COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2024.

Lecture 3

Logistics

- Reminder that there is a weekly quiz, released after class today and due next Monday 8pm.
- Problem Set 1 will be released shortly – hopefully by the end of the week. Sorry for the delay.
- See Piazza for a post to organize homework groups.

Summary

Last Time:

- Review of conditional probability, independence, linearity of expectation and variance.
- Polynomial identity testing and proof of the Schwartz-Zippel Lemma.
- Application of linearity of expectation to randomized Quicksort analysis.

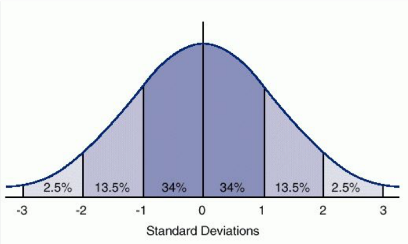
Today:

- Concentation bounds – Markov's and Chebyshev's inequalities.
- The union bound.
- Applications to coupon collecting and statistical estimation.

Concentration Inequalities

Concentration Inequalities

Concentration inequalities are bounds showing that a random variable lies close to it's expectation with good probability. Key tools in the analysis of randomized algorithms.



Markov's Inequality

The most fundamental concentration bound: **Markov's inequality.**

For any non-negative random variable X and any $t > 0$:

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof:

$$\begin{aligned} \mathbb{E}[X] = \sum_{s} \Pr(X = u) \cdot u &\geq \sum_{u \geq t} \Pr(X = u) \cdot u \\ &\geq \sum_{u \geq t} \Pr(X = u) \cdot t \\ &= t \cdot \Pr(X \geq t). \end{aligned}$$

Plugging in $t = \mathbb{E}[X] \cdot s$, $\Pr[X \geq s \cdot \mathbb{E}[X]] \leq 1/s$. The larger the deviation s, the smaller the probability.

Think-Pair-Share: You have a Las Vegas algorithm that solves some decision problem in expected running time T. Show how to turn this into a Monte-Carlo algorithm with worst case running time $3T$ and success probability $2/3$.

Chebyshev's inequality

With a very simple twist, Markov's Inequality can be made much more powerful in many settings.

For any random variable X and any value $t > 0$:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2).$$

X^2 is a nonnegative random variable. So can apply Markov's:

$$\Pr(|X| \geq t) = \Pr(X^2 \geq t^2) \leq \frac{\mathbb{E}[X^2]}{t^2}.$$

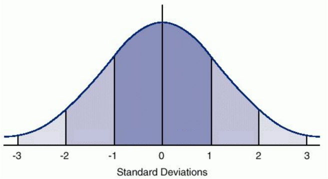
Plugging in the random variable $X - \mathbb{E}[X]$, gives the standard form of **Chebyshev's inequality:**

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2}{t^2} = \frac{\mathrm{Var}(X)}{t^2}.$$

Chebyshev's inequality

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}$$

What is the probability that X falls s standard deviations from it's mean?



$$\Pr(|X - \mathbb{E}[X]| \geq s \cdot \sqrt{\mathrm{Var}[X]}) \leq \frac{\mathrm{Var}[X]}{s^2 \cdot \mathrm{Var}[X]} = \frac{1}{s^2}.$$

Application 2: Statistical Estimation + Law of Large Numbers

Concentration of Sample Mean

Theorem: Let $X_1, \ldots, X_n$ be pairwise independent random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_n$ be their sample average.

For any $\epsilon > 0$, $\Pr[|\bar{X} - \mu| \geq \epsilon\sigma] \leq \frac{1}{n\epsilon^2}$.

- By linearity of expectation, $\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu$.
- By linearity of variance, $\mathbb{E}[\bar{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{\sigma^2}{n}$.
- Plugging into Chebyshev's inequality:

$$\Pr[|\bar{X} - \mu| \geq \epsilon\sigma] \leq \frac{\mathrm{Var}[\bar{X}]}{\epsilon^2\sigma^2} = \frac{1}{n\epsilon^2}.$$

This is the weak law of large numbers.

Concentration of Sample Mean

Application to statistical estimation: There is a large population of individuals. A p fraction of them have a certain property (e.g., 55% of people support decreased taxation, 10% of people are greater than 6' tall, etc.). Want to estimate p from a small sample of individuals.

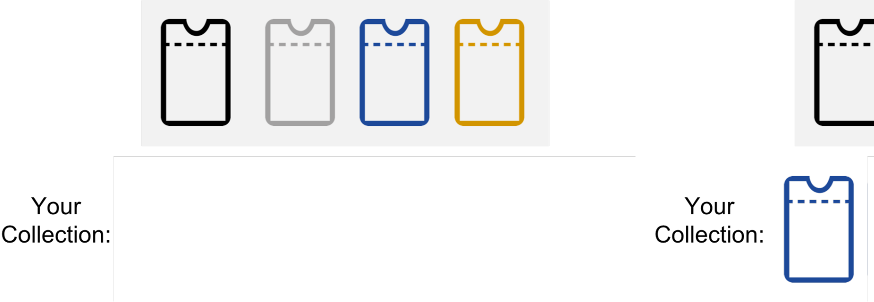
- Sample n individuals uniformly at random, with replacement.
- Let $X_i = 1$ if the i^{th} individual has the property, and 0 otherwise. $X_1, \ldots, X_n$ are i.i.d. draws from $Bern(p)$ – each is 1 with probability p and 0 with probability $1 - p$.

Application to Success Boosting

Think-Pair-Share: You have a Monte-Carlo algorithm with worst case running time T and success probability $2/3$. Show how to obtain, for any $\delta \in (0, 1)$, a Monte-Carlo algorithm with worse case running time $O(T/\delta)$ and success probability $1 - \delta$.

Application 3: Coupon Collecting

There is a set of n unique coupons. At each step you draw a random coupon from this set. How many steps does it take you to collect all the coupons?



Think-Pair-Share: Say you have collected i coupons so far. Let T_{i+1} denote the number of draws needed to collect the $(i+1)^{st}$ coupon. What is $\mathbb{E}[T_i]$?

Coupon Collector Analysis

Think-Pair-Share: Say you have collected i coupons so far. Let T_{i+1} denote the number of draws needed to collect the $(i+1)^{st}$ coupon. What is $\mathbb{E}[\mathrm{T}_i]$?

- T_i is a geometric random variable with success probability $p_i = \frac{n-i}{n}$. I.e., $Pr[\mathrm{T}_i = j] = p_i(1-p_i)^{j-1}$.
- **Exercise:** verify that $\mathbb{E}[\mathrm{T}_i] = 1/p_i = \frac{n}{n-i}$.
- By linearity of expectation, the expected number of draws to collect all the coupons is:

$$\mathbb{E}[\mathrm{T}] = \sum_{i=0}^{n-1} \mathbb{E}[\mathrm{T}_i] = \frac{n}{n} + \frac{n}{n-1} + \ldots \frac{n}{2} + \ldots \frac{n}{1}$$

$$= n \cdot H_n.$$

- By Markov's inequality, $\Pr[\mathrm{T} \geq cn \cdot H_n] \leq$

Coupon Collector Analysis

Can get a tighter tail bound using Chebyshev's inequality in place of Markov's.

- We wrote $\mathsf{T} = \sum_{i=0}^{n-1} \mathsf{T}_i$, which let us compute $\mathbb{E}[\mathsf{T}] = n \cdot H_n$.
- Also have $\mathrm{Var}[\mathsf{T}] = \sum_{i=0}^{n-1} \mathrm{Var}[\mathsf{T}_i]$. Why?
- **Exercise:** show that $\mathrm{Var}[\mathsf{T}_i] = \frac{1-p_i}{p_i^2}$, and recall that $p_i = \frac{n-i}{n}$.
- Putting these together:

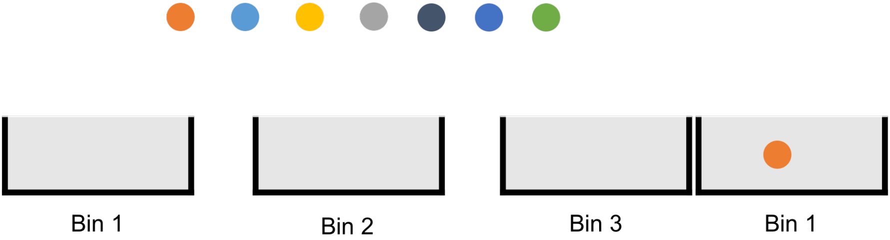
$$\mathrm{Var}[\mathsf{T}] = \sum_{i=0}^{n} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n} \frac{1}{p_i^2} - \sum_{i=0}^{n} \frac{1}{p_i}$$
$$\leq n^2 \cdot \frac{\pi^2}{6} - n \cdot H_n \leq n^2 \cdot \frac{\pi^2}{6}.$$

- Via Chebyshev's inequality, $\Pr[|\mathsf{T} - n \cdot H_n| \geq cn] \leq$

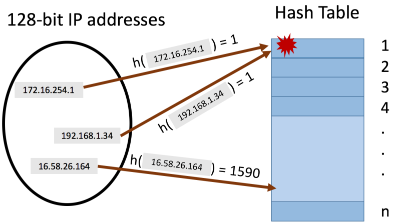
Application 4: Randomized Load Balancing and Hashing, and ‘Ball Into Bins’

Balls Into Bins

I throw m balls independently and uniformly at random into n bins. What is the maximum number of balls any bin?



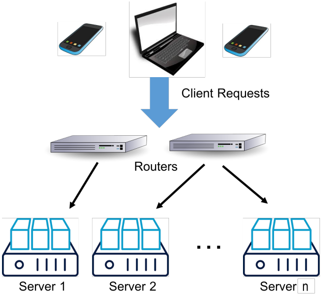
Application: Hash Tables



- **hash function** $h : U \to [n]$ maps elements to indices of an array.
- Repeated elements in the same bucket are stored as a linked list – ‘chaining’.
- Worse-case look up time is proportional to the maximum list length – i.e., the maximum number of ‘balls’ in a ‘bin’.

Note: A ‘fully random hash function’ maps items independently and uniformly at random to buckets. This is a theoretical idealization of practical hash functions.

Application: Randomized Load Balancing



- m requests are distributed randomly to n servers. Want to bound the maximum number of requests that a single server must handle.
- Assignment is often is done via a random hash function so that repeated requests or related requests can be mapped to the same server, to take advantages of caching and other optimizations.

Balls Into Bins Analysis

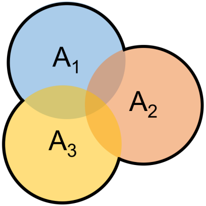
Let $\mathbf{b}_i$ be the number of balls landing in bin i. For n balls into m bins what is $\mathbb{E}[\mathbf{b}_i]$?

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq k\right] = \Pr\left[\bigcup_{i=1}^{n} A_i\right],$$

where A_i is the event that $\mathbf{b}_i \geq k$.

Union Bound: For any random events $A_1, A_2, ..., A_n$,

$$\Pr\left(A_1 \cup A_2 \cup \ldots \cup A_n\right) \leq \Pr(A_1) + \Pr(A_2) + \ldots + Pr(A_n).$$



Exercise: Show that the union bound is a special case of Markov's inequality with indicator random variables.

Balls Into Bins Direct Analysis

Let $\mathbf{b}_i$ be the number of balls landing in bin i. If we can prove that for any i, $\Pr[A_i] = \Pr[\mathbf{b}_i \geq k] \leq p$, then by the union bound:

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq k\right] = \Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq n \cdot p.$$

Claim 1: Assume $m = n$. For $k \geq \frac{c \ln n}{\ln \ln n}$, $\Pr[\mathbf{b}_i \geq k] \leq \frac{1}{n^{c-o(1)}}$.

- $\mathbf{b}_i$ is a binomial random variable with n draws and success probability $1/n$.
$$\Pr[\mathbf{b}_i = j] = \binom{n}{j} \cdot \frac{1}{n^j} \cdot \left(1 - \frac{1}{n}\right)^{n-j}.$$
- We have $\binom{n}{j} \leq \left(\frac{en}{j}\right)^j$, giving $\Pr[\mathbf{b}_i = j] \leq \left(\frac{e}{j}\right)^j \cdot (1 - \frac{1}{n})^{n-j} \leq \left(\frac{e}{j}\right)^j$.
- Summing over $j \geq k$ we have:

$$\Pr[\mathbf{b}_i \geq k] \leq \sum_{j \geq k} \left(\frac{e}{j}\right)^j \leq \left(\frac{e}{k}\right)^k \cdot \frac{1}{1 - e/k}.$$

Balls Into Bins Direct Analysis

We just showed: When $n = m$ (i.e., n balls into n bins)

$$\Pr[\mathbf{b}_i \geq k] \leq \left(\frac{e}{k}\right)^k \cdot \frac{1}{1 - e/k}$$

For $k = \frac{c \ln n}{\ln \ln n}$ we have:

$$\Pr[\mathbf{b}_i \geq k] \leq \left(\frac{\ln \ln n}{\ln n}\right)^{\frac{c \ln n}{\ln \ln n}} \cdot \frac{1}{1 - (e \ln \ln n)/(c \ln n)} = \frac{1}{n^{c - o(1)}}.$$

Upshot: By the union bound, For $k = c\frac{\ln n}{\ln \ln n}$ for sufficiently large c,

$$\Pr\left[\max_{i=1,\ldots,n} \mathbf{b}_i \geq k\right] \leq n \cdot \frac{1}{n^{c-o(1)}} = \frac{1}{n^{c-1-o(1)}}.$$

When throwing n balls in to n bins, with very high probability the maximum number of balls in a bin will be $O\left(\frac{\ln n}{\ln \ln n}\right)$.