CSE 632 (Fall 2019): Analysis of Algorithms II : Randomized Algorithms

Lecture 11 (10/2/2019): Discrepancy

Lecturer: Shi Li *Scriber: Shuyan Chen*

1 Review of Chernoff Bounds

Let X_1,X_2, ... ,X_n be independent random variables, taking values in [0,1]. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then $\forall\delta > 0$, we have $\Pr[X > (1+\delta)\mu] < (\frac{e^\delta}{(1+\delta)^{1+\delta}})^\mu$ and $\Pr[X < (1-\delta)\mu] < (\frac{e^{-\delta}}{(1-\delta)^{1-\delta}})^\mu$.

In this and the next lecture, we shall consider two applications of the Chernoff bound on $\Pr[X > (1+\delta)\mu]$. In particular, we apply the bound for δ tending to 0 and ∞ respectively in the two applications. It is convenient to think of $\frac{e^\delta}{(1+\delta)^{1+\delta}}$ as $e^{-\delta^2/2}$ when δ goes to 0, and as $\frac{1}{\delta^\delta}$ when δ goes to ∞.

2 Problem Description

Assume we have m subsets: $S_1, S_2, ..., S_m$ of $[n]$ (think of that $m = \Theta(n)$). For a coloring $\chi : [n] \to \{-1, 1\}$. Let define discrepancy of S_i w.r.t the coloring χ to be $\mathrm{disc}\chi(S_i) = |\sum_{j\in S_i} X(j)|$. Thus, if we view -1 and 1 as two different colors, then $\mathrm{disc}\chi(S_i)$ is the difference between the numbers of elements in S_i with the two colors. Our goal is to find a coloring χ with small $\max_{i\in[m]} \mathrm{disc}_\chi(S_i)$.

Give an example about discrepancy. Assume $n = 4$ and $m = 6$, we have 4 subsets given in the following table. Let the coloring χ be the following: $\chi_1 = \chi_3 = \chi_4 = -1$ and $\chi_2 = \chi_5 = \chi_6 = 1$. So the $\mathrm{disc}_\chi(S_i)$ given in the table.

Table 1: Discrepancy of subsets.

i	S_i	$\mathrm{disc}_\chi(S_i)$
1	{1,2,3,4}	2
2	{1,5,6}	1
3	{1,2,5,6}	2
4	{3,4,6}	1

Thus χ has maximum discrepancy 2. If we change χ_3 to 1 and χ_5 to -1, then the maximum discrepancy becomes 1.

2.1 Find a good coloring χ

Theorem 1. *There is a coloring χ s.t $max_{i=1}^{m} disc_\chi(S_i) = O(\sqrt{n\log m})$.*

So, if $n = \Theta(m)$, then there is a coloring χ with maximum discrepancy $O(\sqrt{n\log n})$.

Proof. We randomly give each element a $\{\pm1\}$ color and show that with high probability the coloring has a small discrepancy. Define $X_j = \begin{cases} 0 & \text{w.p } \frac{1}{2} \\ 1 & \text{w.p } \frac{1}{2} \end{cases}$. Then the coloring of j will be $\chi_j = 2X_j - 1$.

For a fixed set S_i we have $\mathbb{E}[\sum_{j\in S_i} X_j] = \frac{|S_i|}{2}$. We can define a super set $S'_i \supseteq S_i$ and dummy variables $X_j \in [0, 1]$ for every $j \in S'_i \setminus S_i$. We make sure that X_j for every $j \in S'_i \setminus S_i$ is deterministic and $\sum_{j\in S'_i\setminus S_i} X_j = \frac{n-|S_i|}{2}$. Thus $\mathbb{E}[\sum_{j\in S'_i} X_j] = \frac{n}{2}$.

$$\Pr\left[\sum_{j\in S_i} X_j > \frac{|S_i|}{2} + \frac{\delta n}{2}\right] = \Pr\left[\sum_{j\in S'_i} X_j > \frac{n}{2} + \frac{\delta n}{2}\right] = \Pr\left[\sum_{j\in S'_i} X_j > (1+\delta)\frac{n}{2}\right].$$

Suppose $\delta \in [0,1]$, by using Chernoff bound, we have

If we set $\delta = \sqrt{\frac{6\ln(4m)}{n}}$ and assume $\delta < 1$. Then we have

$$\Pr\left[\sum_{j\in S_i'} X_j > (1+\delta)\frac{n}{2}\right] < e^{-\frac{\delta^2}{3}\cdot\frac{n}{2}} = \frac{1}{4m}.$$

This, implies $\Pr\left[\sum_{j\in S_i} X_j > \frac{|S_i|}{2} + \frac{\delta n}{2}\right] < \frac{1}{4m}$. Similarly, we can prove $\Pr\left[\sum_{j\in S_i} X_j < \frac{|S_i|}{2} - \frac{\delta n}{2}\right] < \frac{1}{4m}$. So we have

$$Pr\left[\Big|\sum_{j\in S_i} X_j - \frac{|S_i|}{2}\Big| > \frac{\delta n}{2}\right] \leq \frac{1}{2m}$$

Lemma 2 (Union Bound)**.** *Suppose there are events E_1, $E2$, ... ,E_m, such that E_i happens with probability p_i for every i in this m. Then with probability at least $1 - \sum_{i=1}^m p_i$, none of the m events happens.*

Applying the union bound, we have

$$Pr\left[\forall i \in [m], \Big|\sum_{j\in S_i} X_j - \frac{|S_i|}{2}\Big| \leq \frac{\delta n}{2}\right] \geq 1 - m\cdot\frac{1}{2m} = \frac{1}{2}.$$

Notice that $\left|\sum_{j\in S_i} X_j - \frac{|S_i|}{2}\right| \leq \frac{\delta n}{2}$ is equivalent to $|\sum_{j\in S_i} \chi_j| \leq \delta n$. Thus, with probability at least $1/2$, we have $\max_{i\in[m]} \mathrm{disc}_\chi(S_i) \leq \delta n = O(\sqrt{n\log m})$. In particular, this implies there exists a coloring χ such that the event happens.

When $\delta \geq 1$ then the discrepancy is at most $n \leq \delta n = O(\sqrt{n\log m})$. $\square$

Notice that the proof not only proves the existence of a good coloring χ, but also gives a randomized algorithm that produces such a coloring χ: We just let χ be a random coloring and check if it has maximum discrepancy at most δn or not; if not we repeat. Since the success probability is at least $1/2$, in expectation we only need to run the procedure twice.