$\boxed{\text{25 Apr 2022}}$ Streaming Distinct Elements
Sketching Frequencies

Announcements
1. Homework 6 to be released Weds, due one week later (plus 2-day grace period)
2. Quiz 4 will be in class, Monday, May 9.

Recap: $h(x) = ax + b \pmod{M}$ $M \geq m$, prime
$(a,b) \in [M]^2$ uniformly random.

Storing a representation of h requires only $2\log_2(M)$ space to store coefficients a and b.

1. $\forall x$ $h(x)$ is unif distrib in $[M]$.
2. $\forall x \neq y$ then $h(x)$, $h(y)$ are independent random variables.

Let $X_{ik} = \begin{cases} 1 & \text{if } h(a_i) \leq k \\ 0 & \text{if } h(a_i) > k \end{cases}$

$Y_k = \sum_{i=1}^{d} X_{ik}$ = # of distinct elements in stream that hash to $\{1, \dots, k\}$.

We proved:
1. $E[X_{ik}] = \frac{k}{M}$
2. $E[Y_k] = \frac{dk}{M}$
3. $\mathrm{Var}[Y_k] < \frac{dk}{M}$.

Algorithm for Distinct Elements

Sample random hash function h as above.

Let $t = \left\lceil \frac{2(1+\varepsilon)}{\varepsilon^2 \delta} \right\rceil$.

Initialize $(Z_1, \ldots, Z_t) = \perp^t$

// $Z_1, \ldots, Z_t$ will store the t smallest distinct values in the set $\{h(a_i) \mid i = 1, \ldots, n\}$ in increasing order.

for $i = 1, \ldots, n$:

  observe token a_i

  compute $z = h(a_i)$

  if $z < Z_t$:

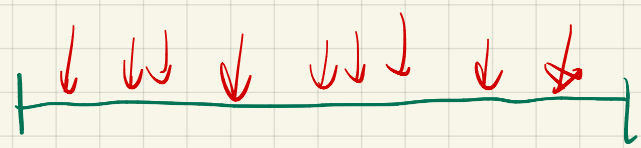
    update $Z_1, \ldots, Z_t$ to preserve the invariant in the comment above.

  endif

endfor

output $\frac{tM}{Z_t}$.

$Z_1 \approx \frac{M}{d}$ $Z_2 \approx \frac{2M}{d}$, ..., $Z_t \approx \frac{tM}{d}$



Analysis of the algorithm

$$Z_t = k \equiv Y_k = t \text{ and } Y_{k-1} < t.$$

So, we resort to analyzing Y_k for various values of k.

Two ways the algo could fail:

- outputs answer $< (1-\varepsilon) d$.

Since answer $= \dfrac{tM}{Z_t}$ this

corresponds to $Z_t > \dfrac{tM}{(1-\varepsilon)d}$

$$\implies Y_k < t \quad \text{for} \quad k = \left\lfloor \frac{tM}{(1-\varepsilon)d} \right\rfloor$$

- outputs answer $> (1+\varepsilon) d$

corresponds to $Z_t < \dfrac{tM}{(1+\varepsilon)d}$

$$\implies Y_\ell \geq t \quad \text{for} \quad \ell = \left\lfloor \frac{tM}{(1+\varepsilon)d} \right\rfloor$$

$$E[Y_\ell] = \frac{d\ell}{M}, \quad Var[Y_\ell] < \frac{d\ell}{M},$$

$$\Pr(Y_\ell \geq t) \overset{\text{Cheby}}{\leq} \frac{d\ell/M}{\left(\frac{d\ell}{M} - t\right)^2}$$

$$\frac{d\ell}{M} = \frac{d}{M} \left\lfloor \frac{t}{1+\varepsilon} \cdot \frac{M}{d} \right\rfloor \leq \frac{t}{1+\varepsilon}$$

$$t - \frac{d\ell}{M} \geq t - \frac{t}{1+\varepsilon} = \frac{\varepsilon t}{1+\varepsilon}.$$

$$\left(t - \frac{d\ell}{M}\right)^2 \geq \frac{\varepsilon^2 t^2}{(1+\varepsilon)^2}.$$

$$t = \left\lceil \frac{2(1+\varepsilon)}{\varepsilon^2 \delta} \right\rceil$$

$$\leq \frac{t/(1+\varepsilon)}{\varepsilon^2 t^2/(1+\varepsilon)^2} = \frac{1+\varepsilon}{\varepsilon^2 t} \leq \frac{\delta}{2}.$$

Similar calculation using Chebyshev shows the probability of answer $< (1-\varepsilon) d$ is also $\leq \frac{\delta}{2}$.

Sketching Model of Computation

Stream $a_1, \ldots, a_n$

Tokens belong to $[m]$

Algo has storage space $s = O(\text{poly}(\log n, \log m))$.

After processing stream, the stored representation is used to answer queries from some set Q of potential queries.

(ε, δ)-PAC property: $\forall\, q \in Q$ $\Pr(\text{answer is } \underbrace{\varepsilon\text{-accurate}}_{??}) \geq 1-\delta$.

Weds lecture: Estimating frequencies of tokens.

Query $q(x)$ for token x asks,
"How many times did x appear in the stream?"

ε-accurate might mean, for example,
answer to $q(x)$ has additive error $\leq \varepsilon n$.