3. Random walks

In the second example in Section 1, we chose one vertex from a graph at random and choose its d nbrs. The key was that this performs well while it uses less randomness than choosing several vertices independent at random.

Instead of choosing one vertex and all of its neighbors, what one can do is the following: Choose one vertex v_0 at random (according to some distribution) and for each $i = 0, \dots, t-1$, choose v_{i+1} uniformly at random among the neighbors of v_i. This yields a walk $v_0 v_1 \dots v_t$ on G, which we call a random walk on G.

What we will aim to show is that if t is large, then the random variable v_t is very close to a random choice of a vertex, and the vertex $v_{t'}$ and v_t have very small dependency, so that we can use these vertices instead of choosing vertices independently at random

Choose a vertex v_0 according to a given distribution π_0 and let π_t be the distribution of v_t on the random walk. It is known that every finite connected non-bipartite graph G, this distribution π_i converges to a limit (called stationary) distribution. Moreover, as G is regular, such a distribution is the uniform distribution on V.

Hence, we want to show that as t increases, the distribution π_i converges to the uniform distribution u. Let a graph be (n, d, α)-graph if it is an n-vertex d-regular graph with $\lambda(G) \leq \alpha d$. Let $\mathbf{p} \in \mathbb{R}^n$ be a probability distribution vector if its coordinates are all non-negative and $\sum_{i=1}^n p_i = 1$. Let $\mathbf{u}$ be the vector corresponding to the uniform distribution, $\mathbf{u} = \frac{1}{n}(1, \dots, 1)$. Let $\hat{A} = \frac{1}{d}A$ be the normalized adjacency matrix. Then we know that when $\mathbf{p}$ is the vector representing the distribution π_i, $\hat{A}\mathbf{p}$ is the distribution π_{i+1}. The end vertex of a random walk is Markov chain with state set V and transition matrix $\hat{A}$.

How do we measure convergence of π_i to $\mathbf{u}$? What we usually want to say is that what would be the $\max_B |\mathbb{P}_p[B] - \mathbb{P}_q[B]|$ where B is taken over all events, which is called the total variation distance. If this is small, then two distribution p and q assign almost same probability to every event.

However, in order to maximize this, (if we think in discrete probability space) we add all atom events having more measure on p than q to B (or the other way around). Then we have

$$\mathbb{P}_p[B] - \mathbb{P}_q[B] = \mathbb{P}_q[\overline{B}] - \mathbb{P}_p[\overline{B}]$$

maximized. One the other hand, adding up this two quantity is exactly

$$\sum_e |\mathbb{P}_p(e) - \mathbb{P}_q(e)| = \|p - q\|_1.$$

So, what we want to measure is the half of ℓ_1-distance between them. Hence, we want to measure the distance between two distribution by its ℓ_1-norm. More precisely, we want to show that $\|\pi_i - u\|_1$ fastly converges to 0.

We first show the following theorem which states that this is true if we consider ℓ_2-norm instead.

Theorem 3.1. *Let G be an (n, d, α)-graph with normalized adjacency matrix $\hat{A}$. Then for any distribution vector $\mathbf{p}$ and any positive integer t, we have*

$$\|\hat{A}^t\mathbf{p} - \mathbf{u}\|_2 \leq \alpha^t \|\mathbf{p} - \mathbf{u}\|_2.$$

Proof. We know that $\mathbf{p} - \mathbf{u}$ is orthogonal to $\mathbf{u}$, hence $A^i\mathbf{p} - \mathbf{u}$ is again orthogonal to $\mathbf{u}$. Hence this shrinks in ℓ_2-norm by a factor of α. Hence,

$$\|\hat{A}^{i+1}\mathbf{p} - \mathbf{u}\|_2 = \|\hat{A}(\hat{A}^i\mathbf{p} - \mathbf{u})\|_2 \leq \alpha\|\hat{A}^i\mathbf{p} - \mathbf{u}\|_2 \leq \alpha^{i+1}.$$

This shows the theorem. □

By using Cauchy-Schwartz, we obtain the following theorem.

Theorem 3.2. *Let G be an (n,d,α)-graph with normalized adjacency matrix $\hat{A}$. Then for any distribution vector $\mathbf{p}$ and any positive integer t, we have*

$$\|\hat{A}^t\mathbf{p}-\mathbf{u}\|_1 \le \alpha^t\sqrt{n}.$$

By using this, we can design dependent sampling resembling independent sampling using not too much randomness.

Assume we sample $t+1$ vertices uniformly at random and compute $f(x,v_i)$. Assume that the bad set B has size βn. If all of them 1 then conclude $x\in\mathcal{L}$, otherwise conclude $x\notin\mathcal{L}$. The number of bits of randomness required for this is $(t+1)\log n$ and the algorithm fails with probability β^{t+1}.

Suppose we are given an (n,d,α)-graph G where a bad vertex set $B\subseteq V(G)$ is given with $|B|=\beta n$. We choose one vertex v_0 uniformly at random from $V(G)$ and perform a random walk from there to obtain $v_0\dots v_t$. We compute $f(x,v_i)$ and conclude $x\in\mathcal{L}$ if $f(x,v_i)=1$ for all i, and conclude $x\notin\mathcal{L}$ otherwise. Our algorithm fails if $v_0,v_1,\dots,v_t\in B$. Note that this uses at most $\log n+t\log d$ random bits, as choosing one random neighbors requires a uniform random choice from a set of size d. Now we show that this algorithm also has also exponentially small failing probability.

Let (B,t) be the event that all vertices $v_0,\dots,v_t$ lie in B.

Theorem 3.3 (Ajtai-Komós-Szemerédi 87, Along-Feige-Wigderson-Zuckerman 95)**.** *Let G be an (n,d,α)-graph and $B\subseteq V$ with $|B|=\beta n$. Then we have*

$$\mathbb{P}[(B,t)]\le(\beta+\alpha)^t.$$

Proof. Let $P=P_B$ be the diagonal matrix where $P_{ij}=1$ if $i=j\in B$ and 0 otherwise. This is an orthogonal projection of vector onto the space whose coordinate belongs to B.

Claim 3. $\mathbb{P}[(B,t)]=\|(P\hat{A})^tP\mathbf{u}\|_1$.

Proof. Note that for $x,y\in B$, the x,y entry of $(P\hat{A})^tP$ counts the sum of probabilities of all the walks from x,y through the vertices in B. By multiplying $\mathbf{u}$, we sum the above entries and multiply by $1/n$. Which is the probability that initial vertex is chosen. This yields the equality. □

So, we claim that the following holds.

$$\|P\hat{A}P\mathbf{v}\|_2\le(\beta+\alpha)\|v\|_2. \tag{3.1}$$

For this, we may assume that $\mathbf{v}$ has support in B, otherwise we replace $\mathbf{v}$ with $P\mathbf{v}$, which shrink the right side while keeping left side. As both sides are linear, we assume $\sum v_i=1$. Then $P\mathbf{v}=\mathbf{v}=\mathbf{u}+\mathbf{z}$ where $\mathbf{z}$ is orthogonal to $\mathbf{u}$. Hence,

$$P\hat{A}P\mathbf{v}=P\hat{A}\mathbf{u}+P\hat{A}\mathbf{z}=P\mathbf{u}+P\hat{A}\mathbf{z}.$$

Hence, we have

$$\|P\hat{A}P\mathbf{v}\|_2\le\|P\mathbf{u}\|_2+\|P\hat{A}\mathbf{z}\|_2. \tag{3.2}$$

Here, we wish to bound $\|P\mathbf{u}\|_2$ and $\|P\hat{A}\mathbf{z}\|_2$.

We know $\|P\mathbf{u}\|=\sqrt{\beta/n}$. Since $\sum v_i=1$ and the support of v has at most βn coordinates, Cauchy-Schwartz yields that

$$\|P\mathbf{u}\|=\|P\mathbf{u}\|\sum v_i\le\|P\mathbf{u}\|\sqrt{\beta n}\|\mathbf{v}\|_2\le\beta\|\mathbf{v}\|_2.$$

Also, $\mathbf{z}$ is orthogonal to $\mathbf{u}$ and it is a linear combination of eigenvectors except the first one. As P is a contraction, this implies that

$$\|P\hat{A}\mathbf{z}\|_2\le\|\hat{A}\mathbf{z}\|_2\le\alpha\|\mathbf{z}\|_2\le\alpha\|\mathbf{v}\|_2.$$

These two inequality with (3.2) implies (3.1).

Now using (3.1), we can prove our Theorem.

$$\|(P\hat{A})^t P\mathbf{u}\|_1 \le \sqrt{n}\|(P\hat{A})^t P\mathbf{u}\|_2 = \sqrt{n}\|(P\hat{A}P)^t\mathbf{u}\|_2 \le \sqrt{n}(\beta+\alpha)^t\|\mathbf{u}\|_2 = (\beta+\alpha)^t.$$

□

There are several variations and strengthenings of this theorem. For example, the exponent above is t instead of $t+1$. We can recover this one difference on the exponent as follows.

Theorem 3.4 (Alon-Feige-Wigderson-Zuckerman 95)**.** *If $\beta > 6\alpha$, then*

$$\beta(\beta+2\alpha)^t \ge \mathbb{P}[(B,t)] \ge \beta(\beta-2\alpha)^t.$$

By adapting the previous proof, one may obtain the following 'time dependent' version of the theorem.

Theorem 3.5. *For every subsete $K \subseteq \{0,\dots,t\}$ and vertex set B of size βn, we have*

$$\mathbb{P}[v_i \in B \text{ for all } i \in K] \le (\beta+\alpha)^{|K|-1}.$$

Also, we can vary $B = B_i$ depending on time i.

Theorem 3.6. *Let $B_0,\dots,B_t$ be vertex sets of density $\beta_0,\dots,\beta_t$ in an (n,d,α)-graph G. Let $v_0,\dots,v_t$ be a random walk on G. Then*

$$\mathbb{P}[v_i \in B_i \text{ for all } i] \le \prod_{i=0}^{t-1}(\sqrt{\beta_i\beta_{i+1}}+\alpha).$$

One thing to note in our algorithm is that we only allow one-sided error. Recall the original setting. If $x \in \mathcal{L}$, then $f(x,r)$ always gives 1 and if $x \notin \mathcal{L}$, then $f(x,r)$ gives 0 if $r \in B$.

What if we have bad sets B for every x, no matter $x \in \mathcal{L}$ or not. In other words, what if $f(x,r)$ can be 0 when $x \in \mathcal{L}$.

Assume that there exists f such that if $x \in \mathcal{L}$, then $f(x,r) = 1$ for at least $0.9\cdot 2^k$ choices of $r \in \{0,1\}^k$ and if $x \notin \mathcal{L}$, then $f(x,r) = 0$ for at least $0.9\cdot 2^k$ choices or $r \in \{0,1\}^k$.

Simple way of overcoming this issue is that we sample random number r independently $2s+1$ times and take the majority vote. Again this requires a large amount of randomness, one question is whether we can do random walk approach again.

Again assume that we have a (n,d,α)-graph G on $n = 2^k$ with $V(G) = \{0,1\}^k$. Assume x is given, and $B = B_x$ is the bad set. We pick a vertex $v_0 \in V(G)$ uniformly at random, and take a random walk $v_0 \dots v_t$ of even length t. We compute $f(x,v_i)$ and take the majority.

This algorithm fails if and only if the majority of the v_i's belong to B. For $K \subseteq \{0,1,\dots,t\}$ of cardinality $|K| \ge (t+1)/2$, by the time-dependent theorem above, we have

$$\mathbb{P}[v_i \in B \text{ for all } i \in K] \le (\beta+\alpha)^{|K|-1} \le (\beta+\alpha)^{(t-1)/2}$$

If $\alpha+\beta$ is small enough, say at most $1/8$, then the union bound yields

$$\mathbb{P}[\text{algorithm fails}] \le 2^t(\beta+\alpha)^{(t-1)/2} \le O(2^{-t/2}).$$

We obtain an exponentially small probability.

We have used ℓ_1, ℓ_2 norms to measure how close the distribution is to the uniform one. Another way of measuring it is to use an entropy of the distribution.

Definition 3.7. *Let $\mathbf{p}$ be a probability distribution on $[n]$. We define*

(1) Shannon entropy: $H(\mathbf{p}) = -\sum_{i\in[n]} p_i \log(p_i)$.
(2) Rényi 2-entropy: $H_2(\mathbf{p}) = -2\log(\|\mathbf{p}\|_2)$.
(3) Min entropy: $H_\infty(\mathbf{p}) = -\log(\|\mathbf{p}\|_\infty)$.

These entropies satisfies some common properties. They are always nonnegative and $= 0$ implies entire probability being concentrated on single element. Also they are at most $\log n$ with equality only for the uniform distribution. For any doubly stochastic matrix X, $\tilde{H}(X\mathbf{p}) \geq \tilde{H}(\mathbf{p})$ and equality holds only for uniform $\mathbf{p}$. (Doubly stochastic matrix can be written as a convex combination of permutation matrixes. Using the concavity of the function $x\log(x), \|x\|_2, \|\|_\infty$, one can show this.)

As we have analyzed ℓ_2 norm, we can analyze 2-entropy and show that this increases in expander.

Write $\mathbf{p} = \mathbf{u} + \mathbf{f}$ where $\mathbf{f} \perp \mathbf{u}$. Let $\mu = \|\mathbf{f}\|/\|\mathbf{p}\| \leq 1$. Then

$$\|\hat{A}\mathbf{p}\|^2 = \|\mathbf{u}\|^2 + \|\hat{A}\mathbf{f}\|^2 \leq ((1-\mu^2) + \alpha^2\mu^2)\|\mathbf{p}\|^2.$$

Hence

$$H_2(\hat{A}\mathbf{p}) \geq H_2(\mathbf{p}) - \log(1-(1-\alpha^2)\mu^2).$$

Hence 2-entropy strictly increase as long as $\mathbf{p}$ is not uniform.