CIS 700: “algorithms for Big Data”

Lecture 1: Intro

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

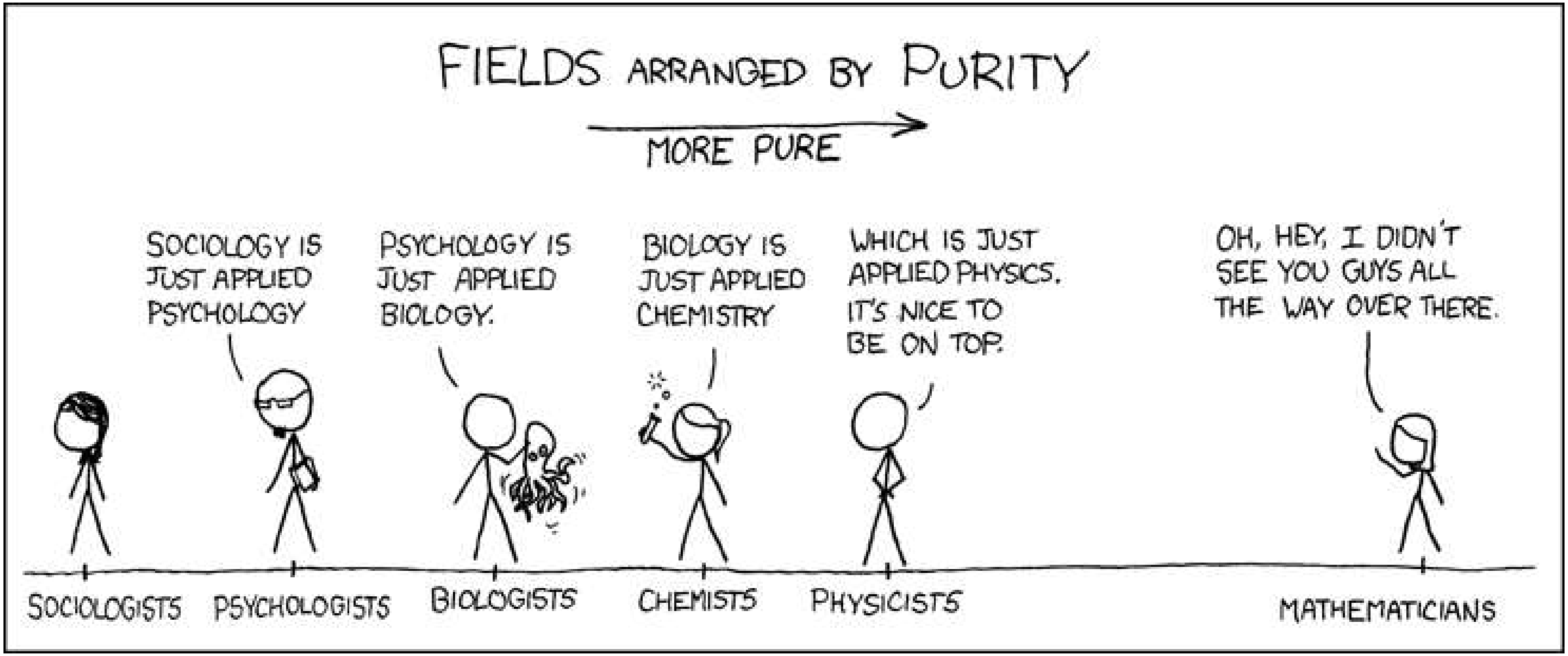
Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

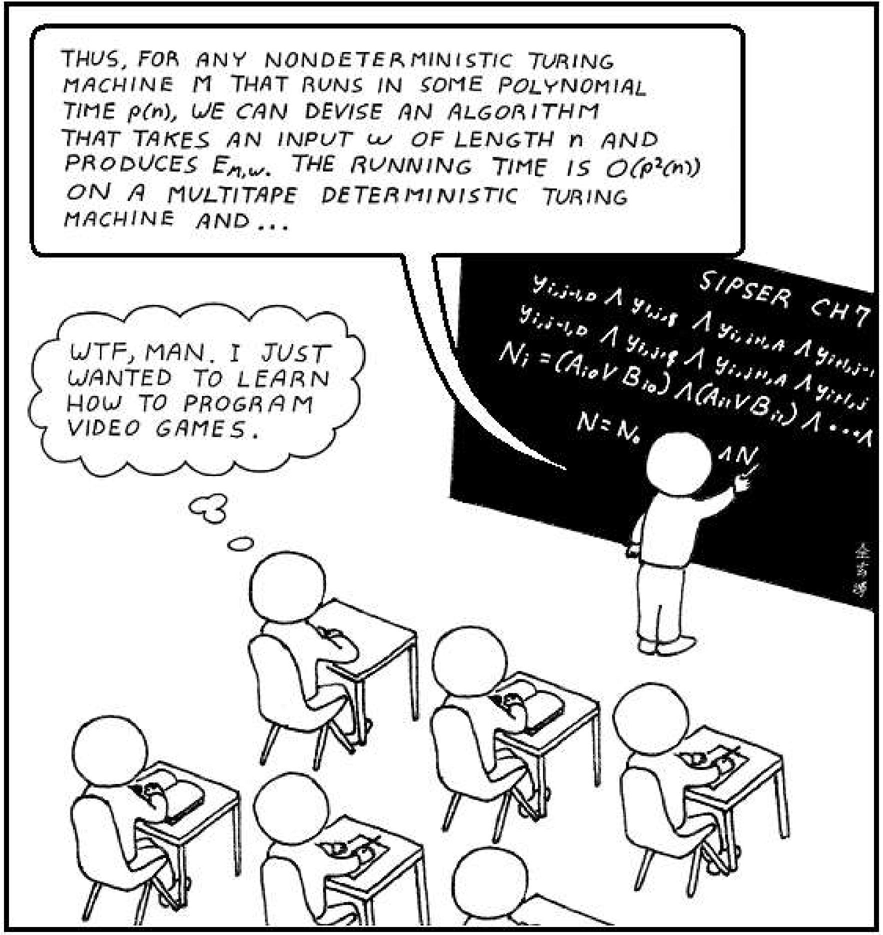
Disclaimers

- A lot of Math!


FIELDS ARRANGED BY PURITY
MORE PURE
SOCIOLOGY IS JUST APPLIED PSYCHOLOGY
PSYCHOLOGY IS JUST APPLIED BIOLOGY.
BIOLOGY IS JUST APPLIED CHEMISTRY
WHICH IS JUST APPLIED PHYSICS. IT'S NICE TO BE ON TOP.
OH, HEY, I DIDN'T SEE YOU GUYS ALL THE WAY OVER THERE.
SOCIOLOGISTS
PSYCHOLOGISTS
BIOLOGISTS
CHEMISTS
PHYSICISTS
MATHEMATICIANS

Disclaimers

- (Almost) no programming!


THUS, FOR ANY NONDETERMINISTIC TURING MACHINE M THAT RUNS IN SOME POLYNOMIAL TIME p(n), WE CAN DEVISE AN ALGORITHM THAT TAKES AN INPUT ω OF LENGTH n AND PRODUCES E_{M,w}. THE RUNNING TIME IS O(p²(n)) ON A MULTITAPE DETERMINISTIC TURING MACHINE AND...
SIPSER CH7
$N_i = (A_{i0} \vee B_{i0}) \wedge (A_{i1} \vee B_{i1}) \wedge \dots \wedge$
$N = N_0 \wedge N$
WTF, MAN. I JUST WANTED TO LEARN HOW TO PROGRAM VIDEO GAMES.

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**): (1), $(\log \mathbf{N})$, $(\quad)$ where $0 < \quad < 1$

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 (~ 1 , with hundreds MB RAM)
    - Fast query access (~ 0.01 per request), “data with queries”

Outline

- Part 1: Streaming Algorithms

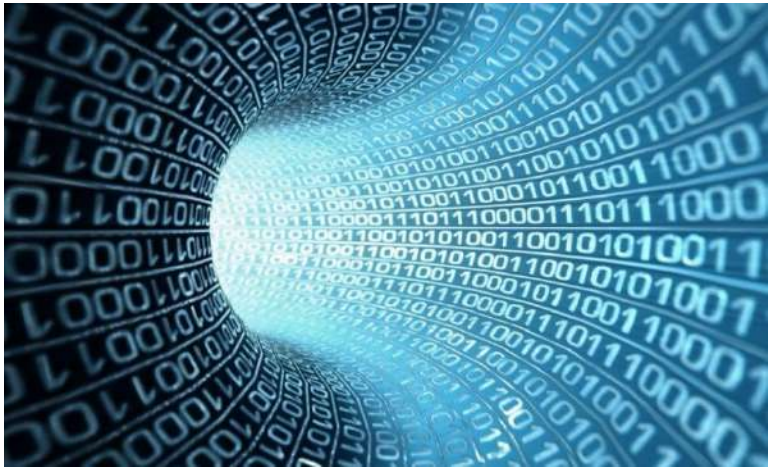
STORM

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 $_1$, …, , arriving one by one:

- (**Easy, “Find a missing player”**)
  - If all $'$ are different and have values between 1 and $+ 1$, which value is missing?
  - You have $(\log\ )$ 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

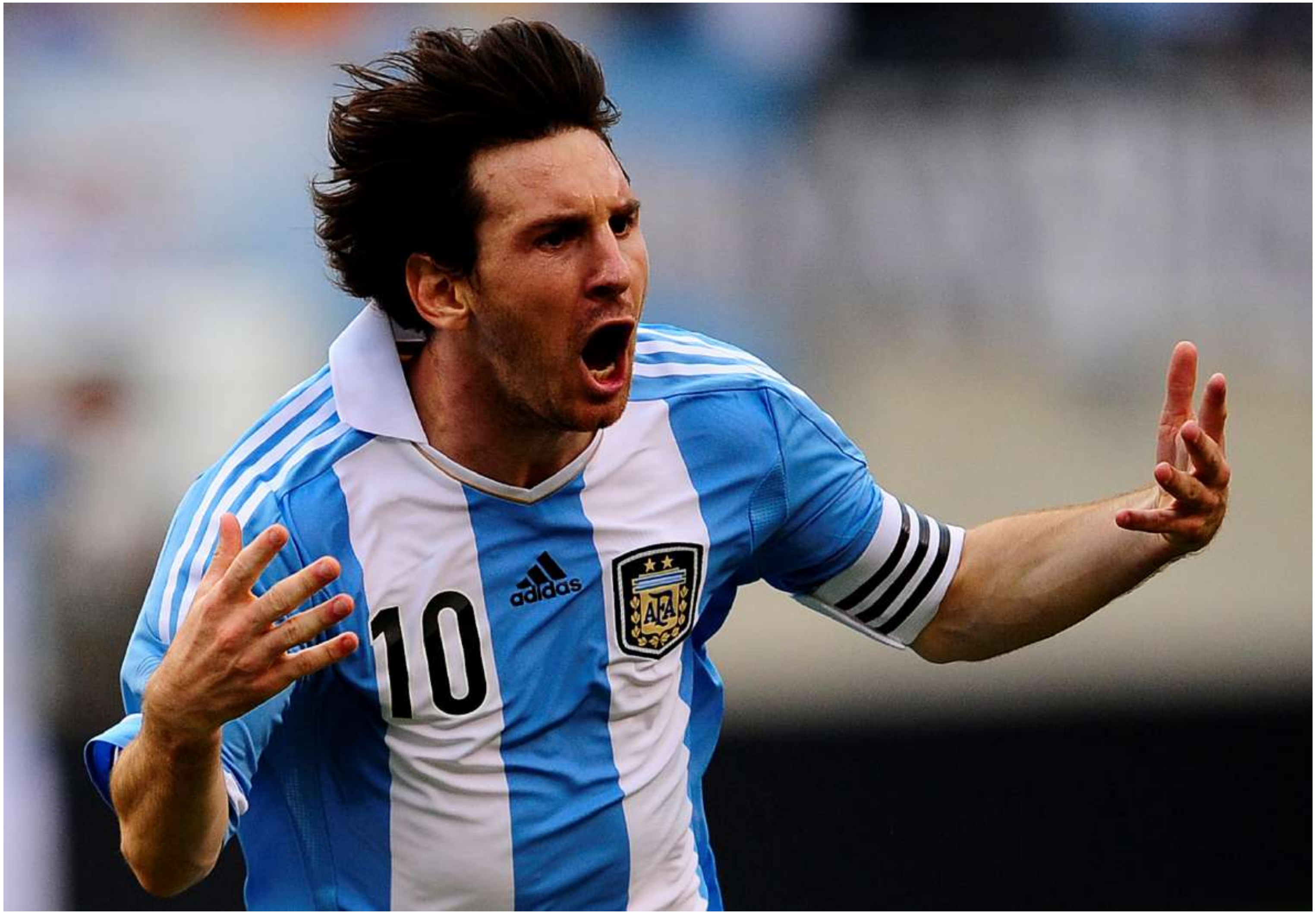
2

6

7

4

Which number was missing?



Puzzle #1

You see a sequence of values $_1$, …, , arriving one by one:

- (**Easy, “Find a missing player”**)
  - If all $'$ are different and have values between 1 and $+ 1$, which value is missing?
  - You have $(\log\ )$ 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 ${}_{1}$, …, , arriving one by one:

- (**Harder, “Keep a random team”**)
  - How can you maintain a uniformly random sample of values out of those you have seen so far?
  - You can store exactly 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 $_1, \ldots,$, arriving one by one:

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

Puzzles

You see a sequence of values $_1, \ldots,$, arriving one by one:

- **(Easy, “Find a missing player”)**
  - If all $'$ are different and have values between 1 and $+1$, which value is missing?
  - You have $(\log)$ space
- **(Harder, “Keep a random team”)**
  - How can you maintain a uniformly random sample of values out of those you have seen so far?
  - You can store exactly items at any time
- **(Very hard, “Count the number of players”)**
  - What is the total number of values up to error $\pm$ $?$

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

- = random variable with values $_1, \ldots,$ $, \ldots$
- Expectation $\mathbb{E}[\quad]$

$$\mathbb{E}[\quad] = \sum_{=1}^{\infty} x_i \cdot \Pr[\boxed{f()} = \quad]$$

- Properties (linearity):

$$\mathbb{E}[\quad] = \mathbb{E}[\quad]$$

$$\mathbb{E}[\quad + \quad] = \mathbb{E}[\quad] + \mathbb{E}[\quad]$$

- Useful fact: if all ≥ 0 and integer then

$$\sum^{\infty}$$

Variance

- Variance $\left[\quad\right] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$

$$\left[\quad\right] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] =$$
$$= \mathbb{E}\left[\ {}^2 - 2 \quad \cdot \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{X}]^2\right]$$
$$= \mathbb{E}[\ {}^2] - 2\mathbb{E}[\quad \cdot \mathbb{E}[\mathbf{X}]] + \mathbb{E}[\mathbb{E}[\mathbf{X}]^2]$$

- $\mathbb{E}[\mathrm{X}]$ is some fixed value (a constant)
- $2\ \mathbb{E}[\quad \cdot \mathbb{E}[\mathbf{X}]] = 2\ \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{X}] = 2\ \mathbb{E}^2[\quad]$
- $\mathbb{E}[\mathbb{E}[\mathbf{X}]^2] = \mathbb{E}^2[\mathbf{X}]$

Independence

- Two random variables and are **independent** if and only if (iff) for every , :

$$\Pr[\ =\ ,\ =\ ] = \Pr[\ =\ ] \cdot \Pr[\ =\ ]$$

- Variables $_1, \ldots,$ are **mutually independent** iff

$$\Pr[\ = {}_1, \ldots, \ =\ ] = \prod_{=1} \Pr[\ =\ ]$$

- Variables $_1, \ldots,$ are **pairwise independent** iff for all pairs i,j

$$\Pr[\ =\ ,\ =\ ] = \Pr[\ =\ ]\Pr[\ =\ ]$$

Conditional Probabilities

- For two events ${}_1$ and ${}_2$:

$$\Pr\left[\ {}_2 \mid \ {}_1\right] = \frac{\Pr[\boxed{f()}\ {}_1 \qquad {}_2]}{\Pr[\boxed{f()}\ {}_1]}$$

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

$$\Pr\left[\ {}_2 = \ {}_2 / \ {}_1 = \ {}_1\right]$$

$$= \frac{\Pr[\boxed{f()}\ {}_1 = \ {}_1 \qquad {}_2 = \ {}_2]}{\Pr\left[\ {}_1 = \ {}_1\right]} \text{ (by definition)}$$

$$= \frac{\Pr\left[\ {}_1 = \ {}_1\right] \ \left[\ {}_2 = \ {}_2\right]}{\Pr[\ {}_1 = \ {}_1]} \text{ (by independence)}$$

Union Bound

For any events ${}_1, \ldots,$:

$$\Pr\left[\;{}_1 \qquad {}_2 \qquad \cdots \qquad \right] \leq \Pr[\;{}_1] + \Pr[\;{}_2] + \cdots$$

$$+ \Pr[\boxed{f()}\;\;]$$

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

Independence and Linearity of Expectation/Variance

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

$$\mathbb{E}\left[\sum_{=1}\right]=\sum_{=1}\mathbb{E}[\quad]$$

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

$$\left[\sum_{=1}\right]=\sum_{=1}\quad[\quad]$$

Part 2: Inequalities

- Markov inequality
- Chebyshev inequality
- Chernoff bound

Markov's Inequality

- For every > 0: $Pr[\ \geq\ \ [\ ]] \leq \frac{1}{}$
- **Proof (by contradiction)** $Pr[\ \geq\ \ [\ ]] > \frac{1}{}$

$$[\ ] = \sum \cdot \Pr[\boxed{f()}\ = \ ]$$ (by definition)

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

Markov’s Inequality

- For every $\; > 0$: $Pr[\;\geq\;[\;]] \leq \frac{1}{\;}$
- **Corollary** $(c' = \;[\;])$ **:**

For every $\;' > 0$: $Pr[\;\geq\;'] \leq \frac{[\;]}{'}$

- **Pro**: always works!
- **Cons**:
  - Not very precise
  - Doesn’t work for the lower tail: $Pr[\;\leq\;[\;]]$

Chebyshev’s Inequality

- For every > 0:

$$\Pr\left[| \quad - \quad [\quad]| \geq \sqrt{\qquad [\quad]}\right] \leq \frac{1}{2}$$

- Proof:

$$\Pr\left[| \quad - \quad [\quad]| \geq \sqrt{\qquad [\quad]}\right]$$

$$= \Pr\left[| \quad - \quad [\quad]|^2 \geq {}^2 \quad [\quad]\right]$$ (by squaring)

$$= \Pr\left[| \quad - \quad [\quad]|^2 \geq {}^2 \; [| \quad - \quad [\quad]|^2]\right]$$ (def. of Var

$$\leq \frac{1}{2}$$ (by Markov’s inequality)

Chebyshev's Inequality

- For every > 0:

$$\Pr\left[\left| \quad - \quad [\;]\right| \geq \sqrt{\quad [\;]}\right] \leq \frac{1}{2}$$

- **Corollary** ($' = \sqrt{\quad [\;]}$):

For every $' > 0$:

$$\Pr\left[\left| \quad - \quad [\;]\right| \geq \;'\right] \leq \frac{[\;]}{'2}$$

Chernoff bound

- Let $_1\ldots$ be independent and identically distributed r.vs with range [0,1] and expectation .

- Then if $=\frac{1}{}\sum$ and $1 > \quad > 0,$

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

Chernoff bound (corollary)

- Let $_1 \dots$ be independent and identically distributed r.vs with range [0, **c**] and expectation .
- Then if $= \frac{1}{}\sum$ and $1 > \quad > 0,$

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

Chernoff v.s Chebyshev

Large values of t is exactly what we need!

Let $_1 \cdots$ be independent and identically distributed r.vs with range [0,1] and expectation . Let $= \frac{1}{} \sum$.

- Chebyshev: $\Pr\left[\left| \quad - \quad \right| \geq \quad \right] = \left(\frac{1}{}\right)$
- Chernoff: $\Pr\left[\left| \quad - \quad \right| \geq \quad \right] = {}^{-\Omega(\ )}$

So is Chernoff always better for us?

- Yes, if we have i.i.d. variables

Answers to the puzzles

You see a sequence of values $_1, \ldots,$, arriving one by one:

- (**Easy**)
  - If all $'$ are different and have values between 1 and $+1$, which value is missing?
  - You have $(\log\ )$ space
  - **Answer**: missing value = $\sum_{=1} - \sum_{=1}$
- (**Harder**)
  - How can you maintain a uniformly random sample of values out of those you have seen so far?

Part 3: Morris’s Algorithm

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

Morris's Algorithm: Alpha-version

Maintains a counter using $log log$ bits

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

Claim: $\left[2 \quad \right] = \quad + 1$

Morris's Algorithm: Alpha-version

Maintains a counter using $log\,log$ bits

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

Claim: $[2\ ] = \ + 1$

- Let the value after seeing items be

$$[2\ ] = \sum_{=0}^{\infty} \Pr[\ _{-1} = \ ]\ [2\ /\ _{-1} = \ ]$$

$$\sum^{\infty} \Pr \quad]\left(\frac{1}{\ }2 + 1\ \left(1 - \frac{1}{\ }\right)2\right)$$

Morris's Algorithm: Alpha-version

Maintains a counter using $log log$ bits

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

Claim: $\left[2^{2}\quad\right] = \frac{3}{2}\;{}^{2} + \frac{3}{2}\quad + 1$

$$\left[2^{2}\quad\right] = \sum_{=0}^{\infty} \Pr[2 \quad - 1 = \quad]\;\left[2^{2}\quad /2 \quad - 1 = \quad\right]$$

$$= \sum^{\infty} \Pr[2 \quad - 1 = \quad]\left(\frac{1}{\;} 4\;{}^{2} + \left(1 - \frac{1}{\;}\right)\;{}^{2}\right)$$

Morris's Algorithm: Alpha-version

Maintains a counter using $log\,log$ bits

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

Morris's Algorithm: Beta-version

Maintains counters $_1, \ldots,$ using $log\, log$ bits for each

- Initialize $'$ to 0, when an item arrives, increase each by 1 independently with probability $\frac{1}{2}$
- Output Z = $\frac{1}{}(\sum_{=1} 2 \quad - 1)$
- $[2 \quad] = \mathrm{n} + 1, \qquad [2 \quad] = \quad (\quad ^2)$

$$[\;] \qquad \left(\frac{1}{}\sum \qquad \right) \qquad (\quad ^2)$$

Morris’s Algorithm: Beta-version

Maintains counters $^{1}, \ldots,$ using $log log$ bits for each

- Output Z = $\frac{1}{}\left(\sum_{=1} 2 \quad - 1\right)$
- $[\ ] = \left(\frac{1}{}\sum_{=1} 2 \quad - 1\right) = \left(\frac{2}{}\right)$
- Claim: If $\geq \frac{}{2}$ then $\Pr\left[\left| \quad - \quad \right| > \quad \right] < 1/3$

$[\ ] \quad \left(\ ^{2}\right) \quad 1$

Morris's Algorithm: Final

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

Morris's Algorithm: Final

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

Maintains counters ^{1}, …, using $log log$ bits for each

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

Morris’s Algorithm: Final Analysis

Claim: $\Pr[| \quad - \quad | > \quad] \leq$

- Let be an indicator r.v. for the event that $| \quad - \quad | \leq$, where is the i-th trial.
- Let $= \sum$.
- $\Pr[| \quad - \quad | > \quad] \leq \Pr\left[\quad \leq \frac{\quad}{2} \right]$

Thank you!

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