Homework #6

Due: 2025-1-9 23:59 | 3 Problems, 50 Pts

Name: 徐靖, ID: 2200012917

Problem 1 (14'). Consider the following algorithm to estimate the frequency of any number in data streams. The numbers in the data stream are in $[n] := \{1, 2, \cdots, n\}$.

Algorithm 1: Estimate the frequency of numbers in data streams

$\boldsymbol{C} \leftarrow 0$ ▷ $\boldsymbol{C}$ is a $t \times k$-dimension matrix

Choose t independent hash functions $h_1, \cdots, h_t : [n] \to [k]$ from a 2-universal hash family

for *j in data streams* **do**

  for $i = 1, 2, \cdots, t$ **do**

    $\boldsymbol{C}[i][h_i(j)] \leftarrow \boldsymbol{C}[i][h_i(j)] + 1$

  end

end

for *a in queries* **do**

  output $\hat{f}(a) = \min_{1 \le i \le t} \boldsymbol{C}[i][h_i(a)]$

end

For any element a, suppose the real frequency of a is $f(a)$. When $k = \lceil \frac{2}{\epsilon} \rceil$, $t = \lceil \log_2 \frac{1}{\delta} \rceil$, prove that for any given a, with probability at least $1 - \delta$, $f(a) \le \hat{f}(a) \le f(a) + \epsilon L$, where L is the length of the data stream. ◀

Answer. Let $X_i = \frac{C[i][h_i(a)] - f(a)}{L}$ and A_i denotes $X_i > \epsilon$. Obviously we have,

$$C[i][h_i(a)] = \sum_{b \in \{b | h_i(a) = h_i(b)\}} f(b) \ge f(a)$$

Thus we know $X_i \ge 0$ and $P(\hat{f}(a) < f(a)) = P(\min_{i \in [t]} X_i < 0) = 0$.

Given that $\forall x, y \in [k], a \ne b \in [n], P(h_i(a) = x, h_i(b) = y) = \frac{1}{k}$, we could get:

$$\begin{aligned}
\sum_{i=1}^{t} C[i][h_i(a)] - f(a) &= \sum_{i=1}^{t} \left(-f(a) + \sum_{b \in \{b | h_i(a) = h_i(b)\}} f(b) \right) \\
&= \sum_{(i,b) \in \{(x,y) | y \ne a, h_x(a) = h_x(b)\}} f(b) \\
&= \sum_{b \in [n] \setminus \{a\}} t P(h_i(a) = h_i(b)) f(b) \\
&= \frac{t}{k}(L - f(a))
\end{aligned}$$

which means $E(X_i) = \frac{L - f(a)}{kL} \le \frac{1}{k}$. Using Markow inequality, we have:

$$P(A_i) = P(X_i > \epsilon) \le \frac{E(X_i)}{\epsilon} \le \frac{1}{\epsilon k}$$

Then,

$$P(\hat{f}(a) > f(a) + \epsilon) = P(\min_{i\in[t]} X_i > \epsilon) = P(\cap A_i) = P(A_i)^t \leq \left(\frac{1}{\epsilon k}\right)^t \leq \delta$$

In conclusion, we have $P(f(a) \leq \hat{f}(a) \leq f(a) + \epsilon L) > 1 - \delta$.

◁

Problem 2 (18'). A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. Find out and prove the threshold for $\mathcal{G}(n, p)$ to be bipartite.

[Hint: The definition of bipartite graph is equivalent to a graph that does not contain any odd-length cycles.] ◀

Answer. We know bipartite graph is equivalent to a graph that does not contain any odd-length cycles. So let X_k denotes the number of k-length cycles in the graph, and $X_S = \sum_{k\in S} X_k$, and index set $I = \{3, 5, 7, \ldots, 2[\frac{n+1}{2}]\}$. Given that,

$$P(G \text{ contain odd-length cycles}) = P(X_I > 0) = \sum_{k\in\mathbb{N}} P(X_I = k) \leq \sum_{k\in\mathbb{N}} kP(X_I = k) = E(X_I)$$

Let S be the set of all places in the graph where a cycle could occur. Explicitly, S_k is the set of all subsets of k vertices ordered up to rotation and orientation of the cycle, and $S = \bigcup_{k\in I} S_k$.

For $s \in S$, define A_s to be the event that a odd-length cycle occurs on s in the random graph . As expectation is linear, we have:

$$E(X_I) = \sum_{s\in S} E(1_{A_S}) = \sum_{k\in I} \sum_{s\in S_k} P(A_s).$$

For $s \in S_k$, the probability that a cycle occurs on s is p^k, as we need each of the k independent edges which form the cycle to be present in our random graph. We want to determine $|S_k|$. The number of ordered sets of size k is $\binom{n}{k}k!$, which overcounts each $S \in S_k$ by $2k$ times.

Hence,

$$|S_k| = \frac{\binom{n}{k}k!}{2^k} = \binom{n}{k}\frac{(k-1)!}{2}.$$

Thus, by the above formula:

$$E(X_I) = \sum_{k\in I} \binom{n}{k}\frac{(k-1)!}{2}p^k.$$

Now, note that $\binom{n}{i}i! = n(n-1)\ldots(n-i+1) \leq n^i$, and we get:

$$E(X_I) \leq \sum_{k\in I} n^k p^k \leq \frac{n^3p^3}{1-n^2p^2}.$$

Thus, as $E(X_I) \to 0$ for $np \to 0$, we have $P(G \text{ has a odd-length cycle}) \leq E(X_I)$, and we have proven that, with high probability , G is bipartite.

For the second part of the proof, we have $0 \leq X_3 \leq X_I$.As demonstrated in class,

$$P(X_I = 0) \leq P(X_3 = 0) \leq \frac{\mathrm{Var}(X_3)}{E^2(X_3)} \sim \frac{n^3p^3 + n^4p^5}{n^6p^6} \to 0 (n \to 0)$$

Hence $\frac{1}{n}$ is the threshold. ◁

Problem 3 (18'). A vertex is called an isolated vertex if it does not have any edges. Prove that, the threshold for $\mathcal{G}(n,p)$ of the existence of isolated vertex is $p = \frac{\ln n}{n}$. ◀

Answer. Let E_i be the event that vertex v_i is isolated in $G_{n,p}$, and let E be the event that at least one vertex is isolated in $G_{n,p}$. I have proved a stronger conclusion: In $G_{n,p}$ with $p = \frac{\ln n + c}{n}$, the probability that there is an isolated vertex converges to $1 - e^{-e^{-c}}$.

First, I want to compute $P(E) = P\left(\bigcup_{i=1}^{n} E_i\right)$. By the inclusion-exclusion principle, we have:

$$P(E) = \sum_{k=1}^{n} (-1)^k \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} P\left(\bigcap_{j=1}^{k} E_{i_j}\right).$$

Using Bonferroni inequalities, we get:

$$P(E) \leq \sum_{k=1}^{l} (-1)^k \sum_{1 \leq i_1 < \cdots < i_k \leq n} P\left(\bigcap_{j=1}^{k} E_{i_j}\right), \quad \text{for odd } l.$$

Now, calculate $P\left(\bigcap_{j=1}^{k} E_{i_j}\right)$, which is the probability that the set of k vertices $v_{i_1}, v_{i_2}, \ldots, v_{i_k}$ are all isolated:

$$P\left(\bigcap_{j=1}^{k} E_{i_j}\right) = (1-p)^{n-k} \cdot \binom{k}{2} = (1-p)^{n-k} \cdot \frac{k(k-1)}{2}.$$

Thus, we have:

$$P(E) \leq \sum_{k=1}^{l} (-1)^k \binom{n}{k} (1-p)^{n-k} \cdot \frac{k(k-1)}{2}, \quad \text{for odd } l.$$

As $n \to \infty$, we have:

$$\binom{n}{k} (1-p)^{n-k} \cdot \frac{k(k-1)}{2} \sim \frac{n^k}{k!} \cdot e^{-ck}.$$

Therefore, the summation becomes:

$$P(E) \sim \sum_{k=1}^{\infty} (-1)^k \frac{e^{-ck}}{k!}.$$

For odd l, we have:

$$\lim_{n\to\infty} P(E) \leq \sum_{k=1}^{\infty}(-1)^k \frac{e^{-c}}{k!} = 1 - e^{-e^{-c}}.$$

Similarly, for even l, we get:

$$\lim_{n\to\infty} P(E) \geq \sum_{k=1}^{\infty}(-1)^k \frac{e^{-c}}{k!} = 1 - e^{-e^{-c}}.$$

Altogether, we conclude that:

$$\lim_{n\to\infty} P(E) = 1 - e^{-e^{-c}}.$$

Back to the original question, we find that when $c \to +\infty$, $P(E) = 1 - e^{-e^{-c}} \to 0$, or the probability of an isolated point appearing is 0. And when $c \to +\infty$, $P(E) \to 1$

So the threshold is $\frac{\ln n}{n}$

$\triangleleft$