Algorithms for Big Data (II)

Chihao Zhang

Shanghai Jiao Tong University

Sept. 27, 2019

Review of Last Lecture

Last time, we met the streaming model.

We studied Morris' algorithm for counting the number of elements in a data stream.

We used Averaging trick and Median trick to boost the quality of Morris' algorithm.

Today we will take a closer look at the mathematical tools needed in the course.

Markov's Inequality

Markov's inequality

For every nonnegative random variable X and every $a \geq 0$, it holds that

$$\mathbf{Pr}\left[X \geq a\right] \leq \frac{\mathbf{E}\left[X\right]}{a}.$$

Proof.

Let $\mathbf{1}_{X \geq a}$ be the indicator random variable such that $\mathbf{1}_{X \geq a}(x) = \begin{cases} 1, & \text{if } x \geq a, \\ 0, & \text{otherwise.} \end{cases}$

Then it holds that $X \geq a \cdot \mathbf{1}_{X \geq a}$. Take the expecation on both sides, we obtain

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

□

Chebyshev's Inequality

Chebyshev's inequality

For every random variable X and every $a \geq 0$, it holds that

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

Proof.

$$\begin{aligned}
\mathbf{Pr}\left[|X - \mathbf{E}[X]| \geq a\right] &= \mathbf{Pr}\left[(X - \mathbf{E}[X])^2 \geq a^2\right] \\
&\leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{a^2} \qquad \text{(Markov's inequality)} \\
&= \frac{\mathbf{Var}[X]}{a^2}.
\end{aligned}$$

□

Chernoff's Bound

Chernoff bound

Let $X_1, \ldots, X_n$ be independent Bernoulli trials with $\mathbf{E}[X_i] = p_i$ for every $i = 1, \ldots, n$. Let $X = \sum_{i=1}^{n} X_i$. Then for every $0 < \varepsilon < 1$, it holds that

$$\mathbf{Pr}\left[|X - \mathbf{E}[X]| > \varepsilon \cdot \mathbf{E}[X]\right] \le 2\exp\left(-\frac{\varepsilon^2 \mathbf{E}[X]}{3}\right).$$

The main tool to prove Chernoff bound is the moment generating function e^{tX} for a random variable X.

It holds that

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

$$= \prod \left(1 - (1 - e^t)p_i\right) \le \prod_{i=1}^{n} e^{-(1-e^t)p_i} = e^{-(1-e^t)\mathbf{E}[X]}.$$

Proof of Chernoff Bound

For every $t > 0$, we have

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

To find an optimal t, we calculate the derivative of above and obtain for $t = \log(1+\varepsilon)$,

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

We can similarly prove that

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

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

Balls-into-Bins

Balls-into-Bins is a simple yet important probabilistic model.

Suppose we throw m ball into n bins uniformly and independently, what is the (expected) maxload of the bins?

When $m = n$, the answer is $\Theta\left(\frac{\log n}{\log\log n}\right)$.

It models an important object, the Hash functions.

Independence

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

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

k-wise Independence

A weaker notion of independence is the k-wise independence.

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

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

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

Examples

Suppose we have n independent bits $X_1, \ldots, X_n \in \{0, 1\}$.

For every $I \in [n]$, define $Y_I = \left(\sum_{j \in I} X_j\right) \mod 2$.

The random bits $\{Y_I\}_{I \subseteq [n]}$ are pairwise independent.

But they are not mutually independent!

Property of Pairwise Independence

Theorem

For pairwise independent $X_1, \ldots, X_n$, 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.

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

□

Hash Functions

In 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.

We will contruct Hash functions with theoretical guarantees.

Universal Hash Function Families

Let $\mathcal{H}$ be a family of functions from $[m]$ to $[n]$ where $m \geq n$.

We call $\mathcal{H}$ k-universal if for every distinct $x_1, \ldots, x_k \in [m]$, we have

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

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

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

Balls-into-Bins with 2-Universal Hash Family

Let X_{ij} be the indicator of the event: i-th ball and j-th ball fall into the same bin.

Let $X = \sum_{1\le i\le j\le m} X_{ij}$ 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 maxload 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 maxload is $1 + \sqrt{2n}$ when $m = n$ with probability at least $1/2$.

Construction of 2-Universal Family

Now we explicitly construct a universal family of Hash functions from $[m]$ to $[n]$.

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

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

The family is

$$\mathcal{H} = \{h_{a,b} : 1 \leq a \leq p-1, 0 \leq b \leq p-1\}.$$

Proof

We show that $\mathcal{H}$ constructed above is indeed 2-universal.

We compute the colliding probability

$$\mathbf{Pr}_{h_{a,b}\in\mathcal{H}}\left[h_{a,b}(x) = h_{a,b}(y)\right]$$

for $x \neq y$.

First, we have if $x \neq y$, then $ax + b \neq ay + b \mod p$.

Moreover $(a, b) \to (ax + b, ay + b)$ is a bijection from $\{1, \dots, p-1\} \times \{0, \dots, p-1\}$ to $\{(u, v) : 0 \le u, v \le p-1, u \neq v\}$.

This is because $\begin{cases} ax + b = u \mod p \\ ay + b = v \mod p \end{cases}$ has a unique solution $\begin{cases} a = \frac{v-u}{y-x} \mod p \\ b = u - ax \mod p. \end{cases}$

Proof (cont'd)

Therefore,

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

The number of (u, v) with $u \neq v$ is $p(p-1)$.

For each u, the number of values of v with $u = v \mod n$ is at most $\lceil p/n \rceil - 1$.

The probabilty is therefore at most

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