CIS 700: "algorithms for Big Data"

Lecture 1: Intro

Slides at http://grigory.us/big-data-class.html

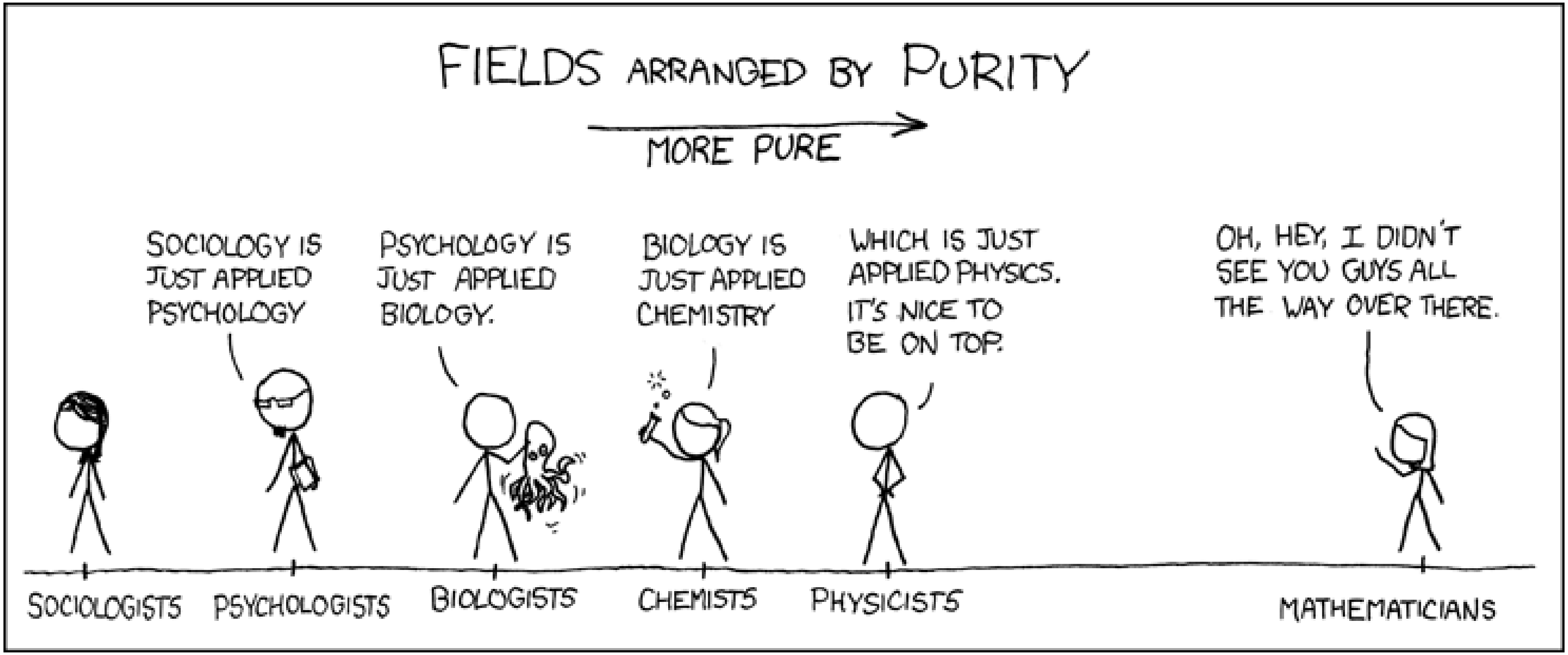
Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

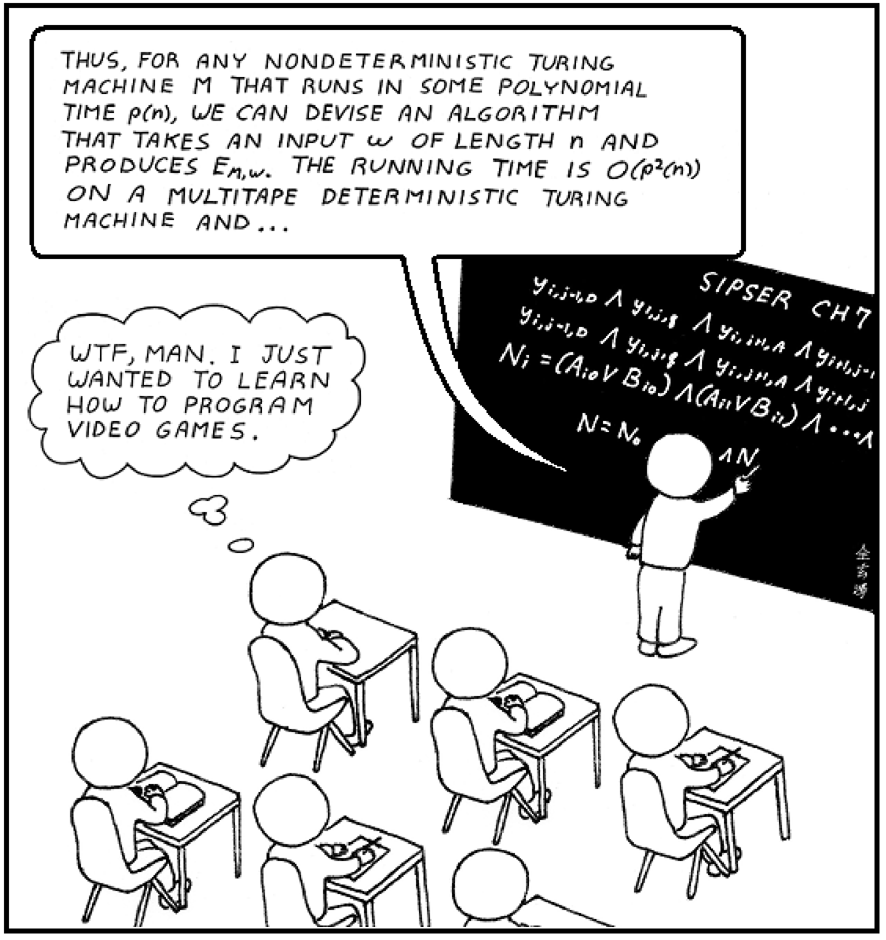
Disclaimers

- A lot of Math!



Disclaimers

- (Almost) no programming!



Class info

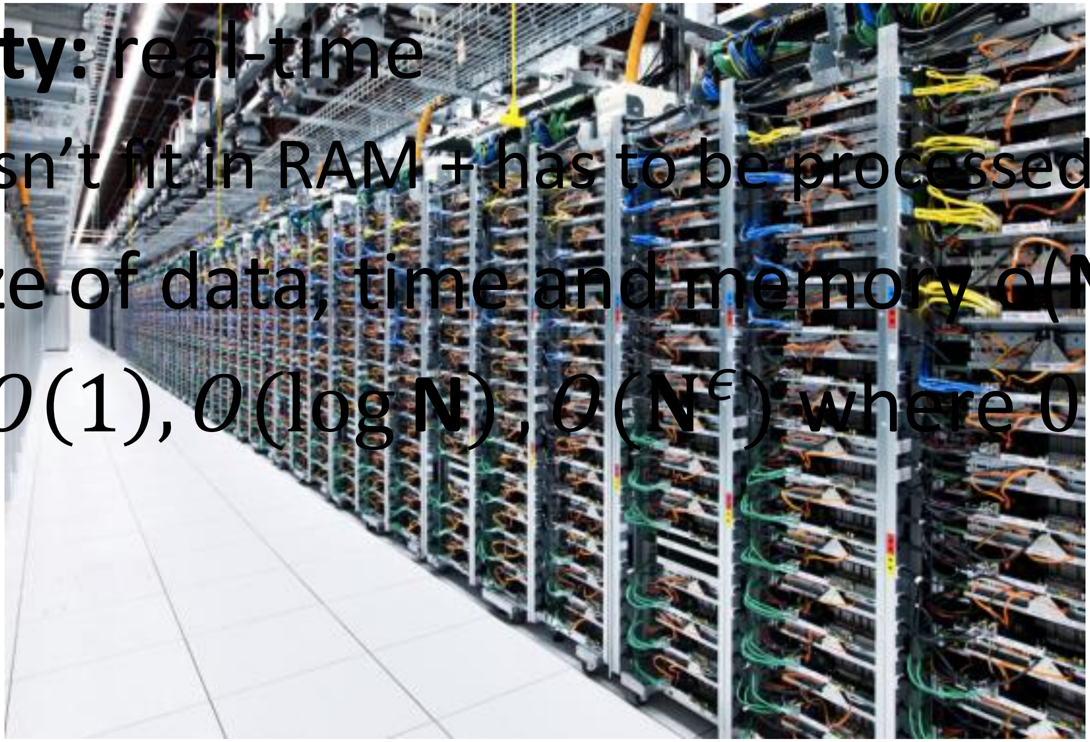
- MW 10:30 – 12:00, Towne 307
- Grading:
  - 1-2 homework assignments (40%)
  - Project (60%)
- Office hours by appointment
- Slides will be posted

What is this class about?

- Not about the band (https://en.wikipedia.org/wiki/Big_Data_(band))

What is this class about?

- The four V's: **volume**, **velocity**, variety, veracity
- **Volume:** "Big Data" = too big to fit in RAM
  - Today 16GB $\approx 100\$$ => "big" starts at terabytes
- **Velocity:** real-time
  - Doesn't fit in RAM + has to be processed on the fly
- **N** = size of data, time and memory o(**N**)
- o(**N**): $O(1)$, $O(\log \mathbf{N})$, $O(\mathbf{N}^{\epsilon})$ where $0 < \epsilon < 1$



Getting hands dirty

- Cloud computing platforms (all offer free trials):
  - Amazon EC2 (1 CPU/12mo)
  - Microsoft Azure ($200/1mo)
  - Google Compute Engine ($200/2mo)
- Distributed Google Code Jam
  - First time in 2015: https://code.google.com/codejam/distributed_index.html
  - Caveats:
    - Very basic aspects of distributed algorithms (few rounds)
    - Small data ($\sim 1\ GB$, with hundreds MB RAM)
    - Fast query access ($\sim 0.01\ ms$ per request), “data with queries”

Outline

- Part 1: Streaming Algorithms

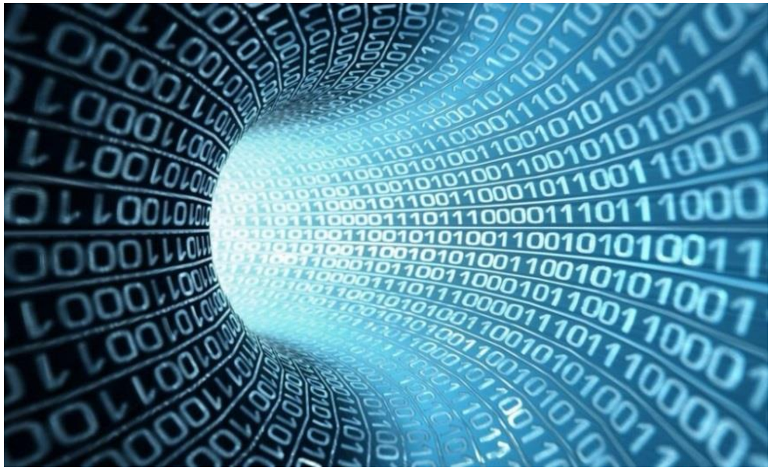
STORM

Highlights:

- Approximate counting
- # Distinct Elements, Hyperloglog
- Median
- Frequency moments
- Heavy hitters
- Graph sketching

Outline

- Part 2: Algorithms for numerical linear algebra



Highlights:

- Dimension reduction
- Nearest neighbor search
- Linear sketching
- Linear regression
- Low rank approximation

Outline

- Part 3: Massively Parallel Algorithms

Hadoop MapReduce

Highlights:

- Computational Model
- Sorting (Terasort)
- Connectivity, MST
- Filtering dense graphs
- Euclidean MST

Outline

- Part 4: Sublinear Time Algorithms

Highlights:

- "Data with queries"
- Sublinear approximation
- Property Testing
- Testing images, sortedness, connectedness
- Testing noisy data

Today

Puzzles

You see a sequence of values $a_1, \dots, a_n$, arriving one by one:

- (**Easy, "Find a missing player"**)
  - If all $a_i's$ are different and have values between 1 and $n+1$, which value is missing?
  - You have $O(\log n)$ space
- Example:
  - There are 11 soccer players with numbers 1, …, 11.
  - You see 10 of them one by one, which one is missing? You can only remember a single number.

1

8

5

11

3

9

2

6

7

4

Which number was missing?

Puzzle #1

You see a sequence of values $a_1, \ldots, a_n$, arriving one by one:

- (**Easy, “Find a missing player”**)
  - If all $a_i's$ are different and have values between 1 and $n + 1$, which value is missing?
  - You have $O(\log n)$ space
- Example:
  - There are 11 soccer players with numbers 1, …, 11.
  - You see 10 of them one by one, which one is missing? You can only remember a single number.

Puzzle #2

You see a sequence of values $a_1, \ldots, a_n$, arriving one by one:

- (**Harder, "Keep a random team"**)
  - How can you maintain a uniformly random sample of S values out of those you have seen so far?
  - You can store exactly S items at any time
- Example:
  - You want to have a team of 11 players randomly chosen from the set you have seen.
  - Players arrive one at a time and you have to decide whether to keep them or not.

Puzzle #3

You see a sequence of values $a_1, \ldots, a_n$, arriving one by one:

- (**Very hard, "Count the number of players"**)
  - What is the total number of values up to error $\pm\,\epsilon n$?
  - You have $O(\log\log n\,/\epsilon^2)$ space and can be completely wrong with some small probability

Puzzles

You see a sequence of values $a_1, \ldots, a_n$, arriving one by one:

- (**Easy, "Find a missing player"**)
  - If all $a_i's$ are different and have values between 1 and $n+1$, which value is missing?
  - You have $O(\log n)$ space
- (**Harder, "Keep a random team"**)
  - How can you maintain a uniformly random sample of S values out of those you have seen so far?
  - You can store exactly S items at any time
- (**Very hard, "Count the number of players"**)
  - What is the total number of values up to error $\pm\epsilon n$?
  - You have $O(\log \log n\, /\epsilon^2)$ space and can be completely wrong with some small probability

Part 1: Probability 101

"The bigger the data the better you should know your Probability"

- Basic Probability:
  - Probability, events, random variables
  - Expectation, variance / standard deviation
  - Conditional probability, independence, pairwise independence, mutual independence

Expectation

- $\boldsymbol{X}$ = random variable with values $x_1, \dots, x_n, \dots$
- Expectation $\mathbb{E}[\boldsymbol{X}]$

$$\mathbb{E}[\boldsymbol{X}] = \sum_{i=1}^{\infty} \mathrm{x_i} \cdot \Pr[\boldsymbol{X} = x_i]$$

- Properties (linearity):

$$\mathbb{E}[c\boldsymbol{X}] = c\mathbb{E}[\boldsymbol{X}]$$
$$\mathbb{E}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}] + \mathbb{E}[\boldsymbol{Y}]$$

- Useful fact: if all $x_i \geq 0$ and integer then

$$\mathbb{E}[\boldsymbol{X}] = \sum_{i=1}^{\infty} \Pr[\boldsymbol{X} \geq i]$$

Variance

- Variance $Var[\boldsymbol{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$

$$Var[\boldsymbol{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] =$$
$$= \mathbb{E}[\boldsymbol{X}^2 - 2\,\mathbf{X} \cdot \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{X}]^2]$$
$$= \mathbb{E}[\boldsymbol{X}^2] - 2\mathbb{E}[\mathbf{X} \cdot \mathbb{E}[\mathbf{X}]] + \mathbb{E}[\mathbb{E}[\mathbf{X}]^2]$$

- $\mathbb{E}[\mathrm{X}]$ is some fixed value (a constant)
- $2\,\mathbb{E}[\mathbf{X} \cdot \mathbb{E}[\mathbf{X}]] = 2\,\mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{X}] = 2\,\mathbb{E}^2[\boldsymbol{X}]$
- $\mathbb{E}[\mathbb{E}[\mathbf{X}]^2] = \mathbb{E}^2[\mathbf{X}]$
- $Var[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}^2] - 2\,\mathbb{E}^2[\boldsymbol{X}] + \mathbb{E}^2[\mathbf{X}] = \mathbb{E}[\boldsymbol{X}^2] - \mathbb{E}^2[\mathbf{X}]$
- Corollary: $Var[c\boldsymbol{X}] = c^2 Var[\boldsymbol{X}]$

Independence

- Two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent** if and only if (iff) for every x, y:
$$\Pr[\boldsymbol{X} = x, \boldsymbol{Y} = y] = \Pr[\boldsymbol{X} = x] \cdot \Pr[\boldsymbol{Y} = y]$$
- Variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are **mutually independent** iff
$$\Pr[\boldsymbol{X}_{\boldsymbol{1}} = x_1, \ldots, \boldsymbol{X}_n = x_n] = \prod_{i=1}^{n} \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i]$$
- Variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are **pairwise independent** iff for all pairs i,j
$$\Pr\left[\boldsymbol{X}_{\boldsymbol{i}} = x_i, \boldsymbol{X}_j = x_j\right] = \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i] \Pr\left[\boldsymbol{X}_{\boldsymbol{j}} = x_j\right]$$

Conditional Probabilities

- For two events E_1 and E_2:

$$\Pr[E_2|E_1] = \frac{\Pr[E_1 \text{ } and \text{ } E_2]}{\Pr[E_1]}$$

- If two random variables (r.vs) are independent

$$\Pr[X_2 = x_2|X_1 = x_1]$$

$$= \frac{\Pr[X_1{=}x_1 \text{ } and \text{ } X_2{=}x_2]}{\Pr[X_1{=}x_1]} \text{ (by definition)}$$

$$= \frac{\Pr[X_1{=}x_1] Pr[X_2{=}x_2]}{\Pr[X_1{=}x_1]}$$ (by independence)

$$= \Pr[X_2 = x_2]$$

Union Bound

For any events $E_1, \ldots, E_k$:

$$\Pr[E_1 or\ E_2\ or\ \ldots or\ E_k] \leq \Pr[E_1] + \Pr[E_2] + \ldots + \Pr[E_k]$$

- **Pro**: Works even for dependent variables!
- **Con**: Sometimes very loose, especially for **mutually** independent events

  $\Pr[E_1\ or\ E_2\ or\ \ldots or\ E_k] = 1 - \prod_{i=1}^{k}(1 - \Pr[E_i])$

Independence and Linearity of Expectation/Variance

- Linearity of expectation (even for dependent variables!):

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} \mathbb{E}[X_i]$$

- Linearity of variance (only for **pairwise independent** variables!)

$$Var\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} Var[X_i]$$

Part 2: Inequalities

- Markov inequality
- Chebyshev inequality
- Chernoff bound

Markov's Inequality

- For every $c > 0$: $\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$
- **Proof (by contradiction)** $\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] > \frac{1}{c}$

$\mathbb{E}[\boldsymbol{X}] = \sum_i i \cdot \Pr[\boldsymbol{X} = i]$ (by definition)

$\geq \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} i \cdot \Pr[\boldsymbol{X} = i]$ (pick only some i's)

$\geq \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} c\mathbb{E}[\boldsymbol{X}] \cdot \Pr[\boldsymbol{X} = i]$ $(i \geq c\mathbb{E}[\boldsymbol{X}])$

$= c\mathbb{E}[\boldsymbol{X}] \sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} \Pr[\boldsymbol{X} = i]$ (by linearity)

$= c\mathbb{E}[\boldsymbol{X}] \Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right]$ (same as above)

$> \mathbb{E}[\boldsymbol{X}]$ (by assumption $\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] > \frac{1}{c}$)

Markov's Inequality

- For every $c > 0$: $\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$
- **Corollary** $(\mathrm{c}' = c\ \mathbb{E}[\boldsymbol{X}])$ **:**

For every $c' > 0$: $\Pr[\boldsymbol{X} \geq c'\ ] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c'}$

- **Pro**: always works!
- **Cons**:
  - Not very precise
  - Doesn't work for the lower tail: $\Pr\left[\boldsymbol{X} \leq c\ \mathbb{E}[\boldsymbol{X}]\right]$

Chebyshev's Inequality

- For every $c > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- Proof:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right]$$

$= \Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 Var[\boldsymbol{X}]]$ (by squaring)

$= \Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 \mathbb{E}[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2]]$ (def. of Var)

$\leq \frac{1}{c^2}$ (by Markov's inequality)

Chebyshev's Inequality

- For every $c > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- **Corollary** ($c' = c\sqrt{Var[\boldsymbol{X}]}$):

For every $c' > 0$:

$$\Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2}$$

Chernoff bound

- Let $X_1 \dots X_t$ be independent and identically distributed r.vs with range [0,1] and expectation μ.
- Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3}\right)$$

Chernoff bound (corollary)

- Let $X_1 \dots X_t$ be independent and identically distributed r.vs with range [0, **c**] and expectation μ.
- Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3\boldsymbol{c}}\right)$$

Chernoff v.s Chebyshev

Large values of t is exactly what we need!

Let $X_1 \ldots X_t$ be independent and identically distributed r.vs with range [0,1] and expectation μ. Let $\boldsymbol{X} = \frac{1}{t}\sum_i X_i$.

- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{t}\right)$
- Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(t)}$

So is Chernoff always better for us?

- Yes, if we have i.i.d. variables.
- No, if we have dependent or only pairwise independent random varaibles.
- If the variables are not identical – Chernoff-type bounds exist.

Answers to the puzzles

You see a sequence of values $a_1, \dots, a_n$, arriving one by one:

- (**Easy**)
  - If all $a_i's$ are different and have values between 1 and $n+1$, which value is missing?
  - You have $O(\log n)$ space
  - **Answer**: missing value = $\sum_{i=1}^{n} i \; - \sum_{i=1}^{n} a_i$
- (**Harder**)
  - How can you maintain a uniformly random sample of S values out of those you have seen so far?
  - You can store exactly S values at any time
  - **Answer**: Store first $a_1, \dots, a_S$. When you see a_i for $i > S$, with probability S/i replace random value from your storage with a_i.

Part 3: Morris's Algorithm

- (**Very hard, "Count the number of players"**)
  - What is the total number of values up to error $\pm \epsilon n$?
  - You have $O(\log \log n / \epsilon^2)$ space and can be completely wrong with some small probability

Morris's Algorithm: Alpha-version

Maintains a counter X using $\log\log n$ bits

- Initialize X to 0
- When an item arrives, increase X by 1 with probability $\frac{1}{2^X}$
- When the stream is over, output $2^X - 1$

Claim: $\mathbb{E}[2^X] = n + 1$

Morris's Algorithm: Alpha-version

Maintains a counter X using $\log\log n$ bits

- Initialize X to 0, when an item arrives, increase X by 1 with probability $\frac{1}{2^X}$

Claim: $\mathbb{E}[2^X] = n + 1$

- Let the value after seeing n items be X_n

$$\mathbb{E}[2^{X_n}] = \sum_{j=0}^{\infty} \Pr[X_{n-1} = j\,]\mathbb{E}[2^{X_n} | X_{n-1} = j]$$

$$= \textstyle\sum_{j=0}^{\infty} \Pr[X_{n-1} = j\,]\left(\frac{1}{2^j}\, 2^{j+1} + \left(1 - \frac{1}{2^j}\right) 2^j\right)$$

$$= \textstyle\sum_{j=0}^{\infty} \Pr[X_{n-1} = j\,]\left(2^j + 1\right) = 1 + \mathbb{E}[2^{X_{n-1}}]$$

Morris's Algorithm: Alpha-version

Maintains a counter X using $\log \log n$ bits

- Initialize X to 0, when an item arrives, increase X by 1 with probability $\frac{1}{2^X}$

Claim: $\mathbb{E}[2^{2X}] = \frac{3}{2}n^2 + \frac{3}{2}n + 1$

$$\mathbb{E}[2^{2X_n}] = \sum_{j=0}^{\infty} \Pr[2^{X_{n-1}} = j\,]\mathbb{E}[2^{2X_n} | 2^{X_{n-1}} = j]$$

$$= \sum_{j=0}^{\infty} \Pr[2^{X_{n-1}} = j\,]\left(\frac{1}{j}\ 4\,j^2 + \left(1\ - \frac{1}{j}\right) j^2\right)$$

$$= \sum_{j=0}^{\infty} \Pr[2^{X_{n-1}} = j\,](j^2 + 3j) = \mathbb{E}[2^{2X_{n-1}}] + 3\mathbb{E}[2^{X_{n-1}}]$$

$$= 3\,\frac{(\mathrm{n}\ - 1)^2}{2} + 3(\mathrm{n} - 1)/2\ \ + 1 + 3\mathrm{n}$$

Morris's Algorithm: Alpha-version

Maintains a counter X using $\log\log n$ bits

- Initialize X to 0, when an item arrives, increase X by 1 with probability $\frac{1}{2^X}$
- $\mathbb{E}[2^X] = \mathrm{n} + 1, Var[2^X] = O(n^2)$
- Is this good?

Morris's Algorithm: Beta-version

Maintains t counters $X^1, \ldots, X^t$ using $\log\log n$ bits for each

- Initialize $X^{i'}s$ to 0, when an item arrives, increase each X^i by 1 independently with probability $\frac{1}{2^{X^i}}$
- Output $\mathrm{Z} = \frac{1}{t}(\sum_{i=1}^{t} 2^{X^i} - 1)$
- $\mathbb{E}[2^{X_i}] = \mathrm{n} + 1, Var[2^{X_i}] = O(n^2)$
- $Var[Z] = Var\left(\frac{1}{t}\sum_{j=1}^{t} 2^{X^j} - 1\right) = O\left(\frac{n^2}{t}\right)$
- Claim: If $t \geq \frac{c}{\epsilon^2}$ then $\Pr[|Z - n| > \epsilon n] < 1/3$

Morris's Algorithm: Beta-version

Maintains t counters $X^1, \ldots, X^t$ using $\log \log n$ bits for each

- Output $\mathrm{Z} = \frac{1}{t}\left(\sum_{i=1}^{t} 2^{X^i} - 1\right)$
- $Var[Z] = Var\left(\frac{1}{t}\sum_{j=1}^{t} 2^{X^j} - 1\right) = O\left(\frac{n^2}{t}\right)$
- Claim: If $t \geq \frac{c}{\epsilon^2}$ then $\Pr[|Z - n| > \epsilon n] < 1/3$
  - $\Pr[|Z - n| > \epsilon\, n] < \frac{Var[Z]}{\epsilon^2 n^2} = O\left(\frac{n^2}{t}\right) \cdot \frac{1}{\epsilon^2 n^2}$
  - If $t \geq \frac{c}{\epsilon^2}$ we can make this at most $\frac{1}{3}$

Morris's Algorithm: Final

- What if I want the probability of error to be really small, i.e. $\Pr[|Z - n| > \epsilon\, n] \leq \delta$?
- Same Chebyshev-based analysis: $t = O\left(\frac{1}{\epsilon^2 \delta}\right)$
- Do these steps $m = O\left(\log\frac{1}{\delta}\right)$ times independently in parallel and output the median answer.
- Total space: $O\left(\frac{\log\log n \cdot \log\frac{1}{\delta}}{\epsilon^2}\right)$

Morris's Algorithm: Final

- Do these steps $m = O\left(\log \frac{1}{\delta}\right)$ times independently in parallel and output the median answer Z^m.

Maintains t counters $X^1, \dots, X^t$ using $\log \log n$ bits for each

- Initialize $X^{i'}s$ to 0, when an item arrives, increase each X^i by 1 independently with probability $\frac{1}{2^{X^i}}$
- Output $\mathrm{Z} = \frac{1}{t}\left(\sum_{i=1}^{t} 2^{X^i} - 1\right)$

Morris's Algorithm: Final Analysis

Claim: $\Pr[|Z^m - n| > \epsilon\, n] \leq \delta$

- Let Y_i be an indicator r.v. for the event that $|Z_i - n| \leq \epsilon n$, where Z_i is the i-th trial.
- Let $Y = \sum_i Y_i$.
- $\Pr[|Z^m - n| > \epsilon n] \leq \Pr\left[Y \leq \frac{m}{2}\right] \leq \Pr\left[|Y - \mathbb{E}[Y]| \geq \frac{m}{6}\right] \leq \Pr\left[|Y - \mathbb{E}[Y]| \geq \frac{\mu}{4}\right] \leq \exp\left(-c\frac{1}{4^2}\frac{2m}{3}\right) < \exp\left(-c\log\frac{1}{\delta}\right) < \delta$

Thank you!

- Questions?
- **Next time**:
  - More streaming algorithms