CSE 632: Analysis of Algorithms II: Randomized Algorithms **Spring 2024**

Lecture 7: Applications of Chernoff Bounds

Lecturer: Zongchen Chen

Lemma 1 (Chernoff Bounds). *Let $X_1, \dots, X_n$ be* independent *random variables taking values in* $[0,1]$. *Let* $Y = \sum_{i=1}^n X_i$ *be their sum with mean* $\mu = \mathbb{E}Y = \sum_{i=1}^n \mathbb{E}X_i$. *Then we have*

$$\forall \delta \in [0,1]: \quad \Pr\left(|Y - \mu| \ge \delta\mu\right) \le 2e^{-\frac{\delta^2 \mu}{3}};$$

$$\textit{Equivalently,} \quad \forall a \in [0,\mu]: \quad \Pr\left(|Y - \mu| \ge a\right) \le 2e^{-\frac{a^2}{3\mu}}.$$

1 Discrepancy

Suppose V is a set of n elements, and $\mathcal{S} = \{S_1, \dots, S_m\}$ is a collection of subsets of V. Then $G = (V, S)$ forms a hypergraph with n vertices and m hyperedges. A 2-coloring of G is a mapping $\chi : V \to \{\text{red}, \text{blue}\}$ which colors every vertex by either red or blue. Our goal is to find a 2-coloring that balances all hyperedges as much as possible, that is, in each hyperedge roughly half of the vertices are red and half are blue. More precisely, we define the *discrepancy* of $\mathcal{S}$ with respect to a 2-coloring χ by

$$\mathrm{disc}(\mathcal{S}, \chi) := \max_{i=1,\dots,m} \Big| \text{\# of red vertices in } S_i - \text{\# of blue vertices in } S_i \Big|.$$

Our goal is to find a 2-coloring $\chi : V \to \{\text{red}, \text{blue}\}$ which makes $\mathrm{disc}(\mathcal{S}, \chi)$ small.

Lemma 2. *There exists a* 2*-coloring* χ *such that*

$$\mathrm{disc}(\mathcal{S}, \chi) = O\left(\sqrt{n \log m}\right).$$

Moreover, we can find such a 2*-coloring in* $O(mn)$ *time with probability at least* $1/2$.

Proof. We shall prove that a uniformly random 2-coloring χ (i.e., each vertex is colored by red or blue with probability $1/2$ independently) satisfies

$$\Pr\left(\mathrm{disc}(\mathcal{S}, \chi) \le \sqrt{6n \log(4m)}\right) \ge \frac{1}{2}.$$

This also gives an algorithm for finding such a 2-coloring with probability $1/2$.

For each $v \in V$, define a Bernoulli random variable X_v such that $X_v = 1$ if v is red, and $X_v = 0$ if v is blue. For each $i = 1, \dots, m$, let $Y_i = \sum_{v \in S_i} X_v$ be the number of red vertices in S_i. Fix some i and consider the discrepancy in the hyperedge S_i. Suppose $k_i = |S_i|$. We observe that $\mathbb{E}Y_i = k_i/2$ and

$$\begin{aligned}
\text{discrepancy in } S_i &= \Big| \text{\# of red vertices in } S_i - \text{\# of blue vertices in } S_i \Big| \\
&= |Y_i - (k_i - Y_i)| = 2|Y_i - \mathbb{E}Y_i|.
\end{aligned}$$

By the Chernoff bound, we have

$$\begin{aligned}
\Pr\left((\text{discrepancy in } S_i) \ge \sqrt{6n \log(4m)}\right) &= \Pr\left(|Y_i - \mathbb{E}Y_i| \ge \sqrt{\frac{3}{2} n \log(4m)}\right) \\
&\le 2e^{-\frac{(3n/2)\log(4m)}{3(k_i/2)}} && \text{(Chernoff bound)} \\
&\le 2e^{-\log(4m)} && (k_i \le n) \\
&= \frac{1}{2m}.
\end{aligned}$$

Notice that to apply the Chernoff bound we need to have $\sqrt{\frac{3}{2}n\log(4m)} \le \mathbb{E}Y_i$, which is $k_i \ge \sqrt{6n\log(4m)}$; we can assume this without loss of generality, since otherwise the discrepancy is trivially upper bounded by the size of the hyperedge.

Finally, an application of the union bound yields

$$\begin{aligned}
\Pr\left(\operatorname{disc}(\mathcal{S},\chi) \ge \sqrt{6n\log(4m)}\right) &= \Pr\left(\exists i : (\text{discrepancy in } S_i) \ge \sqrt{6n\log(4m)}\right) \\
&\le \sum_{i=1}^{m} \Pr\left((\text{discrepancy in } S_i) \ge \sqrt{6n\log(4m)}\right) \\
&\le m \cdot \frac{1}{2m} = \frac{1}{2},
\end{aligned}$$

as claimed. □

2 Mean Estimation

Suppose $X_1, X_2, \ldots$ are i.i.d. samples of a random variable X. Our goal is to estimate $\mathbb{E}X$ from these samples. Moreover, we want to understand how many sample are needed so that our estimate $\hat{X}$ satisfies

$$\Pr\left(|\hat{X} - \mathbb{E}X| \le \varepsilon\mathbb{E}X\right) \ge 1 - \delta,$$

where $\varepsilon, \delta \in (0,1)$ are given (ε is the approximation error and δ is the failure probability).

Simple approach: Sample mean Suppose we have n i.i.d. samples $X_1, \ldots, X_n$. Let $Y = \sum_{i=1}^{n} X_i$ be their sum, so $\mathbb{E}Y = n\mathbb{E}X$. Let $\hat{X} = Y/n$ be the sample mean, so $\mathbb{E}\hat{X} = \mathbb{E}X$. We have

$$\Pr\left(|\hat{X} - \mathbb{E}X| \ge \varepsilon\mathbb{E}X\right) = \Pr\left(|Y - \mathbb{E}Y| \ge \varepsilon\mathbb{E}Y\right).$$

We *cannot* use the Chernoff bound here since X may not be bounded (e.g., X is geometric or Gaussian). Instead, we apply the Chebyshev's inequality:

$$\begin{aligned}
\Pr\left(|Y - \mathbb{E}Y| \ge \varepsilon\mathbb{E}Y\right) &\le \frac{\operatorname{Var}(Y)}{(\varepsilon\mathbb{E}Y)^2} && \text{(Chebyshev's Inequality)} \\
&= \frac{n\operatorname{Var}(X)}{(\varepsilon n\mathbb{E}X)^2} && (\operatorname{Var}(Y) = n\operatorname{Var}(X)) \\
&= \frac{r^2}{\varepsilon^2 n},
\end{aligned}$$

where

$$r = \frac{\sqrt{\operatorname{Var}(X)}}{\mathbb{E}X}.$$

We want the failure probability to be at most δ, so it suffices to take

$$\frac{r^2}{\varepsilon^2 n} \le \delta \quad \Longleftrightarrow \quad n \ge \frac{r^2}{\varepsilon^2\delta}. \tag{1}$$

Namely, the number of samples we need is $n = O(r^2/(\varepsilon^2\delta))$. The linear dependency on $1/\delta$ is bad for many applications. We will give a better approach which achieves $\log(1/\delta)$ dependency for sample complexity.

Better approach: Median of means From previous analysis, we see that for $m = 4r^2/\varepsilon^2$ and $\hat{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, it holds

$$\Pr\left(|\hat{X} - \mathbb{E}X| \geq \varepsilon \mathbb{E}X\right) \leq \frac{r^2}{\varepsilon^2 m} = \frac{1}{4}.$$

We say an estimate $\hat{X}$ is *good* if $|\hat{X} - \mathbb{E}X| \leq \varepsilon \mathbb{E}X$. Therefore,

$$\Pr\left(\hat{X} \text{ is good}\right) \geq \frac{3}{4}.$$

Our plan is to generate many such estimates $\hat{X}$ independently. Since each estimate is good with probability at least 3/4, the majority of these estimates would be good with high probability, and hence we can use the median of these estimates as our final estimator.

Algorithm 1 Median of Means

Input: $n = \ell m$ i.i.d. samples $X_1, \dots, X_n$ where $m = 4r^2/\varepsilon^2$ and $\ell = 24\log(1/\delta)$
  for $k = 1, \dots, \ell$ **do**
    $\hat{X}_k \leftarrow \frac{1}{m}\sum_{i=1}^{m} X_{(k-1)m+i}$;
  end for
    return $\hat{X} \leftarrow$ median of $\{\hat{X}_1, \dots, \hat{X}_\ell\}$

Observe that $\hat{X}_1, \dots, \hat{X}_\ell$ are jointly independent. For each k, we have $\Pr(\hat{X}_k \text{ is good}) \geq 3/4$. We have the following key observation.

Claim 3. *If more than half of the estimates in $\{\hat{X}_1, \dots, \hat{X}_\ell\}$ are good, then the median $\hat{X}$ of them is good.*

It suffices to show that with high probability the fraction of good estimates is larger than 1/2. For each k, define Z_k to be a Bernoulli random variable such that $Z_k = 1$ if $\hat{X}_k$ is good, and $Z_k = 0$ otherwise. Let $Z = \sum_{k=1}^{\ell} Z_k$ be the number of good estimates among all ℓ of them. Note that $Z_1, \dots, Z_\ell$ are independent and

$$\mathbb{E}Z_k = \Pr\left(\hat{X}_k \text{ is good}\right) \geq 3/4.$$

Hence, $\mathbb{E}Z \geq 3\ell/4$. We deduce from the Chernoff bound that

$$\begin{aligned}
\Pr\left(\hat{X} \text{ is } \textit{not} \text{ good}\right) &\leq \Pr\left(Z \leq \frac{\ell}{2}\right) \\
&\leq \Pr\left(Z \leq \frac{2}{3}\mathbb{E}Z\right) \\
&\leq e^{-\frac{(1/3)^2 \mathbb{E}Z}{2}} && \text{(Chernoff Bound)} \\
&\leq e^{-\frac{(1/3)^2 (3\ell/4)}{2}} \\
&= e^{-\ell/24}.
\end{aligned}$$

To make the failure probability at most δ, it suffices to take

$$e^{-\ell/24} \leq \delta \quad \Longleftrightarrow \quad \ell \geq 24\log(1/\delta).$$

In conclusion, the number of samples we need is

$$n = \ell m = 24\log(1/\delta) \cdot \frac{4r^2}{\varepsilon^2} = O\left(\frac{r^2}{\varepsilon^2}\log(1/\delta)\right).$$