0368.4163 Randomness and Computation March 11, 2009

Lecture 7

Lecturer: Ronitt Rubinfeld *Scribe: Dror Marcus, Yaron Ziner*

1 Lecture Outline

1.1 List Of Subjects

- Linear Algebra Review
- Mixing Times
- Mixing & Saving Random Bits
- Mixing & Uniform Generation Of Matching

1.2 General Plan

In the last lecture, the notion of stationary distribution was defined. This lecture will focus on analyzing the time it takes to reach (or at least come fairly close to) such stationary distribution. We will refer to the time taking us to reach the stationary distribution as the *Mixing Time.* To do so we will be using linear algebra, therefore, first we will refresh on some basic definitions and theorems of linear algebra.

We will also classify Markov Chains by the time it takes to reach their respective stationary distributions. Markov Chains with fast mixing times (such as CLIQUE) will be classified as *good* while Markov Chains with slow mixing time (such as PATH or LOLLIPOP) we be classified as *bad.*

Next, a new way for reducing random bits used in a randomized algorithm using random walks will be described.

Getting to the end of the lecture we will begin the subject of uniform generation of matching in a given graph using Markov Chains. We will do so by showing an algorithm for the construction of a *good* Markov Chain Graph (i.e. with fast mixing times).

2 Linear Algebra Review

2.1 Basic Definitions And Theorems

Definition 1 *v is an* eigenvector *of a matrix A with corresponding* eigenvalue λ *if* $Av = \lambda v$.

Definition 2 *The L_2 norm of vector v, marked as $\|v\|_2$, is $\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$*

Definition 3 *A set of vectors, $v^{(1)}, v^{(2)}, ..., v^{(n)}$ are* orthonormal *if $\forall i, j \in [1, n]$:*

$$v^{(i)} \cdot v^{(j)} = \begin{cases} 1 & if\ i = j \text{ (i.e. `normal')} \\ 0 & if\ i \neq j \text{ (i.e. `orthogonal')} \end{cases}$$

Theorem 4 *(Unproven in this scribe) Given an $n \times n$ transition matrix P s.t. P is real and symmetrical, there exists eigenvectors $v^{(1)}, v^{(2)}, .., v^{(n)}$ for which the following holds:*

1. *$v^{(1)}, v^{(2)}, .., v^{(n)}$ form an orthonormal basis.*
2. *$v^{(1)}, v^{(2)}, .., v^{(n)}$ have corresponding eigenvalues s.t. $\lambda_1 = 1, \lambda_2, .., \lambda_n$, s.t. $|\lambda_1| \leq |\lambda_2| \leq .. \leq |\lambda_n|$.*

3. $v^{(1)} = \frac{1}{\sqrt{n}}(1,1,...,1)$.

Observation 5 *If $v^{(1)}, v^{(2)}, .., v^{(n)}$ form an orthonormal basis, then **any** vector w is expressible as*

$$w = \sum \alpha_i v^{(i)}$$

also w's L_2 norm can be expressed as

$$\|w\|_2 = \sqrt{(\sum_i \alpha_i v^{(i)}) \cdot (\sum_j \alpha_j v^{(j)})} = \sqrt{\sum_{i,j} \alpha_i \alpha_j v^{(i)} v^{(j)}} = \sum_i \alpha_i^2$$

2.2 Important Facts About Eigenvectors And Eigenvalues

Given $n \times n$ matrix P with eigenvectors $v^{(1)}, v^{(2)}, .., v^{(n)}$ and corresponding eigenvalues $\lambda_1, \lambda_2, .., \lambda_n$:

- For every scalar α, matrix αP has eigenvectors $v^{(1)}, v^{(2)}, .., v^{(n)}$ with corresponding eigenvalues $\alpha\lambda_1, \alpha\lambda_2, .., \alpha\lambda_n$.

  Proof: Given $v^{(i)}$, $1 \le i \le n$, the following holds - $(\alpha P)v^{(i)} = \alpha(Pv^{(i)}) = \alpha\lambda_i v^{(i)}$.

- Matrix P+I eigenvectors are $v^{(1)}, v^{(2)}, .., v^{(n)}$ with corresponding eigenvalues $\lambda_1+1, \lambda_2+1, .., \lambda_n+1$.

  Proof: Given $v^{(i)}$, $1 \le i \le n$, the following holds - $(I+P)v^{(i)} = (Iv^{(i)}+Pv^{(i)}) = v^{(i)}+\lambda_i v^{(i)} = (1+\lambda_i)v^{(i)}$.

- Matrix P^k eigenvectors are $v^{(1)}, v^{(2)}, .., v^{(n)}$ with corresponding eigenvalues $\lambda_1^k, \lambda_2^k, .., \lambda_n^k$.

  Proof: Given $v^{(i)}$, $1 \le i \le n$, the following holds - $(P^k)v^{(i)} = (P^{k-1})Pv^{(i)} = (P^{k-1})\lambda_i v^{(i)} = \lambda_i(P^{k-1})v^{(i)} = ... = \lambda_i^{k-1}Pv^{(i)} = \lambda_i^k v^{(i)}$.

- If P is a stochastic matrix then, $\forall i \in [1, n]$, $|\lambda_i| \le 1$.

2.3 Example

Let P be the transition matrix for a d_regular undirected graph (Notice that in this case the matrix is double stochastic, i.e., every row and every column sum to one, and has d non zero entries), then:

1. The vector $(\frac{1}{n}, ..., \frac{1}{n})$ is the stochastic distribution and is an eigenvector of P with eigenvalue one i.e., $(\frac{1}{n}, ..., \frac{1}{n})P = 1(\frac{1}{n}, ..., \frac{1}{n})$. Also notice that $|(\frac{1}{n}, ..., \frac{1}{n})|_1 = 1$.
2. The vector $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$ is also an eigenvector of P with eigenvalue one. Also notice that $|(\frac{1}{n}, ..., \frac{1}{n})|_2 = 1$.

3 Mixing Times

Definition 6 *Given $\varepsilon > 0$ the* mixing time *$T(\varepsilon)$ of a Markov Chain A with a unique stationary distribution $\overline{\Pi}$ is the minimum t s.t. $\forall \Pi^{(0)}$, $|\overline{\Pi} - \Pi^{(0)}A^t|_1 < \varepsilon$*

Definition 7 *Markov Chain A is "*rapidly mixing*" if $T(\varepsilon) = poly(log|V|, log\frac{1}{\varepsilon})$, where V stands for the number of states in A.*

Next we will show the connection between properties of eigenvectors described in the previous section and mixing times.

Theorem 8 *Given transition matrix P of an undirected (= symmetric), non bipartite, d_regular, connected graph, with* Π_0 *as the start distribution and* $\overline{\Pi}$ *its stationary distribution then:*

$$\|\Pi_0 P^t - \overline{\Pi}\|_2 \leq |\lambda_2|^t$$

Proof P is real and symmetric hence, its eigenvectors $v^{(1)}, v^{(2)}, .., v^{(n)}$ form an orthonormal basis with eigenvalues $\lambda_1 = 1$, $|\lambda_1| \leq |\lambda_2| \leq .. \leq |\lambda_n|$, so for any vector, in particular the vectors Π_0:

$$\Pi_0 = \sum_{i=1}^{n} \alpha_i v_{(i)},$$

$$|\Pi_0 P^t = \sum_{i=1}^{n} \alpha_i v^{(i)} P^t = \sum_{i=1}^{n} \alpha_i \lambda_i^t v^{(i)}$$

We know that $\lambda_1 = 1$ so

$$\Pi_0 = \alpha_1 v^{(1)} + \sum_{i=2}^{n} \alpha_i \lambda_i^t v^{(i)}$$

Therefore

$$\|\Pi_0 P^t - \alpha_1 v^{(1)}\|_2 = \|\sum_{i=2}^{n} \alpha_i \lambda_i^t v^{(i)}\|_2 = \sqrt{\sum_{i=2}^{n} \alpha_i^2 \lambda_i^{2t}}$$

Where the last transition was possible due to the orthonormality of the eigenvectors $v^{(i)}$. Also notice that $\forall i > 2$, $|\lambda_2|^2 \geq |\lambda_i|^2$. Using this in the previous eq. we get

$$\|\Pi_0 P^t - \alpha_1 v^{(1)}\|_2 \leq |\lambda_i^t| \sqrt{\sum_{i=2}^{n} \alpha_i^2} \leq |\lambda_i^t|$$

The last transition is due to the fact that

$$1 = |\Pi_0|_1 \geq \|\Pi_0\|_2 = \sqrt{\sum_{i=1}^{n} \alpha_i^2} \Rightarrow \sqrt{\sum_{i=2}^{n} \alpha_i^2} \leq 1$$

Now notice that as t increases (indicating more steps taken at the random walk) $|\lambda_2|^t$ is getting smaller, so $|\Pi_0 P^t - \overline{\Pi}\|_2$ is getting closer to zero. Therefore the stationary distribution must be $\overline{\Pi} = \alpha_1 v^{(1)}$ and we get from the last equation

$$\|\Pi_0 P^t - \overline{\Pi}\|_2 \leq |\lambda_2|^t$$

■

4 Reducing Random Bits Using Random Walks

In this part of the lecture we will use our previous result on the eigenvalues of the matrix P and apply it to a new way of reducing the number of random bits needed with amplification of randomized algorithms.

Let's consider a decision problem L, for which we have a randomized algorithm A with one sided-error as follows:

- A's behavior is:

  - If $x \in L$ then $Pr[A(x)\ outputs\ "x \in L"] < \frac{1}{100}$
  - If $x \notin L$ then $A(x)\ alway\ outputs\ "x \notin L"$

- A uses at most $r(n)$ random bits on problems of size n.

An example of such a decision problem is the STCON for which we saw a randomized algorithm who uses random walks that is never mistaken when vertices s, t are not connected, and errs with small probability when s, t are connected.

We saw in previous lectures a few ways to amplify a randomized algorithm's performance. In summation, using these previously discussed methods, in order to achieve an error probability of 2^{-k} we could:

- By running algorithm A for k times. we were able to get an error probability of 2^{-k} by using $O(k \cdot r)$ random bits.

- Using pair-wise independent hash functions we were able to improve the previous result and get an error probability of 2^{-k} using only $O(k + r)$ random bits.

- In this lecture we will use random walks on graph to achieve an amplification scheme which only uses $r + O(k)$ random bits and achieving the optimal constant in front of the r.

The main idea behind the amplification scheme is the use a d-regular graph G (for some const d) with 2^r nodes, each one corresponds the a random assignment w of random bits in algorithm A. There are a few important notes we should mention about the graph G:

- Graph G is a huge graph that could not be constructed in polynomial time. The algorithm will use the graph by computing the adjacency matrix locally only when needed. We thus have a requirement that for a given node G's adjacent edges can be computed efficiently.

- Graph G has no connection to algorithm A nor to its input. The graph is a generic graph suited to amplify any algorithm that confirms to the terms stated above.

- G should have *extractor* property:

  - For G's transition matrix P the following holds: $|\lambda_2| \leq \frac{1}{10}$

In this lecture we will not show how to construct these graphs.

The Scheme:

1. Pick uniformly at random $w \in_R \{0,1\}^r$, a start node in G

2. Repeat the following k times:

   (a) $w \leftarrow$ random neighbor of w

   (b) Run $A(x)$ with w as the random bits.

3. If ever see $" \in L"$ output $"x \in L"$ otherwise output $"x \notin L"$

Since step (1) takes r random bits and step (2.a) takes $k \cdot \log d$ bits, the scheme uses $r + O(k)$ random bits as claimed.

Claim 9 *The scheme presented above will err with probability at most* $(\frac{1}{5})^k$

Proof We divide the analysis into two classes:

If $x \notin L$, than obviously the scheme will never err since $A(x)$ never errors in this case.

If $x \in L$, Let's denote $B = \{w | A(x)\ with\ random\ bits\ w\ is\ incorrect\}$. From the definition of A, it is clear that $|B| \leq \frac{2^r}{100}$

Let's define a diagonal matrix N with

$$N_{ww} = \begin{cases} 0 & \text{if w} \in \text{B} \\ 1 & \text{otherwise} \end{cases}$$

Notice that if q is a distribution on the nodes of G then:

$$|qN|_1 = Pr_{w \in q}[w\ is\ bad\ random\ string\ for\ A(x)]$$

or in general:

$$|q(NP)^k N| = Pr_{w \in q}[w\ and\ the\ next\ k\ steps\ are\ bad\ strings\ for\ A(x)]$$

we can see that in order to prove a bound on this probability we should bound $|q(NP)^k N|$. To do this, we prove the following lemma:

Lemma 10 *For all distributions* Π *we have* $\|\pi PN\|_2 \leq \frac{1}{5}\|\pi\|$

Observation 11 *For every vector* s *we have* $\|sN\|_2 = \sqrt{\sum_{i=1}^{2^r}(s_i N_{ii})^2} \leq \sqrt{\sum_{i=1}^{2^r} s_i^2} = \|s\|_2$

Proof Express π as $\pi = \sum_{i=1}^{2^r} \alpha_i v_i$ (where v_i are eigenvectors of P). Note that $v_1 = (\frac{1}{\sqrt{2^r}}, ..., \frac{1}{\sqrt{2^r}})$.

Let's analyze:

$$\|\pi PN\| = \|\sum_{i=1}^{2^r} \alpha_i v_i PN\|_2 = \|\sum_{i=1}^{2^r} \alpha_i \lambda_i v_i N\|_2 \leq \|\alpha_1 \lambda_1 v_1 N\|_2 + \|\sum_{i=2}^{2^r} \alpha_i \lambda i v_i N\|_2$$

The first step is because v_i are eigenvectors of P, and the second step is due to Cauchy-Schwartz.

If we analyze each term separately we get:

1. $\|\alpha_1 \lambda_1 v_1 N\|_2 = \|\alpha_1 v_1 N\|_2 = |\alpha_1| \|v_1 N\|_2 = |\alpha_1| \cdot \sqrt{\sum_{i \in B}(\frac{1}{\sqrt{2^r}})^2} = |\alpha_1| \sqrt{\frac{|B|}{2^r}} \leq \frac{|\alpha_1|}{10} \leq \frac{\|\pi\|_2}{10}$

   the first step is due to $\lambda_1 = 1$ and the last step is due to $|\alpha_1| \leq \sqrt{\sum \alpha_i^2} = \|\pi\|_2$

2. $\|\sum_{i=2}^{2^r} \alpha_i \lambda_i v_i N\|_2 \leq \|\sum_{i=2}^{2^r} \alpha_i \lambda_i v_i = \sqrt{\sum_{i=2}^{2^r} \alpha_i^2 \lambda_i^2} \leq \frac{1}{10}\sqrt{\sum_{i=2}^{2^r} \alpha_i^2} \leq \frac{\|\pi\|_2}{10}$

   Here we used the fact that $\forall i > 2. |\lambda_2| \geq |\lambda_i|$.

Overall, we got $\|\pi PN\|_2 \leq \frac{\|\pi\|_2}{10} + \frac{\|\pi\|_2}{10} = \frac{\|\pi\|_2}{5}$ ■

How do we use this lemma?

$Pr[A(x)\ is\ incorrect\ for\ all\ w] \leq \left|U_{2^r}(NP)^k\right|_1 \leq \sqrt{2^r} \cdot \|U_{2^r}(PN)^k\|_2 \leq \sqrt{2^r}\|U_{2^r}\| \cdot \frac{1}{5^k} = \frac{1}{5^k}$

At the second step we used the inequality $|v|_1 \leq \sqrt{d}\|v\|_2$ (where d stands for v's dimensionality.) ■

5 Mixing & Uniform Generation Of Matching

We will now show a general construction of a rapidly mixing Markov Chain, and implement this construction on an algorithm for the uniform generation of matching un a given graph. Note that the proof of correctness will follow on the next lecture.

5.1 General Plan

Given graph $G = (V, E)$, Construct (a "huge") markov chain s.t.

1. The states of the Markov Chain, V', corresponds to objects we are trying to uniformly generate (exponential number of states i.e. $|V'| = O(2^{|V|})$)

   E.g. For matching the states will be {M|M is a matching in G}.

2. The Transition correspond to local changes s.t. the degree is polynomial in original G, and we will only need polytime to choose a transition.

   E.g. For matching remove, add edges, or unchanged (self loop).

3. Start at an arbitrary node (we cannot pick a random node).

4. Analyze the stationary distribution (ideally uniform).

5. Find the mixing time. Remember that by definition, in order for the Markov Chain to be rapidly mixing, we need it to be $poly(log|V'|)$ or $poly(V)$), so we are want to show that rapid mixing times ⇒ polytime algorithm.

5.2 Algorithm For Matching

```
1. Pick Edge $e \in_R E$
2. if $e \in M$
       set $M \leftarrow M - \{e\}$
3. else if $M \cup \{e\}$ is matching
       set $M \leftarrow M \cup \{e\}$
4. else
       stay in current state
```

Notice that the graph constructed is:

- Undirected (the graph is reversible - every edge deleted can be added later).
- Connected.
- Not bipartite (can stay at the same state, i.e. containing self loops).
- With degree $|E|$, therefore, the stationary distribution is uniform.