No class on 26.3.

Reservoir Sampling Input stream: $e_1, e_2, \ldots, e_N$

Algorithm Reservoir Sampling (k items without replacement)

Initialize a reservoir array S of size k with the first k items from the stream
Initialize a counter $t \leftarrow k$
for each item x in the stream starting from $(k+1)$-st item do
    $t \leftarrow t+1$
    Draw a random integer j uniformly from $\{1, \ldots, t\}$
    if $j \leq k$ then // happens with probability $\frac{k}{t}$
        $S[j] \leftarrow x$
return S

Theorem After N items have been processed the reservoir S contains a uniform random sample of k items from the stream without replacement.

Proof: We prove by induction on the stream length $t \geq k$ so far that any item e_i $(i \leq t)$ is in the reservoir with probability $\frac{k}{t}$

(Once this claim is established, the theorem follows with $t = N$)

- Base case ($t = k$): The reservoir contains the first k items

  For $i \leq k$, $\Pr[e_i \in S] = 1 = \frac{k}{k} = \frac{k}{t}$ ✓

- Inductive step ($t \to t+1$):

  Assume that after t items, for any $i \leq t$, $\Pr[e_i \in S \text{ at step } t] = \frac{k}{t}$

  Consider $(t+1)$-st item e_{t+1}

  Algorithm includes e_{t+1} in the reservoir with probability $\frac{k}{t+1}$

  $\Pr[e_{t+1} \in S \text{ at step } t+1] = \frac{k}{t+1}$

  Consider item e_i with $i \leq t$

  When does e_i that was in the reservoir, stay in the reservoir? ("survives")

  - New item e_{t+1} not selected to be in the reservoir (prob. $1 - \frac{k}{t+1}$)
  - New item e_{t+1} ~~is~~ selected to be in the reservoir (prob. $\frac{k}{t+1}$), but e_i is not chosen for replacement (with prob. $\frac{k-1}{k}$)

$$\Rightarrow \Pr[e_i \text{ survives in step } t+1 \mid e_i \text{ in the reservoir at step } t] =$$

$$1 - \frac{k}{t+1} + \frac{k}{t+1} \cdot \frac{k-1}{k} = \frac{t+1-k}{t+1} + \frac{k-1}{t+1} = \frac{t}{t+1}$$

$$\Rightarrow \Pr[e_i \in S \text{ at step } t+1] =$$

$$\Pr[e_i \text{ survives step } t+1 / e_i \in S \text{ at step } t] \cdot \Pr[e_i \in S \text{ at step } t]$$

(by IH)

$$= \frac{t}{t+1} \cdot \frac{k}{t} = \frac{k}{t+1} \quad \checkmark$$

Estimating the Median

items $e_1, \ldots, e_N$ (numbers)

median element m: half of items $\geq m$, half of items $\leq m$

ε-approximate median: any element with rank between $(\frac{1}{2} - \varepsilon)N$ and $(\frac{1}{2} + \varepsilon)N$

rank = position when sorted in increasing order by value

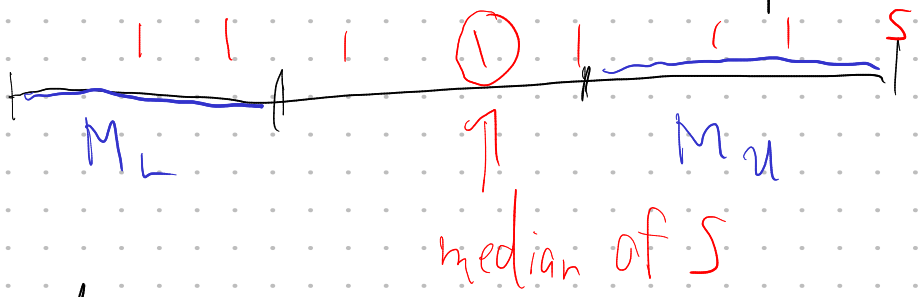
Straightforward idea:
- Take a uniform sample of size k (using reservoir sampling)
- return median of the sample

Space $O(k)$ $k = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ suffices!

Theorem: By choosing a sample S of size $k = O\left(\frac{1}{\varepsilon^2} \cdot \log\frac{1}{\delta}\right)$ in the straightforward algorithm, the median of the sample is an ε-approximate median of all items in the stream (i.e., algorithm is successful) with prob. $\geq 1-\delta$.

Proof:
- M_L: set of smallest $(\frac{1}{2}-\varepsilon)N$ elements in the stream
- M_U: —"— largest —"—

Algorithm fails if median of the sample S is in M_L or M_U



Want to bound $\Pr[\text{median}(S) \in M_L]$ (similar proof then works for M_U)

$\leq \Pr[\underbrace{\text{>half of } S}_{\frac{|S|}{2}+1 \text{ many elements}} \text{ is contained in } M_L] = \Pr\left[|S_L| > \frac{|S|}{2}\right]$

- $S_L := S \cap M_L$ $\quad p := \Pr[x \in S \text{ is also in } M_L] = \frac{|M_L|}{N} = \frac{(\frac{1}{2}-\varepsilon)N}{N} = \frac{1}{2}-\varepsilon$

$E[|S_L|] = p \cdot |S| = p \cdot k = k \cdot (\frac{1}{2}-\varepsilon)$

$$\Pr\left[|S_L| > \frac{|S|}{2}\right] = \Pr\left[|S_L| > \frac{k}{2}\right]$$

$$= \Pr\left[|S_L| - k\left(\tfrac{1}{2} - \varepsilon\right) > \frac{k}{2} - k\left(\tfrac{1}{2} - \varepsilon\right)\right]$$

$$= \Pr\left[|S_L| - \mathbb{E}[|S_L|] > \varepsilon k\right]$$

$$\leq \frac{1}{e^{\frac{2\varepsilon^2 k^2}{k}}} \quad \text{by Hoeffding Bound}$$

$$= \frac{1}{e^{2\varepsilon^2 k}}$$

$$\Pr[\text{algorithm fails}] \leq 2 \cdot \frac{1}{e^{2\varepsilon^2 k}}$$

Want: $2 \cdot \frac{1}{e^{2\varepsilon^2 k}} \lesssim \delta$

$$\Leftrightarrow e^{2\varepsilon^2 k} \gtrsim \frac{2}{\delta}$$

$$\Leftrightarrow 2\varepsilon^2 k \gtrsim \ln \frac{2}{\delta}$$

$$\Leftrightarrow k \gtrsim \frac{1}{2\varepsilon^2} \cdot \ln \frac{2}{\delta}$$

setting $k = \left\lceil \frac{1}{2\varepsilon^2} \cdot \ln \frac{2}{\delta} \right\rceil = \Theta\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$ suffices

HW:
Please do some background reading on unbiased estimators ("erwartungstreue Schätzer")