APPROXIMATION ALGORITHMS

SUMMARIES FOR MULTISETS

RASMUS PAGH
UNIVERSITY OF COPENHAGEN

TODAY

- MOTIVATING EXAMPLE: WEBSITE VISITS
- MISRA-GRIES SUMMARY
- COUNT-MIN SKETCH

DATA SET: BROWSER LOG

2021-05-31 10:06: diku.edu
2021-05-31 10:07: google.com
2021-05-31 10:09: wikipedia.org
⋮

TRIVIAL SOLUTION: USE A DICTIONARY MAPPING SITE NAMES TO COUNTS

1 USER: SMALL DATA

10^9 USERS: BIG DATA

↑ LINEAR SPACE NOT FEASIBLE

QUESTIONS:
- WHAT ARE THE MOST POPULAR WEB SITES?
  - IN NUMBER OF PAGE VIEWS
  - IN TIME SPENT ON PAGES

GOAL: ANSWER SUCH QUESTIONS USING SUBLINEAR SPACE

MISRA-GRIES SUMMARY

- RECEIVES STREAM S OF PAIRS (x, w)
  - ITEM
  - WEIGHT
  - SAME ITEM CAN APPEAR SEVERAL TIMES
- TRUE WEIGHT OF x: $w_x^* = \sum_{(x,w) \in S} w$
- STORES A SUMMARY WITH SPACE FOR k PAIRS (x, w_x)
  - CAN BE MUCH SMALLER THAN # DISTINCT ITEMS
  - NOT NECCESSARILY EQUAL TO TRUE WEIGHT OF x IN S
- NEED TO MAINTAIN SUMMARY UNDER UPDATES
- ABILITY TO MERGE TWO SUMMARIES S_1, S_2 TO FORM SUMMARY OF THE "MULTISET UNION":

$$w_x^* = \sum_{(x,w) \in S_1 \cup S_2} w = \sum_{(x,w) \in S_1} w + \sum_{(x,w) \in S_2} w$$

MISRA-GRIES UPDATES

CONSIDER $k=3$, ALL WEIGHTS 1

INPUT STREAM: a, b, a, c, d, e, a, d, f, a, d

PAUSE AND THINK:
HOW CAN WE MERGE TWO MG SUMMARIES INTO ONE OF SAME SIZE?

T

ITEM	COUNT
a	2
f	0
d	1

NO SPACE IN SUMMARY
- SMALLEST COUNT MUST GO!

SUBTRACT m=1 FROM ALL COUNTS, KICK OUT ZERO COUNT ITEMS

INSERT (x,w):

- IF x∈T, INCREASE COUNTER BY w
- ELSE, IF A SLOT WITH COUNT ZERO EXISTS, WRITE (x,w)
- ELSE, DECREASE ALL COUNTERS BY m=min w (x,w)∈T
  AND THEN INSERT (x,w-m)

MISRA-GRIES ANALYSIS

- INVARIANTS: $w_x \le w_x^*$ ← ASSUMING POSITIVE UPDATES ($w \ge 0$): WE ONLY INCREASE w_x WHEN A PAIR (x, w) IS SEEN.

$$w_x \ge w_x^* - \frac{\|w^*\|_1}{k+1}$$

$\|w^*\|_1 = \sum_x |w_x^*|$ TOTAL SUM OF ALL TRUE WEIGHTS (ABS. VALUE)

LAPTOP 10^7 PAIRS ⇒ ERROR ≤ 1000 — 10^{10} WEBSITE VISITS

OFTEN MUCH BETTER IN PRACTICE

- PROOF:

LET $M = \sum_{(x,w_x) \in T} w_x$, THEN WE CLAIM THAT

$$w_x \ge w_x^* - \frac{\|w^*\|_1 - M}{k+1}$$

TOTAL WEIGHT OF DISCARDED ELEMENTS

WHAT HAPPENS AT UPDATE WITH NEW ELEMENT (x, w)? — NOT IN T ← THE CASE $x \in T$ EASILY HOLDS BY INDUCT.

- NEW VALUE $M' = M - (k+1)m + w$
- FOR $x \in T$: $w_x' = w_x - m \ge w_x^* - \frac{\|w^*\|_1 - M}{k+1} - m = w_x^* - \frac{\|w^*\|_1 - M'}{k+1}$.

↑ INDUCTION HYPOTESIS

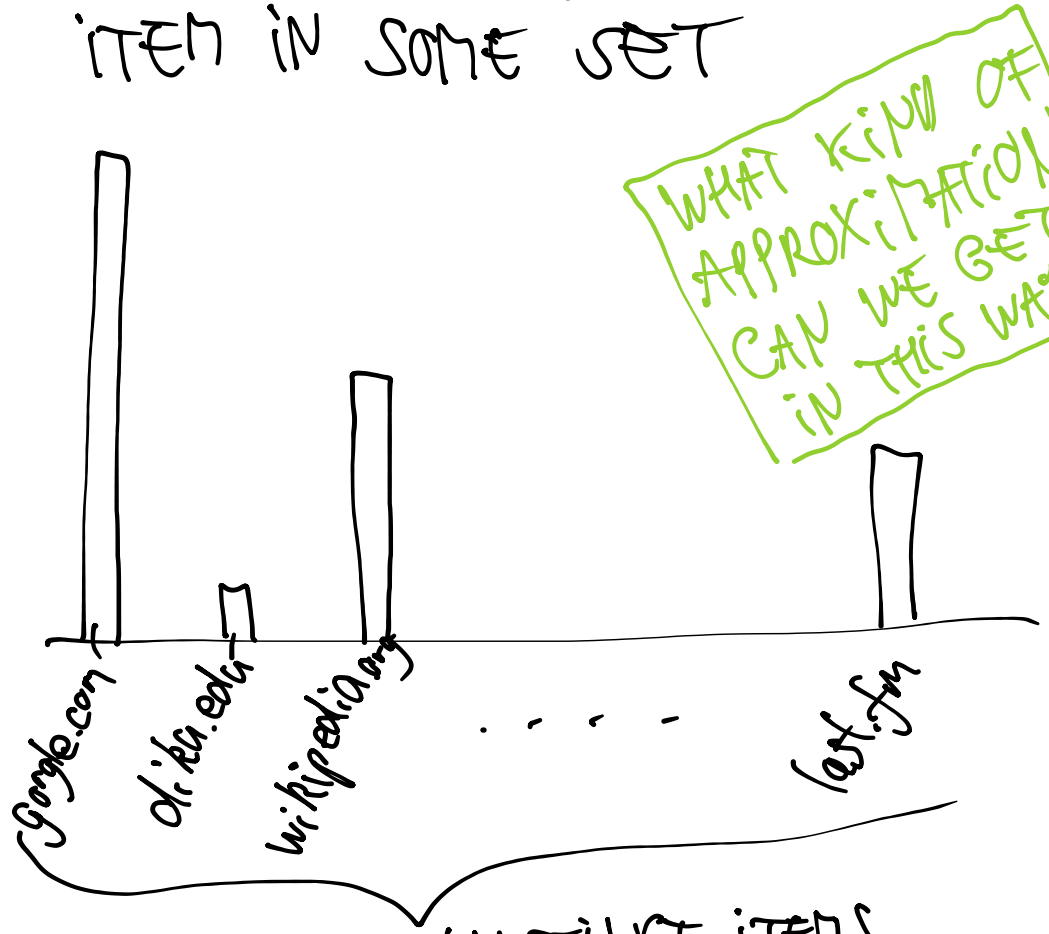
MERGE: SIMILAR ANALYSIS, SEE BOOK

MISRA-GRIES VARIANTS

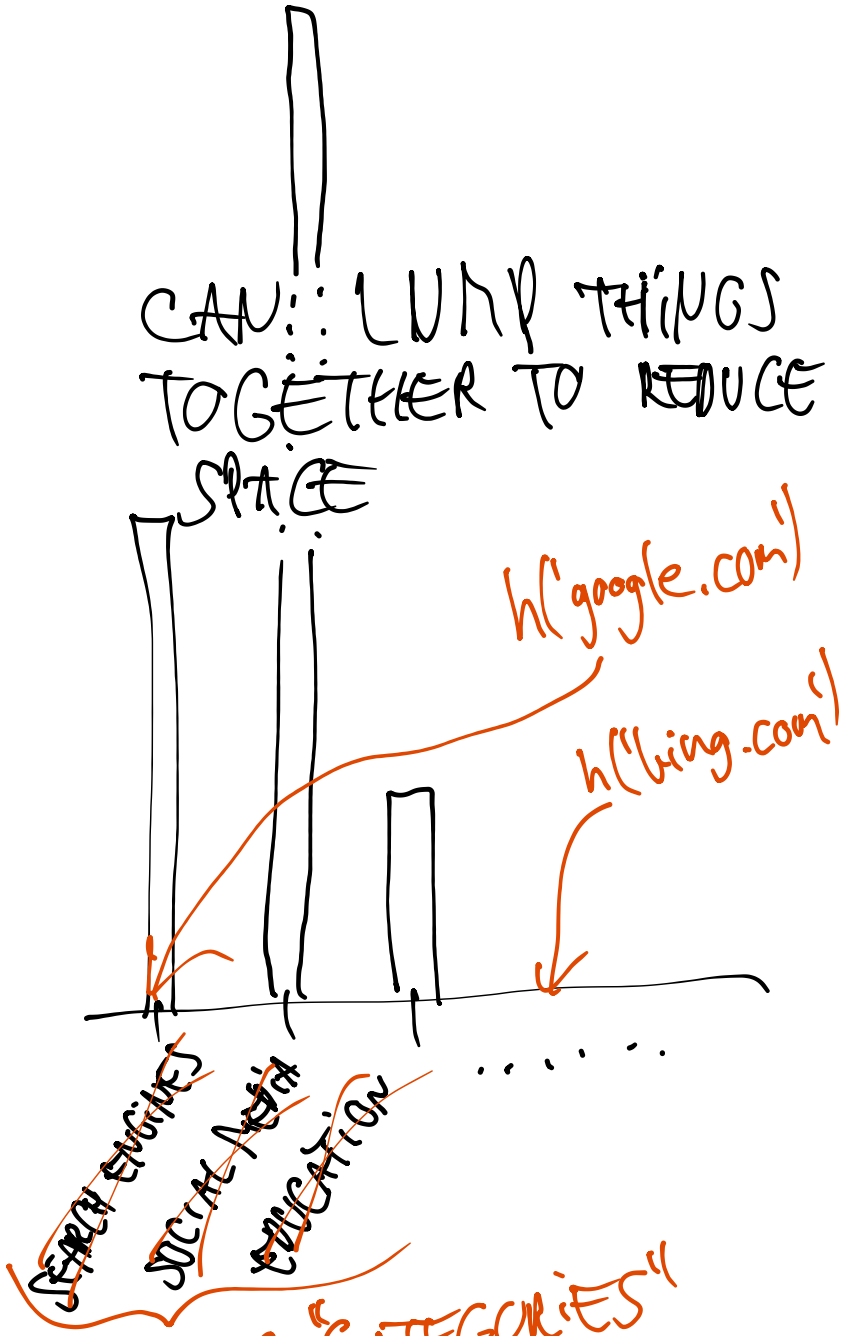
- KEEP TRACK OF HIGHEST WEIGHT SEEN FOR EACH ITEM x IN T. (HEURISTIC TO DECREASE ERROR IN "TYPICAL" CASE)
- KEEP TRACK OF $\|w^*\|_1$ TO SUPPORT UPPER BOUNDS ON w_x^*, USING $w_x^* \leq w_x + \frac{\|w^*\|_1 - M}{k+1}$.
- SUPPORT NEGATIVE UPDATES BY THINKING OF w_x^* AS A DIFFERENCE $w_{x^+}^* - w_{x^-}^*$
  - TRUE SUM OF NEG. UPDATES
  - TRUE SUM OF POS. UPDATES
- "SPACESAVING" ALGORITHM:
  KEEP TRACK OF UPPER BOUNDS RATHER THAN LOWER BOUNDS
  (SEC. 3.3 IN BOOK)
  EQUIVALENT TO MISRA-GRIES

RANDOMIZED HISTOGRAMS

RECORD A NUMBER FOR EACH ITEM IN SOME SET



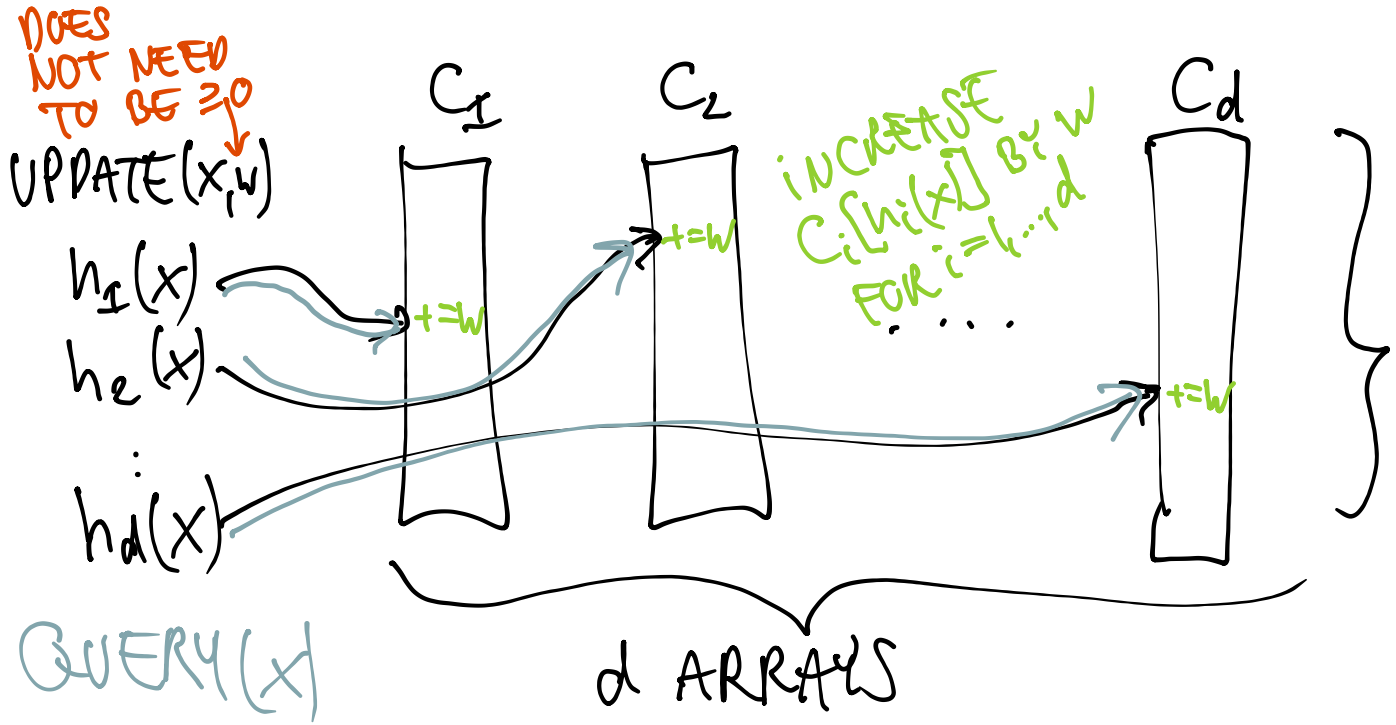
CAN: LUMP THINGS TOGETHER TO REDUCE SPACE



CHOOSING "CATEGORIES" AT RANDOM USING HASHING

COUNT-MIN SKETCH

NB! NOTATION DIFFERENT FROM BOOK



PAUSE AND THINK:

- HOW CAN TWO COUNT-MIN SKETCHES BE MERGED?
- CAN YOU "REMOVE" ONE SKETCH FROM ANOTHER?

SIZE $t = m/d$ EACH

IF YOU KNOW ALL ELEMENTS OF ONE IS IN THE OTHER.

QUERY(x)

RETURN $\min_i C_i[h_i(x)]$

$= w_x^* + \min_i \left(\sum_{y \neq x,\ h(y)=h(x)} w_y^* \right)$

"SMALLEST NOISE"

$$C_i[h_i(x)] = w_x^* + \sum_{y \neq x,\ h_i(y)=h_i(x)} w_y^*$$

w_x^*: TRUE SUM

$\sum_{y \neq x,\ h_i(y)=h_i(x)} w_y^*$: "NOISE"

COUNT-MIN ANALYSIS

ASSUME $w^*_x \geq 0$ FOR ALL x, $\|w^*_x\|_1 = \sum_x w^*_x$

NOT NEEDED, BUT SIMPLIFIES

LET $\hat{w}_x = \min_i C_i[h_i(x)]$ BE THE ESTIMATOR OF w^*_x.

WE HAVE $\hat{w}_x = w^*_x + \min_i \sum_{\substack{y \neq x \\ h_i(y)=h(x)}} w^*_y \geq w^*_x$

BY ASSUMPTION $w^*_y \geq 0$, $\forall y$

EXPECTED ERROR IN $C_i[h_i(x)]$:

$$E\left[\sum_{\substack{y \neq x \\ h_i(y)=h_i(x)}} w^*_y\right] = \sum_{y \neq x} \Pr[h(y)=h(x)] \cdot w^*_y = \sum_{y \neq x} \frac{1}{t} w^*_y \leq \|w^*\|_1 / t.$$

SIMILAR TO ERROR IN MG WITH $k=t$

MARKOV'S INEQUALITY: $\Pr\left[\sum_{\substack{y \neq x \\ h_i(y)=h(x)}} w_y > 2 \cdot \frac{\|w^*\|_1}{t}\right] < \frac{1}{2}$

TWICE EXPECTATION

COUNT-MIN ANALYSIS, CONT.

$$\Pr\left[\hat{w}_x > w_x^* + 2\frac{\|w^*\|_1}{t}\right]$$

$$= \Pr\left[C_i[h_i(x)] > w_x^* + 2\frac{\|w^*\|_1}{t} \text{ FOR } i = 1,\ldots,d\right]$$

$$\overset{\text{INDEP.}}{=} \prod_i \Pr\left[C_i[h_i(x)] > w_x^* + 2\frac{\|w^*\|_1}{t}\right] < \left(\frac{1}{2}\right)^d$$

← CHOOSE $d = \log_2\left(\frac{1}{\delta}\right)$ TO SUCCEED WITH PROB. $1-\delta$

"LINEAR SKETCH"

COMPARISON:	SPACE	ERROR BOUND	ERROR PROB.	SUBTRACTION ELIMINATES INSERT.
COUNT-MIN	$d \cdot t$	$2\|w^*\|_1/t$	2^{-d}	YES
MISRA-GRIES	k	$\|w^*\|_1/(k+1)$	0	NO

DATA STRUCTURE DETERIORATES WITH NUMBER OF SUBTRACTIONS

UNBIASED ESTIMATOR FROM COUNT-MIN

- IDEA: EXPECTED "NOISE" IN $C_i[h_i(x)]$ IS THE SAME AS THE EXPECTED VALUE OF $C_i[j]$ FOR $j \neq h_i(x)$

- ESTIMATOR: $\mathbb{E}\left[C_i[h_i(x)] - C_i[h_i(x)+1 \bmod t]\right] = w_x^*$

- DECREASING ERROR PROBABILITY:

$$\hat{w}_x = \underset{i}{\text{median}} \left(C_i[h_i(x)] - C_i[h_i(x)+1 \bmod t]\right)$$

PARTICULAR IMPLEMENTATION OF THIS IDEA: COUNT SKETCH (SEE BOOK)