Advanced Algorithms (III)

Shanghai Jiao Tong University

Chihao Zhang

March 16th, 2020

Balls-into-Bins

Balls-into-Bins

Throw m balls into n bins uniformly at random

Balls-into-Bins

Throw m balls into n bins uniformly at random

- What is the chance that some bin contains more than one balls? (Birthday paradox)

Balls-into-Bins

Throw m balls into n bins uniformly at random

- What is the chance that some bin contains more than one balls? (Birthday paradox)

- How many balls in the fullest bin? (Max load)

Balls-into-Bins

Throw m balls into n bins uniformly at random

- What is the chance that some bin contains more than one balls? (Birthday paradox)
- How many balls in the fullest bin? (Max load)
- How large is m to hit all bins (Coupon Collector)

Birthday Paradox

Birthday Paradox

In a group of more than 30 people, which very high chances that two of them have the same birthday

Birthday Paradox

In a group of more than 30 people, which very high chances that two of them have the same birthday

$$\begin{aligned}
&\Pr[\text{no same birthday}]\\
&\le 1 \cdot \left(\frac{n-1}{n}\right) \cdot \left(\frac{n-2}{n}\right) \cdots \left(\frac{n-m+1}{n}\right)\\
&= \prod_{i=1}^{m-1} \left(1 - \frac{i}{n}\right) \le \exp\left(-\frac{\sum_{i=1}^{m-1} i}{n}\right) = \exp\left(-\frac{m(m-1)}{2n}\right)
\end{aligned}$$

$$\Pr[\text{no same birthday}] \leq \exp\left(-\frac{m(m-1)}{2n}\right)$$

$$\boxed{\Pr[\text{no same birthday}] \leq \exp\left(-\frac{m(m-1)}{2n}\right)}$$

For $m = 30$, $n = 365$, the probability is less than 0.304

$$\boxed{\Pr[\text{no same birthday}] \leq \exp\left(-\frac{m(m-1)}{2n}\right)}$$

For $m = 30$, $n = 365$, the probability is less than 0.304

For $m = O\left(\sqrt{n}\right)$, the probability can be arbitrarily close to 0.

Max Load

Max Load

Let X_i be the number of balls in the i-th bin

Max Load

Let X_i be the number of balls in the i-th bin

What is $X = \max_{i \in [n]} X_i$? We analyze this when $m = n$

Max Load

Let X_i be the number of balls in the i-th bin

What is $X = \max_{i \in [n]} X_i$? We analyze this when $m = n$

If we can argue that, X_1 is less than k with probability $1 - O\left(\frac{1}{n}\right)$, then by *union bound,*

$\Pr[X \geq k] = O(1)$

Again by union bound, $\Pr[X_1 \geq k] \leq \binom{n}{k} n^{-k} \leq \frac{1}{k!}$

Again by union bound, $\Pr[X_1 \geq k] \leq \binom{n}{k} n^{-k} \leq \frac{1}{k!}$

We apply the Stirling's formula $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$

Again by union bound, $\Pr[X_1 \geq k] \leq \binom{n}{k} n^{-k} \leq \frac{1}{k!}$

We apply the Stirling's formula $k! \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k$

So $\Pr[X \geq k] \leq \frac{1}{k!} \leq \left(\frac{e}{k}\right)^k$

Again by union bound, $\Pr[X_1 \geq k] \leq \binom{n}{k} n^{-k} \leq \frac{1}{k!}$

We apply the Stirling's formula $k! \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k$

So $\Pr[X \geq k] \leq \frac{1}{k!} \leq \left(\frac{e}{k}\right)^k$

We want $\left(\frac{e}{k}\right)^k = O\left(\frac{1}{n}\right)$. Choose $k = O\left(\frac{\log n}{\log \log n}\right)$

Concentration Bounds

Concentration Bounds

We shall develop general tools to obtain “with high probability” results…

Concentration Bounds

We shall develop general tools to obtain “with high probability” results…

These results are critical for analyzing randomized algorithms

Concentration Bounds

We shall develop general tools to obtain "with high probability" results…

These results are critical for analyzing randomized algorithms

This is the main topic in the coming 4-5 weeks

Markov Inequality

Markov Inequality

Markov Inequality

For any *nonnegative* random variable X and $a > 0$,

$$\Pr[X > a] \leq \frac{\mathbf{E}[X]}{a}$$

Markov Inequality

Markov Inequality

For any *nonnegative* random variable X and $a > 0$,

$$\Pr[X > a] \le \frac{\mathbf{E}[X]}{a}$$

Proof.

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}[X \mid X > a] \cdot \Pr[X > a] + \mathbf{E}[X \mid X \le a] \cdot \Pr[X \le a] \\ &\ge a \cdot \Pr[X > a]\end{aligned}$$

Applications

Applications

- A Las-Vegas randomized algorithm with expected $O(n)$ running time terminates in $O(n^2)$ time with probability $1 - O\left(\frac{1}{n}\right)$

Applications

- A Las-Vegas randomized algorithm with expected $O(n)$ running time terminates in $O(n^2)$ time with probability $1 - O\left(\frac{1}{n}\right)$
- In n-balls-into-n-bins problem, $\mathbf{E}[X_i] = 1$. So

$$\Pr\left[X_1 > \frac{\log n}{\log\log n}\right] \le \frac{\log\log n}{\log n}$$

Applications

- A Las-Vegas randomized algorithm with expected $O(n)$ running time terminates in $O(n^2)$ time with probability $1 - O\left(\frac{1}{n}\right)$

- In n-balls-into-n-bins problem, $\mathbf{E}[X_i] = 1$. So

$$\Pr\left[X_1 > \frac{\log n}{\log\log n}\right] \leq \frac{\log\log n}{\log n}$$

This is far from the truth…

Chebyshev's Inequality

Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead of $\mathbf{E}[X]$ for some non-decreasing $f : \mathbb{R} \to \mathbb{R}$

Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead of $\mathbf{E}[X]$ for some non-decreasing $f : \mathbb{R} \to \mathbb{R}$

$$\Pr[X \geq a] = \Pr\left[f(X) \geq f(a)\right] \leq \frac{\mathbf{E}\left[f(X)\right]}{f(a)}$$

Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead of $\mathbf{E}[X]$ for some non-decreasing $f : \mathbb{R} \to \mathbb{R}$

$$\Pr[X \geq a] = \Pr\left[f(X) \geq f(a)\right] \leq \frac{\mathbf{E}\left[f(X)\right]}{f(a)}$$

$f(x) = x^2$ gives the Chebyshev's inequality

Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead of $\mathbf{E}[X]$ for some non-decreasing $f : \mathbb{R} \to \mathbb{R}$

$$\Pr[X \geq a] = \Pr\left[f(X) \geq f(a)\right] \leq \frac{\mathbf{E}\left[f(X)\right]}{f(a)}$$

$f(x) = x^2$ gives the Chebyshev's inequality

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

Coupon Collector

Coupon Collector

Recall the coupon collector problem is to ask

Coupon Collector

Recall the coupon collector problem is to ask

> *"How many ball one needs to throw so that none of the n bins is empty?"*

Coupon Collector

Recall the coupon collector problem is to ask

> *"How many ball one needs to throw so that none of the n bins is empty?"*

We already established that $\mathbf{E}[X] = nH_n \approx n(\log n + \gamma)$

Coupon Collector

Recall the coupon collector problem is to ask

> *"How many ball one needs to throw so that none of the n bins is empty?"*

We already established that $\mathbf{E}[X] = nH_n \approx n(\log n + \gamma)$

The Markov inequality only provides a very weak concentration…

In order to apply Chebyshev's inequality, we need to compute $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

In order to apply Chebyshev's inequality, we need to compute $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

Recall that $X = \sum_{i=0}^{n-1} X_i$ where each X_i follows geometric distribution with parameter $\frac{n-i}{n}$

In order to apply Chebyshev's inequality, we need to compute $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

Recall that $X = \sum_{i=0}^{n-1} X_i$ where each X_i follows geometric distribution with parameter $\dfrac{n-i}{n}$

$X_0, \ldots, X_{n-1}$ are independent, so

In order to apply Chebyshev's inequality, we need to compute $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$

Recall that $X = \sum_{i=0}^{n-1} X_i$ where each X_i follows geometric distribution with parameter $\dfrac{n-i}{n}$

$X_0, \ldots, X_{n-1}$ are independent, so

$$\mathbf{Var}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i]$$

Variance of Geometric Variables

Assume Y follow geometric distribution with parameter p

$$\mathbf{E}[Y^2] = \sum_{i=1}^{\infty} i^2(1-p)^{i-1}p = \frac{2-p}{p^2}$$

$$\mathbf{Var}[Y] = \mathbf{E}[Y^2] - (\mathbf{E}[Y])^2 = \frac{1-p}{p^2}$$

$$\mathbf{Var}[X] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i] = \sum_{i=0}^{n-1} \frac{n \cdot i}{(n-i)^2} \leq n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2}$$

$$= n^2 \left(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} \right) = \frac{\pi^2 n^2}{6}.$$

$$\mathbf{Var}[X] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i] = \sum_{i=0}^{n-1} \frac{n \cdot i}{(n-i)^2} \le n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2}$$

$$= n^2 \left(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} \right) = \frac{\pi^2 n^2}{6}.$$

By Chebyshev's inequality,

$$\Pr[X \ge nH_n + cn] \le \frac{\pi^2}{6c^2}$$

$$\textbf{Var}[X] = \sum_{i=0}^{n-1} \textbf{Var}[X_i] = \sum_{i=0}^{n-1} \frac{n \cdot i}{(n-i)^2} \leq n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2}$$
$$= n^2 \left(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} \right) = \frac{\pi^2 n^2}{6}.$$

By Chebyshev's inequality,

$$\Pr[X \geq nH_n + cn] \leq \frac{\pi^2}{6c^2}$$

The use of Chebyshev's inequality is often referred to as the "second-moment method"

Random Graph

Random Graph

Erdős–Rényi random graph $G(n, p)$

Random Graph

Erdős–Rényi random graph $G(n, p)$

n vertices, each edge appears with probability p independently

Random Graph

Erdős–Rényi random graph $G(n, p)$

n vertices, each edge appears with probability p independently

Given a graph property P, define its *threshold function* $r(n)$ as:

Random Graph

Erdős–Rényi random graph $G(n, p)$

n vertices, each edge appears with probability p independently

Given a graph property P, define its *threshold function* $r(n)$ as:

- if $p \ll r(n)$, $G \sim G(n, p)$ does not satisfy P whp;
- if $p \gg r(n)$, $G \sim G(n, p)$ satisfies P whp.

We will show that the property

$$P = \text{“}G \text{ contains a 4-clique”}$$

has threshold function $n^{-2/3}$

We will show that the property

$$P = \text{“}G \text{ contains a 4-clique”}$$

has threshold function $n^{-2/3}$

For every $S \in \binom{[n]}{4}$, let X_S be the indicator that "$G[S]$ is a clique".

We will show that the property

$$P = \text{“}G \text{ contains a 4-clique”}$$

has threshold function $n^{-2/3}$

For every $S \in \binom{[n]}{4}$, let X_S be the indicator that "$G[S]$ is a clique".

Let $X = \sum_{S \in \binom{[n]}{4}} X_S$, then G satisfies P iff $X > 0$.

Then $\mathbf{E}[X] = \sum_{S \in \binom{[n]}{4}} \mathbf{E}[X_S] \approx \frac{n^4 p^6}{24}$.

Then $\mathbf{E}[X] = \sum_{S \in \binom{[n]}{4}} \mathbf{E}[X_S] \approx \frac{n^4 p^6}{24}$.

If $p \ll n^{-\frac{2}{3}}$, $\mathbf{E}[X] = o(1)$. So by Markov inequality

Then $\mathbf{E}[X] = \sum_{S \in \binom{[n]}{4}} \mathbf{E}[X_S] \approx \frac{n^4 p^6}{24}$.

If $p \ll n^{-\frac{2}{3}}$, $\mathbf{E}[X] = o(1)$. So by Markov inequality

$$\Pr[X \geq 1] \leq \mathbf{E}[X] = o(1)$$

It is not necessary that $\mathbf{E}[X] = \Omega(1)$ implies $\Pr[X > 0] = 1 - o(1)$. (Why?)

It is not necessary that $\mathbf{E}[X] = \Omega(1)$ implies $\Pr[X > 0] = 1 - o(1)$. (Why?)

We require some control over $\mathbf{Var}[X]$

It is not necessary that $\mathbf{E}[X] = \Omega(1)$ implies $\Pr[X > 0] = 1 - o(1)$. (Why?)

We require some control over $\mathbf{Var}[X]$

By Chebyshev's inequality,

It is not necessary that $\mathbf{E}[X] = \Omega(1)$ implies $\Pr[X > 0] = 1 - o(1)$. (Why?)

We require some control over $\mathbf{Var}[X]$

By Chebyshev's inequality,

$$\Pr[X = 0] \le \Pr[\,|X - \mathbf{E}[X]\,| \ge E[X]] \le \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} = \frac{\mathbf{E}[X^2]}{\mathbf{E}[X]^2} - 1$$

It is not necessary that $\mathbf{E}[X] = \Omega(1)$ implies $\Pr[X > 0] = 1 - o(1)$. (Why?)

We require some control over $\mathbf{Var}[X]$

By Chebyshev's inequality,

$$\Pr[X = 0] \leq \Pr[\,|X - \mathbf{E}[X]\,| \geq E[X]] \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} = \frac{\mathbf{E}[X^2]}{\mathbf{E}[X]^2} - 1$$

A sufficient condition is $\mathbf{E}[X^2] = (1 + o(1)) \cdot \mathbf{E}[X]^2$

$$
\begin{aligned}
&\mathbf{E}[X^2] - \mathbf{E}[X]^2 \\
&= \mathbf{E}\Big[\Big(\sum_{S \in \binom{[n]}{4}} X_S\Big)^2\Big] - \Big(\mathbf{E}\Big[\sum_{S \in \binom{[n]}{4}} X_S\Big]\Big)^2 \\
&= \sum_{S,T \in \binom{[n]}{4} : |S \cap T| = 2} \Big(\mathbf{E}[X_S \cdot X_T] - \mathbf{E}[X]\mathbf{E}[X_T]\Big) + \\
&\quad \sum_{S,T \in \binom{[n]}{4} : |S \cap T| = 3} \Big(\mathbf{E}[X_S \cdot X_T] - \mathbf{E}[X_S]\mathbf{E}[X_T]\Big) + \\
&\quad \sum_{S \in \binom{[n]}{4}} \Big(\mathbf{E}[X_S^2] - \mathbf{E}[X_S]^2\Big) \\
&\le n^6 p^{11} + n^5 p^9 + n^4 p^6 = o(\mathbf{E}[X]^2)
\end{aligned}
$$