Algorithms for Big Data (II) (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Guoliang Qiu

Last modified on Oct 6, 2020

Today we first complete the proofs of the concentration inequalities used last time. Then I will introduce the Balls-into-Bins model, a simple yet useful model which can be used to analyze the Hash function, an important algorithmic tool used in streaming algorithms.

1 Concentration Inequality

We first complete the proof of Markov inequality, Chebyshev inequality and Chernoff bound we mentioned last time in this section.

1.1 Markov's Inequality

Lemma 1 (Markov's inequality). *For every non-negative random variable X and every $a > 0$, it holds that*

$$\mathbf{Pr}\left[X \ge a\right] \le \frac{\mathbf{E}\left[X\right]}{a}.$$

Proof. Let $\mathbf{1}[X \ge a]$ be the indicator random variable such that $\mathbf{1}[X \ge a] = \begin{cases} 1, \text{if } x \ge a, \\ 0, \text{otherwise.} \end{cases}$ Then it holds that $X \ge a \cdot \mathbf{1}[X \ge a]$. Take the expectation on both sides, we obtain

$$\mathbf{E}\left[X\right] \ge a \cdot \mathbf{E}\left[\mathbf{1}[X \ge a]\right] = a \cdot \mathbf{Pr}\left[X \ge a\right].$$

□

1.2 Chebyshev Inequality

Lemma 2 (Chebyshev Inequality). *For every random variable X and every $a \ge 0$, it holds that*

$$\mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \ge a\right] \le \frac{\mathbf{Var}\left[X\right]}{a^2}.$$

Proof.

$$\begin{aligned} \mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \ge a\right] &= \mathbf{Pr}\left[(X - \mathbf{E}\left[X\right])^2 \ge a^2\right] \\ &\le \frac{\mathbf{E}\left[(X - \mathbf{E}\left[X\right])^2\right]}{a^2} \quad \text{(According to the Markov's inequality)} \\ &= \frac{\mathbf{Var}\left[X\right]}{a^2}. \end{aligned}$$

□

1.3 Chernoff Bound

Lemma 3 (Chernoff Bound). *Let $X_1, \cdots, X_n$ be independent Bernoulli trials with $\mathbf{E}[X_i] = p_i$ for every $i = 1, \cdots, n$. Let $X = \sum_{i=1}^n X_i$. Then for every $0 < \epsilon < 1$, it holds that*

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq \epsilon \cdot \mathbf{E}[X]] \leq 2\exp\left(-\frac{\epsilon^2 \mathbf{E}[X]}{3}\right)$$

The main tool to prove Chernoff bound is the moment generating function $\mathbf{E}\left[e^{tX}\right]$ for a random variable X.

Proof. It holds that

$$\begin{aligned}\mathbf{E}\left[e^{tX}\right] = \mathbf{E}\left[e^{t\sum_{i=1}^n X_i}\right] &= \prod_{i=1}^n \mathbf{E}\left[e^{tX_i}\right] \\ &= \prod_{i=1}^n ((1+p_i) + p_i e^t) \\ &= \prod_{i=1}^n (1 - (1-e^t)p_i) \\ &\leq \prod_{i=1}^n e^{-(1-e^t)p_i} = e^{-(1-e^t)\mathbf{E}[X]}\end{aligned}$$

For every $t > 0$, we have

$$\mathbf{Pr}[X \geq (1+\epsilon)\mathbf{E}[X]] = \mathbf{Pr}\left[e^{tX} \geq e^{t(1+\epsilon)\mathbf{E}[X]}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t(1+\epsilon)\mathbf{E}[X]}} \leq \frac{e^{-(1-e^t)\mathbf{E}[X]}}{e^{t(1+\epsilon)\mathbf{E}[X]}}.$$

The next step is to find an optimal t to minimize the last term in the above line. By calculating the its derivative, we can determine that $t = \log(1+\epsilon)$ is the minimizer. Plugging this into

$$\mathbf{Pr}[X \geq (1+\epsilon)\mathbf{E}[X]] \leq \left(\frac{e^\epsilon}{(1+\epsilon)^{1+\epsilon}}\right)^{\mathbf{E}[X]} \leq e^{-\epsilon^2 \mathbf{E}[X]/3}.$$

We can similarly prove that

$$\mathbf{Pr}[X \leq (1-\epsilon)\mathbf{E}[X]] \leq e^{-\epsilon^2 \mathbf{E}[X]/2}.$$

Combining the bounds for both lower and upper tails, we finish the proof. □

2 Balls-into-Bins Model

The Balls-into-Bins model is a simple yet important probabilistic model, especially from the randomized algorithm analysis perspective. In today's lecture, we focus on the m-balls-into-n-bins model. It models an important object, the hash functions which are central technical tools in streaming algorithms.

First assume $m = n$. Suppose we throw n balls into n bins uniformly and independently, what is the (expected) max load (the number of balls in the fullest bin) of bins? We can check that the max load is $\frac{\log n}{\log\log n} \cdot (1 + o(1))$ with probability $1 - o(1)$. For simplicity, we show that for some c, $\mathbf{Pr}\left[X > \frac{c \log n}{\log\log n}\right] \leq \frac{1}{2}$.

For each $i = 1, \ldots, n$, let the random variable X_i denote the balls in the i-th bin and let $X = \max X_i$. Suppose we fix $k = \frac{c \log n}{\log \log n}$ for some c, then by the union bound and the Stirling's formula, we have

$$\begin{aligned} \mathbf{Pr}[X > k] &= \mathbf{Pr}[\exists i, X_i > k] \leq n \cdot \mathbf{Pr}[X_i > k] \\ &\leq n \cdot \binom{n}{k} \cdot n^{-k} \leq \frac{n}{k!} \leq n \cdot (\frac{e}{k})^k \leq \frac{1}{2} \end{aligned}$$

where $n \cdot (\frac{e}{k})^k \leq \frac{1}{2}$ follows from the fact that

$$\begin{aligned} \log n + \frac{c \log n}{\log \log n}(1 + \log\log\log n - \log c - \log\log n) &= \log n(-c + 1 - \frac{c \log c}{\log \log n} + \frac{c \log\log\log n}{\log \log n}) \\ &\leq -\log 2 \ \text{ (for sufficiently large } n \text{ and } c). \end{aligned}$$

3 Independence

In this section, we discuss the notion of "independence" in probability theory. (Notice that we only discuss the discrete random variables.)

- A set of random variables $X_1, \cdots, X_n$ are *mutually independent* if for every index set $I \subseteq [n]$ and values $\{x_i\}_{i \in I}$,

  $$\mathbf{Pr}\left[\bigwedge_{i \in I} X_i = x_i\right] = \prod_{i=1}^{n} \mathbf{Pr}[X_i = x_i].$$

  Obviously, the mutual independence is a very strong condition as it requires the property of "being independent" for every subset of variables $I \subseteq [n]$. We can relax the requirement and only ask for independence for those $I \subseteq [n]$ with $|I| \leq k$. This is called k-wise independence.

- A set of random variables $X_1, \cdots, X_n$ are k-wise independent if for every index set $I \subseteq [n]$ with $|I| \leq k$, and values $\{x_i\}_{i \in I}$,

  $$\mathbf{Pr}\left[\bigwedge_{i \in I} X_i = x_i\right] = \prod_{i=1}^{n} \mathbf{Pr}[X_i = x_i].$$

  We call $X_1, \cdots, X_n$ pairwise independent if they are 2-wise independent.

It is clear that mutual independence implies k-wise independence, however, the opposite direction is not correct. Suppose we have two independent random variables $X, Y \in_R \{0, 1\}$, and a random variable $Z = X \oplus Y$. We know that the random variable Z is also uniformly distributed on $\{0, 1\}$ and these three random variables are pairwise independent but not mutually independent.

3.1 Property of Pairwise Independence

We know the linearity property of variance holds for independent random variables. The independence here can be pairwise independence.

Theorem 4. *For pairwise independent $X_1, \cdots, X_n$, we have*

$$\mathbf{Var}\left[X_1 + \cdots, X_n\right] = \mathbf{Var}\left[X_1\right] + \cdots + \mathbf{Var}\left[X_n\right].$$

Proof.

$$\begin{aligned}\mathbf{Var}\left[X_1 + \cdots + X_n\right] &= \sum_{i=1}^{n} \mathbf{E}\left[X_i^2\right] + \sum_{1 \le i < j \le n} \mathbf{E}\left[X_i X_j\right] - \left(\sum_{i=1}^{n} (\mathbf{E}\left[X_i\right])^2 + 2 \sum_{1 \le i < j \le n} \mathbf{E}\left[X_i\right] \cdot \mathbf{E}\left[X_j\right]\right) \\ &= \sum_{i=1}^{n} \left(\mathbf{E}\left[X_i^2\right] - (\mathbf{E}\left[X_i\right])^2\right) = \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right].\end{aligned}$$

□

4 Hash Functions

In the model of balls-into-bins, we distribute balls uniformly and independently. This can be implemented using Hash functions. Hash functions are important data structures that have been widely used in computer science. In this section, we will construct Hash functions with theoretical guarantees.

4.1 Universal Hash Function Families

Let $\mathcal{H}$ be a family of functions from $[m]$ to $[n]$ where $m \ge n$. We call $\mathcal{H}$ *k-universal* if for every distinct $x_1, \cdots, x_k \in [m]$, we have

$$\mathbf{Pr}_{h \in \mathcal{H}}\left[h(x_1) = h(x_2) = \cdots = h(x_k)\right] \le \frac{1}{n^{k-1}}.$$

Moreover, we call $\mathcal{H}$ *strongly k-universal* if for every distinct $x_1, \cdots, x_k \in [m]$ and $y_1, \cdots, y_k \in [n]$, we have

$$\mathbf{Pr}_{h \in \mathcal{H}}\left[\bigwedge h(x_i) = y_i\right] = \frac{1}{n^k}$$

4.2 Balls-into-Bins with 2-Universal Hash Family

We already see in Section 2 that if each ball can be uniformly and independently thrown into a bin, the max load of n bins is $O(\frac{\log}{\log \log n})$ with high probability. Now we assume the independence is not perfect, say we distribute the m balls using a pairwise universal Hash family $\mathcal{H}$. This is equivalent to the following:

- First draw a random function $h \in \mathcal{H}$;
- For each $i \in [m]$, throw the ball i into bin $h(i)$.

Let X_{ij} be the indicator of the event: i-th ball and j-th ball fall into the same bin and let $X = \sum_{1 \le i < j \le m} X_{ij}$ be the total number of collisions. Then

$$\mathbf{E}\left[X\right] = \sum_{1 \le i < j \le m} \mathbf{E}\left[X_{ij}\right] \le \binom{m}{2}\frac{1}{n} < \frac{m^2}{2n}.$$

Assume the max load is Y, which causes $\binom{Y}{2} \le X$ collisions. Then

$$\mathbf{Pr}\left[\binom{Y}{2} \ge \frac{m^2}{n}\right] \le \mathbf{Pr}\left[X \ge \frac{m^2}{n}\right] \le \frac{1}{n}.$$

Therefore, $\mathbf{Pr}\left[Y - 1 \ge m\sqrt{2/n}\right] \le \frac{1}{2}$. The max load is $1 + \sqrt{2n}$ when $m = n$ with probability at least $\frac{1}{2}$. The bound is much worse than the one we obtained using perfect randomness!

4.3 Construction of 2-Universal Family

In this section, we explicitly construct a universal family of Hash functions from $[m]$ to $[n]$.

Let $p \ge m$ be a prime and let

$$h_{a,b}(x) = ((ax + b) \mod p) \mod n.$$

The family is $\mathcal{H} = \{h_{a,b} : 1 \le a \le p - 1, 0 \le b \le p - 1\}$.

In the follow, we verify that $\mathcal{H}$ constructed above is indeed 2-universal, i.e.,

Lemma 5.

$$\mathbf{Pr}_{h_{a,b} \in \mathcal{H}}\left[h_{a,b}(x) = h_{a,b}(y)\right] \le \frac{1}{n}, \quad \textit{for } x \ne y$$

Proof. First, we have if $x \ne y$, then $ax + b \ne ay + b \mod p$. Otherwise, the fact that $ax + b = k_1 p + c$ and $ay + b = k_2 p + c$ where $0 \le k_1, k_2 \le m$ implies $(x - y)a = (k_1 - k_2)p$, which holds only when $x = y$. Moreover, $(a, b) \longrightarrow (ax + b, ay + b)$ is a bijection from $\{1, \cdots, p-1\} \times \{0, \cdots, p-1\}$ to $\{(u, v) : 0 \le u, v \le p-1, u \ne v\}$. This is because $\begin{cases} ax + b = u \mod p \\ ay + b = v \mod p \end{cases}$ has the unique solution $\begin{cases} a = \dfrac{v - u}{y - x} \mod p \\ b = u - ax \mod p \end{cases}$.

Therefore,

$$\mathbf{Pr}_{h_{a,b}(x) \in \mathcal{H}}\left[h_{a,b}(x) = h_{a,b}(y)\right] = \mathbf{Pr}_{(u,v) \in \mathbb{F}_p^2}\left[u = v \mod n\right]$$

The number of (u, v) with $u \ne v$ is $p(p - 1)$. For each u, the number of values of v with $u = v \mod n$ is at most $\lceil \frac{p}{n} \rceil - 1$. The probability is therefore at most

$$\frac{p(\lceil \frac{p}{n} \rceil - 1)}{p(p - 1)} \le \frac{1}{n}.$$

□