Randomized Methods in Computation*
Lecture 10: Random Walks on Expanders

Notes taken by Omer Angel, Dani Halevi, Amos Gilboa

Spring 2001

Summary: The purpose of this lecture is to show that taking an l step random walk in an expander graph is in a way similar to choosing l vertices at random. The advantage of using a random walk on an expander over uniform independent sampling is that it requires only $\log_2 d$ random bits for each step, whereas when you choose l different vertices at random each vertex requires $\log_2 N$ random bits. As a warm up, we'll show that in $O(\log N)$ random steps on an N-vertex expander beginning at any vertex we'll reach a roughly uniform distribution on the vertices.

We also recall the relation between two definitions of expanders and describe some known constructions of expanders.

1 Two Definitions

Recall the combinatorial and algebraic definitions of expansion:

Definition 1 *A sequence of graphs $\{G_N\}$ has combinatorial expansion C if any of vertices S with $|S| < |V|/2$ has $|\Gamma^-(S)| > C|S|$ where $\Gamma^-(S)$ denotes the neighbours of a set S that are not in S.*

Definition 2 *A sequence of graphs $\{G_N\}$ has algebraic expansion with gap $\epsilon > 0$ if for all n the second eigenvalue of G_n satisfies $|\lambda_2| \leq 1 - \epsilon$.*

remark. We assume that the eigenvalues are ordered by non-increasing order, i.e.

$$1 = \lambda_1 > |\lambda_2| \geq \ldots |\lambda_n|$$

(Recall that we proved that for the normalized adjacency matrix of any connected regular and non-bipartite graph $\lambda_1 = 1$ and $|\lambda_i| < 1$ for all $i > 1$.)

The following theorem implies that the combinatorial definition of expansion based on boundary sizes and the algebraic definition based on eigenvalue separation are equivalent. In fact it gives a quantitative translation between the two notions. (This Theorem is not proved here).

Theorem 3 *Algebraic expansion $\Longleftrightarrow$ Combinatorial expansion. In particular:*

- *If $\{G_N\}$ are expanders w.r.t. the algebraic definition with constant ϵ then they are expanders w.r.t. the combinatorial definition with constant $C \geq \frac{2\epsilon}{1+2\epsilon}$.*

* Lecture Notes for a course given by Oded Goldreich at the Weizmann Institute of Science, Israel.

- *If $\{G_N\}$ are d-regular expanders w.r.t. the combinatorially definition with constant C then they are expanders w.r.t the algebraic definition with $\epsilon \geq \frac{C^2}{(4+2C^2)d}$.*

Note. These inequalities are not tight. One can see it by beginning with expansion parameter C, substituting in the second inequality to derive ϵ, and substituting this ϵ in the first inequality to derive

$$C' \geq \frac{C^2}{C^2 + (2 + C^2)d}$$

Since this way we got $C' < C$, it follows that at least one of the inequalities is not always tight.

Using the second inequality, we have a practical and easy way to conclude that a graph does not have an expansion C, just calculate ϵ and check whether $\epsilon \geq \frac{C^2}{(4+2C^2)d}$, if this is not the case - obviously the graph does not have expansion C. On the other hand if $\epsilon \leq \frac{C}{2-2C}$ then the graph certainly has expansion C. However, as we just observed the inequalities are not tight, so if we find that ϵ is in the interval $[\frac{C^2}{(4+2C^2)d}, \frac{C}{2-2C}]$ then we can not determine from the theorem whether or not the graph has expansion C. Thus this theorem is more important as a qualitative statement.

Just as there is a relation between the combinatorial expansion of a graph G and the expansion of its power G^k (i.e. given by $1 + C_{G^k} > (1 + C_G)^k$) there is also a similar relation between the algebraic expansions of G and G^k. This is given by:

Proposition 1.1 $\lambda_2(G^k) = \lambda_2(G)^k$

Proof: The normalized adjacency matrix of G^k is given by $A' = A^k$ (i.e. A raised to power k). If v is an eigenvector of A with eigenvalue λ then

$$A'v = A^k v = \lambda^k v$$

So the eigenvalues of G' are the k'th powers of the eigenvalues of G. Since A has a basis of eigenvectors there are no addition eigenvalues of A'. ■

2 Known Constructions of Expanders

Following are two known ways to build big expander graphs. These graphs are strongly constructible (i.e. given a vertex, we can find it's neighbours in time poly(n), where n is the number of bits needed to represent a vertex).

- For any N such that $N = m^2$ for some integer m, we construct an expander in the following way: the set of vertices will be the set of all pairs (x, y) with $x, y \in \{0, \ldots, m-1\}$. The neighbours of a vertex (x, y) are the pairs

  $$\{(x, y \pm (2x + \sigma)) : \sigma \in \{0, 1\}\} \cup \{(x \pm (2y + \sigma), y) : \sigma \in \{0, 1\}\}$$

  where operations are modulo m. Each vertex can be represented in $2 \log m$ bits its neighbours can be found in time poly$(\log m)$.

  This graph is 8-regular and has $\lambda \approx 0.9$.

- For primes p, q primes such that $p, q \equiv 1 \pmod 4$ and $(\frac{p}{q}) = 1$ and for any integer k there is an N-vertex expander with $N = \frac{q^k(q^{2k}-1)}{2}$. The vertices are 2×2 matrices over the field with q^k elements ($GF(q^k)$). This graph is $(p+1)$-regular and has $\lambda \approx \frac{2}{\sqrt{d}}$.

  These graphs are less convenient to work with, because of the constraints on p, q (and so on N). However, it is still easy to find the neighbours of any given vertex. They also have the advantage of having λ close to optimal, as seen from:

Theorem 4 *For any d-regular graph $\lambda_2 \geq \frac{2}{\sqrt{d}}$*

3 Random Walks on Expanders

Throughout this section λ will denote an upper bound on $|\lambda_2|$.

3.1 Mixing Time of Random Walks

Let $X_0, X_1, \ldots, X_l$ be a random walk on an expander, starting at X_0 (i.e. X_0 is given either deterministically or by some distribution, and X_{i+1} is chosen uniformly at random from the neighbours of X_i). We show that for $l = O(\log N)$ the distribution of X_l tends to the uniform distribution on V, regardless of the starting point (or its distribution). In a way this number of steps is optimal. This is due to the fact that the diameter of an expander graph is logarithmic in N (because the graph has constant degree). Thus we can't expect to get from one vertex to some other vertices in less than $O(\log n)$ steps.

Theorem 5 *Let $X_0, X_1, \ldots, X_l$ be a random walk on an expander graph starting at X_0. For every δ there is an $l = O(\log \frac{1}{\delta})$ such that for every vertex v*

$$\left| \Pr(X_l = v) - \frac{1}{N} \right| < \delta$$

Note. The important fact here is not the dependence on δ which is the same for every graph (not just expanders). The difference between expanders and other graphs is that the constant implied by the $O()$ notation does not depend on the size of the graph. As the proof will show, the dependence of l on the second eigenvalue λ_2 is $l = \frac{\log 1/\delta}{\log 1/\lambda_2}$. If the bound on the eigenvalues $\lambda = 1 - \epsilon$ is close to 1 then we get

$$l \approx \frac{\log 1/\delta}{\log 1/(1-\epsilon)} \approx \frac{\log 1/\delta}{\epsilon}$$

For a fixed δ, the same l will work for all graphs with a given bound λ. Typically we will want to use $\delta \leq 1/2N$. For that value of δ we have

$$\frac{1}{2N} < \Pr(X_l = v) < \frac{3}{2N}$$

So that the probability of being at any vertex is within a constant factor of the uniform distribution. For $\delta = 1/2N$, we need to have $l = O(\log N)$, so we get:

Corollary 6 *If $\{G_N\}$ are expander graphs then after $l = O(\log N)$ random steps we are within constant factor of the uniform distribution.*

Proof: [Proof of Theorem 5] Let X be a random vertex in G with probability vector v (i.e. v_i is the probability that $X = i$). Let Y be a uniformly chosen neighbour of X. We claim that the probability vector of Y is given by Av where A is the normalized adjacency matrix of G. To see this write the Bayesian equation:

$$\begin{aligned}
\Pr(Y = i) &= \sum_j \Pr(Y = i | X = j) \Pr(X = j) \\
&= \sum_j A_{ij} v_j \\
&= (Av)_i
\end{aligned}$$

Applying this recursively we see that if the distribution of X_0 is given by a probability vector v then the distribution of X_l is given by $A^l v$.

Next we write v as the sum of the uniform probability vector $\pi = (N^{-1}, \ldots, N^{-1})^T$ and another vector $v_\perp = v - \pi$. Since the sum of the coordinates in v is 1 and the same is true of π we see that the sum of the coordinates of $v_\perp$ is 0. This shows that $v_\perp$ is perpendicular to π justifying the notation $v_\perp$.

A is a real symmetric matrix and as such has perpendicular eigenvectors. Thus $v_\perp$ is spanned by the remaining eigenvectors of A (other then π). We can now estimate the difference in probabilities:

$$\begin{aligned}
\|A^l v - \pi\|_\infty &\leq \|A^l v - \pi\|_2 \\
&= \|A^l (v_\perp + \pi) - \pi\|_2 \\
&= \|A^l v_\perp + A^l \pi - \pi\|_2 \\
&= \|A^l v_\perp\|_2
\end{aligned}$$

Here we used the fact that π is an eigenvector of A with eigenvalue 1. Since all eigenvalues of A except 1 are bounded by λ the norm of A as an operator on the space perpendicular to π is $|\lambda_2|$. Since $v_\perp$ is perpendicular to π we have:

$$\begin{aligned}
\|A^l v_\perp\|_2 &\leq \lambda^l \|v_\perp\|_2 \\
&\leq \lambda^l \|v\|_2 \\
&\leq \lambda^l \|v\|_1 \\
&= \lambda^l
\end{aligned}$$

since $v_\perp$ is the projection of v on the space perpendicular to π it has smaller $\|\cdot\|_2$ norm then v and v is a probability vector.

If $\lambda^l < \delta$ then we conclude that for every vertex $|\Pr(X_l = v) - N^{-1}| < \delta$. This happens whenever

$$l > \frac{\log 1/\delta}{\log 1/\lambda} = O(\log \frac{1}{\delta})$$

■

It is interesting to note the combinatorial approach to estimating the mixing time of a random walk. The combinatorial definition of an expander says that there are no small cuts in the graph. Specifically it gives a relation between the size of a set and the number of edges leaving it. (The requirement that $|S| < |V|/2$ is not important since if S contains half the vertices we can remove one from S and find that S still has many neighbours.) There are estimates of the mixing time of the random walk on a graph in terms of the sizes of cuts in the graph and vice versa, but they will not be shown here. We will note, however, that absence of small cuts is a necessary requirement for fast mixing. If a graph is composed of two large vertex sets with a small number of edges (e.g., a single edge) between them then the random walk on the graph started in one set will take a long time to reach the other set.

3.2 A random walk yields a sequence of "good samples"

Consider the following problem: In a large set V there is a large subset G of good elements (satisfying some condition) and we wish to find one of them. By large we will mean that the size of G is a constant fraction of the size of V. Call the set of bad elements B and denote its density by $\rho(B) = |B|/|V|$. The set G (and B) are given by an oracle allowing us to check whether some x is in G or in B.

Deterministic algorithms have the drawback that they may take very long for some particularly bad sets B (It is possible to see most of V before finding a point in G). Some randomness must be used to avoid the possibility of terrible sets (at the cost of introducing a probability of error). The trivial random algorithm is to randomly choose elements in V and checking if they are in G. If we make l samples independently then the probability of not finding one in G is $\rho(B)^l$.

This is close to the optimal success probability. The optimal success probability is achieved by also ensuring that the samples are distinct, but since $l \ll |V|$ the improvement is negligible. The drawback here is that this requires the use of many ($\log |V|$) random bits for each sample used.

Expanders can be used to improve this. The algorithm we consider is as follows: Using an expander $G = (V, E)$, Generate l samples in V by taking $X_1 \in_R V$ uniformly at random and $X_2, \dots, X_l$ are a random walk on G started at X_1. Use the X_i's as samples, while checking if $X_i \in B$. The algorithm is successful if not all X_i fall in B.

Since at each step we only need to choose randomly one of d neighbours (if G is d-regular) then the number of random bits needed is $\log |V| + (l-1)\log d$. On the other hand we will see that the probability of success is comparable to that of the trivial random algorithm (with the same number of vertex checked).

Theorem 7 *Let $G = (V, E)$ is d-regular with $|\lambda_2| = \lambda < 1$ and $B \subset V$ has density $\rho = \frac{|B|}{|V|}$. Choose $X_1 \in_R V$ uniformly at random and let $X_2, \dots, X_l$ be a random walk on G starting at X_1.*

$$\Pr(\forall i\, X_i \in B) \le \sqrt{\rho} \cdot (\sqrt{\rho} + \lambda)^{l-1}$$

Note this estimate is not quite the same as for independent samples. First of all, even if λ is very small the bound is only $\rho(B)^{l/2}$ compared to $\rho(B)^l$ for independent samples. While this is indeed a difference, the most harm it can do is force us to use twice as many random steps as independent samples, hence the damage is at most multiplication by a constant of complexity and random bits needed. In addition, it is possible to prove the stronger estimate

$$\Pr(\forall i\, X_i \in B) \le \rho(B) \cdot (\rho(B) + \lambda(1 - \rho(B))^{l-1}$$

which when λ is close to 0 is close to $\rho(B)^l$.

If λ is not very small then the bound is even worse. This drawback can be alleviated by reducing λ by using instead of G the power G^k. By taking k large enough we can make λ smaller. However, we don't gain much by doing this since each random step on G^k takes as many random bits as k

steps on G. Using G^k is equivalent to doing k times as many steps and only checking if $X_i \in Good$ on every k'th step. Since we are taking these steps anyway, usually there is no loss in checking the intermediate vertices as well, i.e. using the random walk on G and not G^k.

One situation where using G^k rather then G may be useful is when the cost of checking whether or not some X is good or bad is not negligible. Then it may be better to take a few more random steps between checks to increase the independence of the positions checked.

Proof: Let $P = P_B$ be a projection on the space of vectors supported in B, i.e. P is a matrix with $P_{ij} = 1$ if $i = j \in B$ and 0 otherwise. If v is a distribution vector then Pv is the residual distribution vector of the distribution v conditioned on being in the set B.

We claim that the probability we wish to estimate is given by: $\|(PA)^{l-1}P\pi\|_1$ where as before, $\pi = (1/N, \ldots, 1/N)^T$ is the uniform distribution. To see this, we need to understand the actions of A and P on probability vectors. The action of A is known: for a probability vector v, Av is the probability vector after a random step has been taken.

The action of P is to nullify all the coordinates outside the set B. This transforms a probability vector into the residual probability vector of the same distribution, but conditioned on being in B. Thus $P\pi$ is the residual probability of the uniform distribution conditioned to be in B. $AP\pi$ is the residual distribution after a random step has been takes. $PAP\pi$ is the residual probability conditioned on the random step remaining in B. Repeating this we see that $(PA)^{l-1}P\pi$ is the residual probability vector of the random initial point and all $l-1$ steps being in B. Since we don't care where in B we end up we need to sum the coordinates of this vector, hence the required probability is indeed given by $\|(PA)^{l-1}P\pi\|_1$.

To estimate this we will use the following lemma:

Lemma 3.1 *For any non negative vector v:*

$$\|PAv\|_2 \leq (\sqrt{\rho} + \lambda) \cdot \|v\|_2$$

Given this lemma we proceed as follows: using the bound on the ratio between $\|v\|_1$ and $\|v\|_2$ and applying the lemma l times:

$$\begin{aligned}
\|(PA)^{l-1}P\pi\|_1 &\leq \sqrt{N} \cdot \|(PA)^{l-1}P\pi\|_2 \\
&\leq \sqrt{N} \cdot (\sqrt{\rho} + \lambda)^{l-1}\|P\pi\|_2 \\
&= \sqrt{N} \cdot (\sqrt{\rho} + \lambda)^{l-1}\sqrt{\frac{\rho}{N}} \\
&= \sqrt{\rho} \cdot (\sqrt{\rho} + \lambda)^{l-1}
\end{aligned}$$

Since $\|P\pi\|_2 = \sqrt{|B|/N^2} = \sqrt{\rho/N}$.

■

Proof: [Proof of Lemma 3.1] P. The idea of the proof is that A shrinks all components of a vector except the uniform distribution component, whereas P will shrinks the uniform component without increasing anything else. Together they reduce all parts of a vector.

To prove the above formally, break up v as before: $v = v_1 + v_\perp$ where v_1 is a constant vector, i.e. a multiple of the uniform distribution vector, and $v_\perp$ is orthogonal to the uniform distribution vector. In the rest of the proof all norms are Euclidean.

$$\|PAv\| = \|PAv_1 + PAv_\perp\| \leq \|PAv_1\| + \|PAv_\perp\|$$

Next we note how PA effects each part of v. Assume $v_1 = (a, a, \ldots, a)^T$. Then we have $\|v_1\| = \sqrt{N}a$. Since v_1 is an eigenvector of A with eigenvalue 1:

$$\|PAv_1\| = \|Pv_1\| = \sqrt{|B| \cdot a^2} = \sqrt{\rho N} \cdot a = \sqrt{\rho} \cdot \|v_1\|$$

Next we look at the effect on $v_\perp$. P's effect is to multiply some coordinates by 0 without changing the others, so P can only shrink a vector. Since $v_\perp$ is perpendicular to the constant vector, it is spanned by the remaining eigenvectors of A, all with eigenvalues at most λ. This implies that

$$\|PAv_\perp\| \leq \|Av_\perp\| \leq \lambda \|v_\perp\|$$

Combining the two inequalities together with

$$||v_1||, ||v_\perp|| \leq ||v||$$

We see that

$$\|PAv\| \leq \|PAv_1\| + \|PAv_\perp\| \leq \sqrt{\rho}\|v_1\| + \lambda\|v_\perp\| \leq (\sqrt{\rho} + \lambda)\|v\|$$

as required. ■