COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco
University of Massachusetts Amherst. Spring 2024.
Lecture 3
• 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:

- Concentration bounds – Markov’s and Chebyshev’s inequalities.
- The union bound.
- Applications to coupon collecting and statistical estimation.
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:

• Concentration bounds – Markov’s and Chebyshev’s inequalities.
• The union bound.
• Applications to coupon collecting and statistical estimation.
Concentration Inequalities
Concentration inequalities are bounds showing that a random variable lies close to its expectation with good probability. Key tools in the analysis of randomized algorithms.
The most fundamental concentration bound: **Markov’s inequality.**
Markov’s Inequality

The most fundamental concentration bound: Markov’s inequality.

For any non-negative random variable \(X\) and any \(t > 0\):

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

$$\mathbb{E}[X] = \sum_{u \in \mathcal{X}} \Pr(X = u) \cdot u$$
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:

$$\mathbb{E}[X] = \sum_{u} \Pr(X = u) \cdot u \geq \sum_{u \geq t} \Pr(X = u) \cdot u$$
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:

$$
\mathbb{E}[X] = \sum_s \Pr(X = u) \cdot u \geq \sum_{u \geq t} \Pr(X = u) \cdot u
$$

$$
\geq \sum_{u \geq t} \Pr(X = u) \cdot t
$$

$$
= \Pr[X \geq t] \cdot \sum_{u \geq t} \Pr(X = u)
$$

$$
= \Pr[X \geq t] \cdot \Pr(X \geq t)
$$
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:

$$\mathbb{E}[X] = \sum_s \Pr(X = u) \cdot u \geq \sum_{u \geq t} \Pr(X = u) \cdot u$$

$$\geq \sum_{u \geq t} \Pr(X = u) \cdot t$$

$$= t \cdot \Pr(X \geq t).$$
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:

$$\mathbb{E}[X] = \sum_s \Pr(X = u) \cdot u \geq \sum_{u \geq t} \Pr(X = u) \cdot u \geq \sum_{u \geq t} \Pr(X = u) \cdot t = t \cdot \Pr(X \geq t).$$

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.
Markov’s Inequality

\[ \text{BPP} \subseteq \text{BPP} \]

**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 \( \frac{2}{3} \).

\[
\Pr (\text{runtime} \geq 3 \cdot \mathbb{E}[\text{runtime}]) \leq \frac{1}{3}
\]

\[
\Pr (\text{runtime} \geq 3T) \leq \frac{1}{3}
\]

After \( 3T \) steps: terminate and guess \( \perp \) succeeds at least \( 2/3 \) of the time.
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 $t > 0$:

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

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

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\text{Var}(X)}{t^2}.$$
Chebyshev’s inequality

\[ \Pr(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}[X]}{t^2} \]
Chebyshev’s inequality

Pr(|X − μ| ≥ t) ≤ \frac{Var[X]}{t^2}

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

Pr(|X − μ| ≥ s \cdot \sqrt{Var[X]}) ≤ \frac{Var[X]}{s^2 \cdot Var[X]} = \frac{1}{s^2}. 

s. d.
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 $\text{Var}[X_i] = \sigma^2$. Let $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be their sample average.

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

$n \to \infty \quad \Pr[\neg J] \to 0$
Concentration of Sample Mean

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

$$
\text{Var} (\bar{X}) = \text{Var} \left( \frac{1}{n} \sum X_i \right) \\
= \frac{1}{n^2} \sum \text{Var} (X_i) \\
= \frac{1}{n} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}
$$
Concentration of Sample Mean

**Theorem:** Let \( X_1, \ldots, X_n \) be pairwise independent random variables with \( \mathbb{E}[X_i] = \mu \) and \( \text{Var}[X_i] = \sigma^2 \). Let \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) 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, \( \mathbb{E}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{\sigma^2}{n} \).
- Plugging into Chebyshev’s inequality:

\[
\Pr[|\bar{X} - \mu| \geq \epsilon \sigma] \leq \frac{\text{Var}[\bar{X}]}{\epsilon^2 \sigma^2} = \frac{1}{n\epsilon^2}.
\]
Theorem: Let $X_1, \ldots, X_n$ be pairwise independent random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$. 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, $\mathbb{E}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{\sigma^2}{n}$.
- Plugging into Chebyshev’s inequality:

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

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.
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.
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.
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 $\text{Bern}(p)$ – each is 1 with probability $p$ and 0 with probability $1 - p$. 
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 $1 - p$.

• $\mathbb{E}[X_i] = p$ and $\text{Var}[X_i] = p(1 - p)$.

• Thus, letting $\bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\mathbb{E}[\bar{p}] = p$ and $\text{Var}[\bar{p}] = \frac{p(1-p)}{n} \leq \frac{p}{n}$.

By Chebyshev’s inequality $\Pr[|\bar{p} - p| \geq \epsilon] \leq \frac{p}{\epsilon^2 n}$.

Upshot: If we take $n = \frac{p}{\epsilon^2 \delta}$ samples, then with probability at least $1 - \delta$, $\bar{p}$ will be a $\pm \epsilon$ estimate to the true proportion $p$. A prototypical sublinear time algorithm.
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 $1 - p$.
- $\mathbb{E}[X_i] = p$ and $\text{Var}[X_i] = p(1 - p)$.
- Thus, letting $\bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\mathbb{E}[\bar{p}] = p$ and $\text{Var}[\bar{p}] = \frac{p(1-p)}{n} \leq \frac{p}{n}$.
- By Chebyshev’s inequality $\Pr[|p - \bar{p}| \geq \epsilon] \leq \frac{p}{\epsilon^2 n}$. 

Upshot: If we take $n = \frac{p}{\epsilon^2 \delta}$ samples, then with probability at least $1 - \delta$, $\bar{p}$ will be a $\pm \epsilon$ estimate to the true proportion $p$. A prototypical sublinear time algorithm.
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 \( \text{Bern}(p) \) – each is 1 with probability \( p \) and 0 with probability \( 1 - p \).
- \( \mathbb{E}[X_i] = p \) and \( \text{Var}[X_i] = p(1 - p) \).
- Thus, letting \( \bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i \), \( \mathbb{E}[ar{p}] = p \) and \( \text{Var}[\bar{p}] = \frac{p(1-p)}{n} \leq \frac{p}{n} \).
- By Chebyshev’s inequality \( \Pr[|p - \bar{p}| \geq \epsilon] \leq \frac{p}{\epsilon^2 n} \).

**Upshot:** If we take \( n = \frac{p}{\epsilon^2 \delta} \) samples, then with probability at least \( 1 - \delta \), \( \bar{p} \) will be a \( \pm \epsilon \) estimate to the true proportion \( p \). A prototypical sublinear time algorithm.
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 $1 - \delta$. 

\[
\begin{align*}
\text{run n times} & \quad n = O(1/\delta) \\
\text{return majority} & \\
\downarrow & \\
X_1, \ldots, X_n & = 1 \text{ if correct} \\
& = 0 \text{ if incorrect} \\
\text{Ex: } P &= 2/3 \\
p & > \frac{1}{2} \\
|p - \bar{p}| & < \frac{1}{6} \\
\frac{p}{n} & \leq \delta \\
\frac{1}{n} \sum_{i=1}^{n} x_i & > \frac{1}{2} \\
p & > \frac{1}{2} \\
\frac{1}{n} & \sum_{i=1}^{n} x_i \\
\text{majority} & > \frac{1}{2} \\
\text{BPP} & < 2/3
\end{align*}
\]
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?
Coupon Collector Problem

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 $E[T_i]$?
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?
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?
Coupon Collector Problem

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?

Your Collection:
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?
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 \( \mathbb{E}[T_i] \)?

\[
\mathbb{E}[T_i] = \frac{\Omega}{n-i}
\]
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 $\mathbb{E}[T_i]$?
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 $\mathbb{E}[T_i]$?

- $T_i$ is a \textit{geometric random variable} with success probability $p_i = \frac{n-i}{n}$. I.e., $Pr[T_i = j] = p_i(1 - p_i)^{j-1}$.
- \textbf{Exercise:} verify that $\mathbb{E}[T_i] = 1/p_i = \frac{n}{n-i}$.

\[
\mathbb{E}[T_i] = p_i \cdot 1 + (1-p_i) \cdot (\mathbb{E}[T_i] + 1) \\
\mathbb{E}[T_i] = p_i + 1 - p_i + (1-p_i) \cdot \mathbb{E}[T_i] \\
p_i \cdot \mathbb{E}[T_i] = 1 \\
\mathbb{E}[T_i] = \frac{1}{p_i}
\]
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 $\mathbb{E}[T_i]$?

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

$$\mathbb{E}[T] = \sum_{i=0}^{n-1} \mathbb{E}[T_i]$$
**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 \( \mathbb{E}[T_i] \)?

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

\[
\mathbb{E}[T] = \sum_{i=0}^{n-1} \mathbb{E}[T_i] = \frac{n}{n} + \frac{n}{n-1} + \cdots \frac{n}{2} + \cdots \frac{n}{1}
\]

\[
= n \cdot \left( \frac{1}{n} + \frac{1}{n-1} + \cdots + 1 \right)
\]

\[
= n \cdot \left( \frac{1}{n} + \frac{1}{n-1} + \cdots + 1 \right)
\]
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 $\mathbb{E}[T_i]$?

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

$$\mathbb{E}[T] = \sum_{i=0}^{n-1} \mathbb{E}[T_i] = \frac{n}{n} + \frac{n}{n-1} + \cdots \frac{n}{2} + \cdots \frac{n}{1} = n \cdot H_n = \mathcal{O}(n \log n)$$
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 $\mathbb{E}[T_i]$?

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

$$
\mathbb{E}[T] = \sum_{i=0}^{n-1} \mathbb{E}[T_i] = \frac{n}{n} + \frac{n}{n-1} + \cdots \frac{n}{2} + \cdots \frac{n}{1} = n \cdot H_n.
$$

- By Markov's inequality, $\Pr[T \geq cn \cdot H_n] \leq \frac{1}{c}$. 


Coupon Collector Analysis

Can get a tighter tail bound using Chebyshev’s inequality in place of Markov’s.
Coupon Collector Analysis

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

- We wrote $T = \sum_{i=0}^{n-1} T_i$, which let us compute $E[T] = n \cdot H_n$.
- Also have $\text{Var}[T] = \sum_{i=0}^{n-1} \text{Var}[T_i]$. Why?
Can get a tighter tail bound using Chebyshev’s inequality in place of Markov’s.

- We wrote $T = \sum_{i=0}^{n-1} T_i$, which let us compute $\mathbb{E}[T] = n \cdot H_n$.
- Also have $\text{Var}[T] = \sum_{i=0}^{n-1} \text{Var}[T_i]$. Why?
- **Exercise:** show that $\text{Var}[T_i] = \frac{1-p_i}{p_i^2}$, and recall that $p_i = \frac{n-i}{n}$.
Coupon Collector Analysis

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

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

$$
\text{Var}[T] = \sum_{i=0}^{n} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n} \frac{1}{p_i^2} - \sum_{i=0}^{n} \frac{1}{p_i}
$$
Can get a tighter tail bound using Chebyshev’s inequality in place of Markov’s.

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

$$\text{Var}[T] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{1}{p_i^2} - \sum_{i=0}^{n-1} \frac{1}{p_i}$$

$$\sum_{x=0}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}$$

$$\sum_{i=0}^{n-1} \frac{1}{p_i^2} \leq n^2 \cdot \frac{\pi^2}{6} - n \cdot H_n$$

$$\sum_{i=0}^{n-1} \frac{1}{(n-i)^2} = n^2 \sum_{i=1}^{\infty} \frac{1}{i^2} \leq \pi^2$$
Coupon Collector Analysis

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

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

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

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

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

$$\text{Var}[T] = \sum_{i=0}^{n} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n} \frac{1}{p_i^2} - \sum_{i=0}^{n} \frac{1}{p_i}$$

$$\mathbb{E}[T] \leq n^2 \cdot \frac{\pi^2}{6} - n \cdot H_n \leq n^2 \cdot \frac{\pi^2}{6}.$$ 

• Via Chebyshev’s inequality, $\Pr[|T - n \cdot H_n| \geq cn] \leq \frac{\mathbb{E}[T]}{cn^2}$.
Application 4: Randomized Load Balancing and Hashing, and ‘Ball Into Bins’
I throw $m$ balls independently and uniformly at random into $n$ bins. What is the maximum number of balls any bin?
I throw $m$ balls independently and uniformly at random into $n$ bins. What is the maximum number of balls any bin?
I throw $m$ balls independently and uniformly at random into $n$ bins. What is the maximum number of balls any bin?
I throw $m$ balls independently and uniformly at random into $n$ bins. What is the maximum number of balls any bin?
I throw $m$ balls independently and uniformly at random into $n$ bins. What is the maximum number of balls any bin?
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** \( h : U \rightarrow [n] \) 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’.
Application: Hash Tables

• **hash function** $h : U \rightarrow [n]$ 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 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 $b_i$ be the number of balls landing in bin $i$. For $n$ balls into $m$ bins what is $\mathbb{E}[b_i]$? 

$$= \frac{n}{m}$$
Balls Into Bins Analysis

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

\[
\Pr \left[ \max_{i=1,\ldots,m} b_i \geq k \right] = \Pr \left[ \bigcup_{i=1}^{m} A_i \right],
\]

where $A_i$ is the event that $b_i \geq k$.

\[
\begin{array}{cccc}
\hline
0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & \emptyset \\
\hline
0 & 0 & \emptyset & \emptyset \\
\hline
\end{array}
\]
Let $b_i$ be the number of balls landing in bin $i$. For $n$ balls into $m$ bins what is $\mathbb{E}[b_i]$?

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

where $A_i$ is the event that $b_i \geq k$.

**Union Bound:** For any random events $A_1, A_2, \ldots, A_n$,

$$\Pr (A_1 \cup A_2 \cup \ldots \cup A_n) \leq \Pr(A_1) + \Pr(A_2) + \ldots + \Pr(A_n).$$
Balls Into Bins Analysis

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

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

where $A_i$ is the event that $b_i \geq k$.

**Union Bound:** For any random events $A_1, A_2, \ldots, A_n$,

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

**Exercise:** Show that the union bound is a special case of Markov’s inequality with indicator random variables.
Balls Into Bins Direct Analysis

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

$$\Pr \left[ \max_{i=1,\ldots,n} b_i \geq k \right] = \Pr \left[ \bigcup_{i=1}^{n} A_i \right] \leq n \cdot p.$$
Balls Into Bins Direct Analysis

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

$$\Pr\left[\max_{i=1,\ldots,n} b_i \geq k\right] = \Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq n \cdot p. \leq \frac{1}{n^2}$$

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

$$K = \frac{3 \ln n}{\ln \ln n} \quad \Pr[b_i \geq K] \leq \frac{1}{n^3}$$
Let $b_i$ be the number of balls landing in bin $i$. If we can prove that for any $i$, $Pr[A_i] = Pr[b_i \geq k] \leq p$, then by the union bound:

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

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

- $b_i$ is a **binomial random variable** with $n$ draws and success probability $1/n$.

  $$Pr[b_i = j] = \binom{n}{j} \cdot \frac{1}{n^j} \cdot \left(1 - \frac{1}{n}\right)^{n-j}.$$
Let $b_i$ be the number of balls landing in bin $i$. If we can prove that for any $i$, $\Pr[A_i] = \Pr[b_i \geq k] \leq p$, then by the union bound:

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

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

- $b_i$ is a binomial random variable with $n$ draws and success probability $1/n$.

  $$\Pr[b_i = j] = \binom{n}{j} \cdot \frac{1}{n^j} \cdot \left(1 - \frac{1}{n}\right)^{n-j} \leq 1$$

- We have $\binom{n}{j} \leq \left(\frac{en}{j}\right)^{j}$, giving $\Pr[b_i = j] \leq \left(\frac{e}{j}\right)^{j} \cdot \left(1 - \frac{1}{n}\right)^{n-j} \leq \left(\frac{e}{j}\right)^{j}$.
Let $b_i$ be the number of balls landing in bin $i$. If we can prove that for any $i$, $\Pr[A_i] = \Pr[b_i \geq k] \leq p$, then by the union bound:

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

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

- $b_i$ is a binomial random variable with $n$ draws and success probability $1/n$.

$$\Pr[b_i = j] = \binom{n}{j} \cdot \frac{1}{n^j} \cdot \left(1 - \frac{1}{n}\right)^{n-j}.$$  

- We have $\binom{n}{j} \leq \left(\frac{en}{j}\right)^j$, giving $\Pr[b_i = j] \leq \left(\frac{e}{j}\right)^j \cdot \left(1 - \frac{1}{n}\right)^{n-j} \leq \left(\frac{e}{j}\right)^j$.

- Summing over $j \geq k$ we have:

$$\Pr[b_i \geq k] \leq \sum_{j \geq k} \left(\frac{e}{j}\right)^j \leq \left(\frac{e}{k}\right)^k \cdot \frac{1}{1 - e/k}.$$
We just showed: When $n = m$ (i.e., $n$ balls into $n$ bins)

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

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

$$\Pr [b_i \geq k] \leq \left( \frac{\ln \ln n}{\ln n} \right)^{\frac{c \ln n}{\ln \ln n}} \cdot \frac{1}{1 - (e \ln \ln n)/(c \ln n)}$$
We just showed: When $n = m$ (i.e., $n$ balls into $n$ bins)

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

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

$$Pr[b_i \geq k] \leq \left(\frac{\ln \ln n}{\ln n}\right)^{\frac{c \ln n}{\ln \ln n}} \cdot \frac{1}{1 - (e \ln \ln n)/(c \ln n)} = \frac{1}{n^{c-o(1)}}.$$
We just showed: When \( n = m \) (i.e., \( n \) balls into \( n \) bins)

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

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

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

Upshot: By the union bound, for sufficiently large \( c \),

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

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