Sublinear Algorihms for Big Data

Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

Part 0: Introduction

- Disclaimers
- Logistics
- Materials
- …

Name

Correct:

- Grigory
- Gregory (easiest and highly recommended!)

Also correct:

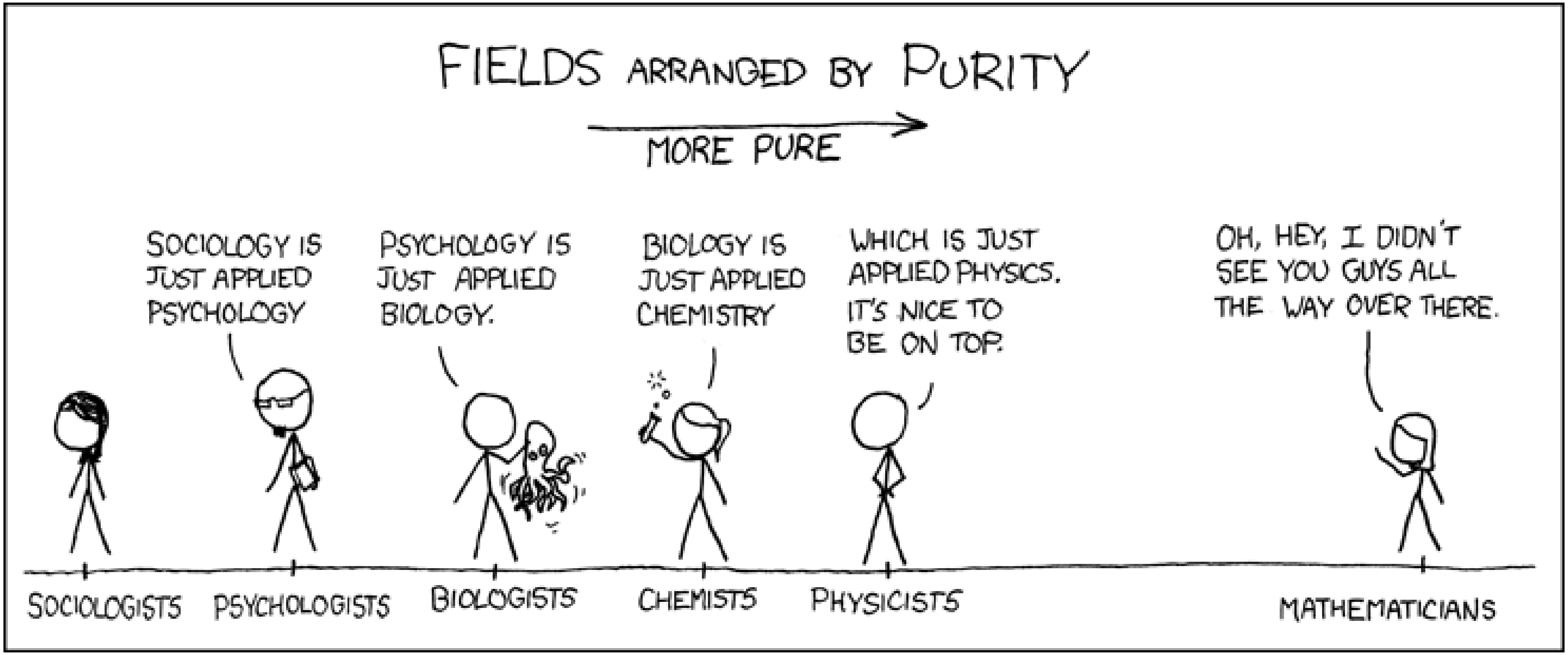
- Dr. Yaroslavtsev (I bet it’s difficult to pronounce)

Wrong:

- Prof. Yaroslavtsev (Not any easier)

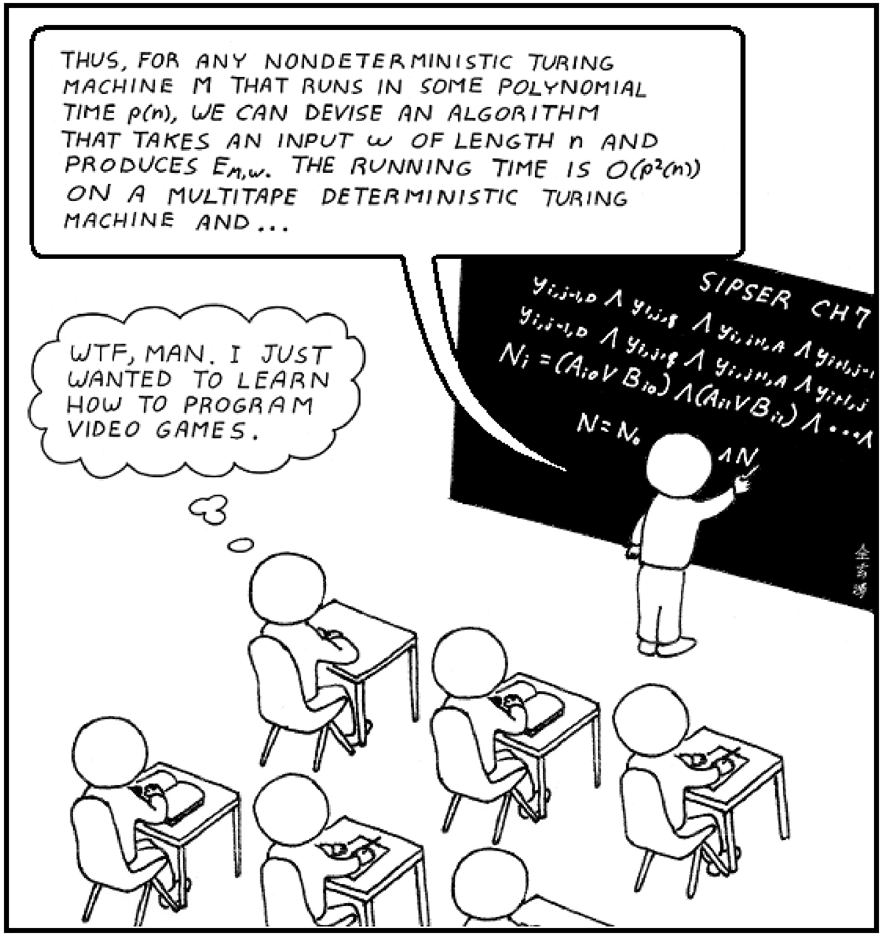
Disclaimers

- A lot of Math!



Disclaimers

- No programming!



Disclaimers

- 10-15 times longer than “Fuerza Bruta”, soccer game, milonga…

Big Data

- Data
- Programming and Systems
- **Algorithms**
- **Probability and Statistics**

Sublinear Algorithms

n = size of the data, we want $o(n)$, not $O(n)$

- Sublinear Time
  - Queries
  - Samples
- Sublinear Space
  - Data Streams
  - Sketching
- Distributed Algorithms
  - Local and distributed computations
  - MapReduce-style algorithms

Why is it useful?

- Algorithms for big data used by big companies (ultra-fast (randomized algorithms for approximate decision making)
  - Networking applications (counting and detecting patterns in small space)
  - Distributed computations (small sketches to reduce communication overheads)
- Aggregate Knowledge: startup doing streaming algorithms, acquired for $150M
- Today: Applications to soccer

Course Materials

- Will be posted at the class homepage:

  http://grigory.us/big-data.html

- Related and further reading:
  - **Sublinear Algorithms** (MIT) by Indyk, Rubinfeld
  - **Algorithms for Big Data** (Harvard) by Nelson
  - **Data Stream Algorithms** (University of Massachusetts) by McGregor
  - **Sublinear Algorithms** (Penn State) by Raskhodnikova

Course Overview

- Lecture 1
- Lecture 2
- Lecture 3
- Lecture 4
- Lecture 5

3 hours = 3 x (45-50 min lecture + 10-15 min break).

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
- 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 Spanish: Hola, Gracias, Bueno, Por favor, Bebida, Comida, Jamon, Queso, Gringo, Chica, Amigo, …
- Basic Probability:
  - Probability, events, random variables
  - Expectation, variance / standard deviation
  - Conditional probability, independence, pairwise independence, mutual independence

Expectation

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

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

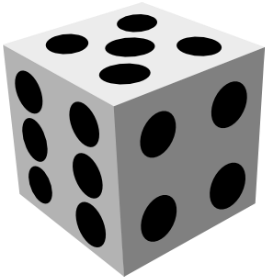
- Properties (linearity):

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

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

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

Expectation



- Example: dice has values $1, 2, \ldots, 6$ with probability 1/6

$$\mathbb{E}[\text{Value}] = \sum_{i=1}^{6} i \cdot \Pr[Value = i]$$

$$= \frac{1}{6}\sum_{i=1}^{6} i = \frac{21}{6} = 3.5$$

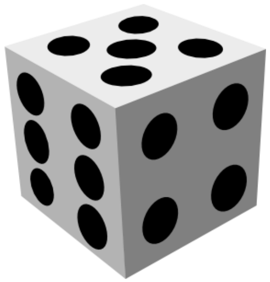
Variance

- 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: $Var[c\boldsymbol{X}] \ = \ c^2 Var[\boldsymbol{X}]$

Variance



- Example (Variance of a fair dice):

$$\mathbb{E}[Value] = 3.5$$

$$Var[Value] = \mathbb{E}[(Value - \mathbb{E}[Value])^2]$$

$$= \mathbb{E}[(Value - 3.5)^2]$$

$$= \sum_{i=1}^{6}(i - 3.5)^2 \cdot Pr\,[Value = i]$$

$$= \frac{1}{6}\sum_{i=1}^{6}(i - 3.5)^2$$

$$= \frac{1}{6}\,[\left(1 - 3.5\right)^2 + \left(2 - 3.5\right)^2 + \left(3 - 3.5\right)^2$$

$$+ (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2]$$

$$= \frac{1}{6}[6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25\,]$$

$$= \frac{8.75}{3} \approx 2.917$$

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 $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are **mutually independent** iff
$$\Pr[\boldsymbol{X}_\mathbf{1} = x_1, \dots, \boldsymbol{X}_n = x_n] = \prod_{i=1}^{n} \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i]$$
- Variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are **pairwise independent** iff for all pairs i,j
$$\Pr\left[\boldsymbol{X}_{\boldsymbol{i}} = x_i, \boldsymbol{X}_j = x_j\right] = \Pr[\boldsymbol{X}_{\boldsymbol{i}} = x_i] \Pr\left[\boldsymbol{X}_{\boldsymbol{j}} = x_j\right]$$

Independence: Example

Independent or not?

- Event E_1 = Argentina wins the World Cup
- Event E_2 = Messi becomes the best striker

Independent or not?

- Event E_1 = Argentina wins against Netherlands in the semifinals
- Event E_2 = Germany wins against Brazil in the semifinals

Independence: Example

- Ratings of mortgage securities
  - AAA = 1% probability of default (over X years)
  - AA = 2% probability of default
  - A = 5% probability of default
  - B = 10% probability of default
  - C = 50% probability of default
  - D = 100% probability of default
- You are a portfolio holder with 1000 AAA securities?
  - Are they all independent?
  - Is the probability of default $(0.01)^{1000} = 10^{-2000}$?

Conditional Probabilities

- For two events E_1 and E_2:

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

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

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

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

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

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

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

Union Bound: Example

Events “Argentina wins the World Cup” and “Messi becomes the best striker” are **not independent**, but:

Pr[“Argentina wins the World Cup” or “Messi becomes the best striker”] $\leq$

Pr[“Argentina wins the World Cup”] +

Pr[“Messi becomes the best striker”]

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

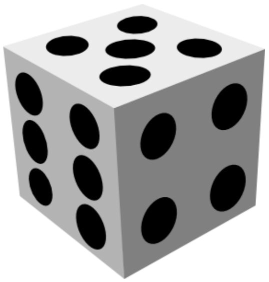
Markov's Inequality

- 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: $\Pr\left[\boldsymbol{X} \leq c\ \mathbb{E}[\boldsymbol{X}]\right]$

Markov Inequality: Example



Markov 1: For every $c > 0$:

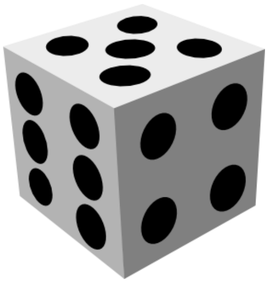
$$\Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

- Example:

$$\Pr\left[Value \geq 1.5 \cdot\ \mathbb{E}[Value]\right] = \Pr[Value \geq 1.5 \cdot 3.5] =$$
$$\Pr[Value \geq 5.25] \leq \frac{1}{1.5} = \frac{2}{3}$$

$$\Pr\left[Value \geq 2 \cdot\ \mathbb{E}[Value]\right] = \Pr[Value \geq 2 \cdot 3.5]$$
$$= \Pr[Value \geq 7] \leq \frac{1}{2}$$

Markov Inequality: Example



Markov 2: For every $c > 0$:

$$\Pr[\boldsymbol{X} \geq c\,] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c}$$

- Example:

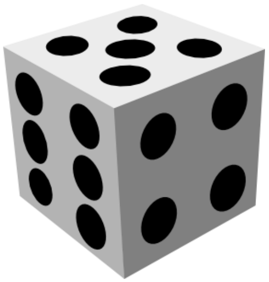
$$\Pr[Value \geq 4] \leq \frac{\mathbb{E}[Value]}{4} = \frac{3.5}{4} = 0.875\ (= 0.5)$$

$$\Pr[Value \geq 5] \leq \frac{\mathbb{E}[Value]}{5} = \frac{3.5}{5} = 0.7 \quad (\approx 0.33)$$

$$\Pr[Value \geq 6] \leq \frac{\mathbb{E}[Value]}{6} = \frac{3.5}{6} \approx 0.58 \quad (\approx 0.17)$$

$$\Pr[Value \geq 3] \leq \frac{\mathbb{E}[Value]}{3} = \frac{3.5}{3} \approx 1.17 \quad (= 1)$$

Markov Inequality: Example



Markov 2: For every $c > 0$:

$$\Pr[\boldsymbol{X} \geq c\ ] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c}$$

- $\Pr[Value \leq z] = \Pr[(7\ -\ Value) \geq z]$:

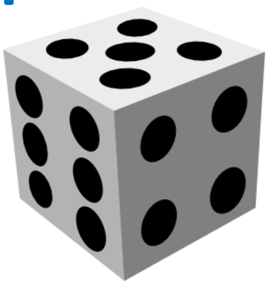
$$\Pr[Value \leq 3] \leq \frac{\mathbb{E}[7 - Value]}{4} = \frac{3.5}{4} = 0.875\ (= \mathbf{0.5})$$

$$\Pr[Value \leq 2] \leq \frac{\mathbb{E}[7 - Value]}{5} = \frac{3.5}{5} = 0.7 \quad (\approx \mathbf{0.33})$$

$$\Pr[Value \leq 1] \leq \frac{\mathbb{E}[7 - Value]}{6} = \frac{3.5}{6} \approx 0.58 \ \ (\approx \mathbf{0.17})$$

$$\Pr[Value \leq 4] \leq \frac{\mathbb{E}[7 - Value]}{3} = \frac{3.5}{3} \approx 1.17 \qquad (= \mathbf{1})$$

Markov + Union Bound: Example



Markov 2: For every $c > 0$:

$$\Pr[\boldsymbol{X} \geq c\,] \leq \frac{\mathbb{E}[\boldsymbol{X}]}{c}$$

- Example:

$$\Pr[Value \geq 4 \; or \; Value \leq 3] \leq$$
$$\Pr[Value \geq 4] + Pr[Value \leq 3] \; = 2 \cdot 0.875 = 1.75$$
$$(\boldsymbol{= 1})$$

$$\Pr[Value \geq 5 \; or \; Value \leq 2] \leq 2 \cdot 0.7 = 1.4$$
$$(\boldsymbol{\approx 0.66})$$

$$\Pr[Value \geq 6 \; or \; Value \leq 1] \leq 2 \cdot 0.58 \approx 1.16$$
$$(\boldsymbol{\approx 0.33})$$

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:

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

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

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

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

Chebyshev's Inequality

- For every $c > 0$:

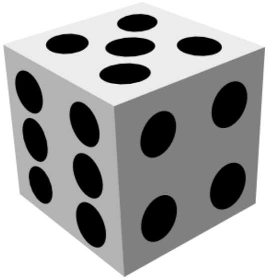
$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

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

For every $c' > 0$:

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

Chebyshev: Example



- For every $c' > 0$: $\Pr[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2}$

$$\mathbb{E}[Value] = 3.5; \;\; Var\,[Value] \approx 2.91$$

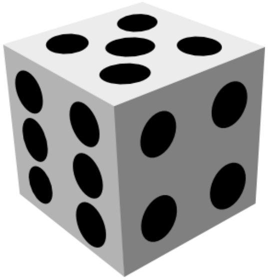
$$\Pr[Value \geq 4 \; or \; Value \leq 3] =$$

$$Pr[|Value - 3.5| > 0.5] \leq \frac{2.91}{0.5^2} \approx 11.64 \; (= \mathbf{1})$$

$$\Pr[Value \geq 5 \; or \; Value \leq 2] \leq \frac{2.91}{1.5^2} \approx 1.29 \quad (\approx \mathbf{0.66})$$

$$\Pr[Value \geq 6 \; or \; Value \leq 1] \leq \frac{2.91}{2.5^2} \approx 0.47 \quad (\approx \mathbf{0.33})$$

Chebyshev: Example



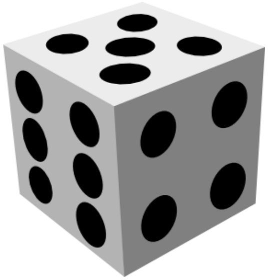
- Roll a dice 10 times:

$$Value_{10} = \text{Average value over 10 rolls}$$
$$\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$$

- X_i = value of the i-th roll, $\boldsymbol{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$
- Variance ($=$ by linearity for **independent** r.vs):

$$Var[\boldsymbol{X}] = Var\left[\frac{1}{10}\sum_{i=1}^{10} X_i\right] = \frac{1}{100} Var\left[\sum_{i=1}^{10} X_i\right]$$
$$= \frac{1}{100}\sum_{i=1}^{10} Var[X_i] \approx \frac{1}{100} \cdot 10 \cdot 2.91 = 0.291$$

Chebyshev: Example



- Roll a dice 10 times:

$$Value_{10} = \text{Average value over 10 rolls}$$
$$\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$$

- $Var[Value_{10}] = 0.291$ (if n rolls then 2.91 / n)
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq \frac{0.291}{0.5^2} \approx 1.16$
- $\Pr[Value_n \geq 4 \; or \; Value_n \leq 3] \leq \frac{2.91}{n \cdot 0.5^2} \approx \frac{11.6}{n}$

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 $1 > \delta > 0$,

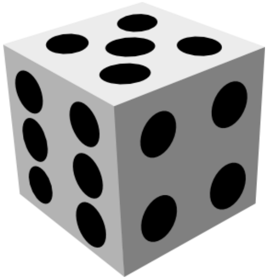
$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3}\right)$$

Chernoff bound (corollary)

- Let $X_1 \ldots X_t$ 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: Example



- $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3c}\right)$
- Roll a dice 10 times:

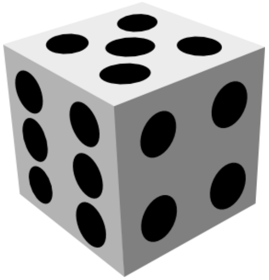
  $Value_{10}$ = Average value over 10 rolls

  $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] = ?$

  - $X = Value_{10}, \; t = 10, \; \mathrm{c} = 6$
  - $\mu = \mathbb{E}[X_i] = 3.5$
  - $\delta = \frac{0.5}{3.5} = \frac{1}{7}$
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq 2\exp\left(-\frac{3.5 \cdot 10}{3 \cdot 6 \cdot 49}\right) =$

$2\exp\left(-\frac{35}{882}\right) \approx 2 \cdot 0.96 = 1.92$

Chernoff: Example



- $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3c}\right)$
- Roll a dice 1000 times:

  $Value_{1000}$ = Average value over 1000 rolls

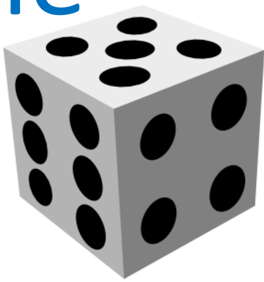
  $\Pr[Value_{1000} \geq 4 \; or \; Value_{1000} \leq 3] = ?$

  – $X = Value_{1000}, \; t = 1000, \; \mathrm{c} = 6$

  – $\mu = \mathbb{E}[X_i] = 3.5$

  – $\delta = \frac{0.5}{3.5} = \frac{1}{7}$
- $\Pr[Value_{10} \geq 4 \; or \; Value_{10} \leq 3] \leq$ $2\exp\left(-\frac{3.5 \cdot 1000}{3 \cdot 6 \cdot 49}\right) = 2\exp\left(-\frac{3500}{882}\right) \approx$ $2 \cdot \exp(-3.96) \approx 2 \cdot 0.02 = 0.04$

Chernoff v.s Chebyshev: Example



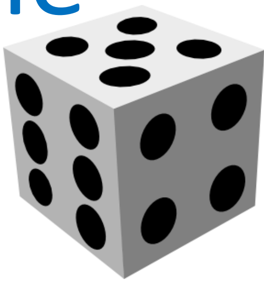
Let $\sigma = Var[X_i]$:

- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2} = \frac{\sigma}{t\, c'^2}$
- Chernoff: $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu t \delta^2}{3c}\right)$

If t is very big:

- Values $\mu, \sigma, \delta, c, c'$ are all constants!
  - Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{t}\right)$
  - Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(t)}$

Chernoff v.s Chebyshev: Example



Large values of t is exactly what we need!

- 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 $a_1, \ldots, a_n$, 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 $a_1, \ldots, a_S$. 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

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

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

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

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

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

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

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

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

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 $X^1, \ldots, X^t$ 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 $\mathrm{Z} = \frac{1}{t}\left(\sum_{i=1}^{t} 2^{X^i} - 1\right)$

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 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
  - Testing distributions