

CS-GY 9223 I: Lecture 2

Chernoff Bounds + Sketching and Streaming

NYU Tandon School of Engineering, Prof. Christopher Musco

Central question in randomized algorithms: How well does a random variable X concentrate around its expectation $\mathbb{E}[X]$?

Two Concentration bounds:

Markov's Inequality $\Pr[X > k\mathbb{E}[X]] \leq \frac{1}{k}$

- Requires that $X > 0$ always.

Chebyshev's Inequality $\Pr[|X - \mathbb{E}[X]| > k\sigma] \leq \frac{1}{k^2}$

- Here $\sigma^2 = \text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

Applications: Space efficient hash table design, understanding randomized load balancing, many more.



New York University Tandon School of Engineering
Computer Science and Engineering
CS-GY 9223I: Lecture 1 Coursework

Problem 1: Hash collisions are useful?

Your company is considering paying for a cloud service that provides CAPTCHA-like visual puzzles for verifying that users are human. The company providing the service claims to have a larger database of unique puzzles than any competitors, but you don't trust the salesperson.

- The company has provided you with an API end-point which returns puzzles uniformly and independently at random from their database. Using this endpoint, describe a simple randomized estimator for the number of puzzles in the database, n .
- The company claims their database has 1,000,000 unique CAPTCHAs in it. Using your estimator, roughly how many queries do you need to verify their claim with good probability (e.g. 9/10)? You should need far less than 1,000,000 queries!
- More generally, how many samples are required to estimate the true number of CAPTCHAs, n , in the database up to additive error $\pm en$, with good probability?

IN-CLASS EXERCISE

Part (a): n = size of the database
 m = number of queries (random capture)

$$\boxed{X_1, \dots, X_n}$$

$$C \rightarrow n$$

$$C = \sum_{i=1}^m \sum_{j \neq i} \mathbb{I}[X_i = X_j]$$

↓
collisions

$$\mathbb{I} \begin{cases} 0 & \text{with prob } 1 - \frac{1}{n} \\ \underline{\underline{1}} & \text{w/ prob } \frac{1}{n} \end{cases}$$

$$\mathbb{E} C = \sum_{i=1}^m \sum_{j \neq i} \mathbb{E} \mathbb{I}[X_i = X_j]$$

$$\boxed{\mathbb{E} C = \frac{m(m-1)}{2n}}$$

↓ m
↓ $(m-1)$

$\frac{1}{n}$

$$\mathbb{E} \mathbb{I} = (1 - \frac{1}{n}) \cdot 0 + \frac{1}{n} \cdot 1 = \frac{1}{n}$$

IN-CLASS EXERCISE

Parts (b)/(c):
$$EC = \frac{m(m-1)}{n}$$

$$\boxed{n = \frac{m(m-1)}{EC}} \quad C$$

$$\underline{n} \approx \underline{\tilde{n}} = \frac{m(m-1)}{2C}$$

If C is close to EC ,
 \tilde{n} is close to n .

$$\boxed{(1-\epsilon)n \leq \tilde{n} \leq (1+\epsilon)n} \rightarrow \text{want}$$

$$(1-\epsilon) \frac{1}{EC} \leq \frac{1}{C} \leq (1+\epsilon) \frac{1}{EC} \rightarrow \text{want}$$

$$\boxed{\frac{1}{(1+\epsilon)}} EC \leq C \leq EC \boxed{\frac{1}{(1-\epsilon)}}$$

Successes to prove: $\boxed{(1-\frac{\epsilon}{2}) EC \leq C \leq (1+\frac{\epsilon}{2}) EC}^5$

IN-CLASS EXERCISE

Parts (b)/(c): If $\boxed{(1 - \frac{\epsilon}{2}) \mathbb{E}C \leq C \leq (1 + \frac{\epsilon}{2}) \mathbb{E}C}$ then

$$\tilde{n} = \frac{m(m-1)}{2C}$$

$$C \in [n - \epsilon n, n + \epsilon n] \quad \checkmark$$

(Chebyshev's inequality). $\text{Var}[C]$.

$$\text{Var} \left[\sum_{i=1}^m \sum_{j>i} \mathbb{1}[X_i = X_j] \right] = \sum_{i=1}^m \underbrace{\sum_{j>i} \text{Var}[\mathbb{1}[X_i = X_j]]}_{\text{I}}$$

$$\text{Var}[I] = \mathbb{E}[I^2] - \mathbb{E}[I]^2$$

$$\frac{1}{4} - \frac{1}{4}^2 \leq \frac{1}{4}$$

$$\text{Var}[C] \leq \frac{m(m-1)}{2} \cdot \frac{1}{4}$$

IN-CLASS EXERCISE

Parts (b)/(c): $E[C] = \frac{n(n-1)}{2n}$ $Var[C] \leq \frac{n(n-1)}{2n} \leq n^2/2n$

$Pr(|C - E[C]| \geq \frac{1}{\sqrt{10}} \sqrt{\frac{n^2}{2n}}) \leq 1/10$ \rightarrow Chebyshev's

$Pr(|C - E[C]| \geq \epsilon E[C]) \leq 1/10$ \rightarrow what I want to prove

$\frac{1}{\sqrt{10}} \sqrt{\frac{n^2}{2n}} \leq \frac{\epsilon n(n-1)}{2n}$ $k = 1/\sqrt{10}$

Choose n
such that

$$\frac{1}{\sqrt{10}} \frac{1}{\sqrt{2}} \leq \frac{\epsilon n}{2\sqrt{n}}$$

$$\frac{2\sqrt{n}}{\sqrt{10}\sqrt{2}} \leq n$$

$$n = O\left(\frac{\sqrt{n}}{\epsilon}\right)$$

$$\tilde{n} = (1 \pm \epsilon)n$$

$$O(\sqrt{n})$$

IN-CLASS EXERCISE

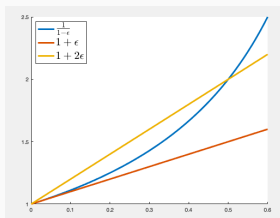
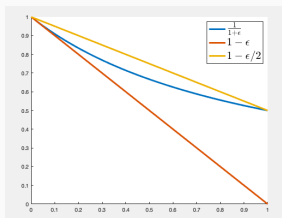
For small ϵ , $(1 - \epsilon) \approx 1/(1 + \epsilon)$ and $(1 + \epsilon) \approx 1/(1 - \epsilon)$.

Useful identities:

$$1 - \epsilon \leq \frac{1}{1 + \epsilon} \leq 1 - \epsilon/2 \quad \text{for all } 0 \leq \epsilon \leq 1$$

$$1 + \epsilon \leq \frac{1}{1 - \epsilon} \leq 1 + 2\epsilon \quad \text{for all } 0 \leq \epsilon \leq 1/2$$

```
1 - eps = 0:.01:1;
2 - figure(); hold();
3 - plot(eps, 1./(1+eps));
4 - plot(eps, 1-eps);
```



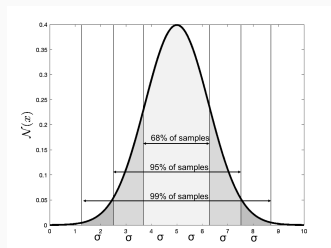
IN CLASS PROBLEM

Fun facts:

- Known as the “mark-and-recapture” method in ecology.
- Can also be used by webcrawlers to estimate the size of the internet, a social network, etc.



Motivating question: Is Chebyshev's Inequality tight?



68-95-99 rule for Gaussian bell-curve. $X \sim N(0, \sigma^2)$

Chebyshev's Inequality:

$$\Pr(|X - \mathbb{E}[X]| \geq 1\sigma) \leq 100\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 2\sigma) \leq 25\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 3\sigma) \leq 11\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 4\sigma) \leq 6\%.$$

Truth:

$$\Pr(|X - \mathbb{E}[X]| \geq 1\sigma) \approx 32\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 2\sigma) \approx 5\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 3\sigma) \approx 1\%$$

$$\Pr(|X - \mathbb{E}[X]| \geq 4\sigma) \approx .01\%$$

GAUSSIAN CONCENTRATION

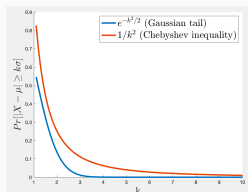
For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\Pr[X = \mu \pm x] = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

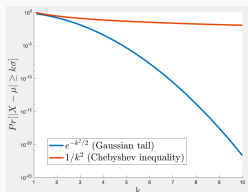
Lemma (Gaussian Tail Bound)

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\Pr[|X - \mathbb{E}X| \geq k\sigma] \leq O(e^{-k^2/2}).$$



Standard y-scale.



Logarithmic y-scale.

Give an example of a random variable X with variance σ^2 for which Chebyshev's inequality is tight.

$$\Pr[|X - \mathbb{E}X| \geq k\sigma] \leq \frac{1}{k^2}.$$

Describe random variable X such that

$$\Pr[|X - \mathbb{E}[X]| \geq k\sigma] = \frac{1}{k^2} \quad \text{or} \quad = \frac{1}{10k^2} \dots$$

Takeaway: Gaussian random variables concentrate much tighter around their expectation than variance alone predicts.

Why does this matter for algorithm design?

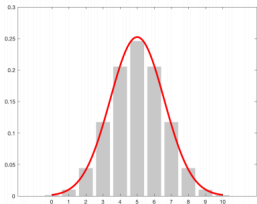
CENTRAL LIMIT THEOREM

Theorem (CLT – Informal)

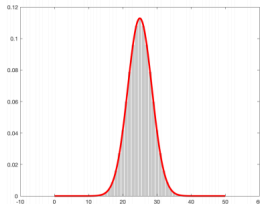
Any sum of *independent, (identically distributed)* r.v.'s X_1, \dots, X_n with mean μ and finite variance σ^2 converges to a Gaussian r.v. with mean $n \cdot \mu$ and variance $n \cdot \sigma^2$, as $n \rightarrow \infty$.

$$S = \sum_{i=1}^n X_i \implies \mathcal{N}(n \cdot \mu, n \cdot \sigma^2).$$

$$E\{S\} = n \cdot \mu \quad \text{Var}\{S\} = n \cdot \sigma^2$$



(a) Distribution of # of heads after 10 coin flips, compared to a Gaussian.



(b) Distribution of # of heads after 50 coin flips, compared to a Gaussian.

Definition (Mutual Independence)

Random variables X_1, \dots, X_n are mutually independent if, for all possible values v_1, \dots, v_n ,

$$\Pr[X_1 = v_1, \dots, X_n = v_n] = \Pr[X_1 = v_1] \cdot \dots \cdot \Pr[X_n = v_n]$$

Strictly stronger than pairwise independence.

$$\Pr[X_i = v_i] \Pr[X_j = v_j] = \Pr[X_i = v_i, X_j = v_j]$$

for all i, j

IN-CLASS EXERCISE

You have access to a coin and want to determine if it's ϵ -close to unbiased. To do so, you flip the coin repeatedly and check that the ratio of heads flips is between $1/2 - \epsilon$ and $1/2 + \epsilon$. If it is not, you reject the coin as overly biased.

- (a) How many flips n are required so that, with probability $(1 - \delta)$, you do not accidentally reject a truly unbiased coin? Your solution will depend on ϵ and δ .

For this problem, you can assume the CLT holds exactly for a sum of independent random variables – i.e., that this sum looks exactly like a Gaussian random variable.

Lemma (Gaussian Tail Bound)

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\Pr[|X - \mathbb{E}X| \geq k\sigma] \leq O(e^{-k^2/2}).$$

IN-CLASS EXERCISE

$$H = \sum_{i=1}^n X_i$$

$$X_i = \begin{cases} 1 & \text{w/ prob } 1/2 \\ 0 & \text{w/ prob } 1/2 \end{cases}$$

$$\left[\Pr\left[\left| H - \frac{n}{2} \right| \geq \underline{\underline{\epsilon n}} \right] \leq \underline{\underline{\delta}} \right]$$

reject a good coin

$$E[H] = n/2$$

$$\text{Var}[H] = \sum_{i=1}^n \underbrace{\text{Var}[X_i]}_{1/4} = n/4$$

$$H \approx N(n/2, n/4)$$

variance
P
r.v.

$$\Pr\left[\left| H - \frac{n}{2} \right| \geq k\sqrt{n/4} \right] \leq O\left(e^{-k^2/2}\right) \leq \underline{\underline{\delta}}$$

$$k^2/2 \geq \log(1/\delta)$$

$$\Pr\left[\left| H - \frac{n}{2} \right| \geq \underbrace{\sqrt{\log(1/\delta)} \sqrt{n/4}}_{\leq \epsilon n} \right] \leq \delta$$

$$k = \sqrt{2 \log(1/\delta)}$$

$$\sqrt{\log(1/\delta)} \sqrt{\frac{1}{4}} \leq \epsilon \sqrt{n}$$

$$\sqrt{n} \geq O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$$

$$n \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$$

These back-of-the-envelope calculations can be made rigorous! **Lots of different “versions” of bound which do so.**

- Chernoff bound
- Bernstein bound
- Hoeffding bound
- ...

Different assumptions on random variables (e.g. binary, bounded, i.i.d), different forms (additive vs. multiplicative error), etc. **Wikipedia is your friend.**

Theorem (Bernstein Inequality)

Let X_1, X_2, \dots, X_n be independent random variables with each $X_i \in [-1, 1]$. Let $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = \text{var}[X_i]$. Let $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. Then, for $k \leq \frac{1}{2}\sigma$, $S = \sum_i X_i$ satisfies

$$\Pr[|S - \mu| > k\sigma] \leq 2 \exp\left(-\frac{k^2}{4}\right).$$

Sample Application: Flip random coin n times. As long as $n \geq O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$,

$$\Pr[|\#\text{ heads} - n/2| \geq \epsilon n] \leq \delta$$

Pay very little for higher probability – if you increase the number of coin flips by $2x$, δ goes from $1/10 \rightarrow 1/100 \rightarrow 1/0000$

Theorem (Chernoff Bound)

Let X_1, X_2, \dots, X_n be independent $\{0, 1\}$ -valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Then the sum $S = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

$$\Pr[X \geq (1 + \epsilon)\mu] \leq e^{\frac{-\epsilon^2 \mu}{3+3\epsilon}}.$$

Any guess for how these bounds are proven?

Higher order moments. $\frac{\mathbb{E}[(X - \mathbb{E}X)^2]}{\mathbb{E}[(X - \mathbb{E}X)^4]}$

LOAD BALANCING

As in the previous lecture, we want to use concentration bounds to study the randomized load balancing problem. n jobs are distributed randomly to n servers using a hash function. Let S_i be the number of jobs sent to server i . What's the smallest B for which we can prove:

$$\Pr[\max_i S_i \geq B] \leq 1/10$$



Recall: Suffices to prove that, for any i , $\Pr[S_i \geq B] \leq 1/10n$:

$$\begin{aligned} \Pr[\max_i S_i \geq B] &= \Pr[S_1 \geq B \text{ or } \dots \text{ or } S_n \geq B] \\ &\leq \Pr[S_1 \geq B] + \dots + \Pr[S_n \geq B] \quad (\text{union bound}). \end{aligned}$$

What do you expect the answer to be?

LOAD BALANCING

Theorem (Chernoff Bound)

Let X_1, X_2, \dots, X_n be independent $\{0, 1\}$ -valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Then the sum $S = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

$$\Pr[X \geq (1 + \epsilon)\mu] \leq e^{\frac{-\epsilon^2 \mu}{3 + 3\epsilon}}.$$

$$\underline{S_i} = \sum_{j=1}^n \mathbb{1}(\text{ball } j \text{ lands in bin } i) \quad p_i = 1/n$$

X_1, \dots, X_n

$\epsilon = 6 \log(n)$

$$\mathbb{E}[S_i] = \mu = n \cdot 1/n = \underline{\underline{1}} \quad O(\log(n))$$

$$\underline{\Pr[S_i \geq (1 + 6 \log n) \mathbb{1}] \leq e^{-\frac{6^2 \log^2(n) \mathbb{1}}{3 + 3 \cdot 6 \log(n)}} \leq e^{-\frac{36 \log n}{21}} \leq \frac{1}{10^2} \leq \left(\frac{1}{n}\right)^{36/21} \approx \frac{1}{n^{1.522}}$$

POWER OF TWO CHOICES

Power of 2 Choices: Instead of assigning job to a random server, choose 2 random servers and assign to the least loaded. With probability $1/10$ the maximum load is bounded by:

- (a) $O(\log n)$
- (b) $O(\sqrt{\log n})$
- (c) $O(\log \log n)$
- (d) $O(1)$

Power of 3 choices? $O(\log \log n / \log(3))$ $O(1)$

Abstract architecture of a streaming algorithm:

- Given a dataset $D = d_1, \dots, d_n$ with n pieces of data, we want to output $f(D)$ for some function f .
- Maintain state S_t with $\ll |D|$ space at each time step t .
- **Update phase:** Receive d_1, \dots, d_n in sequence, update $S_t \leftarrow U(S_{t-1}, d_t)$.
- **Process phase:** Using S_n , compute approximation to $f(D)$.

Typical setup for training models in machine learning, required for large scale data monitoring (e.g. processing sensor data, time series, seismic data, satellite imagery, etc.)



DISTINCT ELEMENTS PROBLEM

Input: $d_1, \dots, d_n \in \mathcal{U}$ where \mathcal{U} is a huge universe of items.

Output: Number of distinct inputs.

Example: $f(1, 10, 10, 4, 9, 1, 1, 4) \rightarrow 4$

Applications:

- **In practice:** Google (Sawzall, Dremel, PowerDrill), Yahoo, Twitter, Facebook Presto, etc. etc.
- Distinct users hitting a webpage.
- Distinct values in a database column (e.g. for estimating the size of group by queries)
- Number of distinct queries to a search engine.
- Distinct motifs in DNA sequence.

DISTINCT ELEMENTS PROBLEM

Input: $d_1, \dots, d_n \in \mathcal{U}$ where \mathcal{U} is a huge universe of items.

Output: Number of distinct inputs.

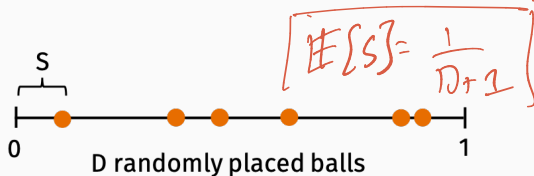
Example: $f(1, 10, 10, 4, 9, 1, 1, 4) \rightarrow 4$

Flajolet–Martin (simplified):

- Choose random hash function $h : \mathcal{U} \rightarrow [0, 1]$.
- $S = \infty$
- For $i = 1, \dots, n$
 - $S \leftarrow \min(S, h(d_i))$
- Return: $\frac{1}{S} - 1$

What is $\mathbb{E}S$?

$\Pr[\text{ball}_{D+1} \text{ is closest to } 0] = \mathbb{E}[S]$



Let D equal the number of distinct elements in our stream.

Lemma

$$\mathbb{E}S = \frac{1}{D+1}.$$

$$\mathbb{E}S = \frac{1}{D+1}$$

Estimate: $\tilde{D} = \frac{1}{S} - 1.$

If $|S - \mathbb{E}S| \leq \frac{\epsilon}{4} \cdot \mathbb{E}S$, then:

$$(1 - \epsilon)D \leq \tilde{D} \leq (1 + \epsilon)D.$$

To show concentration, need a variance bound for S .

Lemma

$$\text{Var}[S] = \mathbb{E}[S^2] - \mathbb{E}[S]^2 = \frac{2}{(D+1)(D+2)} - \frac{1}{(D+1)^2} \leq \frac{1}{(D+1)^2}.$$

Proof:

$$\begin{aligned} \mathbb{E}[S^2] &= \int_0^1 \Pr[S^2 \geq \lambda] d\lambda \\ &= \int_0^1 \Pr[S \geq \sqrt{\lambda}] d\lambda \\ &= \int_0^1 (1 - \sqrt{\lambda})^D d\lambda \\ &= \frac{2}{(D+1)(D+2)} \end{aligned}$$

Exercise: Why?

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu.$
- $\text{Var}[S] = \mu^2$
- Want to bound $\Pr[|S - \mu| \leq \epsilon\mu] \leq \delta.$
- Won't get a good bound with one estimator alone...

Trick of the trade: Repeat many independent trials and use a Chernoff/Bernstein bound.

1 estimator:

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu.$
- $\text{Var}[S] = \mu^2$

k estimators:

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu.$
- $\text{Var}[S] = (\mu^2 \cdot k)/k^2 = \mu^2/k$
- By Bernstein, $\Pr[|S - \mathbb{E}S| \leq c\mu/\sqrt{k}] \leq e^{-c^2}.$

Setting $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ gives:

$$\Pr[|S - \mu| \leq \epsilon\mu] \leq \delta.$$

Total space complexity: $O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ to estimate distinct elements up to error ϵ with success probability $1 - \delta$.

1 estimator:

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu.$
- $\text{Var}[S] = \mu^2$

k estimators:

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu.$
- $\text{Var}[S] = (\mu^2 \cdot k)/k^2 = \mu^2/k$
- By Chebyshev, $\Pr[|S - \mathbb{E}S| \geq c\mu/\sqrt{k}] \leq \frac{1}{c^2}.$

Setting $c = 1/\sqrt{\delta}$ and $k = O\left(\frac{1}{\epsilon^2\delta}\right)$ gives:

$$\Pr[|S - \mu| \geq \epsilon\mu] \leq \delta.$$

Total space complexity: $O\left(\frac{1}{\epsilon^2\delta}\right)$ to estimate distinct elements up to error ϵ with success probability $1 - \delta$.