APPROXIMATION ALGORITHMS

RANDOM SAMPLING & PRIORITY SAMPLING

RASMUS PAGH
UNIVERSITY OF COPENHAGEN

TODAY

- WARM-UP: POLLS
- RESERVOIR SAMPLING
- TAIL INEQUALITIES AND ESTIMATORS
- WEIGHTED RANDOM SAMPLING
- PRIORITY SAMPLING

POLLS

CASE 1: - 2000 CALLS ARE MADE TO RANDOM PHONE NOS
- 1000 PEOPLE RESPOND
- 60% SAY THEY WILL VOTE FOR CURRENT GOVT.
  40% SAY THEY WILL VOTE FOR OPPOSITION

HOW MANY WILL VOTE FOR GOVT. IN POPULATION AT LARGE?

CASE 2: A NEWSPAPER CALLS 100 COMPANIES AND ASK ABOUT THE PROBABILITY THEY WILL HIRE IN 2021
- PROBABILITIES $p_1, \ldots, p_{100}$ ARE REPORTED

WHAT FRACTION OF COMPANIES DO WE EXPECT WILL HIRE?

RE. CASE 1 NEED TO KNOW RESPONSE RATE (SAMPLING PROBABILITY) FOR THE SETS $S_{Govt.}$ OF GOVT. VOTERS AND S_{Opp} OF OPP. VOTERS

VOTE \ RESPONSE	YES	NO
GOVT.	30% 30%	20% 30%
OPP.	20% 20%	30% 20%

POSSIBLE DISTRIBUTION

ANOTHER POSSIBLE DISTRIBUTION

CAN DISTINGUISH IF WE KNOW HOW TO WEIGH EACH ANSWER

RANDOM VOTER X

SET A: VOTERS WHO ANSWER

$$Pr[x \in A \mid x \in S_{Govt}] = Pr[x \in A \cap S_{Govt}] / Pr[x \in S_G]$$

RATIO BETWEEN OBS. AND ANSWER

CAN BE OBSERVED

WHAT WE WANT

RE. CASE 2

$X_i = \begin{cases} 1 & \text{if COMPANY } i \text{ HIRES} \\ 0 & \text{OTHERWISE} \end{cases}$

$Y_i = \begin{cases} 1 & \text{if COMPANY } i \text{ IS SAMPLED} \\ 0 & \text{OTHERWISE} \end{cases}$

$$E[X_i] = \Pr[\text{COMPANY } i \text{ HIRES}] = p_i$$

$$E[\#\text{HIRES AMONG 100}] = E\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} E[X_i] = \sum_{i=1}^{100} p_i$$

SUPPOSE n COMPANIES IN TOTAL:

$$E[\text{TOTAL \#HIRES}] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \Pr[X_i Y_i = 1] \,/\, \Pr[Y_i = 1]$$

$$= \left(\sum_{i=1}^{100} p_i\right) \Big/ \left(\frac{100}{n}\right) = \frac{n}{100} \cdot \sum_{i=1}^{100} p_i .$$

FISCHER-YATES SHUFFLE

RESERVOIR SAMPLING

SPACE s

MAINTAIN: - COUNTER n
- RANDOM PERMUTATION OF n ELEMENTS

UNDER INSERTION OF NEW ELEMENTS

INITIALLY $n = 2$

ARRAY $A = [87, 42]$

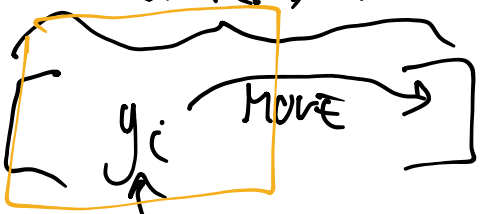
MOVE TO NEW ENTRY

CHOOSE RANDOM POSITION

INSERTIONS: 42, 87, ... x_{n+1}

KEEP ONLY FIRST s ENTRIES IN SHUFFLE.

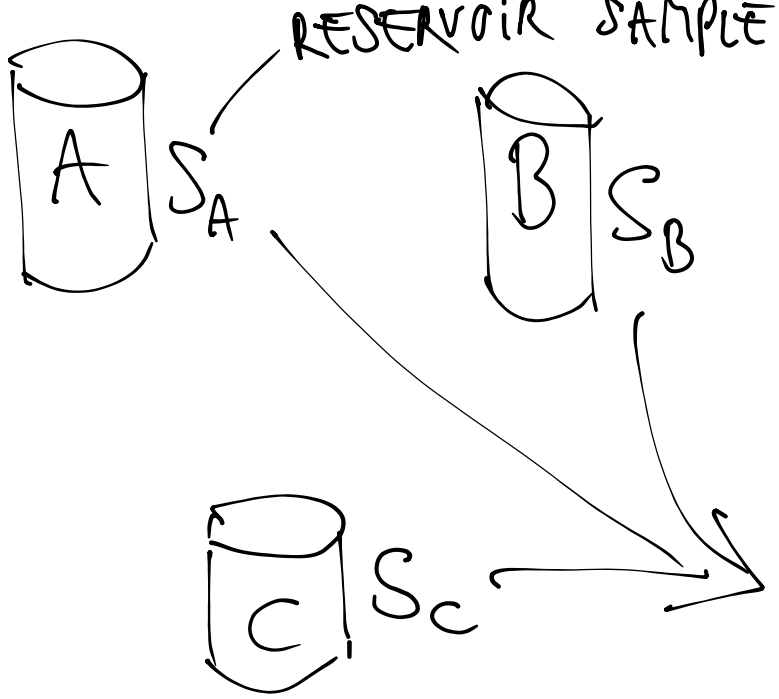
EXTEND TO $n+1$



RANDOM INDEX i IN $\{1, \ldots, n+1\}$

INVARIANT: A IS A RANDOM PERMUTATION OF $x_1, \ldots, x_n$

SAMPLING DISTRIBUTED DATA



COMPUTERS WITH DATA SETS A, B, C. (DISJOINT)

PAUSE AND THINK:
HOW CAN S_A, S_B, S_C BE COMBINED INTO A RESERVOIR SAMPLE OF $A \cup B \cup C$?

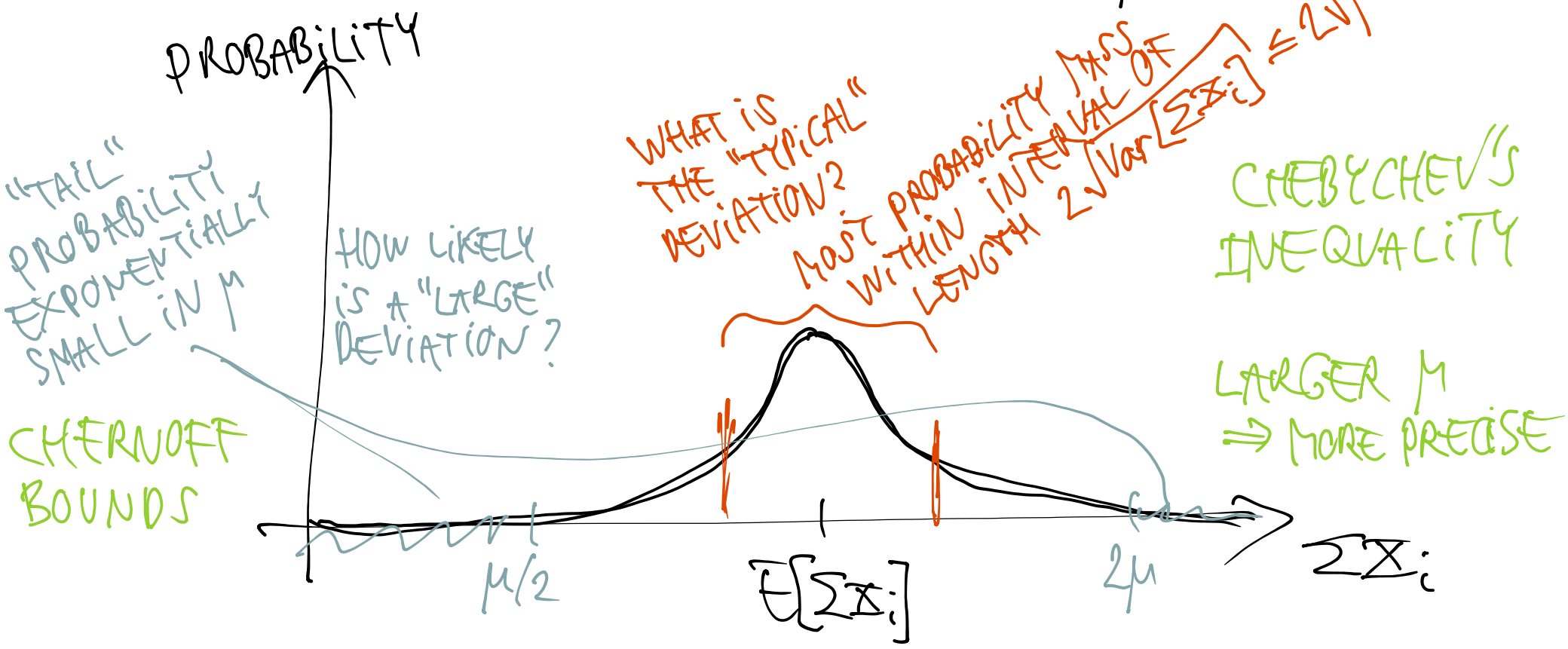
SIZE s

WANT $\Pr[x \in S_{A \cup B \cup C}] = \frac{s}{|A \cup B \cup C|}$.

TAIL INEQUALITIES

SETTING: HAVE RANDOM (INDEPENDENT) $X_1, \ldots, X_n \in \{0,1\}$, WANT TO KNOW $\mu = \mathbb{E}[\sum_i X_i]$

HOW CLOSE IS THE OBSERVED SUM TO μ?



TAIL BOUNDS & APPROXIMATION

CHEBYCHEV'S INEQUALITY:

$$\Pr[|X - E[X]| > k] \leq \mathrm{Var}(X)/k^2$$

(EXAMPLE)
APPROXIMATION GUARANTEE:

IF $\mathrm{Var}(X) \leq E[X]$, X IS BETWEEN $\frac{1}{2}E[X]$ AND $2E[X]$ WITH PROBABILITY $1 - 4/E[X]$.

CHERNOFF BOUND:

IF $X = \sum_{i=1}^{n} X_i$, FOR INDEP. $X_i \in \{0,1\}$

$$\Pr[X \leq (1-\varepsilon)E[X]] \leq \exp\left(-\frac{\varepsilon^2}{2}E[X]\right)$$

$$\Pr[X \geq (1+\varepsilon)E[X]] \leq \exp\left(\frac{-\varepsilon^2}{4}E[X]\right)$$

APPROXIMATION GUARANTEE:

IF $E[X] \geq 4\ln(2/\delta)/\varepsilon^2$ THEN

$X \in [(1-\varepsilon)E[X], (1+\varepsilon)E[X]]$

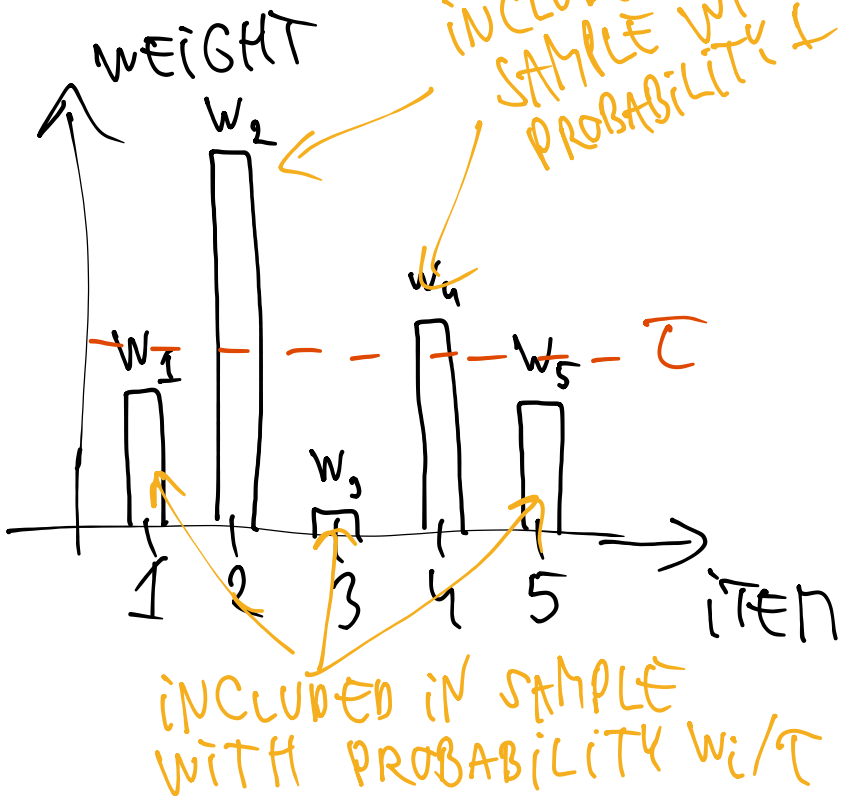
WITH PROBABILITY $\geq 1 - \delta$

WEIGHTED RANDOM SAMPLING

- ITEM i COMES WITH WEIGHT $w_i \geq 0$
- WANT TO SAMPLE HIGHER WEIGHT ITEMS MORE FREQUENTLY

PROPORTIONAL TO WEIGHT

EXAMPLE



$S=1$: SAMPLE i WITH PROB. $\frac{w_i}{\sum_{j=1}^{n} w_j}$

$\tau = \sum w_i$

NORMALIZING FACTOR

$S=5$: SAMPLE ALL !

$\tau = 0$

GENERAL S:

FIND $\tau \geq 0$ SUCT THAT

$$\sum_{i=1}^{n} \min(1, w_i/\tau) = S$$

SAMPLE i WITH PROB. $p_i = \min(1, w_i/\tau)$

ESTIMATORS FROM WEIGHTED SAMPLES

INTERESTED IN $\mu = \sum_{i \in Q} w_i$ ← SET NOT KNOWN BEFORE SAMPLING

EFFICIENT UPDATES, MERGING: SEE BOOK (COMPLICATED)

CAN COMPUTE $\sum_{i \in S \cap Q} w_i$ FROM SAMPLE.

HORVITZ-THOMPSON ESTIMATOR

UNBIASED ESTIMATOR $\hat{\mu} = \sum_{i=S \cap Q} w_i / p_i = \sum_{i \in Q} 1(i \in S) \cdot w_i / p_i$

$$E[\hat{\mu}] = \sum_{i \in Q} \Pr[i \in S] \, w_i / p_i = \sum_{i \in Q} p_i w_i / p_i = \mu$$

FACT: WEIGHTED SAMLING PROPORTIONAL TO WEIGHT GIVES THE SMALLEST POSSIBLE VARIANCE AMONG ALL SAMPLING METHODS USING SPACE S

PRIORITY SAMPLING

A SIMPLER WAY OF DOING WEIGHTED SAMPLING

PRIORITY OF ITEM x IS COMPUTED AS FOLLOWS:

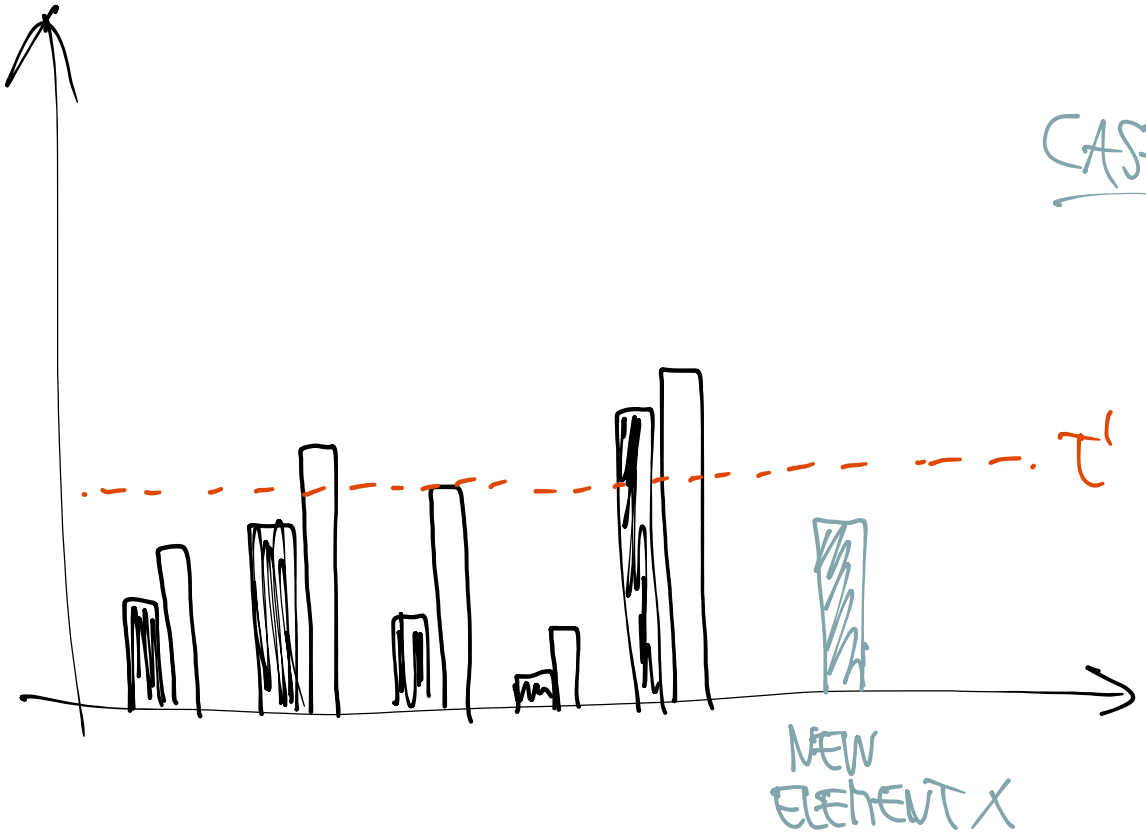
- SAMPLE $\alpha_x \in (0,1)$ UNIFORMLY
- LET $q_x = w_x / \alpha_x$

SAMPLE: ITEMS WITH s HIGHEST PRIORITIES } MAINTAIN IN A PRIORITY QUEUE
THRESHOLD τ EQUALS $(s+1)$ST HIGHEST PRIORITY. }

ESTIMATOR: $\hat{\mu} = \sum_{x \in Q \cap S} \max(w_x, \tau)$

$$E[\hat{\mu}] = \sum_{x \in Q} w_x \Leftarrow E[\max(w_x, \tau) \mid \text{sth HIGHEST PRIORITY FOR ELEM. } y \neq x \text{ IS } \tau]$$
$$\simeq w_x$$

ILLUSTRATION



WEIGHT

PRIORITY (≥WEIGHT)

CASE 1: $w_x \geq \tau'$: IN SAMPLE WITH PROB. 1

CASE 2: $w_x < \tau'$: IN SAMPLE IFF $w_x/\alpha_x > \tau'$

↑

PROBABILITY w_x/τ

PRIORITY SAMPLING IS DUE TO DUFFIELD, LUND & THORUP

KNOWN TO BE OPTIMAL UP TO A SINGLE SAMPLE.