SDSC3001 Assignment 2

2024.10.31

Q1

1. (**20 points**) Suppose $A \in \mathbb{R}^{t \times d}$ ($d \gg t$) is a random projection matrix where each entry A_{ij} is independently sampled from a standard normal distribution $\mathcal{N}(0, 1)$. Let a_i^T be the i-th row of A. Prove that the rows of A are nearly orthogonal to each other, that is, $\Pr\{\forall i \neq j, \frac{a_i^T a_j}{d} \leq \epsilon\} \geq 1 - \frac{t^2}{2\epsilon^2 d}$. (**Hint**: use the Chebyshev's inequality)

Q1

Solution:

Denote $x = (x_1, x_2, ..., x_d)$ and $y = (y_1, y_2, ..., y_d)$, where $x_i, y_i \sim \mathcal{N}(0, 1)$ for all i. Then by Chebyshev's inequality:

$$
\begin{aligned}
Pr(\frac{x^T y}{d} \geq \epsilon) \leq \frac{Var(\frac{x^T y}{d})}{\epsilon^2} &= \frac{\sum_{i=1}^{d} Var(x_i y_i)}{d^2 \epsilon^2} \\
&= \frac{Var(x_1 y_1)}{d \epsilon^2} \\
&= \frac{1}{d \epsilon^2}
\end{aligned}
$$

Q1

Solution:

By Union Bound:

$$
\begin{aligned}
Pr(\exists i \neq j, \frac{a_i^T a_j}{d} \geq \epsilon) &\leq \sum_{i \neq j} Pr(\frac{a_i^T a_j}{d} \geq \epsilon) \\
&\leq C_t^2 \frac{1}{d\epsilon^2} \\
&\leq \frac{t^2}{2d\epsilon^2}
\end{aligned}
$$

where $C_t^2 = \binom{t}{2}$ is the number of combination. Then:

$$
Pr(\forall i \neq j, \frac{a_i^T a_j}{d} \leq \epsilon) \geq 1 - \frac{t^2}{2d\epsilon^2}
$$

Q2

2. (**30 points**) Given an undirected graph $G = \langle V, E \rangle$, we hope to partition V into two disjoint sets V_1 and V_2 such that
$cut\,(V_1, V_2) = |\{(u, v) \mid u \in V_1, v \in V_2, (u, v) \in E\}|$ is large. Suppose we randomly, uniformly, and independently assign each node $u \in V$ to V_1 or V_2.

2.1 (6 **points**) Let X_{uv} be an indicator for the edge $(u, v) \in E$. $X_{uv} = 1$ if (u, v) is a cut edge and $X_{uv} = 0$ otherwise. Show that the indicators X_{uv}'s for all edges are not mutually independent.

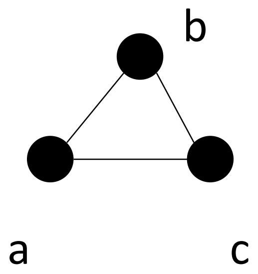
2.2 (9 **points**) Show that the indicators X_{uv}'s for all edges are pairwise independent.

2.3 (15 **points**) Prove that if we generate k random and independent partitions of V, with probability at least $1 - \frac{1}{k}$, among the k random partitions we have a partition V_1 and V_2 such that $cut(V_1, V_2) \geq \frac{|E| - \sqrt{|E|}}{2}$.

Q2

2.1 (6 **points**) Let X_{uv} be an indicator for the edge $(u, v) \in E$. $X_{uv} = 1$ if (u, v) is a cut edge and $X_{uv} = 0$ otherwise. Show that the indicators X_{uv}'s for all edges are not mutually independent.

Solution:



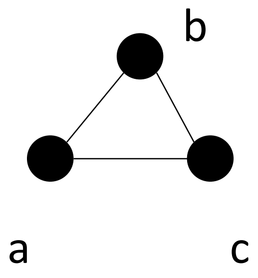
$$Pr(X_{uv} = 0) = Pr(X_{uv} = 1) = \frac{1}{2}$$

$$Pr(X_{bc} = 1 | X_{ab} = 1, X_{ac} = 1) = 0 \neq \frac{1}{2}$$

Q2

2.2 (9 **points**) Show that the indicators X_{uv}'s for all edges are pairwise independent.

Solution:



$$Pr(X_{ac} = 1 | X_{ab} = 0) = Pr(X_{ac} = 0 | X_{ab} = 0) = \frac{1}{2}$$

$$Pr(X_{ac} = 1 | X_{ab} = 1) = Pr(X_{ac} = 0 | X_{ab} = 1) = \frac{1}{2}$$

Q2

2.3 (15 **points**) Prove that if we generate k random and independent partitions of V, with probability at least $1 - \frac{1}{k}$, among the k random partitions we have a partition V_1 and V_2 such that $cut(V_1, V_2) \geq \frac{|E| - \sqrt{|E|}}{2}$.

Solution:

Denote a set of edges $E = \{e_1, e_2, ..., e_m\}$ and k partitions $\{(V_1^{(1)}, V_2^{(1)}), (V_1^{(2)}, V_2^{(2)}) ..., (V_1^{(k)}, V_2^{(k)})\}$.

Let $X^{(i)} = \sum_{j=1}^{m} X_{e_j}^{(i)}$, where $X_{e_j}^{(i)} = X_{e_j}(V_1^{(i)}, V_2^{(i)})$. We have $E(X^{(i)}) = \frac{m}{2}$.

Q2

Solution:

We should prove the following inequality:

$$Pr\big(\max\{X^{(1)}, X^{(2)}, ...X^{(k)}\} \geq \frac{m-\sqrt{m}}{2}\big) \geq 1 - \frac{1}{k}$$

Denote $Z = \frac{1}{k}\sum_{i=1}^{k} X^{(i)} \leq \max\{X^{(1)}, X^{(2)}, ..., X^{(k)}\}$, we can prove:

$$Pr\big(Z \geq \frac{m-\sqrt{m}}{2}\big) \geq 1 - \frac{1}{k}$$
$$\Leftrightarrow Pr\big(Z - \frac{m}{2} \leq \frac{-\sqrt{m}}{2}\big) \leq \frac{1}{k}$$

where Z is the mean number of cut edges in each partition.

$$\boxed{E(Z) = \frac{m}{2}}$$

Q2

Solution:

Target: $Pr(Z - \frac{m}{2} \leq \frac{-\sqrt{m}}{2}) \leq \frac{1}{k}$

Proof: By using Chebyshev's inequality and setting $\epsilon = \frac{\sqrt{m}}{2}$, we have:

$$Pr(Z - E(Z) \leq -\epsilon) \leq \frac{Var(Z)}{\epsilon^2}$$

$$\Leftrightarrow Pr(Z - \frac{m}{2} \leq -\frac{\sqrt{m}}{2}) \leq \frac{4Var(Z)}{m}$$

$\boxed{E(Z) = \frac{m}{2}}$

$$\Leftrightarrow Pr(Z - \frac{m}{2} \leq -\frac{\sqrt{m}}{2}) \leq \frac{1}{k}$$

$\boxed{Var(Z) = \frac{m}{4k}}$

Q2

Solution:

$$Var(Z) = Var(\frac{1}{k}\sum_{i=1}^{k} X^{(i)}) = \frac{Var(X^{(1)})}{k}$$

$$\begin{aligned}
Var(X^{(1)}) &= Var(\sum_{j=1}^{m} X_{e_j}^{(1)}) \\
&= E((\sum_{j=1}^{m} X_{e_j}^{(1)})^2) - E[(\sum_{j=1}^{m} X_{e_j}^{(1)})]^2 \\
&= E(\sum_{j=1}^{m} (X_{e_j}^{(1)})^2) + 2\sum_{i<j} E(X_{e_i}^{(1)} X_{e_j}^{(1)}) - (\frac{m}{2})^2 \\
&= \frac{m}{2} + 2C_m^2 \cdot \frac{1}{2} \cdot \frac{1}{2} - (\frac{m}{2})^2 \\
&= \frac{m}{4}
\end{aligned}$$

So we get: $\boxed{Var(Z) = \frac{m}{4k}}$

Q3

3. **(50 points)** Write code to compute PageRank values of nodes in the DBLP network used in Assignment 1. **You are required to upload your code for this question**. Set $\alpha = 0.15$.

3.0 Implement the power iteration method. Initialize the PageRank vector as $\pi^{(0)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$ where n is the number of nodes. Let $\pi^{(t)}$ be the PageRank vector obtained after the t-th iteration and $\pi_v^{(t)}$ be the PageRank value of node v after the t-th iteration. The power iteration method can be regarded as applying the following updating rule for iterations.

$$\pi_v^{(t)} = (1-\alpha)\left(\sum_{u \in N(v)} \pi_u^{(t-1)} \times \frac{1}{d_u}\right) + \frac{\alpha}{n}$$

Apply the power iterations and return $\pi = \pi^{(t)}$ as the PageRank vector if the t-th iteration is the first time that $\forall v, \left|\pi_v^{(t)} - \pi_v^{(t-1)}\right| \leq 10^{-9}$. π is regarded as the ground truth PageRank vector in the following questions.

Q3

3.1 (25 **points**) Implement the Monte Carlo method. Simulate M random walks as follows. (1) Randomly pick a node as the starting point. (2) At each step, stop with probability α and with probability $1-\alpha$, jump to a random neighbor of the current node. Let f_v be the number of random walks terminated at the node v. Use $\frac{f_v}{M}$ to estimate the PageRank value of v. Denote by π_v the PageRank value of v computed by the power iteration method. The difference between π and the PageRank vector approximated by the Monte Carlo method can be regarded as $\sum_v \left|\pi_v - \frac{f_v}{M}\right|$. Vary M and report the values of $\sum_v \left|\pi_v - \frac{f_v}{M}\right|$ when $M = 2n,\ 4n,\ 6n,\ 8n,\ 10n$.

3.2 (15 **points**) In the above Monte Carlo method, we only use the stopping node to approximate PageRank which is wasteful as all the non-stopping nodes in random walks are ignored. Let s_v be the number of times that v appears in the M random walk. Use $\frac{\alpha s_v}{M}$ to estimate the PageRank value of v. Report the values of $\sum_v \left|\frac{\alpha s_v}{M} - \pi_v\right|$ when $M = 2n,\ 4n,\ 6n,\ 8n,\ 10n$.

3.3 (10 **points**) Show that $\frac{\alpha s_v}{M}$ is an unbiased estimation of π_v (the ground truth PageRank value of v), that is, $E\left[\frac{\alpha s_v}{M}\right] = \pi_v$.

3.0 Implement the power iteration method. Initialize the PageRank vector as $\pi^{(0)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$ where n is the number of nodes. Let $\pi^{(t)}$ be the PageRank vector obtained after the t-th iteration and $\pi_v^{(t)}$ be the PageRank value of node v after the t-th iteration. The power iteration method can be regarded as applying the following updating rule for iterations.

$$\pi_v^{(t)} = (1-\alpha)\left(\sum_{u \in N(v)} \pi_u^{(t-1)} \times \frac{1}{d_u}\right) + \frac{\alpha}{n}$$

Apply the power iterations and return $\pi = \pi^{(t)}$ as the PageRank vector if the t-th iteration is the first time that $\forall v, \left|\pi_v^{(t)} - \pi_v^{(t-1)}\right| \leq 10^{-9}$. π is regarded as the ground truth PageRank vector in the following questions.

```
iteration:  73
3.814575012893011e-09
iteration:  74
3.1457518718232364e-09
iteration:  75
2.5943392343618293e-09
iteration:  76
2.1404505567411334e-09
iteration:  77
1.7660575941852868e-09
iteration:  78
1.4577264963265498e-09
iteration:  79
1.203278289893443e-09
iteration:  80
9.93633356130934e-10
```

3.1 (25 **points**) Implement the Monte Carlo method. Simulate M random walks as follows. (1) Randomly pick a node as the starting point. (2) At each step, stop with probability α and with probability $1-\alpha$, jump to a random neighbor of the current node. Let f_v be the number of random walks terminated at the node v. Use $\frac{f_v}{M}$ to estimate the PageRank value of v. Denote by π_v the PageRank value of v computed by the power iteration method. The difference between π and the PageRank vector approximated by the Monte Carlo method can be regarded as $\sum_v \left|\pi_v - \frac{f_v}{M}\right|$. Vary M and report the values of $\sum_v \left|\pi_v - \frac{f_v}{M}\right|$ when $M = 2n,\ 4n,\ 6n,\ 8n,\ 10n.$

M	Result
2n	0.5421
4n	0.3826
6n	0.3112
8n	0.2702
10n	0.2415

3.2 (15 **points**) In the above Monte Carlo method, we only use the stopping node to approximate PageRank which is wasteful as all the non-stopping nodes in random walks are ignored. Let s_v be the number of times that v appears in the M random walk. Use $\frac{\alpha s_v}{M}$ to estimate the PageRank value of v. Report the values of $\sum_v \left|\frac{\alpha s_v}{M} - \pi_v\right|$ when $M = 2n,\ 4n,\ 6n,\ 8n,\ 10n.$

M	Result
2n	0.2617
4n	0.1847
6n	0.1513
8n	0.1305
10n	0.1164

Q3

3.3 (10 **points**) Show that $\frac{\alpha s_v}{M}$ is an unbiased estimation of π_v (the ground truth PageRank value of v), that is, $E\left[\frac{\alpha s_v}{M}\right] = \pi_v$.

Solution:

Denote $S = [s_1, s_2, ..., s_n]^T$ as the result in M random walks and S^1 the result in 1 random walk. We have:

$$E(S) = ME(S^1)$$

$$\text{Let } a_v^{(i)} = \begin{cases} 1, & if\ the\ i^{th}\ node\ in\ the\ random\ walk\ is\ v \\ 0, & otherwise \end{cases}$$

$$\begin{aligned} S^1 &= \left[\sum_{i=1}^{\infty} a_1^{(i)}, \sum_{i=1}^{\infty} a_2^{(i)}, ..., \sum_{i=1}^{\infty} a_n^{(i)}\right]^T \\ &= \left[a_1^{(1)}, a_2^{(1)}, ..., a_n^{(1)}\right]^T + \left[a_1^{(2)}, a_2^{(2)}, ..., a_n^{(2)}\right]^T + ... \end{aligned} \tag{27}$$

Q3

Solution:

$$E(S^1) = \left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T + (\alpha - 1)P\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T + (\alpha - 1)^2 P^2\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T + ...$$

$$= (I + (\alpha - 1)P + (\alpha - 1)^2 P^2 + ...)\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T$$

(geometric sequence)

$$= (I - (\alpha - 1)P)^{-1}\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T$$

$$\pi = (1 - \alpha)P\pi + \alpha\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T$$

$$\Longrightarrow \pi = \alpha(I - (1 - \alpha)P)^{-1}\left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]^T$$

$$\Longrightarrow \pi = \alpha E(S^1)$$

$$\Longrightarrow \pi = E(\frac{\alpha S}{M})$$