PRINCETON UNIVERSITY F'02 COS 597D: A THEORIST'S TOOLKIT

Lecture 8: Recycling random bits & High dimensional geometry

Lecturer: *Sanjeev Arora* Scribe:*Nir Ailon*

We finish up one topic about expanders and random walks: how to recycle random bits. Then we move on to high-dimensional geometry, specifically, an analysis of the convergence rate of a random walk inside a convex body.

1 Recycling Random Bits

Assume we have a randomized polynomial time algorithm deciding on a language L, that given an input x answers as follows:

$$\begin{cases} Pr(\text{Answer "yes"}) \geq \frac{1}{2} & \text{if } x \in L \\ Pr(\text{Answer "yes"}) = 0 & \text{if } x \notin L \end{cases} \tag{1}$$

(Note: the class of such languages is called *co-RP*). We would like to get an amplification of the success probability by repeating the algorithm k times. This would replace the probability $\frac{1}{2}$ with probability $1 - \frac{1}{2^k}$. Assuming that the algorithm requires r random bits for input x, a naive approach would require rk random bits. A result by [AKS],[CW],[IZ] (87-89) says that we can get this using only $\mathcal{O}(r+k)$ bits. This seems surprising at first, but from an information theoretic point of view there is a simple intuition: we pay r bits and get 1 "yes"/"no" bit back, so we can pocket the change. The idea is as follows:

Take, say, a degree 5 expander on 2^r nodes (note: there are explicit such constructions that allow computation of $\Gamma(v)$ in poly(r) time for any node v). Recall that one of the equivalent definitions of an expander graph is that there is an eigenvalue gap in the adjacency matrix of the graph. Start with a random node (requires r random bits) and perform a random walk of length k, each time taking the bits corresponding to the visited node as the fresh r random bits. In each step, we need $\mathcal{O}(1)$ random bits, and there are a total of k steps, resulting in a total of $\mathcal{O}(r+k)$ random bits.

To prove that this works, assume $x \in L$ and denote by C the set of random strings of length r that would cause the algorithm to answer the correct answer, "yes". So $\frac{|C|}{2^r} \geq \frac{1}{2}$. You will show in the homework that a random walk of length k avoids C with probability at most $\mathcal{O}(e^{-k})$.

2 High Dimensional Geometry

Some useful generalizations of geometric objects to higher dimensional geometry:

- The *n-cube* in $\mathbf{R}^n$: $\{(x_1 ... x_n) : 0 \leq x_i \leq 1\}$. To visualize this in $\Re^4$, think of yourself as looking at one of the faces, say $x_1 = 1$. This is a cube in $\Re^3$ and if you were

able to look in the fourth dimension you would see a parallel cube at $x_1 = 0$. The visualization in $\Re^n$ is similar.

The volume of the n-cube is 1.

- The unit *n-ball* in $\mathbf{R}^n$: $B_n := \{(x_1...x_n) : \sum x_i^2 \leq 1\}$. Again, to visualize the ball in $\Re^4$, imagine you have sliced through it with a hyperplane, say $x_1 = 1/2$. This slice is a ball in $\Re^3$ of radius $\sqrt{1 - 1/2^2} = \sqrt{3}2$. Every parallel slice also gives a ball.

  The volume of B_n is $\frac{\pi^{n/2}}{(n/2)!}$ (assume n even if the previous expression bothers you), which is $\frac{1}{n^{\Theta(n)}}$.

2.1 An approximate way to think about B_n

A good approximation to picking a random point on the surface of B_n is by choosing random $x_i \in \{-1, 1\}$ independently for $i = 1..n$ and normalizing to get $\frac{1}{\sqrt{n}}(x_1, ..., x_n)$. To get a point inside the ball, it is necessary to pick the distance from $\bar{0}$ randomly. Note that the distance is not distributed uniformly: the density at radius r is proportional to r^{n-1}.

Remark: An exact way to pick a random point on the surface of B^n is to choose x_i from the normal distribution for $i = 1..n$, and to normalize: $\frac{1}{l}(x_1, ..., x_n)$, where $l = (\sum_i x_i^2)^{1/2}$.

2.2 Funny facts

1. Scribe's contribution: The volume of the unit n-ball tends to 0 as the dimension tends to ∞.

2. For any $c > 1$, a $(1 - \frac{1}{c})$ - fraction of the volume of the n-ball lies in a strip of width $\mathcal{O}(\sqrt{\frac{c \log n}{n}})$. A strip of width a is B_n intersected with $\{(x_1, ..., x_n) | x_1 \in [-\frac{a}{2}, \frac{a}{2}]\}$.

3. If you pick 2 vectors on the surface of B_n independently, then with probability $> 1 - \frac{1}{n}$,

$$|\cos(\Theta_{x,y})| = \mathcal{O}(\frac{\sqrt{\log n}}{n}),$$

   where $\Theta_{x,y}$ is the angle between x and y. In other words, the 2 vectors are almost orthogonal w.h.p. To prove this, we use the following lemma:

LEMMA 1

Suppose a is a unit vector in $\mathbf{R}^n$. Let $x = (x_1, ..., x_n) \in R^n$ be chosen from the surface of B_n by choosing each coordinate at random from $\{1, -1\}$ and normalizing by factor $\frac{1}{\sqrt{n}}$. Denote by X the random variable $a \cdot x = \sum a_i x_i$. Then:

$$Pr(|X| > t) < e^{-nt^2}$$

PROOF: We have:

$$\mu = E(X) = E(\sum a_i x_i) = 0$$

$$\sigma^2 = E[(\sum a_i x_i)^2] = E[\sum a_i a_j x_i x_j] = \sum a_i a_j E[x_i x_j] = \sum \frac{a_i^2}{n} = \frac{1}{n}.$$

Using the Chernoff bound, we see that,

$$Pr(|X| > t) < e^{-(\frac{t}{\sigma})^2} = e^{-nt^2}$$

□

Corollary 2

If two unit vectors x, y are chosen at random from $\mathbf{R}^n$, then

$$Pr\left(|cos(\theta_{x,y})| > \sqrt{\frac{-\log \epsilon}{n}}\right) < \epsilon$$

Now, to get fact (3), put $\epsilon = \frac{1}{n}$.

3 Random Walks in Convex Bodies

Definition of a convex body: A set $K \subset \mathbf{R}^n$ is convex if $\forall x, y \in K, \lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)y \in K.$$

In other words, for any two points in K, the line segment connecting them is in the set.

examples: Balls, cubes, polytopes (=set of feasible solutions to an LP), ellipsoides etc.

Convex bodies are very important in theory of algorithms. Many optimization problems can be formulated as optimizing convex functions in convex domains. A special case is linear programming.

Today's Goal: To generate a random point in a convex body. This can be used to approximate the volume of the body and to solve convex programs (we saw how to do the latter in Vempala's talk at the theory lunch last week). We emphasize that the description below is missing many details, though they are easy to figure out.

First we need to specify how the convex body is represented:

- $0 \in K$.
- K is contained in a cube of side R with center at 0.
- A unit cube is contained in K.
- $R \leq n^2$.
- there exists a "separation oracle" that given any point $x \in \mathbf{R}^n$ returns "yes" if $x \in K$ and if $x \notin K$ then returns a separating hyperplane $\{y|a^T y = b\}$ s.t. x and K are on opposite sides (i.e. $a^T x < b, a^T z > b \ \forall z \in K$). Note 1: a (closed) convex body can actually be characterized by the property that it can be separated from any external point with a separating hyperplane. Note 2: In the linear program case, giving a separating hyperplane is equivalent to specifying a violated inequality.

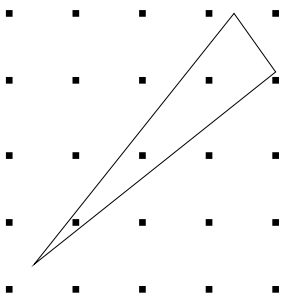

Figure 1: The bottom-left most grid point in the triangle is isolated.

The idea is to approximate the convex body K with a fine grid of scale $\delta < \frac{1}{n^2}$. The volume of K is approximately proportional to the number of grid points. There is some error because we are approximating the convex body using a union of small cubes, and the error term is like

$$(\text{number of grid cubes that touch the surface of } K) \times (\text{volume of small cube}),$$

which is $\ll$ volume of K since the minimum crosssection of K (at least 1) is much larger than the dimensions of the grid cube.

Consider the graph G_K whose vertices are grid notes contained in K, and there is an edge between every pair of grid neighbors. To a first approximation, the graph is regular: almost every node has degree $2n$. The exceptions are grid points that lie close to the surface of K, since not their $2n$ neighbors may lie inside K. But as noted, such grid points form a negligible fraction. Thus generating a random point of G_K is a close approximation to generating a random point of K.

However, we only know one node in G_K, namely, the origin 0. Thus one could try the following idea. Start a random walk from 0. At any step, if you're on a grid point x, stay at x with probability 1/2, else randomly pick a neighbor y of x. If $y \in K$, move to y, else stay at x. If this walk is ergodic, then the stationary distribution will be close to uniform on G_K, as noted. The hope is that this random walk is rapidly mixing, so running it for poly(n) time does yield a fairly unbiased sample from the stationary distribution.

Unfortunately, the walk may not be ergodic since G_K may not be connected (see figure 1).

To solve this problem, we smoothen K, namely, we replace K with a ρ-neighborhood of K defined as

$$K' = \bigcup_{x \in k} B_n(x, \rho),$$

where $B_n(x, \rho)$ is a closed ball centered at x with radius ρ, and $\rho \approx \frac{1}{n}$ (see figure 2). It can be checked that this negligibly increases the volume while ensuring the following two facts:

- K' is convex, and a separation oracle for K' can be easily built from the separation oracle for K, and,
- Any 2 grid points in K are connected via a path of grid points in K'.

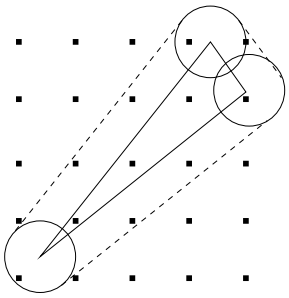

Figure 2: The dashed convex body is obtained by smoothening the solid triangle.

Now let $G_{K'}$ be the graph on grid nodes contained in K'. It is connected, and almost all its nodes are in K (since the volume increase in going from K to K' was minimal).

We go back to estimating the mixing time. Recall that the "mixing time" is the time t s.t. starting from any distribution x_0, after t steps the distribution x_t satisfies

$$\|x_t - \pi\|_2 \leq \frac{1}{4}.$$

Using the Jerrum-Sinclair inequality from the previous lecture, we have

$$\|x_t - \pi\|_2 \leq (1 - \frac{\phi^2}{2})^t \sqrt{N},$$

where ϕ is the Cheeger number attached to the graph and N is the size of the graph (number of grid points in K'). So we reduce to the problem of estimating the Cheeger constant.

[Aside: In case you ever go through the literature, you will see that many papers use the "conductance" of the graph instead of the "Cheeger" number. The conductance of a graph G is

$$\min_{S \subset V} \frac{|E(S, \bar{S})| \mathrm{Vol}(V)}{\mathrm{Vol}(S)\mathrm{Vol}(\bar{S})}.$$

Easy exercise: the conductance is within factor 2 of the Cheeger number.]

To estimate the Cheeger number of the grid graph, we use a result by Lovasz and Simonovits. Let U be a subset of nodes of $G_{K'}$. Consider the union of the corresponding set of grid cubes; this is a measurable subset (actually it is even better behaved; it is a finite union of convex subsets) of K. Call this set S. Let ∂S be the boundary of S. Then it is not hard to see that:

$$\mathrm{Vol}_{n-1}(\partial S) \leq E(S, \bar{S}) \leq 2n \mathrm{Vol}_{n-1}(\partial S),$$

where $E(S, \bar{S})$ is the number of grid edges crossing between S and $\bar{S}$. The [LS] result is:

THEOREM 3
If K is a convex body containing a unit cube, then for any measurable $S \subset K$,

$$Vol_{n-1}(\partial S) \geq \frac{2}{D} \min\{Vol_n(S), Vol_n(\bar{S})\},$$

where D is the diameter of K.

Recall that the diameter of our body is poly(n) since we started with a body whose diameter was n^2 and then we placed a grid of size $1/n^2$ or so. Combining all the above information, we get that

$$\phi \geq \frac{1}{\text{poly}(n)}.$$

Therefore, the mixing time is $\mathcal{O}(\frac{1}{\phi^2}\log N) = \mathcal{O}(\text{poly}(n))$.

4 Dimension Reduction

Now we describe a central result of high-dimensional geometry (at least when distances are measured in the ℓ_2 norm). Problem: Given n points $z^1, z^2, ..., z^n$ in $\Re^n$, we would like to find n points $u^1, u^2, ..., u^n$ in $\Re^m$ where m is of low dimension (compared to n) and the metric restricted to the points is almost preserved, namely:

$$\|z^i - z^j\|_2 \leq \|u^i - u^j\|_2 \leq (1+\epsilon)\|z^j - z^j\|_2 \ \forall i, j. \tag{2}$$

The following main result is by Johnson & Lindenstrauss :

THEOREM 4
In order to ensure (2), $m = \mathcal{O}(\frac{\log n}{\epsilon^2})$ suffices.

Note: In class, we used the notation of k vectors $z^1 ... z^k$ in $\Re^n$, but we can always embed the k vectors in a k-dimensional space, so here I assume that $n = k$ and use only n.
PROOF:

Choose m vectors $x^1, ..., x^m \in \Re^n$ at random by choosing each coordinate randomly from $\{\sqrt{\frac{1+\epsilon}{m}}, -\sqrt{\frac{1+\epsilon}{m}}\}$. Then consider the mapping from $\Re^n$ to $\Re^m$ given by

$$z \longrightarrow (z \cdot x^1, z \cdot x^2, \ldots, z \cdot x^m).$$

In other words $u^i = (z^i \cdot x^1, z^i \cdot x^2, ..., z^i \cdot x^m)$ for $i = 1, \ldots, k$. We want to show that with positive probability, $u^1, ..., u^k$ has the desired properties. This would mean that there exists at least one choice of $u^1, ..., u^k$ satisfying inequality 2. To show this, first we write the expression $\|u^i - u^j\|$ explicitly:

$$\|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} (z_l^i - z_l^j) x_l^k \right)^2.$$

Denote by z the vector $z^i - z^j$, and by u the vector $u^i - u^j$. So we get:

$$\|u\|^2 = \|u^i - u^j\|^2 = \sum_{k=1}^{m} \left(\sum_{l=1}^{n} z_l x_l^k \right)^2.$$

Let X_k be the random variable $(\sum_{l=1}^{n} z_l x_l^k)^2$. Its expectation is $\mu = \frac{1+\epsilon}{m}\|z\|^2$ (can be seen similarly to the proof of lemma 1). Therefore, the expectation of $\|u\|^2$ is $(1+\epsilon)\|z\|^2$. If we show that $\|u\|^2$ is concentrated enough around its mean, then it would prove the theorem. More formally, we state the following Chernoff bound lemma:

LEMMA 5
There exist constants $c_1 > 0$ and $c_2 > 0$ such that:

1. $Pr[\|u\|^2 > (1+\beta)\mu] < e^{-c_1\beta^2 m}$
2. $Pr[\|u\|^2 < (1-\beta)\mu] < e^{-c_2\beta^2 m}$

Therefore there is a constant c such that the probability of a "bad" case is bounded by:

$$Pr[(\|u\|^2 > (1+\beta)\mu) \vee (\|u\|^2 < (1-\beta)\mu)] < e^{-c\beta^2 m}$$

Now, we have $\binom{n}{2}$ random variables of the type $\|u_i - u_j\|^2$. Choose $\beta = \frac{\epsilon}{2}$. Using the union bound, we get that the probability that any of these random variables is not within $(1 \pm \frac{\epsilon}{2})$ of their expected value is bounded by

$$\binom{n}{2} e^{-c\frac{\epsilon^2}{4}m}.$$

So if we choose $m > \frac{8(\log n + \log c)}{\epsilon^2}$, we get that with positive probability, all the variables are close to their expectation within factor $(1 \pm \frac{\epsilon}{2})$. This means that for all i,j:

$$(1-\frac{\epsilon}{2})(1+\epsilon)\|z^i - z^j\|^2 \leq \|u^i - u^j\|^2 \leq (1+\frac{\epsilon}{2})(1+\epsilon)\|z^i - z^j\|^2$$

Therefore,

$$\|z_i - z_j\|^2 \leq \|u^i - u^j\|^2 \leq (1+\epsilon)^2\|z^i - z^j\|^2,$$

and taking square root:

$$\|z^i - z^j\| \leq \|u^i - u^j\| \leq (1+\epsilon)\|z^i - z^j\|,$$

as required.

It remains to prove lemma 4. We prove the first part. Let $\alpha^2 = \frac{1+\epsilon}{m}$, so $\mu = \alpha^2 m\|z\|^2$ and we get the following equation:

$$\begin{aligned}
P := Pr[\|u\|^2 > (1+\beta)\mu] &= Pr[\|u\|^2 > (1+\beta)\alpha^2 m\|z\|^2] \\
&= Pr[\|u\|^2 - (1+\beta)\alpha^2 m\|z\|^2 > 0] \\
&= Pr[t(\|u\|^2 - (1+\beta)\alpha^2 m\|z\|^2) > 0] \quad \forall t > 0 \\
&= Pr[\exp(t(\|u\|^2 - (1+\beta)\alpha^2 m\|z\|^2)) > 1] \\
&\leq E[\exp(t(\|u\|^2 - (1+\beta)\alpha^2 m\|z\|^2))] \quad \text{(Markov)}
\end{aligned} \tag{3}$$

We calculate the last expectation:

$$\begin{aligned}
P &\leq E[\exp(t(\|u\|^2))]\exp(-t(1+\beta)\alpha^2 m\|z\|^2) \quad \text{(constant goes out)} \\
&= E[\exp(t(\sum_{k=1}^{m}(\sum_{l=1}^{n} z_l x_l^k)^2))]\exp(-t(1+\beta)\alpha^2 m\|z\|^2) \\
&= E[\exp(t(\sum_{k}(\sum_{l} z_l^2 (x_l^j)^2)) + t(\sum_{k}(\sum_{l\neq h} z_l z_h x_l^k x_h^k)))]\exp(-t(1+\beta)\alpha^2 m\|z\|^2) \\
&= E[\exp(t\alpha^2 m\|z\|^2 + t(\sum_{k}(\sum_{l\neq h} z_l z_h x_l^k x_h^k)))]\exp(-t(1+\beta)\alpha^2 m\|z\|^2)
\end{aligned} \tag{4}$$

The last step used the fact that $(x_l^k)^2 = \alpha^2$ and $\sum z_l^2 = \|z\|^2$. So continuing, we get:

$$P \leq E[\exp\left(t(\sum_k(\sum_{l\neq h} z_l z_h x_l^k x_h^k))\right)] \exp\left(-t\beta\alpha^2 m\|z\|^2\right) \tag{5}$$

The set of variables $\{x_l^k x_h^k\}_{l\neq h}$ are pairwise independent. Therefore the above expectation can be rewritten as a product of expectations:

$$P \leq \left(\prod_k \prod_{l\neq h} E[\exp(t z_l z_h x_l^k x_h^k)]\right) \exp(-t\beta\alpha^2 m\|z\|^2) \tag{6}$$

we notice that

$$E[\exp(t z_l z_h x_l^k x_h^k)] = \frac{1}{2}\exp(t z_l z_h \alpha^2) + \frac{1}{2}\exp(-t z_l z_h \alpha^2) < \exp(t^2 z_l^2 z_h^2 \alpha^4)$$

(the last inequality is easily obtained by Taylor expanding the exponent function). Plugging that in (6), we get:

$$\begin{aligned} P &< \left(\prod_k \prod_{l\neq h} \exp\left(t^2 z_l^2 z_h^2 \alpha^4\right)\right) \exp\left(-t\beta\alpha^2 m\|z\|^2\right) \\ &= \left(\prod_{l\neq h} \exp\left(t^2 z_l^2 z_h^2 \alpha^4\right)\right)^m \exp\left(-t\beta\alpha^2 m\|z\|^2\right) \\ &= \exp\left(mt^2 \sum_{l\neq h} z_l^2 z_h^2 \alpha^4 - t\beta\alpha^2 m\|z\|^2\right) \end{aligned} \tag{7}$$

Using simple analysis of quadratic function we see that the last expression obtains its minimum when

$$t = \frac{\beta\|z\|^2}{2\alpha^2 \sum_{l\neq h} z_l^2 z_h^2}.$$

Substituting for t, we get:

$$P < \exp\left(-\beta^2 m \frac{\|z\|^4}{4\sum_{l\neq h} z_l^2 z_h^2}\right) \tag{8}$$

Finally, the expression

$$\delta(z) = \left(\frac{\|z\|^4}{4\sum_{l\neq h} z_l^2 z_h^2}\right)$$

is bounded below by a constant c_1. To prove this, first note that $\delta(z) = \delta(\gamma z)$ for any $\gamma \neq 0$. So it is enough to consider the case $\|z\| = 1$. Then, using Lagrange multipliers technique, for example, we get that $\delta(z)$ obtains its minimum when $z_l = \frac{1}{\sqrt{n}}$ for each $l = 1..n$. Plugging this in the expression for $\delta(z)$ we see that it is bounded above by a constant c_1 that does not depend on n. This completes the proof. □