Lecture #4 (AA). 1/21/21.

Perfect Hashing

Recap: Dict: given $S \subseteq U$, query "is $x \in S$".

$n \triangleq |S|$

$h: U \to [m]$. random / pseudo-random (can

$C_x \triangleq \#\{y \neq x, y \in S \text{ s.t. } h(x) = h(y)\}$

$C = \sum_{x \in S} C_x$.

Claim (last time): $\mathbb{E}[C] \leq n^2/m$.

Hence can construct a hash table of size $m = 4n^2$ where each bucket size ≤ 1.

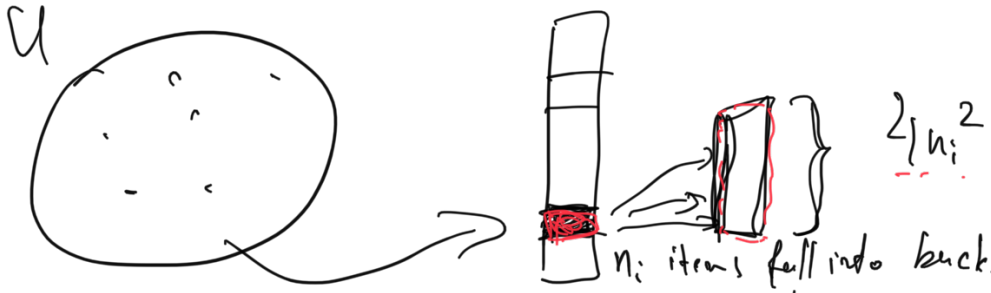
larger than before.

query time $= O(1)$ always.

Thm [P.M.]: can get d.s. space $O(n)$, q.t. $= O(1)$ always. Construction time $= O(n)$ in expectation

Pf: just combine "standard" hashing

+ $4n^2$ solution.



$h: U \to [m]$ think $m \approx n$.

Issue: some buckets may be large $\gg 1$.

$n_i \triangleq$ size of bucket $i = |h^{-1}(i) \cap S|$.

Add 2^{nd} level using P.H. quadratic-sp sol.

Algo:
- pick random $h: U \to [m]$, $m = n$.
- n_i def as above.
- for each $i \in [m]$,
  - build a 2^{nd} level h.f., say $h_i: U \to [2n_i^2]$ such that no collisions in 2^{n_i}

also from past lecture

Analysis. 1. Correctness immediate.

2. Query time: $O(1)$ always by construction.

3. Space: $O(n) + O\left(\sum_i n_i\right)$.

$O(n)$: 1st lev. h.f. — $O\left(\sum_i n_i\right)$: 2nd level h.f.'s.

Claim: $\mathbb{E}\left[\sum_i n_i^2\right] = O(n)$.

pf: $\mathbb{E}\left[\sum_i n_i^2\right] \le \mathbb{E}\left[n_1 + \#\text{collisions in bucket } 1 + \cdots + n_2 + \#\text{collisions in bucket } 2 + \cdots\right]$

$= \mathbb{E}\left[C + \sum_i n_i\right] = n + \mathbb{E}[C]$

$n_i^2 = n_i(n_i-1) + n_i = n_i + 2\binom{n_i}{2}$ ← # collisions in bucket i

$2\binom{n_i}{2} + n_i = n_i^2$.

$= h + n^2/m = 2n$. ⊠

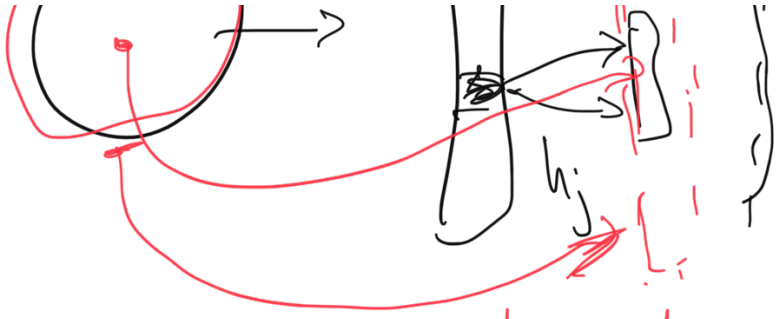
read thm.

Obs: space is $O(n)$ in expectation only.

by Markov, $O(n)$ with prob $\ge 90\%$

⇒ can repeat construction $O(1)$ exp. # tim

until space is $O(n)$.



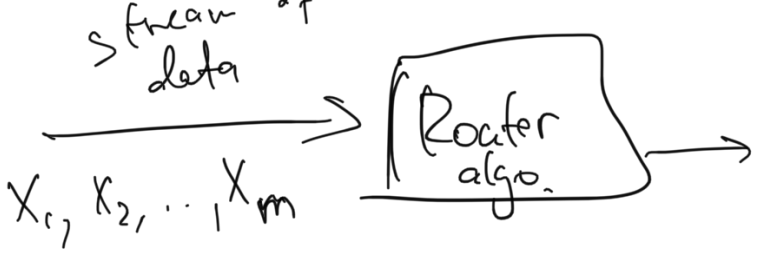
h' depends on S.

Remark : - in all this construction, used $\mathbb{E}$ to analyze.

$\Rightarrow$ enough to use universal hash. f. (both 1st & 2nd)

(II) Sketching & Streaming algos.

Model : streaming model.



size of sketch $\ll$ size of data.

Few problems to consider.

Problem: counting # distinct elements

$x_1, \ldots x_m \in [n]$.

Performance measure (complexity) : space used by [drawing] to be able to answer query @ end stream

1. $m \cdot \lg n$ space (store everything). $O(d \cdot \lg n)$. just store the distinct items.

2. via hashing:

$O(n) \rightarrow$ bit-table

$O(d) \rightarrow$ hash table of items seen so far.

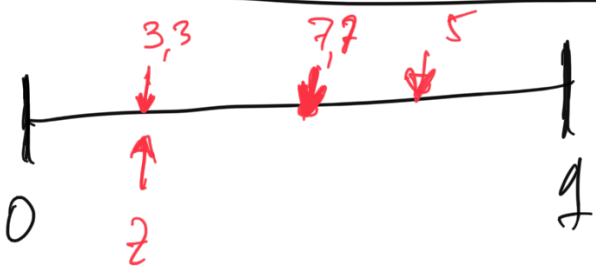
$d = \#$ distinct items.

Thm: for $\epsilon < 1/2$, $d > \Omega(1/\epsilon^2)$.
can get space $O\left(\frac{\lg n}{\epsilon^2}\right)$ space

Algo [Flajolet-Martin '85]:
uses a hash function $h : [n] \rightarrow [0,1]$. fully random.

- Init: $z = 1$.
- @ item i: $z := \min\{z, h(i)\}$.

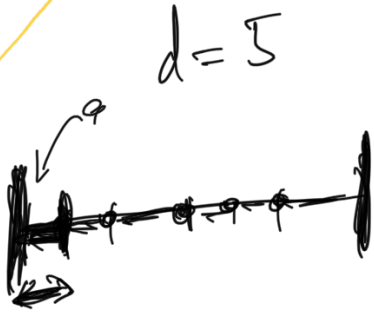
- Estimator: $\frac{1}{z} - 1$.



Obs: z depends only on d.

$\mathbb{E}\left[\frac{1}{z} - 1\right] = +\infty$.

Claim: $\mathbb{E}[z] = \frac{1}{d+1}$.



pf: z = min of d random vars $\in [0,1]$.

Pick $a \in [0,1]$ at random.

$\Pr[a < z] =$ z

$\Pr[$ choose $d+1$ r.v. $\in [0,1]$, last chosen one is smallest $]$

$= 1/d+1$.

$\mathbb{E}[z] = \frac{1}{d+1}$. ☒

In general $\mathbb{E}[\frac{1}{Z}] \neq 1/\mathbb{E}[Z]$.

Claim: $\mathrm{Var}[Z] \leq Z/d^2$.

pf. omitted.

$\sqrt{\mathrm{Var}}$.

By Chebyshev: $Z = \frac{1}{d} \pm O(\frac{1}{d})$ with $\geq 90\%$ p

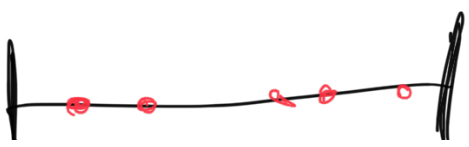
Possible to get $1 \pm \varepsilon$ factor approx by running $k = O(1/\varepsilon^2)$ FM algo's iid.

Algo to get $1 \pm \varepsilon$ approx.

Bottom-k Algo:

- keep $Z_1, \ldots Z_k := 1$ (counters.)
- @ item i in stream compute $h(i)$.
  maintain $Z_1 < Z_2 < \ldots < Z_k$ smallest k hash values seen in the stream.
- Estimator: $\frac{k}{Z_k}$.

$k = 2$



$1 \; z_1 \; z_2$

Goal: to prove $\hat{d} = \frac{k}{z_k}$ is within factor $1 \pm \varepsilon$ of d with $\geq 90\%$ prob.

Claim: $\Pr\left[\hat{d} > d \cdot (1+\varepsilon)\right] \leq 0.05$ ← (for $k = \Theta(1/\varepsilon^2)$).

π:

Pf: $\hat{d} > d \cdot (1+\varepsilon) \iff \frac{k}{z_k} > d(1+\varepsilon)$

$\iff z_k < \frac{k}{d(1+\varepsilon)}$.

$\iff$ there are at least k items (in stream) whose $h(\cdot)$ is $< \frac{k}{d(1+\varepsilon)}$

$\pi = \Pr\left[\text{choose } d \text{ r.v.} \in [0,1], \text{ at least } k \text{ of them} < \frac{k}{d(1+\varepsilon)}\right]$

X_i = indic. r.v. of event i^{th} item has $h(\cdot) < k/d(1+\varepsilon)$.

$= \Pr[\sum X_i \geq k]$

$\Pr[X_i = 1] = \Pr[\nearrow] = \frac{k}{d(1+\varepsilon)}$ (assuming ≥ 1).

$= \Pr[X_i = 1] \cdot 1 + \Pr[X_i = 0] \cdot 0$

$\mathrm{Var}[X_i] \leq \mathbb{E}[X_i^2] \leq \frac{k}{d(1+\varepsilon)} < \frac{k}{d}$.

$\mathbb{E}[\sum X_i] = d \cdot \frac{k}{d(1+\varepsilon)} = k/(1+\varepsilon)$.

$\mathrm{Var}[\sum X_i] = d \cdot \mathrm{Var}[X_i] = k$.

By Chebyshev: $\Pr\left[\sum X_i - \frac{k}{1+\varepsilon} \geq \sqrt{20k}\right] \leq \frac{1}{20} = 0.05$

$\sum X_i > \frac{k}{1+\varepsilon} + \sqrt{20k}$ $\quad$ Set k s.t. $\boxed{\quad} < k$.

$< k$

for $k = 40 \cdot 1/\varepsilon^2$ for c large enough.

TF $\Pr[\sum X_i > k] \leq 0.05$. ☒

$k > \frac{k}{1+\varepsilon} + \sqrt{20k}$

$k \geq k(1-\varepsilon+\varepsilon^2) + \sqrt{20k}$

$k(\varepsilon - \varepsilon') \geq 120k$

$\sqrt{k} \geq \sqrt{20}/(\varepsilon - \varepsilon^2) \Rightarrow k \geq \frac{4}{\varepsilon^2}$

using $\varepsilon < 1/2$.