COMPSCI 614: Randomized Algorithms with Applications to Data Science

Prof. Cameron Musco
University of Massachusetts Amherst. Spring 2024
Lecture 5

Logistics

- Problem Set 2 is due next Wednesday 2/21 at 11:59pm.
- Next week we do not have class on Thursday, so I will move my office hours to **Tuesday at 11:30am**.

Summary

Last Time:

- Practice questions on applications of linearity of expectation and variance from quiz.
- Balls-into-bins analysis showing max load of $O(\sqrt{n})$ with Chebyshev's inequality.
- Start on exponential concentration bounds for sums of bounded independent random variables.

Today:

- Finish up exponential concentration bounds.
- Applications to balls-into-bins and linear probing analysis.
- Maybe start on hashing/finger printing?

Exponential Concentration Bounds

The Chernoff Bound

Chernoff Bound (simplified version): Consider independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$ and let $X = \sum_{i=1}^n X_i$. Let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^n X_i]$. For any $\delta \geq 0$

$$\Pr\left(X \geq (1+\delta)\mu\right) \leq \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}}$$

Chernoff Bound (alternate version): Consider independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$ and let $X = \sum_{i=1}^n X_i$. Let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^n X_i]$. For any $\delta \geq 0$

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right).$$

As δ gets larger and larger, the bound falls off exponentially fast.

Balls Into Bins Via Chernoff Bound

Recall that $\mathbf{b}_i$ is the number of balls landing in bin i, when we randomly throw n balls into n bins.

- $\mathbf{b}_i = \sum_{i=1}^{n} \mathbf{I}_{i,j}$ where $\mathbf{I}_{i,j} = 1$ with probability $1/n$ and 0 otherwise. $\mathbf{I}_{i,1}, \ldots \mathbf{I}_{i,n}$ are independent.
- Apply Chernoff bound with $\mu = \mathbb{E}[\mathbf{b}_i] = 1$:

$$\Pr[\mathbf{b}_i \geq k] \leq \frac{e^k}{(1+k)^{(1+k)}}.$$

- For $k \geq \frac{c \log n}{\log \log n}$ we have:

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

Upshot: We recover the right bound for balls into bins.

Bernstein Inequality

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ each with magnitude bounded by $M1$ and let $X = \sum_{i=1}^n X_i$. Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i]$. For any $t \geq 0 s \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq s\sigma\right) \leq 2\exp\left(-\frac{s^2}{4}\right).$$

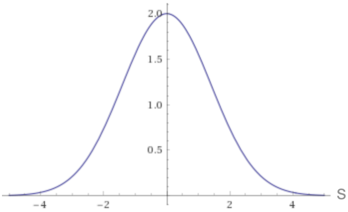
Assume that $M = 1$ and plug in $t = s \cdot \sigma$ for $s \leq \sigma$.

Compare to Chebyshev's: $\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq s\sigma\right) \leq \frac{1}{s^2}$.

- An exponentially stronger dependence on s!

Interpretation as a Central Limit Theorem

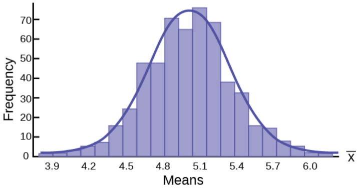
Simplified Bernstein: Probability of a sum of independent, bounded random variables lying $\geq s$ standard deviations from its mean is $\approx \exp\left(-\frac{s^2}{4}\right)$. Can plot this bound for different s:



- Looks like a Gaussian (normal) distribution – can think of Bernstein's inequality as giving a quantitative version of the central limit theorem.
- The distribution of the sum of bounded independent random variables can be upper bounded with a Gaussian distribution.

Central Limit Theorem

Stronger Central Limit Theorem: The distribution of the sum of *n bounded* independent random variables converges to a Gaussian (normal) distribution as *n* goes to infinity.



- The Gaussian distribution is so important since many random variables can be approximated as the sum of a large number of small and roughly independent random effects. Thus, their distribution looks Gaussian by CLT.

Sampling for Approximation

I have an $n \times n$ matrix with entries in $[0, 1]$. I want to estimate the sum of entries. I sample s entries uniformly at random with replacement, take their sum, and multiply it by n^2/s. How large must s be so that this method returns the correct answer, up to error $\pm\epsilon \cdot n^2$ with probability at least $1 - 1/n$?

(a) $O(n^2)$ (b) $O(n/\epsilon)$ (c) $O(\log n/\epsilon)$ (d) $O(\log n/\epsilon^2)$

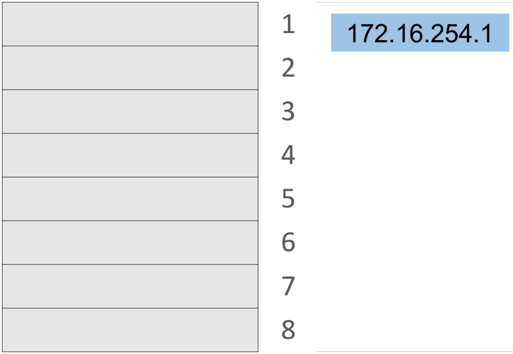
Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ each with magnitude bounded by $M1$ and let $X = \sum_{i=1}^n X_i$. Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Application: Linear Probing

Linear Probing

Linear probing is the simplest form of open addressing for hash tables. If an item is hashed into a full bucket, keep trying buckets until you find an empty one.

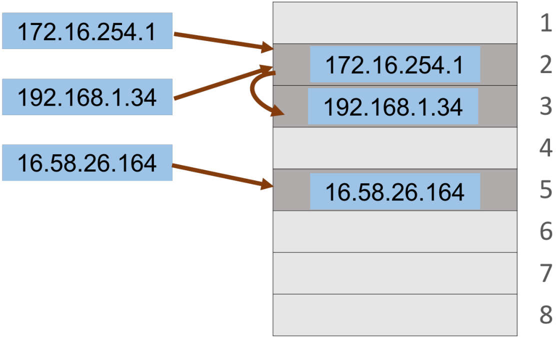


Simple and potentially very efficient – but performance can degrade as the hash table fills up.

Linear Probing Expected Runtime

Theorem: If the hash table has n inserted items and $m \geq 2n$ buckets, then linear probing requires $O(1)$ expected time per insertion/query.

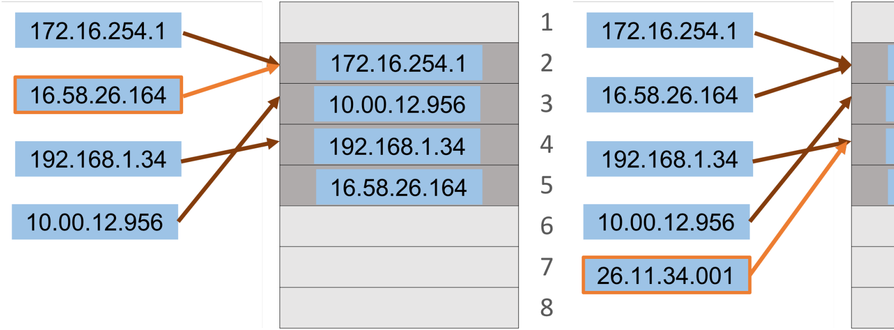
Definition: For any interval $I \subset [m]$, let $\mathbf{L}(I) = |\{x : \mathbf{h}(x) \in I\}|$ be the number of items hashed to the interval. We say I is full if $\mathbf{L}(I) \geq |I|$.



Which intervals in this table are full?

Analysis via Full Intervals

Claim Let $\mathsf{T}(x)$ denote the number of steps required for an insertion/query operation for item x. If $\mathsf{T}(x) > k$, there are at least k full intervals of different lengths containing $\mathsf{h}(x)$.



Let $\mathbf{I}_j = 1$ if $\mathsf{h}(x)$ lies in some length-j full interval, $\mathbf{I}_j = 0$ otherwise. Operation time for x is can be bounded as $\mathsf{T}(x) \leq \sum_{j=1}^{n} \mathbf{I}_j$.

Expectation Analysis

$\mathbf{I}_j = 1$ if $\mathbf{h}(x)$ lies in some length-j full interval, $\mathbf{I}_j = 0$ otherwise. Expected operation time for any x is:

$$\mathbb{E}[\mathbf{T}(x)] \leq \sum_{j=1}^{n} \mathbb{E}[\mathbf{I}_j].$$

Observe that $\mathbf{h}(x)$ lies in at most 1 length-1 interval, 2 length-2 intervals, etc. So we can upper bound this expectation by:

$$\mathbb{E}[\mathbf{T}(x)] \leq \sum_{j=1}^{n} j \cdot \Pr[\text{any length-}j\text{ interval is full}].$$

A length-j interval is full if the number of items hashed into it, $\mathbf{L}(I)$ is at least j. Note that when $m \geq 2n$, $\mathbb{E}[\mathbf{L}(I)] = j/2$. Applying a Chernoff bound with $\delta = 1/2$, $\mu = \mathbb{E}[\mathbf{L}(I)] = j/2$:

$$\Pr[\mathbf{L}(I) \geq j] \leq \Pr[|\mathbf{L}(I) - \mu| \geq \delta \cdot \mu]$$
$$\leq 2e^{-\frac{(1/2)^2 \cdot j/2}{2+1/2}} = 2e^{-c \cdot j}.$$

Finishing the Analysis

Expected operation time for any x is:

$$\begin{aligned}
\mathbb{E}[\mathsf{T}(x)] &\leq \sum_{j=1}^{n} j \cdot \Pr[\text{any length-}j\text{ interval is full}] \\
&\leq \sum_{j=1}^{n} j \cdot 2e^{-c \cdot j} \\
&= O(1).
\end{aligned}$$

This matches the expected operation cost of chaining when $m \geq 2n$.
In practice, linear probing is typically much faster.