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 it's 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 $\sigma^2 = \mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

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

IN-CLASS EXERCISE

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.

(a) 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.

(b) 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!

(c) More generally, how many samples are required to estimate the true number of CAPTCHAs, n, in the database up to additive error $\pm\epsilon n$, with good probability?

Part (a):

Parts (b)/(c):

Parts (b)/(c):

Parts (b)/(c):

IN-CLASS EXERCISE

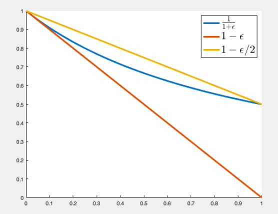
For small ϵ, $(1-\epsilon) \approx 1/(1+\epsilon)$ and $(1+\epsilon) \approx 1/(1-\epsilon)$.

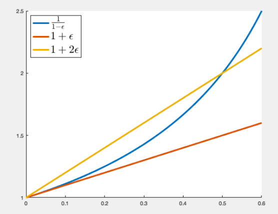
Useful identities:

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

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

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



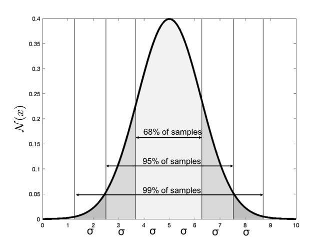


IN CLASS PROBLEM

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:

$$\Pr\left(|X - \mathbb{E}[X]| \geq 1\sigma\right) \leq 100\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 2\sigma\right) \leq 25\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 3\sigma\right) \leq 11\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 4\sigma\right) \leq 6\%.$$

Truth:

$$\Pr\left(|X - \mathbb{E}[X]| \geq 1\sigma\right) \approx 32\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 2\sigma\right) \approx 5\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 3\sigma\right) \approx 1\%$$
$$\Pr\left(|X - \mathbb{E}[X]| \geq 4\sigma\right) \approx .01\%$$

GAUSSIAN CONCENTRATION

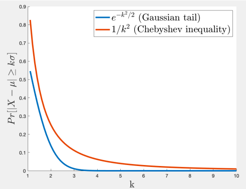
For $X \sim \mathcal{N}(\mu, \sigma^2)$:

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

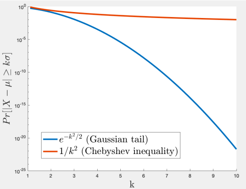
Lemma (Guassian Tail Bound)

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.

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

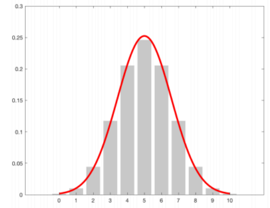
Takeaway: Gaussian random variables concentrate much tighter around their expectation than variance alone predicts.

Why does this matter for algorithm design?

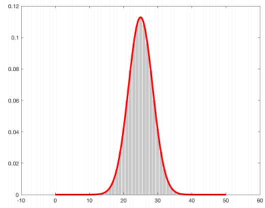
Theorem (CLT – Informal)

*Any sum of **independent**, **(identically distributed)** r.v.'s $X_1, \ldots, X_n$ 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 $n \to \infty$.*

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



(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 $X_1, \ldots, X_n$ are <u>mutually independent</u> if, for all possible values $v_1, \ldots, v_n$,

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

Strictly stronger than pairwise independence.

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.

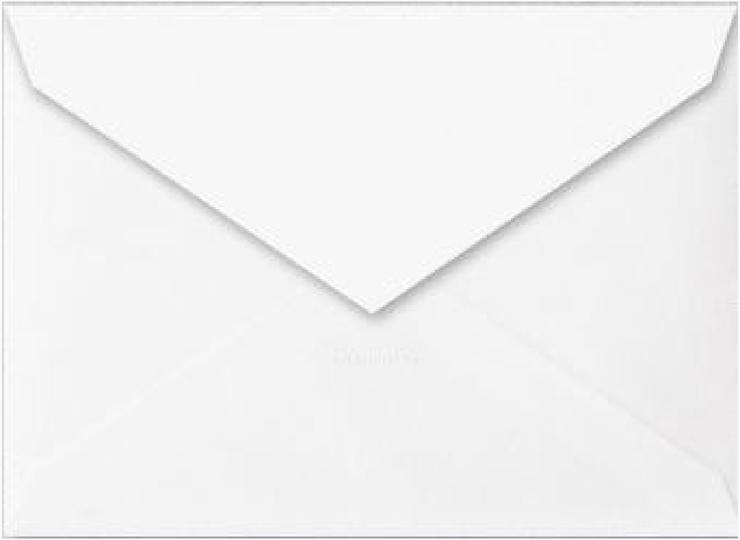
(a) How many flips n are required so that, with probability $(1 - \delta)$, you do not accidentally reject a truly unbiased coin? You solution with depend on ϵ and δ.

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 (Guassian Tail Bound)

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

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



These back-of-the-envelop calculations can be made rigorous! Lots of different "versions" of bound which do so.

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

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

Theorem (Bernstein Inequality)

Let $X_1, X_2, \ldots, X_n$ be independent random variables with each $X_i \in [-1, 1]$. Let $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = var[X_i]$. 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

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

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)

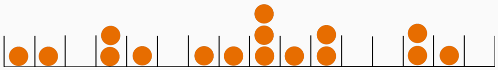
Let $X_1, X_2, \ldots, X_n$ be independent $\{0,1\}$-valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Then the sum $S = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

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

Any guess for how these bounds are proven?

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:

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



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

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

What do you expect the answer to be?

Theorem (Chernoff Bound)

Let $X_1, X_2, \ldots, X_n$ 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

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

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? $O(\log \log n / \log(3))$

Abstract architecture of a streaming algorithm:

- Given a dataset $D = d_1, \ldots, d_n$ 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 $d_1, \ldots, d_n$ 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.)

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

Output: Number of distinct inputs.

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

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.

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

Output: Number of distinct inputs.

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

Flajolet–Martin (simplified):

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

What is $\mathbb{E}S$?

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

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

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

To show concentration, need a variance bound for S.

Lemma

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

Proof:

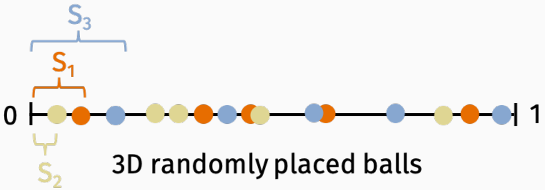
$$\begin{aligned}
\mathbb{E}[S^2] &= \int_0^1 \Pr[S^2 \geq \lambda] d\lambda && \text{Exercise: Why?} \\
&= \int_0^1 \Pr[S \geq \sqrt{\lambda}] d\lambda \\
&= \int_0^1 (1 - \sqrt{\lambda})^D d\lambda \\
&= \frac{2}{(D+1)(D+2)}
\end{aligned}$$

www.wolframalpha.com/input/?i=integral+from+0+to+1+of+%281-sqrt%28x%29%29%5ED

- $\mathbb{E}[S] = \frac{1}{D+1} = \mu$.
- $\text{Var}[S] = \mu^2$
- 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 Chebyshev bound or Chernoff/Bernstein bound.

Using independent hash functions, maintain k independent sketches $S_1, \ldots, S_k$.



Flajolet–Martin:

- Choose k random hash function $h_1, \ldots, h_k : \mathcal{U} \to [0, 1]$.
- $S_1 = \infty, \ldots, S_k = \infty$
- For $i = 1, \ldots, n$
  - $S_j \leftarrow \min(S, h_j(d_i))$ for all $j \in 1, \ldots, k$.
- $S = (S_1 + \ldots + S_k)/k$
- Return: $\frac{1}{S} - 1$

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 $1 - \delta$.