrow \pi$ as $t \to \infty$.

This proves that the lazy random walk converges to π in a connected graph for any initial distribution.

Convergence Rate

Suppose $\lambda_2 = 1 - \varepsilon$. This is same as saying $\alpha_2 = 1 - 2\varepsilon$ for the normalized adjacency matrix $\mathcal{A}$.

Recall that $p_t = W^t p_0 = \Pi + D^{\frac{1}{2}} \sum_{i \geq 2} c_i \lambda_i^t v_i \leq \Pi + (1-\varepsilon)^t D^{\frac{1}{2}} \sum_{i \geq 2} c_i v_i$.

So $\|p_t - \Pi\| \leq (1-\varepsilon)^t \|D^{\frac{1}{2}} \sum_{i \geq 2} c_i v_i\| \leq (1-\varepsilon)^t \|D^{1/2}\| \|\sum_{i \geq 2} c_i v_i\| \leq (1-\varepsilon)^t \max_v \sqrt{d(v)} \|\sum_{i \geq 2} c_i v_i\|$.

Recall that $D^{-\frac{1}{2}} p_0 = \sum_{i \geq 1} c_i v_i$, and so $\|D^{-\frac{1}{2}} p_0\|^2 = \|\sum_{i \geq 1} c_i v_i\|^2 = \sum_{i \geq 1} c_i^2 > \sum_{i \geq 2} c_i^2 = \|\sum_{i \geq 2} c_i v_i\|^2$ as v_i are orthonormal.

Note that $\|D^{-\frac{1}{2}} p_0\| \leq \frac{1}{\min \sqrt{d(v)}}$, and so $\|p_t - \Pi\| \leq (1-\varepsilon)^t \frac{\max_v \sqrt{d(v)}}{\min_v \sqrt{d(v)}}$

Therefore, when $t = \Omega(\frac{1}{\varepsilon} \log n)$, $\|p_t - \Pi\|$ is inverse polynomial in n.

So, when there is a constant gap, the random walk will converge to Π in $\Theta(\log n)$ steps.

By Cheeger's inequality, the random walk will converge to Π in $O(\log n / \phi^2(G))$ steps where $\phi(G)$ is the conductance.

Random sampling

One of the most important applications of random walk is to do random sampling.

Consider the card shuffling problem for example. We have a deck of 52 cards and our goal is to obtain a random permutation of the cards. Let's say in every step we pick a random card and put it on the top of the deck. Will we get a random permutation if we repeat this step for sufficiently many steps? If so, how many steps are enough to guarantee an almost random permutation? These are exactly the first two basic questions above in this special case. To study these questions, one way is to study the corresponding graph where we have one vertex for each permutation, and there is an edge from a vertex u to a vertex v if we can obtain the permutation corresponding to v by moving one card to the top from the permutation corresponding to u. (Note that in this example we have a directed graph, we can modify it to an undirected graph but we won't bother to do it.) In this setting, the first question can be answered if we show that the stationary distribution is the uniform distribution. The second question can be answered by bounding the mixing time of the random walk. One can then study the performance of different shuffling rules, for example it was proven that seven steps of "riffle" shuffling are enough.

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 distribution. For example, can you generate a random perfect matching in a graph efficiently? To do so, one 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 two). Show that there is a unique stationary distribution, and all perfect matchings are equally likely in the stationary distribution. This step is usually easy. The difficult step is to show that the random walk converges quickly. We can analyze it by the results that we have just discussed, by studying the second eigenvalue or the conductance of the graph, or by some other techniques such as coupling. This is also where the Cheeger's inequality is very useful, by relating the second eigenvalue and the conductance.

It is interesting to note that 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 can't bound the conductance directly and instead we bound the second eigenvalue (since the constructions are usually algebraic) and use Cheeger to bound the conductance.

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

Derandomization

Suppose we have a randomized algorithm with error probability $1/100$ by reading n random bits.

This means that among the 2^n n-bit strings, only $2^n/100$ are "bad" strings.

To amplify the success probability, one can pick k random n-bit strings, then the error probability is at most $(1/100)^k$ with kn random bits.

We show how to exponentially decrease the error probability while using only $n + ck$ bits for a constant c.

Construct a d-regular expander graph with 2^n vertices and $\varepsilon \leq 1/10$.

In the first step, we use an n-bit random string, with error probability $\leq 1/100$.

In the subsequent steps, instead of picking independent n-bit strings, we do a $(k-1)$-step random walk and use

the strings corresponding to the vertices in the random walk as "random" n-bit strings.

After we try the k "random" strings, we use the majority answer as our answer.

What is the error probability of this algorithm?

We output the wrong answer if the wrong answer is the majority.

Since the number of bad strings is at most $2^n/100$, the error probability is at most $\left(\frac{2}{\sqrt{5}}\right)^k$ as we will show

   (by letting X to be the number of bad strings in the theorem in the next section).

The number of random bits used is $n + (k-1)\log_2 d$ since each random neighbor can be chosen with $\log_2 d$ bits.

Note that it works for two-sided error randomized algorithms as well.

This is just one example, expander graphs have many applications in derandomization.

Random walks in expander graphs   (optional)

Let $G = (V, E)$ be a d-regular graph. Let's assume $\max\{\alpha_2, |\alpha_n|\} \le \varepsilon \le \frac{1}{10}$ where α_i is the eigenvalue of $\mathcal{A}$.

Let $X \subseteq V$ with $|X| \le |V|/100$.

Let v_0 be the initial random vertex, and $v_1, v_2, ..., v_t$ be the vertices produced by the t steps of the random walk.

Let $S = \{i : v_i \in X\}$. We choose v_0 as a uniformly random vertex, each vertex is of probability $1/n$.

Theorem $\Pr(|S| > t/2) \le \left(\frac{2}{\sqrt{5}}\right)^{t+1}$.

First we set up the matrix formulation of the problem.

The initial distribution is $\pi = \vec{1}/n$.

Let χ_X and $\chi_{\bar{X}}$ be the characteristic vectors of X and $\bar{X}$, where $\bar{X} = V - X$.

Let I_X be the diagonal matrix with a 1 in the i-th diagonal entry if $i \in X$, and similarly $I_{\bar{X}}$.

Let p be a probability distribution. Then $I_X p$ is the probability vector on X.

Then $q = W I_X \pi$ is the probability vector where the initial random vertex is in S, where $W = \mathcal{A}$.

Then, the probability that the walk is in X at precisely the time steps in S is

   $\vec{1}^T I_{z_t} W I_{z_{t-1}} W I_{z_{t-2}} W \dots I_{z_2} W I_{z_1} W \pi$, where $z_i = X$ if $i \in S$ and $z_i = \bar{X}$ if $i \notin S$.

We will prove that this probability is at most $(1/5)^{|S|}$.

This will imply that $\Pr(|S| > t/2) \le \sum_{|S| > t/2} \Pr(\text{the walk is in } X \text{ at precisely the times in } S) \leftarrow$ union bound

$$\le 2^{t+1} (1/5)^{(t+1)/2} = (2/\sqrt{5})^{t+1}$$

Recall that $\|M\| = \max\limits_x \|Mx\|/\|x\| = \max\limits_x x^T M x / x^T x$ for symmetric M. You can check that $\|I_x\| = \|I_{\bar{x}}\| = \|W\| = 1$.

We will prove that $\|I_x W\| \leq 1/5$, and this would imply that the above probability is at most $(1/5)^{|S|}$.

To see this, $\vec{1}^T I_{z_t} W I_{z_{t-1}} W \dots I_{z_2} W I_{z_1} W \pi = \vec{1}^T (I_{z_t} W)(I_{z_{t-1}} W) \dots (I_{z_2} W)(I_{z_1} W) \pi$

$$\leq \|\vec{1}^T\| \, \|(I_{z_t} W)(I_{z_{t-1}} W) \dots (I_{z_2} W)(I_{z_1} W) \pi\| \quad \text{Cauchy-Schwartz}$$

$$\leq \|\vec{1}^T\| \left(\prod_{i=1}^{t} \| I_{z_i} W\| \right) \|\pi\| \quad \text{triangle inequality}$$

$$\leq \|\vec{1}^T\| (1/5)^{|S|} \|\pi\| \quad \text{as } \|I_{z_i} W\| \leq 1/5 \text{ if } z_i = X \text{ and}$$
$$\|I_{z_i} W\| \leq \|I_{z_i}\| \|W\| = 1 \text{ if } z_i = \bar{X}$$

$$\leq (1/5)^{|S|} \quad \text{as } \|\vec{1}^T\| = \sqrt{n} \quad \text{and} \quad \|\pi\| = 1/\sqrt{n}.$$

It remains to prove that $\|I_x W\| \leq 1/5$.

Let x be any nonzero vector and write $x = c_1 v_1 + \dots + c_n v_n$ where $v_1 = \vec{1}/\sqrt{n}$ and $c_1 = \langle x, v_1 \rangle = \sum x_i / \sqrt{n}$.

$\|I_x W x\|^2 = \|I_x W (c_1 v_1 + \dots + c_n v_n)\|^2 \leq \|I_x \sum\limits_{i \geq 1} c_i \alpha_i v_i\|^2 \leq 2\| I_x c_1 \alpha_1 v_1\|^2 + 2\| I_x \sum\limits_{i \geq 2} c_i \alpha_i v_i\|^2$

$$\leq 2\| I_x \left(\tfrac{\sum x_i}{n} \vec{1} \right)\|^2 + 2\|I_x\|^2 \| \sum\limits_{i \geq 2} c_i \alpha_i v_i\|^2$$

$$= 2|X| \left(\tfrac{\sum x_i}{n} \right)^2 + 2 \sum\limits_{i \geq 2} \| c_i \alpha_i v_i\|^2 \quad \text{by orthogonality}$$

$$\leq 2|X| \cdot \tfrac{\|x\|^2}{n} + 2 \sum\limits_{i \geq 2} \| c_i \alpha_i v_i\|^2 \quad \text{by Cauchy Schwartz}$$
$$\text{as } (\sum x_i)^2 \leq (\sum x_i^2)(\sum 1^2)$$

$$\leq \tfrac{1}{50} \|x\|^2 + 2 \sum\limits_{i \geq 2} c_i^2 \alpha_i^2$$

$$\leq \tfrac{1}{50} \|x\|^2 + 2\varsigma^2 \sum\limits_{i \geq 2} c_i^2 \leq \tfrac{1}{50}\|x\|^2 + \tfrac{1}{50} \|x\|^2$$

Thus $\|I_x W x\| \leq \tfrac{1}{5} \|x\|^2$, finishing the proof.

References   The presentation mostly follows the course notes of "spectral graph theory" by Dan Spielman.