Algebraic Methods in Computer Science	3.12.14
Lecture: 6	
Lecturer: *Amir Shpilka*	Scribe: *Geva Kipper*

Random Walks on Graphs

1 Motivation - Randomized Algorithms

We want to reduce the error of a randomized algorithm. We know we can get an exponential reduction by applying the same algorithm over k independent random inputs, but this will require generating $O(k \cdot n)$ random bits, which is a lot.

What we can do instead is build some d-regular graph $G = (V, E)$ whose vertices are the possible random inputs, start with some random vertex $v_1 \in V$, and randomly walk on the graph for k steps $v_1 \to v_2 \to \ldots \to v_k$, running the algorithm on each v_i we get to.

Performing this will require much less randomness - just enough to pick a starting vertex and a random neighbor at each step, which is $O(r + k \cdot$
$Pr((d))$.

Obviously these v_i are not independent, but we will show that testing them still greatly reduces the error. In order to do this we will analyze random walks on d-regular graphs, and seek to answer questions such as:

- How long do we have to walk for the distribution to converge to the uniform distribution?
- How unlikely is it that during our walk we always remain in some subset of the vertices, $B \subset V$?

2 Convergence to the Uniform Distribution

Definition 2.1. *Let M be a real matrix. We define the spectral norm of M:* $\|M\| = \max \frac{\|Mx\|}{\|x\|}$.

Notice that: $\|M_1 M_2\| \leq \|M_1\| \cdot \|M_2\|$, $\|c \cdot M\| = |c| \cdot \|M\|$ and $\|M_1 + M_2\| \leq \|M_1\| + \|M_2\|$. Also notice that when M is diagonalizable, $\|M\|$ is equal to the maximum eigenvalue of M in absolute value.

Take A to be the adjacency matrix of our graph G. Since A is symmetric we know it is diagonalizable and has eigenvalues $\lambda_1 = d \geq \lambda_2 \geq \ldots \geq \lambda_n$. Take $\lambda = \max(\lambda_2, |\lambda_n|)$. Let $\bar{u} = \frac{1}{n}\vec{1}$ denote the uniform distribution vector.

Claim 2.1.

$$\lambda = \max_{x \perp \vec{1}} \frac{\|Ax\|}{\|x\|} = \max_{\pi} \frac{\|A\pi - A\bar{u}\|}{\|\pi - \bar{u}\|}.$$

A slight variant of the first equality appeared in exercise 2. To prove the second equality, we will prove both $\leq$ and $\geq$ directions.

Proof. Suppose π is a distribution, meaning $\sum \pi_i = 1$, $\forall i\, \pi_i \geq 0$. Then $\langle \pi - \bar{u}, \vec{1} \rangle = \langle \pi, \vec{1} \rangle - \langle \bar{u}, \vec{1} \rangle = \sum \pi_i - \sum \bar{u}_i = 1 - 1 = 0$. So $\pi - \bar{u} \perp \vec{1}$, and therefore $\max_{x \perp \vec{1}} \frac{\|Ax\|}{\|x\|} \geq \frac{\|A\pi - A\bar{u}\|}{\|\pi - \bar{u}\|}$. Since this is true for any distribution π, we get $\max_{x \perp \vec{1}} \frac{\|Ax\|}{\|x\|} \geq \max_{\pi} \frac{\|A\pi - A\bar{u}\|}{\|\pi - \bar{u}\|}$.

Now, let x be an arbitrary vector satisfying $\langle x, \vec{1} \rangle = 0$. Then, for a small enough α, the vector $y = \bar{u} - \alpha x$ will have only positive coordinates. Notice that it holds that $\sum y_i = \langle y, \vec{1} \rangle = \langle \bar{u}, \vec{1} \rangle - \alpha \overbrace{\langle x, \vec{1} \rangle}^{=0} = \langle \bar{u}, \vec{1} \rangle = 1$, which means that y is a distribution vector. We thus get

$$\max_{\pi} \frac{\|A\pi - A\bar{u}\|}{\|\pi - \bar{u}\|} \geq \frac{\|Ay - A\bar{u}\|}{\|y - \bar{u}\|} = \frac{\|A(-\alpha x)\|}{\|-\alpha x\|} = \frac{\|Ax\|}{\|x\|}.$$

Since this was true for any $x \perp \vec{1}$, we get $\max_{x \perp \vec{1}} \frac{\|Ax\|}{\|x\|} \leq \max_{\pi} \frac{\|A\pi - A\bar{u}\|}{\|\pi - \bar{u}\|}$, which completes the proof. □

We now wish to analyze the convergence rate of random walks on graphs.

Theorem. *Let π_k be the induced probability distribution on the vertices after starting at vertex v_0 and taking k steps of the random walk, i.e., $(\pi_k)_i = \Pr(v_k = i)$. Then:*

$$\|\pi_k - \bar{u}\| \leq \left(\frac{\lambda}{d}\right)^k$$

Proof. We will analyze the random walk algebraically. Let us look at a single step. Suppose at the i'th step our current distribution over the vertices is π. We can think of our walk until now as selecting a vertex u randomly with distribution π, and the next step will be to randomly select a neighbor of u and move to it. The probability we end up at the vertex v after one more step is equal to:

$$\Pr(v_{i+1} = v) = \sum_{(u,v) \in V} \Pr(v_i = u) \cdot \Pr\left(v_{i+1} = v | v_i = u\right) = \sum_{(u,v) \in V} \pi_u \cdot \frac{1}{d} = (\frac{1}{d} A \cdot \pi)_v$$

Denote $\bar{A} = \frac{1}{d} A$, $\bar{\lambda_i} = \frac{\lambda_i}{d}$, $\bar{\lambda} = \frac{\lambda}{d}$, and notice that these are the eigenvalues of $\bar{A}$. By what we just showed, we have:

$$\Pr(v_{i+1} = v) = (\frac{1}{d} A \cdot \pi)_v = (\bar{A}\pi)_v$$

So if our current distribution is π, on the next step it will be equal to $\bar{A}\pi$.

Using our previous observations, we can give a bound on the distance from the uniform distribution after the next step:

$$\|\bar{A}\pi - \bar{u}\| = \|\bar{A}\pi - \bar{A}\bar{u}\| = \frac{1}{d}\|A\pi - A\bar{u}\| \overset{Claim\ 2.1}{\leq} \frac{\lambda}{d}\|\pi - \bar{u}\| = \bar{\lambda}\|\pi - \bar{u}\|$$

This is true for any distribution π. Using induction, we get that after k steps:

$$\|\pi_k - \bar{u}\| = \|\bar{A}^k \pi_0 - \bar{u}\| \leq \bar{\lambda}^k \cdot \|\pi_0 - \bar{u}\|$$

And if we started with a single vertex as our distribution, meaning $\pi_0 = e_i$, then $\|\pi_0 - \bar{u}\|^2 = \sum {\pi_0}_i^2 - \frac{2}{n} \sum \pi_{0i} + \frac{1}{n} = \sum {\pi_0}_i^2 - \frac{1}{n} \leq \sum \pi_{0i} - \frac{1}{n} \leq 1 - \frac{1}{n}$ which means $\|\pi_0 - \bar{u}\| \leq 1$, giving:

$$\|\bar{A}^k \pi - \bar{u}\| \leq \bar{\lambda}^k \cdot \|\pi_0 - \bar{u}\| \leq \bar{\lambda}^k = \left(\frac{\lambda}{d}\right)^k.$$

□

What we would really like is to give a bound on the distance between any coordinate of $\bar{A}^k\pi$ and $\frac{1}{n}$, but such a bound can easily be obtained from the result we showed regarding $\|\ \|_2$.
Notice that we showed in exercise 1 that $\bar{\lambda} = 1$, equivalently, $\lambda = d$, if and only if G is either bipartite or not connected. In such cases our result is not interesting, and that is to be expected since it is clear that we do not always get convergence to $\bar{u}$ in such cases (actually, when starting at a single vertex we never converge to the uniform distribution). In all other cases this result means that for any given graph we will converge to $\bar{u}$ at an exponential rate, being faster when $\bar{\lambda} = \frac{\lambda}{d}$ is smaller. This implies that expander graphs having $d \gg \lambda$, are graphs which will have fast convergence rate for random walks. This raises the natural question of how small can $\bar{\lambda}$ be.

Claim 2.2. *For a d-regular graph on n vertices it holds that*

$$\bar{\lambda} \leq 1 - \frac{1}{\text{poly}(n, d)}.$$

The quantity $1 - \bar{\lambda}$ is called the 'spectral gap', and it determines how quickly does the random walk mix. The existence of the above-mentioned bound guarantees that the mixing time of a random walk is at most polynomial. This completes our third alternative definition for expander graphs. To summarize, up to exact setting of parameters the three definitions are equivalent.

- Edge / Vertex expansion property - $h(G)$ is large.
- Large spectral gap - $\frac{\lambda}{d}$ is small.
- Convergence of random walks in $O(\log n)$ steps.

3 Reducing One-sided Errors

Suppose we have some randomized algorithm for a language L which has a one-sided error, meaning for inputs not in L the algorithm might output "yes" with some probability. As explained in the Section 1, we're about to do a random walk over a graph whose vertices V are the possible random inputs for the algorithm. An error happens when the random input that the algorithm gets belongs to some set of bad inputs, $B \subset V$. During our random walk, if we encounter a random input for which the algorithm says 'no', we know for sure that our input does not belong to L. If during our entire walk we did not get to any random input for which the algorithm said no then we accept the input. Thus, the probability that we make a mistake after a random walk of length k is equal to the probability that all steps of the random walk stayed inside the set B. The following theorem gives an upper bound of this probability.

Theorem. *Let $G = (V, E)$ be a d-regular graph on n vertices with normalized second-largest eigenvalue $\bar{\lambda}$. Let $B \subset V$ be of size $|B| \leq \beta \cdot n$. Let $v_1 \to v_2 \to \ldots \to v_k$ be a random walk on G where v_1 is chosen uniformly. Then,*

$$\Pr(\forall i :\ v_i \in B) \leq (\beta + \bar{\lambda}(1 - \beta))^k.$$

Proof. We want to describe the problem algebraically. Let $\bar{A} = \frac{1}{d}A_G$.
Let P be the projection matrix of B, meaning P is diagonal with entry $p_{i,i}$ being 1 if $i \in B$ and zero otherwise. Some initial observations:
$\Pr(v_1 \in B) = \|P \cdot \bar{u}\| = \|\frac{1}{n} \cdot \vec{1}_B\| = \beta$, where $\vec{1}_B$ is the indicator vector for B.
$(\bar{A}P\bar{u})_i = \Pr(v_1 \in B \wedge v_2 = i)$, which leads to: $\|P\bar{A}P \cdot \bar{u}\|_1 = \Pr(v_1 \in B \wedge v_2 \in B)$.

By induction on k, we get that the probability that the random never leaves B in the first k-1 steps and ends at vertex v is:

$$\Pr\left(v_k = v \wedge \forall i < k : \; v_i \in B\right) = \left(\bar{A}(P\bar{A})^{k-2} \cdot P\bar{u}\right)_v$$

Which in turn gives:

$$\Pr(\forall i : \; v_i \in B) = \|(P\bar{A})^{k-1} \cdot P\bar{u}\|_1$$

To understand $\|(P\bar{A})^{k-1} \cdot P\bar{u}\|_1$, we start by doing analysis in $\| \|_2$ and then switch back to $\| \|_1$. Notice that since $P^2 = P$, we get $(P\bar{A})^{k-1} \cdot P = (P\bar{A}P)^{k-1} = P(P\bar{A}P)^{k-1}P$. This allows us to get the following:

$$(*) \quad \Pr(\forall i : \; v_i \in B) = \|P(P\bar{A}P)^{k-1} \cdot P\bar{u}\|_1 = \langle (P\bar{A}P)^{k-1} \cdot P\bar{u}, \vec{1}_B \rangle \leq$$
$$\underset{\text{cauchy-schwartz}}{\leq} \|(P\bar{A}P)^{k-1} \cdot P\bar{u}\| \cdot \sqrt{|B|} \leq \|(P\bar{A}P)\|^{k-1} \cdot \|P\bar{u}\| \cdot \sqrt{|B|}$$

Now that we're working in $\| \|_2$, we look at the spectral norm $\|P\bar{A}P\|$. Let $\bar{J}$ be the normalized matrix of the complete graph, $\bar{J}_{i,j} = \frac{1}{n}$.

Claim 3.1. *$\bar{A}$ is 'almost' $\bar{J}$:* $\bar{A} = (1-\bar{\lambda}) \cdot \bar{J} + \bar{\lambda} \cdot E$, *where* $\|E\| \leq 1$.
We think of it as being almost $\bar{J}$ when $\bar{\lambda}$ is small.

Proof. Notice that if $v_1 = \vec{1}/\sqrt{n}, v_2, \ldots, v_n$ are orthonormal eigenvectors corresponding to the $\bar{\lambda_i}$, we can write $\bar{A} = \sum \bar{\lambda_i} v_i^t v_i$. Equality of these two linear transformations can be easily verified for all v_i, and since they form a basis for $\mathbb{R}^n$ the equality holds.
Notice $v_1^t v_1 = \bar{J}$, $\bar{\lambda_1} = 1$, so we can rearrange this to get:

$$\bar{A} = \sum \bar{\lambda_i} v_i^t v_i = (1-\bar{\lambda})\bar{J} + \bar{\lambda}\bar{J} + \sum_{i \neq 1} \bar{\lambda_i} v_i^t v_i = (1-\bar{\lambda})\bar{J} + \bar{\lambda}(\bar{J} + \sum_{i \neq 1} \frac{\bar{\lambda_i}}{\bar{\lambda}} v_i^t v_i)$$

Taking $E = \bar{J} + \sum_{i \neq 1} \frac{\bar{\lambda_i}}{\bar{\lambda}} v_i^t v_i = 1 \cdot v_1^t v_1 + \sum_{i \neq 1} \frac{\bar{\lambda_i}}{\bar{\lambda}} v_i^t v_i$, its form indicates that E is orthogonally diagonalizable, and its eigenvalues $\{1, \frac{\bar{\lambda_2}}{\bar{\lambda}}, \ldots, \frac{\bar{\lambda_n}}{\bar{\lambda}}\}$ are all smaller than 1 due to the definition of $\bar{\lambda}$. As we previously mentioned this means $\|E\| \leq 1$ as required, completing the proof. □

The latter claim will allow us to give a bound on $\|P\bar{A}P\|$:

Claim 3.2. $\|P\bar{A}P\| \leq \beta + \bar{\lambda}(1-\beta)$

Proof. Using claim 3.1:

$$\|P\bar{A}P\| = \|P\left((1-\bar{\lambda})\bar{J} + \bar{\lambda}E\right)P\| \leq (1-\bar{\lambda})\|P\bar{J}P\| + \bar{\lambda}\|PEP\|$$

Notice $\|P\| \leq 1$, $\|E\| \leq 1$. We shall use the rather simple bound $\|PEP\| \leq \|P\| \cdot \|E\| \cdot \|P\| \leq 1$.
We also need to give an upper bound for the other part of the sum, namely $\|P\bar{J}P\|$.
Notice that $(\bar{J}P)v = \frac{1}{n}(\sum_{i \in B} v_i) \cdot \vec{1}$, so $(P\bar{J}P)v = \frac{1}{n}(\sum_{i \in B} v_i) \cdot \vec{1}_B$. This gives:

$$\|P\bar{J}Pv\| = \frac{1}{n}(\sum_{i \in B} v_i) \cdot \|\vec{1}_B\| = \frac{\sqrt{|B|}}{n}\langle \vec{1}_B, v \rangle \overset{\text{cauchy-schwartz}}{\leq}$$
$$\leq \frac{\sqrt{|B|}}{n}\sqrt{|B|}\|v\| = \frac{B}{n}\|v\| = \beta\|v\|$$

Yielding $\|P\bar{J}P\| \leq \beta$. Putting this into the previous inequality we get:

$$\|P\bar{A}P| \leq (1-\bar{\lambda})\|P\bar{J}P\| + \bar{\lambda}\|PEP\| \leq (1-\bar{\lambda})\beta + \bar{\lambda} = \beta + \bar{\lambda}(1-\beta)$$

Which is exactly the inequality we wanted: $\|P\bar{A}P| \leq \beta + \bar{\lambda}(1-\beta)$. □

Using the last claim we are able to finish the proof by giving a bound on the spectral norm in (*):

$$\Pr(\forall i:\ v_i \in B) \leq \|(P\bar{A}P)\|^{k-1} \cdot \|P\bar{u}\| \cdot \sqrt{|B|} = \|(P\bar{A}P)\|^{k-1}\sqrt{\frac{|B|}{n^2}}\sqrt{|B|} =$$
$$= \|(P\bar{A}P)\|^{k-1}\beta \leq (\beta + \bar{\lambda}(1-\beta))^{k-1}\beta \underset{\bar{\lambda}(1-\beta)>0}{\leq} (\beta + \bar{\lambda}(1-\beta))^k$$

Completing the proof of Theorem 1: $\Pr(\forall i:\ v_i \in B) \leq (\beta + \bar{\lambda}(1-\beta))^k$. □

4 Chernoff Bound for Random Walks

4.1 Background

Suppose we have n independent random variables $X_1, \ldots, X_n$, all of which get values in $[0,1]$. It is rather intuitive that if we measure them all and take the average we shall most likely get a result very close to the expectation of the average.
The Chernoff bound in probability theory gives us a very good assurance of how likely we are to be close to the expected value. More formally, it states that if X_i has e.v μ_i then:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > \epsilon\right) \leq e^{-\epsilon^2 n}$$

In the finite case, we can think of the random variables X_i as functions f_i over some finite sample space Ω, where measuring X_i means plugging a randomly chosen value from Ω into f_i and looking at the result.
Restating the above bound with these new definitions, taking $v_1, \ldots, v_k \in \Omega$ to be the random sample points, it would translate to:

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} f_i(v_i) - \frac{1}{k}\sum_{i=1}^{k}\mu_i\right| > \epsilon\right) \leq e^{-\epsilon^2 n}$$

Suppose our sample space Ω is taken to be the set of vertices of some graph G, and we don't want to be forced to generate many random values in Ω. We could measure the functions f_i with the sample points being vertices of a random walk on G.
Obviously these are not independent sample points. We will show that this is still "very random" by showing a variant of the Chernoff bound. We will not get the same bound, since we have to pay in some way for the lack of randomness, but we will still get that the probability shrinks exponentially as n goes to infinity.
Note that this is a generalization of what we saw in the previous section, where the function we measured, $f_i = f$ was the indicator function for some subset B of vertices. We then estimated the probability of only walking on vertices $v \in B$, meaning $f(v) = 1$, which implies a large deviation from the expected value of f : $\mathbb{E}(f) = \frac{|B|}{n}$.
The proof has many similarities to the proof of the ordinary Chernoff bound, so we shall start with a brief review of the main ideas behind that proof.

4.2 Ordinary Chernoff bound

We will mention some key elements in the proof of the the ordinary Chernoff bound.
First notice that it suffices to show that $\Pr(\frac{1}{n}\sum_{i=1}^{n} X_i > \frac{1}{n}\sum_{i=1}^{n}\mu_i + \epsilon) \leq e^{-\epsilon^2 n}$. The other complementary result can be obtained by taking $Y_i = 1 - X_i$, and applying the previous result on $\{Y_i\}$.
If we think of Markov's inequality: $\Pr(X > a) \leq \frac{\mathbb{E}(X)}{a}$, or Chebichev's inequality: $\Pr\left((X - \mathbb{E}X)^2 > k^2\right) \leq \frac{\mathbb{E}((X-\mathbb{E}X)^2)}{k^2} = \frac{Var(X)}{k^2}$, we can see that Markov's uses the 1st moment of the r.v X, and Chebichev's simply applies Markov using the second moment of X. In the proof of Chernoff's bound, we want to capture all of X's moments. To achieve this we use a trick which allows us to switch to the r.v $Y = e^X$.
Another trick used for final optimizations is that we actually look at δX_i for some constant δ, like so:

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i > \frac{1}{n}\sum_{i=1}^{n}\mu_i + \epsilon\right) = \Pr\left(\sum_{i=1}^{n}\delta X_i > \delta(\sum_{i=1}^{n}\mu_i + n\epsilon)\right) = \Pr\left(e^{\sum_{i=1}^{n}\delta X_i} > e^{\delta(\sum_{i=1}^{n}\mu_i + n\epsilon)}\right)$$

Using markov's inequality on the non-negative r.v $Y = e^{\sum_{i=1}^{n}\delta X_i}$, and the independence of $X_1, \ldots, X_n$, we get:

$$\Pr\left(e^{\sum_{i=1}^{n}\delta X_i} > e^{\delta(\sum_{i=1}^{n}\mu_i + n\epsilon)}\right) \leq \frac{\mathbb{E}(\prod e^{\delta X_i})}{e^{\delta n\epsilon}\cdot\prod e^{\delta\mu_i}} \overset{(1)}{=} \frac{\prod \mathbb{E}(e^{\delta X_i})}{e^{\delta n\epsilon}\cdot\prod e^{\delta\mu_i}}$$

It can then be shown that for some good selection of δ this is smaller than $e^{-\epsilon^2 n}$, as required.

4.3 Random walks variant

We want to prove the following variant of the Chernoff bound:

Theorem. *Let $G = (V, E)$ with second normalized eigenvalue $\bar{\lambda}$, $\{f_i : V \to [0,1]\}$, $\mu_i = \frac{1}{n}\sum_{v\in V} f_i(v)$. If $v_1 \to v_2 \to \ldots \to v_k$ is a random walk on G then:*

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} f_i(v_i) - \frac{1}{k}\sum_{i=1}^{k}\mu_i\right| > \epsilon\right) \leq e^{-c\cdot\epsilon^2 k(1-\bar{\lambda})}$$

Proof. This proof starts with exactly the same steps from the proof of the ordinary bound. Trying to modify that proof for this new case, we face the problem that our samples are no longer independent, meaning equality (1) does not hold.
We will need another way, an algebraic way, to estimate the numerator $\mathbb{E}(\prod e^{\delta f_i(v_i)})$.
For convenience we think of the vertices of the graph, V, being the set $\{1, \ldots, n\}$.
Define for each f_i a diagonal matrix P_i - having $e^{\delta f_i(j)}$ in the j'th entry of its diagonal.
Using the notation $\bar{u} = \frac{1}{n}\vec{1} = (\frac{1}{n}, \ldots, \frac{1}{n})$, notice that $\|P_1\bar{u}\|_1 = \sum_{j=1}^{n}\frac{1}{n}e^{\delta f_1(j)} = \mathbb{E}_{v\in V}e^{\delta f_1(v)}$. Taking this further, with $\bar{A} = \frac{1}{d}A_G$ being the normalized adjacency matrix of G:

$$(\bar{A}P_1\bar{u})_j = \frac{1}{nd}\sum_{(i,j)\in E} e^{\delta f_1(i)} \implies (P_2\bar{A}P_1\bar{u})_j = \frac{1}{nd}\sum_{(i,j)\in E} e^{\delta f_2(j)}\cdot e^{\delta f_1(i)}$$

This implies $\|P_2\bar{A}P_1\bar{u}\|_1 = \mathbb{E}_{v_1\to v_2}(e^{\delta f_2(j)}\cdot e^{\delta f_1(i)})$, and continuing via induction allows us to express the subject of interest as a norm of a vector:

$$\mathbb{E}(\prod_{i=1}^{k} e^{\delta X_i}) = \|P_k\bar{A}P_{k-1}\ldots\bar{A}P_1\bar{u}\|_1$$

To estimate this norm, we shall think of this as an iterative process. Define the vector at the l'th step, $z_l = P_l\bar{A}\dots P_2\bar{A}P_1\bar{u}$, $z_0 = \bar{u}$. In this notation we are interested in $\|z_k\|_1$. We split vectors into two components: $z = z^{\|} + z^{\perp}$, where $z^{\|} = \langle z, \frac{1}{\sqrt{n}}\vec{1}\rangle \cdot \frac{1}{\sqrt{n}}\vec{1} = \langle z, \vec{1}\rangle\bar{u}$, the projection of z on $\vec{1}$. The idea is that as our iterative process progresses the perpendicular component will have very little influence.
Denote $m_i = \max_{j<i} \|z_j^{\|}\|_2$. The rest of the proof will consist of 2 claims.

Claim 4.1. $\forall i: \ \|z_i^{\perp}\|_2 \le \frac{4\delta}{1-\bar{\lambda}} \cdot m_i$

We will use this claim to prove a second claim:

Claim 4.2. $\forall i: \ \|z_i^{\|}\|_2 \le e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}} \cdot m_i$

We will see the proofs for these two claims in the next lesson.
Assuming that we have proved them, we now show that we can finish the proof of the theorem using claim 4.2 First, notice that using induction with claim 4.2 we get: $\|z_k^{\|}\|_2 \le \prod_{i=1}^{k} e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}} \cdot \|\bar{u}\|_2 = \prod_{i=1}^{k} e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}} \cdot \frac{1}{\sqrt{n}}$.
So we got a bound on $\|z_k^{\|}\|_2$, but we want a bound on $\|z_k\|_1$. Using the fact that $\langle z_k^{\perp}, \vec{1}\rangle = 0$, and as usual the Cauchy-Schwartz inequality to switch between $\|\ \|_1$ and $\|\ \|_2$:

$$\|z_k\|_1 = \langle z_k, \vec{1}\rangle = \langle z_k^{\|}, \vec{1}\rangle + \overbrace{\langle z_k^{\perp}, \vec{1}\rangle}^{=0} = \langle z_k^{\|}, \vec{1}\rangle \overset{\text{c.s}}{\le} \|z_k^{\|}\|_2 \cdot \|\vec{1}\|_2 = \prod_{i=1}^{k} e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{n}$$

In conclusion: $\|z_k\|_1 \le \prod_{i=1}^{k} e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}}$. This is the bound we needed to "fix" equality (1):

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} f_i(v_i) - \frac{1}{k}\sum_{i=1}^{k}\mu_i\right| > \epsilon\right) \le \frac{\overbrace{\prod \mathbb{E}(e^{\delta f_i(v_i)})}^{=\|z_k\|_1}}{e^{\delta k\epsilon}\cdot\prod e^{\delta\mu_i}} \le \frac{\prod e^{\delta\mu_i + \frac{9\delta^2}{1-\bar{\lambda}}}}{e^{\delta k\epsilon}\cdot\prod e^{\delta\mu_i}} = \frac{e^{\frac{9\delta^2\cdot k}{1-\bar{\lambda}}}}{e^{\delta k\epsilon}} = e^{\frac{9\delta^2\cdot k}{1-\bar{\lambda}} - \delta k\epsilon}$$

To get the best bound possible we want the RHS to be small. We proved it is true for any δ, so we insert the δ which minimizes $f(\delta) = \frac{9\delta^2\cdot k}{1-\bar{\lambda}} - \delta k\epsilon$.

$$f(\delta)' = 0 \implies \frac{18\delta\cdot k}{1-\bar{\lambda}} - k\epsilon = 0 \implies \delta = \frac{1-\bar{\lambda}}{18}\cdot\epsilon$$

Assigning this $\delta = \frac{1-\bar{\lambda}}{18}\cdot\epsilon$, we have proved our theorem:

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} f_i(v_i) - \frac{1}{k}\sum_{i=1}^{k}\mu_i\right| > \epsilon\right) \le e^{\frac{9\delta^2\cdot k}{1-\bar{\lambda}} - \delta k\epsilon} = e^{-\frac{1}{36}\epsilon^2 k(1-\bar{\lambda})}$$

□