473 | 10/9/18

Sampling ⇐
Streaming

Median selection

$B[1 \ldots n]$

$R(i)$: ith smallest element of B.

Q: $R(\frac{n}{2})$?

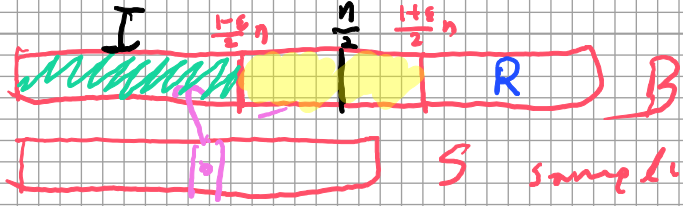
$\varepsilon \in \left(0, \frac{1}{2}\right)$ $\quad \varepsilon \doteq 0.1$

Compute $X \in \mathbb{B}$ s.t.

$$P\left[R\left((1-\varepsilon)\frac{n}{2}\right) \leq X \leq R\left(\frac{1+\varepsilon}{2} n\right)\right] \geq 1-\delta$$

$\delta \in (0,1)$ "bad prob". $\quad \delta \doteq \frac{1}{n^{10}}$

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



$$p = \left(\frac{1-\varepsilon}{2}\eta\right)^k$$

$$= \frac{1-\varepsilon}{2}$$

Bad: more than $k/2$ elements of the sample are in I.

$X_i = 1 \iff$ ith sample is from $\mathbb{I}$

$Y = \sum_{i=1}^{k} X_i$ total number of samples from $\mathbb{I}$

$P[X_i = 1] = \frac{1-\varepsilon}{2} = p$

$Y \geq \frac{k}{2}$

$\mu = E[Y] = \sum E[X_i] = \frac{1-\varepsilon}{2} k$

$$P_p\left[X \geq \frac{k}{2}\right] \leq P_p\left[X \geq \frac{(1+\varepsilon)(1-\varepsilon)}{2} k\right]$$

$$(1+\varepsilon)(1-\varepsilon) = 1-\varepsilon^2 \leq 1$$

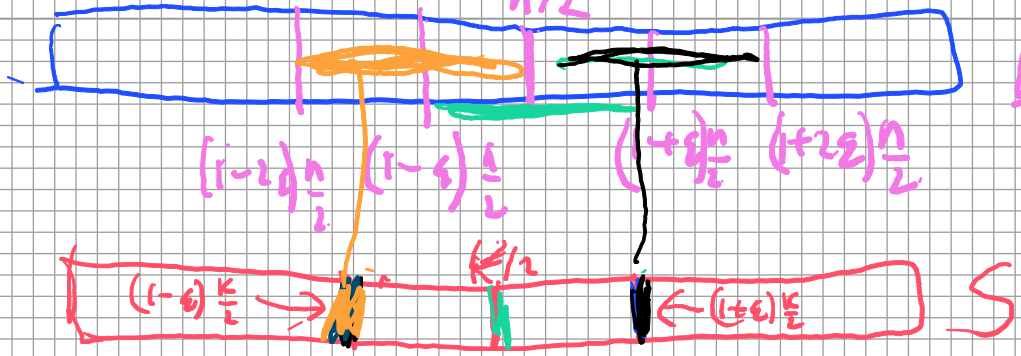
$$= P_p\left[X \geq (1+\varepsilon) M\right]$$

$$\Pr\left[Y \geq (1+\varepsilon)\mathbb{E}[Y]\right]$$

Chernoff's inequality

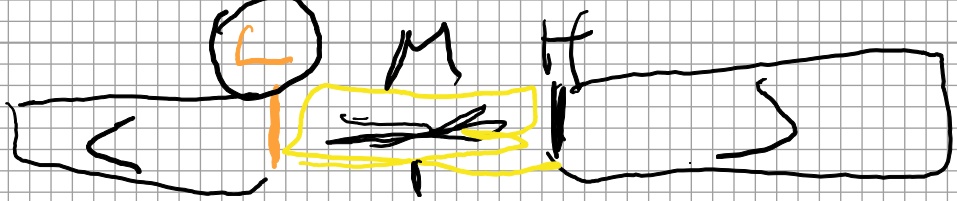
$$\leq \exp\left(-\frac{\varepsilon^2 n}{3}\right) \leq \frac{\delta}{2}$$

$$n \geq \frac{1-\varepsilon}{2} k \geq \frac{k}{4} \qquad k \geq \frac{12}{\varepsilon^2} \ln \frac{2}{\delta}$$



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

$$\Pr[\text{good}] \geq 1 - \delta$$



B $\left((1-\varepsilon)\frac{n}{2}\right.$ $\left((1+2\varepsilon)\frac{n}{2}\right.$

$|M| \leq 4\varepsilon n < \eta$

sort M return desired element.

$\varepsilon = \frac{1}{n^{1/4}}$ $\quad$ $\delta = \frac{1}{n^{10}}$

$k = O\left(\frac{1}{\varepsilon} \log \frac{1}{\delta}\right) = O(\sqrt{n} \log n)$

$O(\sqrt{n} \log n)$

$|M| = 4\varepsilon n = O(n^{3/4})$

$O(\sqrt{n} \log^2 n)$

2γ

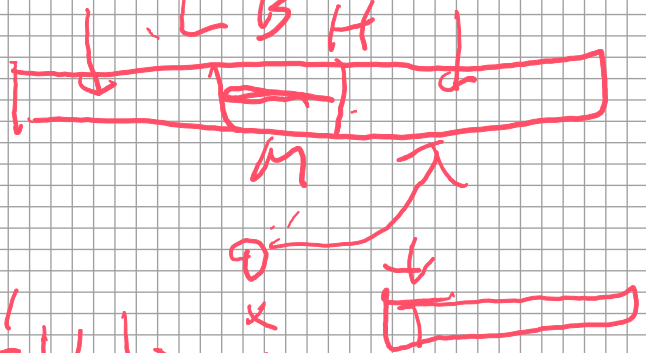
$O(n^{3/4} \log n)$

$$2n + O(\sqrt{n}\log^2 n) + O(n^{3\log_4 6}\log n)$$

$$2n + o(n) \cdot$$

$$\downarrow$$

$$1.5n + o(n)$$



$$\frac{1}{2}\left(1+\frac{1}{2}\right)^2 < \frac{3}{2}$$

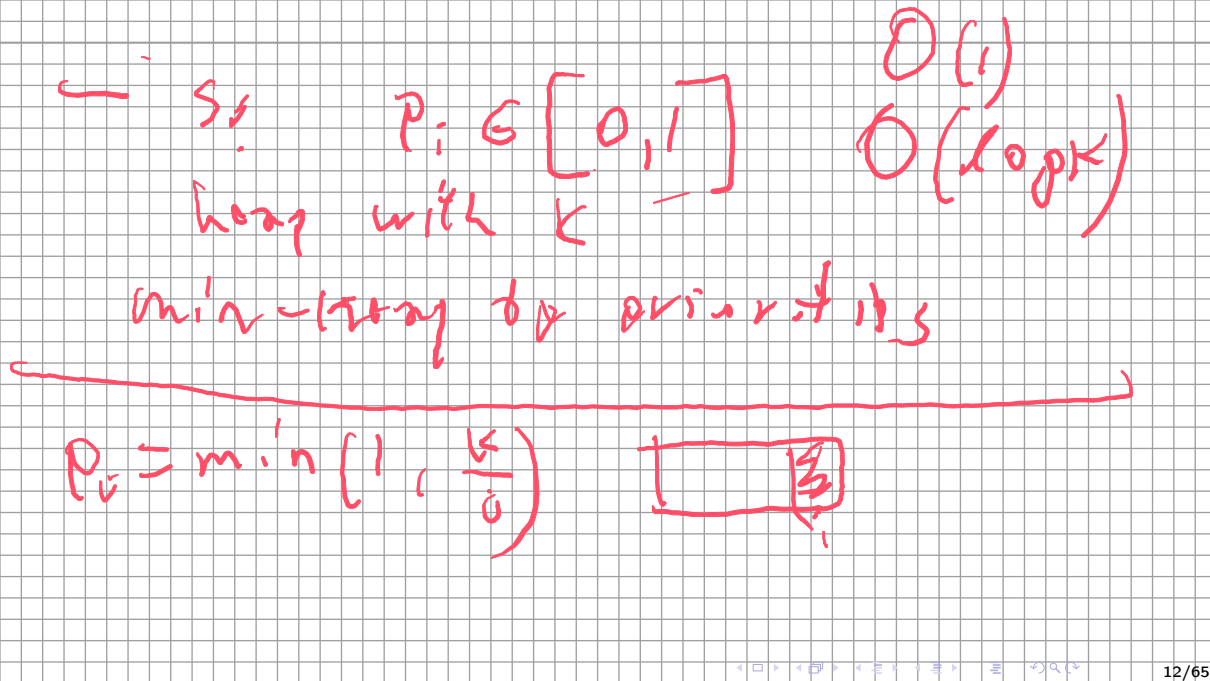
Random sampling from a stream

$s_1, s_2, \ldots$

k

Q: Maintain a random sample of size k of the stream

$O(k)$ space.


$p_i \in [0,1]$
$O(1)$
$O(\log k)$
loop with k
min-heap of p_i with i
$p_i = \min\left(1, \frac{k}{i}\right)$

Heavy hitters

$s_1, s_2, \ldots, s_n$,

report elements that appear at least εn times.

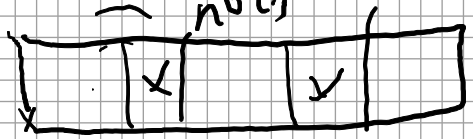
$k = O\left(\frac{\log n}{\varepsilon}\right)$ random sample

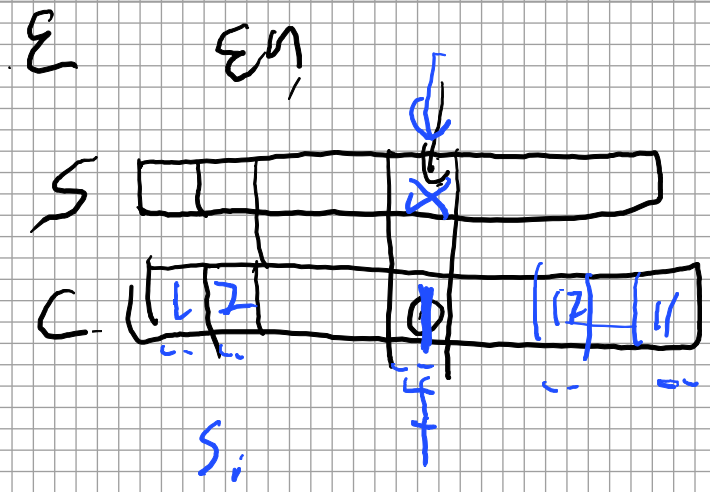
$$P\left[\begin{array}{l} x \text{ is a heavy center} \\ \text{but not in } S \end{array}\right]$$

$$\leq \left(1-\varepsilon\right)^{k} \leq \left(1-\varepsilon\right)^{O\left(\frac{\log n}{\varepsilon}\right)}$$

$$\leq \frac{1}{n^{O(1)}}$$

$$\sum_i \eta_i$$





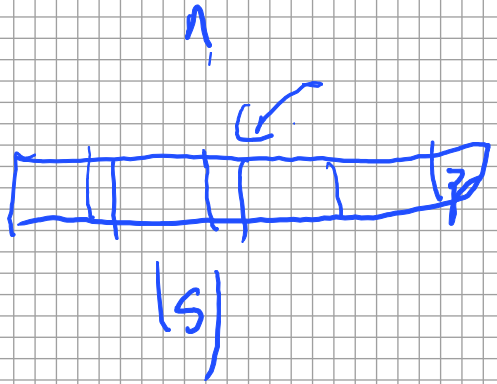
$S \geq \left\lceil \frac{1}{\varepsilon} \right\rceil$

Real count

alg count

z

$C(z)$



$(|\varepsilon) = \varepsilon_y$

alg $count^{\sim}(\Sigma) \geq rep\text{-}count(\Sigma) - \varepsilon n$