APPROXIMATION ALGORITHMS

SUMMARIES FOR ORDERED DATA

RASMUS PAGH
UNIVERSITY OF COPENHAGEN

TODAY

- MOTIVATING EXAMPLE
- SAMPLING SOLUTION
- Q-DIGEST
- DYADIC COUNT-MIN

} DATA FROM FIXED DOMAIN

- KLL SUMMARY (ANY ORDERED DOMAIN)

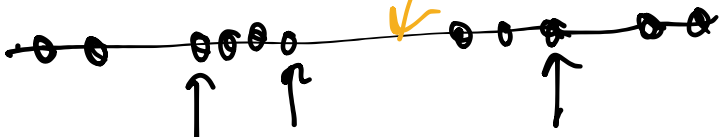
MOTIVATING EXAMPLE

CONSIDER A SET S OF TIME STAMPS (E.G. LINK CLICKS)
WANT TO APPROXIMATE NUMBER OF TIMESTAMP IN A GIVEN INTERVAL.

CLASSIC APPROACH:

$$\text{rank}(x) = |\{y \in S \mid y \leq x\}|$$

QUANTILES



CAN APPROXIMATE RANK FROM QUANTILES UP TO ERROR εN

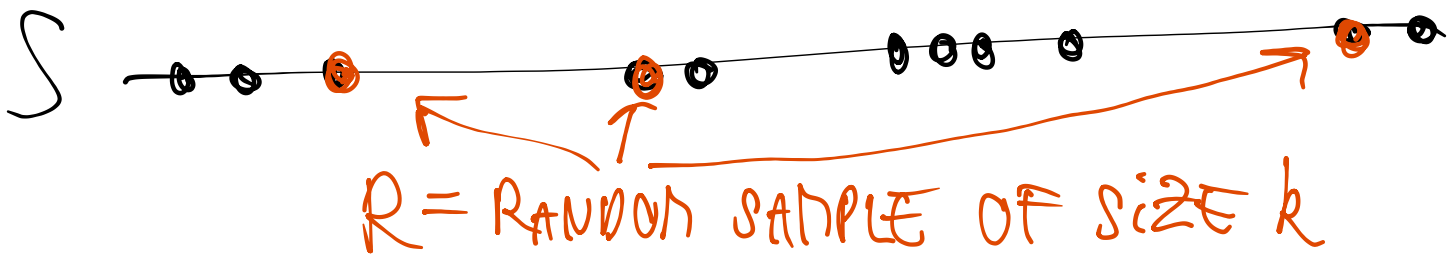
STORE EVERY εN'TH ELEMENT
("QUANTILES", E.G. 10%, 20%, ..., 90%)

REQUIRES SPACE N TO COMPUTE, $\frac{1}{\varepsilon}$ TO STORE

TWO SETTINGS

- NEXT → $S \in \{0, \ldots, U-1\}$, $U \in \mathbb{N}$
- LATER → S IS ANY ORDERED SET

WHY NOT JUST USE SAMPLING?



$$E[|\{y \in R \mid y \le x\}|] = \frac{k}{N}|\{y \in S \mid y \le x\}| = \frac{k}{N}\,\mathrm{rank}(x)$$

$$\Downarrow$$

$$\mathrm{rank}(x) \approx \frac{N}{k}|\{y \in R \mid y \le x\}|$$

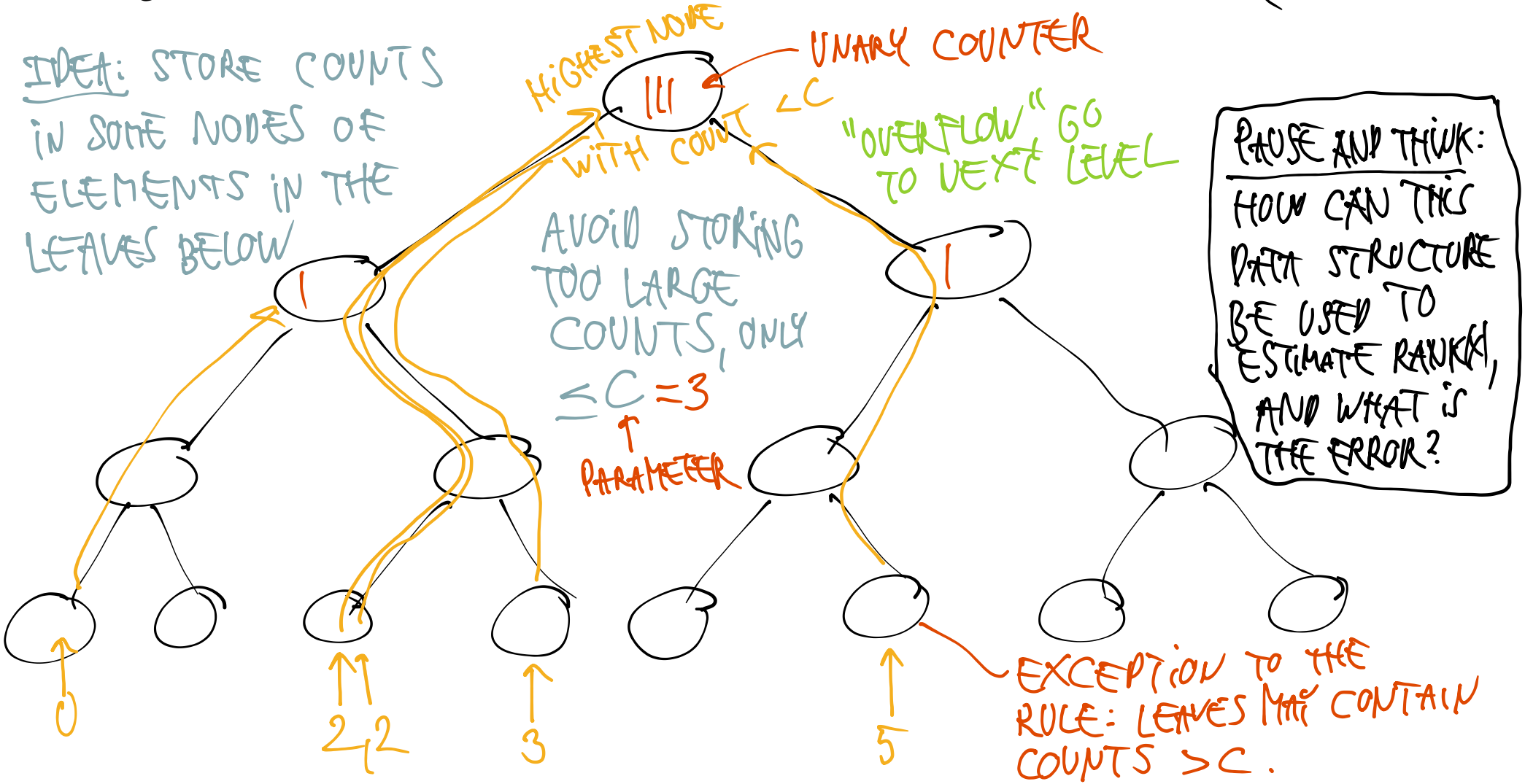
EXERCISE:
DERIVE THE SIZE OF R NEEDED TO MAKE THE ERROR $\le \varepsilon N$ FOR ALL X

Q-DIGEST

$S \subseteq \{0, \ldots, U-1\}$

ASSUME U IS A POWER OF 2 (WITHOUT LOSS OF GEN)

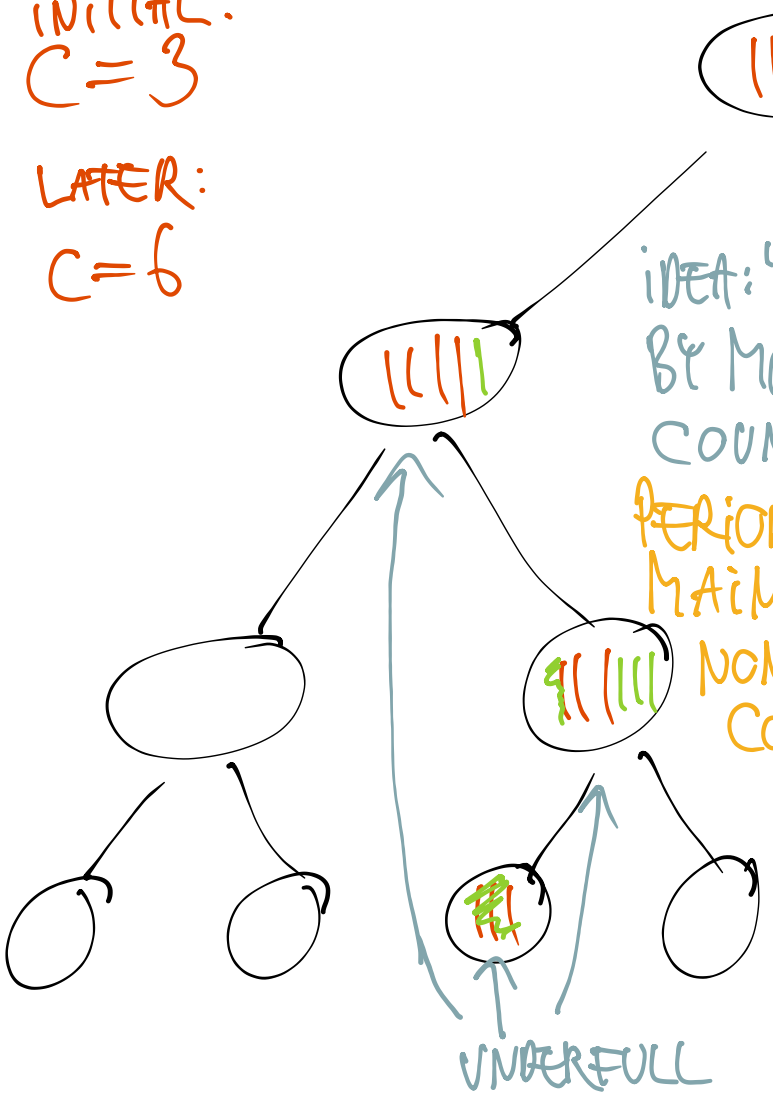
COMPLETE BINARY TREE WITH U LEAVES $(U = 8)$



REDUCING THE SIZE OF Q-DIGEST

INITIAL:
$C = 3$

LATER:
$C = 6$



IDEA: "COMPRESS" BY MOVING COUNTS UP PERIODICALLY TO MAINTAIN $O(N/C)$ NONZERO COUNTS

NEED CAPACITY $C \geq \varepsilon N / \log U$ TO ENSURE ERROR $\leq \varepsilon N$

ONLY $O((\log U)/\varepsilon)$ NODES WITH NONZERO COUNTS IF MOST OF THE CAPACITY IS USED

$O(N/C)$ NODES IN TOTAL.

AFTER COMPRESS: ALL NODES WITH COUNT $< C$ HAVE A PARENT WITH COUNT C

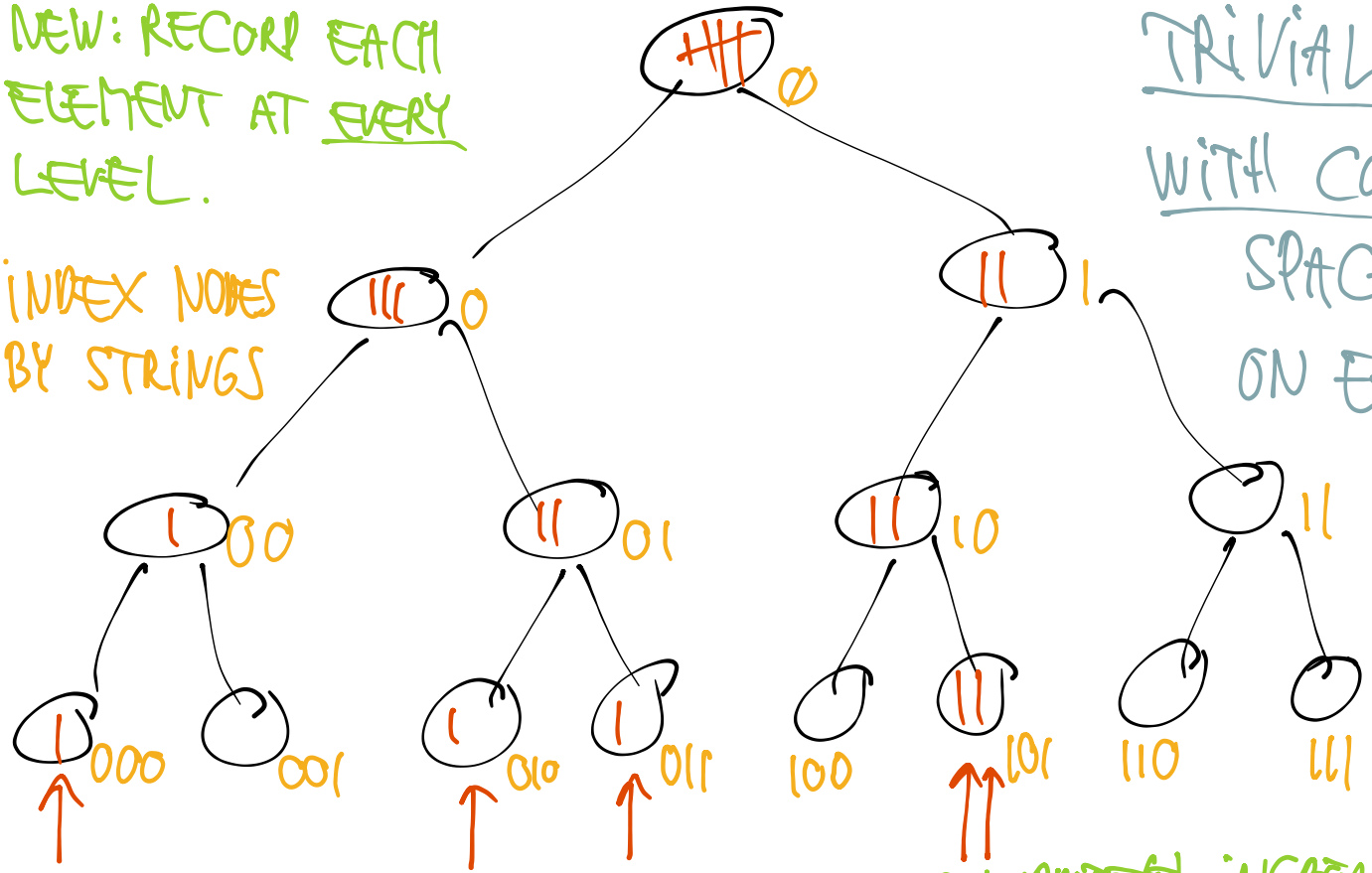
DYADIC COUNT-MIN SKETCH

← IN BOOK COUNT SKETCH IS USED, BUT IT WORKS THE SAME WAY

DIFFERENT APPROACH TO APPROXIMATING COUNTS IN NODES OF COMPLETE BINARY TREE: USE A SKETCH!

NEW: RECORD EACH ELEMENT AT EVERY LEVEL.

INDEX NODES BY STRINGS



TRIVIAL SPACE: # DISTINCT STRINGS

WITH COUNT-MIN: CAN CHOOSE SPACE TO GET ERROR $\leq \frac{\varepsilon N}{\log U}$ ON EACH ESTIMATE WITH PROBABILITY $1 - \delta/\log U$.

SPACE: $\|w^*\| \leq N \log U$

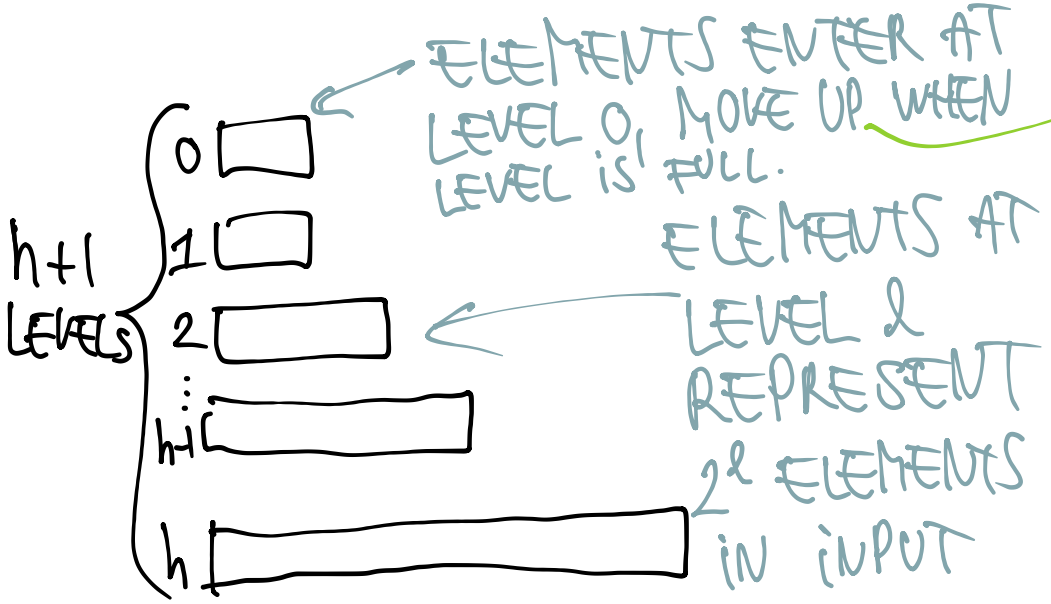
$O\left(\frac{1}{\varepsilon} \log^2 U \log\left(\frac{\log U}{\delta}\right)\right)$

(CAN BE IMPROVED)

ON UPDATE(5), INCREASE COUNTS OF 101, 10, 1, ε (ALL PREFIXES OF BIN. REP.)

KARNIN-LANG-LIBERTY (KLL)

- WORKS WITH ANY ORDERED DOMAIN (NO "log n" DEPENDENCE)
- DATA STRUCTURE IS A LIST OF ORDERED LISTS — "BUFFERS"



FOR SOME $c \in (\frac{1}{2}, 1)$
EACH LEVEL GROWS BY
$\approx$ A FACTOR $1/c$
(EXCEPT LOWEST LEVELS WHICH HAVE SIZE 2)

COMPRESS BY EXAMPLE:

2	3	5	7	11	13	17	19

CHOOSE BETWEEN

ODD-NUMBERED

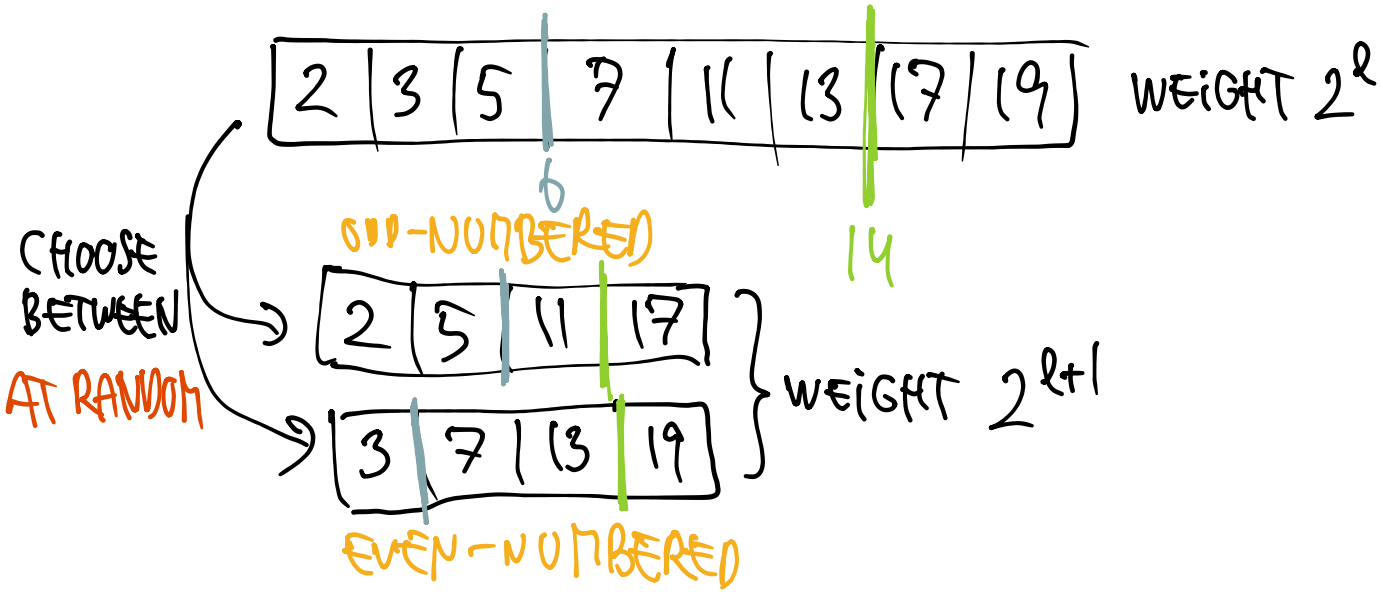
2	5	11	17

3	7	13	19

EVEN-NUMBERED

HOW COMPRESS AFFECTS RANK

COMPRESS BY EXAMPLE:



CASE 1: RANK IS "EVEN", $2i \cdot 2^\ell$ IN ORIGINAL LIST

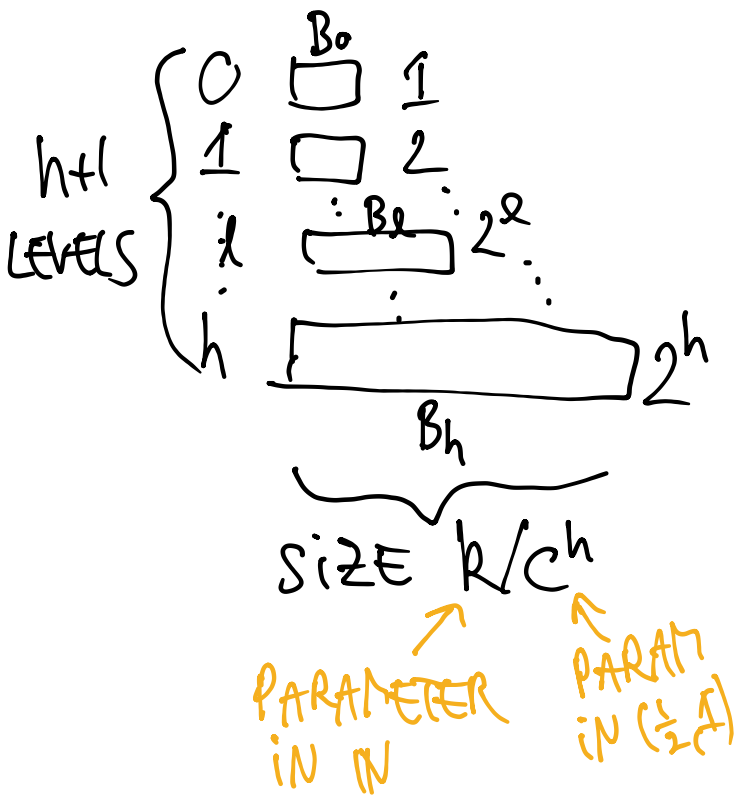
$\rightsquigarrow$ RANK BECOMES $i \cdot 2^{\ell+1}$ IN BOTH ODD- AND EVEN-NUMBERED CASES

CASE 2: RANK IS "ODD", $(2i+1)2^\ell$ IN ORIGINAL LIST

$\rightsquigarrow$ RANK IS $i \cdot 2^{\ell+1}$ IN EVEN-NUMBERED CASE, AND $(i+1) \cdot 2^{\ell+1}$ IN ODD-NUMBERED CASE

} CORRECT ~~ON AVERAGE EXPECTATION~~

KLL, THE GENERAL PICTURE



UNBIASED ESTIMATOR FOR RANK(x):

$$\tilde{r}(x) = \sum_{i=0}^{h} |\{y \in B_i \mid y \le x\}| \cdot 2^i$$

$$\tilde{r}(x) - \text{RANK}(x) = \sum_{l=0}^{h-1} \sum_{i=1}^{m_l} 2^l X_{i,l}$$

COMPACTION OPERATIONS AT LEVEL l

$X_{i,l} = \begin{cases} +1 & \text{IF POS. ERR} \\ -1 & \text{IF NEG. ERR} \\ & \text{IN COMPACTION} \end{cases}$

ADDITIVE CHERNOFF-HOEFFDING BOUND:

GIVEN INDEPENDENT $X_1, \ldots, X_n$ WITH $|X_i| \le a_i$, FOR $X = \sum_i X_i$ AND $\forall [X] = 0$ IT HOLDS THAT

$$\Pr[|X| > k] \le 2\exp\left(\frac{-k^2}{2\sum_i a_i^2}\right).$$

(SPECIAL CASE OF FACT 1.4 FROM BOOK)

$$\tilde{r}(x) - \text{RANK}(x) = \sum_{l=0}^{h-1} \sum_{i=1}^{m_l} 2^l X_{i,l} = X$$

"a_i"

COMPACTION OPERATIONS

$$\sum a_i^2 = \sum_{l=0}^{h-1} \sum_{i=1}^{m_l} 2^{2l} = \sum_{l=0}^{h-1} m_l \cdot 2^{2l} = O\left(\sum_{l=0}^{h-1} \left(\frac{2}{c}\right)^{h-l} \cdot 2^{2l}\right) = O(2^{2h})$$

CLAIM 1

DOMINATED BY LAST TERM

$$\Pr[|X| > \varepsilon N] \le 2\exp\left(-\frac{(\varepsilon N)^2}{2^{2h+1}}\right)$$

CLAIM 2 →

$$\le 2\exp\left(-\frac{(\varepsilon N)^2}{4(N/k)^2}\right)$$

$$= 2\exp(-k^2/(4\varepsilon^2))$$

ENOUGH TO CHOOSE $k = 2\sqrt{\log(1/\delta)}/\varepsilon$ TO GET ERROR PROB. δ.

ADDITIVE CHERNOFF-HOEFFDING BOUND:

GIVEN INDEPENDENT $X_1, \ldots, X_n$ WITH $|X_i| \le a_i$, FOR $X = \sum_i X_i$ AND $\mathbb{E}[X] = 0$ IT HOLDS THAT

$$\Pr[|X| > k] \le 2\exp\left(-\frac{k^2}{2 \cdot \sum_i a_i^2}\right).$$

PAUSE AND THINK:

WHY ARE ALL RANKS ACCURATE WITHIN $\pm\varepsilon N$ WITH PROB. δ/ε? (HINT: UNION BOUND)

CLAIM 1:

$$m_\ell \le O\left(\left(\frac{2}{c}\right)^{h-\ell}\right)$$

$h-\ell=0$. NO COMPACTIONS AT LEVEL h.

INDUCTION STEP. (INFORMAL)

SUPPOSE $m_\ell = O\left(\left(\frac{2}{c}\right)^{h-\ell}\right)$, THEN

THE OUTPUT OF ~~THESE~~ COMPACTIONS HAS SIZE $m_\ell \cdot \underbrace{c^{h-\ell}k}_{\text{BUFFER } \ell \text{ SIZE}}/2 = \underbrace{m_\ell \cdot \frac{c}{2}}_{\text{BOUND ON \#COMPACTIONS AT LEVEL } \ell+1} \cdot \underbrace{c^{h-(\ell+1)}k}_{\text{BUFFER } \ell+1 \text{ SIZE}}$

$$m_{\ell-1} \le m_\ell \cdot \frac{c}{2} \overset{\text{IND. HYP}}{\le} O\left(\left(\frac{2}{c}\right)^{h-(\ell-1)}\right)$$

CLAIM 2

$$2^{2h} \le 4(N/k)^2$$

- LEVEL $h-1$ ELEMENTS CORRESPOND TO 2^{h-1} INPUT ELEM.
- AT LEAST kc INPUT ELEMENTS HAVE BEEN AT LEVEL $h-1$

$$kc \cdot 2^{h-1} \le N$$

$$\Downarrow$$

$$2^h \le \frac{2}{c} N/k$$

$$\Downarrow$$

$$2^{2h} \le 16 (N/k)^2$$

↑ CAN BE IMPROVED TO 4, SEE BOOK

KLL WRAP-UP

- SPACE USAGE:

$$\sum_{l=0}^{h} c^{h-l} k = c^h k \sum_{l=0}^{h} c^{-l} = O(k)$$

GEOMETRICALLY INCREASING, DOMINATED BY LAST TERM

THEORETICAL SPACE USAGE CAN BE IMPROVED TO $O(\frac{1}{\varepsilon} \log\log(\frac{1}{\varepsilon}))$ BUT MUCH MORE COMPLICATED AND NOT MERGEABLE

- MERGING:

GIVEN TWO KLL SUMMARIES CAN MERGE BY MERGING LEVEL $l = 0, 1, 2, \ldots, h$ (IN THIS ORDER), DOING COMPACTIONS AS NEEDED.

- TIME USAGE: IN EXPECTATION, EACH INPUT IS THROWN OUT AFTER BEING IN 2 COMPACTIONS
  $\rightsquigarrow$ EXPECTED CONSTANT TIME.