COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2023.

Lecture 6

Logistics

- Problem Set 1 is due tomorrow at 11:59pm in Gradescope. Separate submissions for core-competency problems and challenge problems.
- Quiz 3 is due Monday at 8pm.

Last Time

Last Class:

- Higher moment bounds and exponential concentration bounds
- Bernstein inequality

sample converges to true mean.

This Class: law of large numbers

- Connection between exponential concentration bounds and the central limit theorem. — distribution of sample mean converges to a normal dist.
- The Chernoff bound.
- Bloom filters: random hashing to maintain a large set in small space.

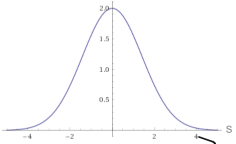
Interpretation as a Central Limit Theorem

Bernstein Inequality (Simplified): Consider independent random variables $X_1, \ldots, X_n$ falling in [-1,1]. Let $\mu = \mathbb{E}[\sum X_i]$, $\sigma^2 = \mathrm{Var}[\sum X_i]$, and $s \leq \sigma$. Then:

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

Can plot this bound for different s:

prob.



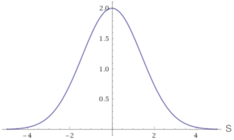
deviation

Interpretation as a Central Limit Theorem

Bernstein Inequality (Simplified): Consider independent random variables $X_1, \ldots, X_n$ falling in [-1,1]. Let $\mu = \mathbb{E}[\sum X_i]$, $\sigma^2 = \mathrm{Var}[\sum X_i]$, and $s \leq \sigma$. Then:

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

Can plot this bound for different s:



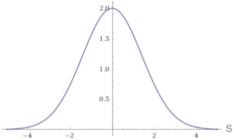
Looks a lot like a Gaussian (normal) distribution.

Interpretation as a Central Limit Theorem

Bernstein Inequality (Simplified): Consider independent random variables $X_1, \ldots, X_n$ falling in [-1,1]. Let $\mu = \mathbb{E}[\sum X_i]$, $\sigma^2 = \mathrm{Var}[\sum X_i]$, and $s \leq \sigma$. Then:

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

Can plot this bound for different s:



Looks a lot like a Gaussian (normal) distribution.

$$\mathcal{N}(0, \sigma^2) \text{ has density } p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}$$

Gaussian Tails

$\mathcal{N}(0, \sigma^2)$ has density $p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}$.

Gaussian Tails

$$\mathcal{N}(0, \sigma^2) \text{ has density } p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}.$$

Exercise: Using this can show that for X $\sim \mathcal{N}(0, \sigma^2)$: for any $s \geq 0$,

$$\Pr\left(|X| \geq s \cdot \sigma\right) \leq 2e^{-\frac{s^2}{2}}.$$

Gaussian Tails

$$\mathcal{N}(0, \sigma^2) \text{ has density } p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}.$$

Exercise: Using this can show that for $X \sim \mathcal{N}(0, \sigma^2)$: for any $s \geq 0$,

$$\Pr\left(|X| \geq s \cdot \sigma\right) \leq 2e^{-\frac{s^2}{2}}.$$

Essentially the same bound that Bernstein's inequality gives!

Gaussian Tails

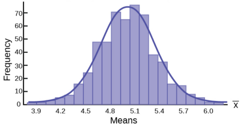
$$\mathcal{N}(0,\sigma^2) \text{ has density } p(s\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{s^2}{2}}.$$

Exercise: Using this can show that for $X \sim \mathcal{N}(0,\sigma^2)$: for any $s \geq 0$,

$$\Pr\left(|X| \geq s \cdot \sigma\right) \leq 2e^{-\frac{s^2}{2}}.$$

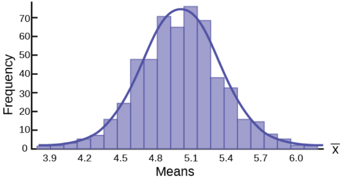
Essentially the same bound that Bernstein's inequality gives!

Central Limit Theorem Interpretation: Bernstein's inequality gives a quantitative version of the CLT. The distribution of the sum of *bounded* independent random variables can be upper bounded with a Gaussian (normal) distribution.



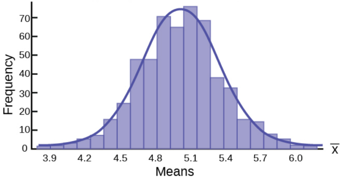
Central Limit Theorem

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.



Central Limit Theorem

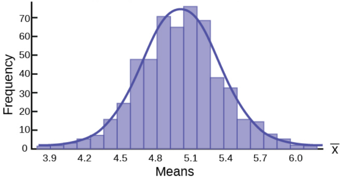
Stronger Central Limit Theorem: The distribution of the sum of n *bounded* independent random variables converges to a Gaussian (normal) distribution as n goes to infinity.



- Why is the Gaussian distribution is so important in statistics, science, ML, etc.?

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.



- Why is the Gaussian distribution is so important in statistics, science, ML, etc.?
- 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.

A useful variation of the Bernstein inequality for binary (indicator) random variables is:

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

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

$\exp(-\delta\mu)$

$\delta = .1$

0 : .6
1 : .4

binomial : $X_1, \ldots X_n$ are identically distributed

A useful variation of the Bernstein inequality for binary (indicator) random variables is:

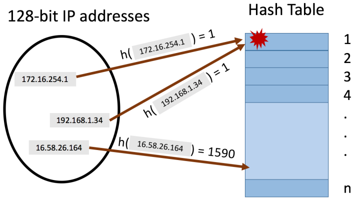
Chernoff Bound (simplified version): Consider independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$. For any $\delta \geq 0$

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

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

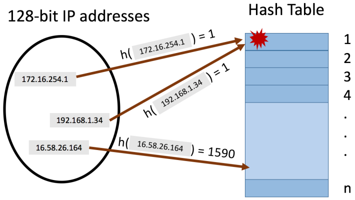
$\mathrm{Var}(X) \leq \mathbb{E}X^2 = \mathbb{E}X$

Return to Random Hashing



We hash m values $x_1, \ldots, x_m$ using a random hash function into a table with $n = m$ entries.

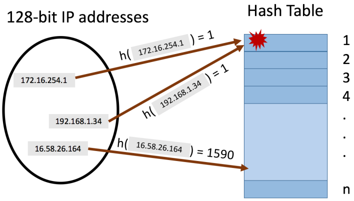
Return to Random Hashing



We hash m values $x_1, \ldots, x_m$ using a random hash function into a table with $n = m$ entries.

- I.e., for all $j \in [m]$ and $i \in [m]$, $\Pr(\mathbf{h}(x_j) = i) = \frac{1}{m}$ and hash values are chosen independently.

Return to Random Hashing



We hash m values $x_1, \ldots, x_m$ using a random hash function into a table with $n = m$ entries.

- I.e., for all $j \in [m]$ and $i \in [m]$, $\Pr(\mathbf{h}(x_j) = i) = \frac{1}{m}$ and hash values are chosen independently.

What will be the maximum number of items hashed into the same location?

Maximum Load in Randomized Hashing

Let S_i be the number of items hashed into position i and $\mathsf{S}_{i,j}$ be 1 if x_j is hashed into bucket i ($\mathsf{h}(x_j) = i$) and 0 otherwise.

$$\mathbb{E}[S_{i,j}] = \frac{1}{m}$$

$$S_i = \sum_{j=1}^{m} S_{ij}$$

$$\mathbb{E}S_i = \frac{m}{m} = 1$$

> m: total number of items hashed and size of hash table. $x_1, \ldots, x_m$: the items.
> h: random hash function mapping $x_1, \ldots, x_m \to [m]$.

Maximum Load in Randomized Hashing

Let S_i be the number of items hashed into position i and $\mathsf{S}_{i,j}$ be 1 if x_j is hashed into bucket i ($\mathsf{h}(x_j) = i$) and 0 otherwise.

$$\mathbb{E}[\mathsf{S}_i] = \sum_{j=1}^{m} \mathbb{E}[\mathsf{S}_{i,j}] = m \cdot \frac{1}{m} = 1$$

m: total number of items hashed and size of hash table. $x_1, \ldots, x_m$: the items.
h: random hash function mapping $x_1, \ldots, x_m \to [m]$.

Maximum Load in Randomized Hashing

Let S_i be the number of items hashed into position i and $\mathsf{S}_{i,j}$ be 1 if x_j is hashed into bucket i ($\mathsf{h}(x_j) = i$) and 0 otherwise.

$$\mathbb{E}[\mathsf{S}_i] = \sum_{j=1}^{m} \mathbb{E}[\mathsf{S}_{i,j}] = m \cdot \frac{1}{m} = 1 = \mu.$$

m: total number of items hashed and size of hash table. $x_1, \ldots, x_m$: the items.
h: random hash function mapping $x_1, \ldots, x_m \to [m]$.

Maximum Load in Randomized Hashing

Let S_i be the number of items hashed into position i and $\mathsf{S}_{i,j}$ be 1 if x_j is hashed into bucket i ($\mathsf{h}(x_j) = i$) and 0 otherwise.

$\mu = \frac{1}{2}$

$$\mathbb{E}[\mathsf{S}_i] = \sum_{j=1}^{m} \mathbb{E}[\mathsf{S}_{i,j}] = m \cdot \frac{1}{m} = 1 = \mu.$$

$\delta = 3$

By the Chernoff Bound: for any $\delta \geq 0$,

$\mathbb{E} S_i$ $\quad$ $\mathbb{E} S_i = 1$

$$\Pr(\mathsf{S}_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{j=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta \cdot \mu\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right)$$

$$\leq \Pr\left(S_i \geq \mathbb{E} S_i + \tfrac{1}{2} + \delta\right) \leq \Pr\left(|S_i - \mathbb{E} S_i| \geq \tfrac{1}{2} + \delta\right) = \Pr\left(\ldots \geq (1+2\delta)\mu\right)$$

m: total number of items hashed and size of hash table. $x_1, \ldots, x_m$: the items.
h: random hash function mapping $x_1, \ldots, x_m \to [m]$.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

Set $\underline{\delta = 20 \log m.}$ Gives:

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1+\delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

Set $\delta = 20\log m$. Gives:

$$\underbrace{\Pr(\mathsf{S}_i \geq 20\log m + 1)}_{\approx 20\log m} \leq 2\exp\left(-\frac{(20\log m)^2}{2+20\log m}\right).$$

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1+\delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

Set $\delta = 20\log m$. Gives:

$$\frac{20^2(\log m)^2}{22\log m} \quad 20$$

$$\Pr(\mathsf{S}_i \geq 20\log m + 1) \leq 2\exp\left(-\frac{(20\log m)^2}{2+20\log m}\right) \leq 2\exp(-18\log m) \leq \frac{2}{m^{18}}.$$

$m \geq 10$

$m = e^{10}$

$\log m = 10$

201

$$\frac{2}{10^{18}}$$

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

$$\Pr(\mathsf{S}_i \geq 1+\delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathsf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

Set $\delta = 20\log m$. Gives:

$$\Pr(\mathsf{S}_i \geq 20\log m + 1) \leq 2\exp\left(-\frac{(20\log m)^2}{2+20\log m}\right) \leq \exp(-18\log m) \leq \frac{2}{m^{18}}.$$

Apply Union Bound:

$$\Pr(\max_{i\in[m]} \mathsf{S}_i \geq 20\log m + 1) = \Pr\left(\bigcup_{i=1}^{m}(\mathsf{S}_i \geq 20\log m + 1)\right) \leq \sum_{i=1}^{m} \Pr(S_i \geq 20\log m + 1)$$

$$\leq m \cdot \frac{2}{m^{18}} \leq \frac{2}{m^{17}}$$

m: total number of items hashed and size of hash table. S_i: number of items hashed to bucket i. $\mathsf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

$$\Pr(\mathbf{S}_i \geq 1+\delta) \leq \Pr\left(\left|\sum_{i=1}^{n} \mathbf{S}_{i,j} - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{\delta^2}{2+\delta}\right).$$

Set $\delta = 20\log m$. Gives:

$$\Pr(\mathbf{S}_i \geq 20\log m + 1) \leq 2\exp\left(-\frac{(20\log m)^2}{2+20\log m}\right) \leq \exp(-18\log m) \leq \frac{2}{m^{18}}.$$

Apply Union Bound:

$$\Pr(\max_{i\in[m]} \mathbf{S}_i \geq 20\log m + 1) = \Pr\left(\bigcup_{i=1}^{m} (\mathbf{S}_i \geq 20\log m + 1)\right)$$

$$\leq \sum_{i=1}^{m} \Pr(\mathbf{S}_i \geq 20\log m + 1) \leq m \cdot \frac{2}{m^{18}} = \frac{2}{m^{17}}.$$

m: total number of items hashed and size of hash table. $\mathbf{S}_i$: number of items hashed to bucket i. $\mathbf{S}_{i,j}$: indicator if x_j is hashed to bucket i. δ: any value ≥ 0.

Maximum Load in Randomized Hashing

Upshot: If we randomly hash m items into a hash table with m entries the maximum load per bucket is $O(\log m)$ with very high probability.

Maximum Load in Randomized Hashing

Upshot: If we randomly hash m items into a hash table with m entries the maximum load per bucket is $O(\log m)$ with very high probability.

- So, even with a simple linked list to store the items in each bucket, worst case query time is $O(\log m)$.

Maximum Load in Randomized Hashing

Upshot: If we randomly hash m items into a hash table with m entries the maximum load per bucket is $O(\log m)$ with very high probability.

- So, even with a simple linked list to store the items in each bucket, worst case query time is $O(\log m)$.
- Using Chebyshev's inequality could only show the maximum load is bounded by $O(\sqrt{m})$ with good probability (good exercise).

Maximum Load in Randomized Hashing

Upshot: If we randomly hash m items into a hash table with m entries the maximum load per bucket is $O(\log m)$ with very high probability.

- So, even with a simple linked list to store the items in each bucket, worst case query time is $O(\log m)$.
- Using Chebyshev's inequality could only show the maximum load is bounded by $O(\sqrt{m})$ with good probability (good exercise).
- The Chebyshev bound holds even with a pairwise independent hash function. The stronger Chernoff-based bound can be shown to hold with a *k-wise independent hash function* for $k = O(\log m)$.