Streaming Algorithms

Stream: \( A = \langle a_1, a_2, \ldots, a_m \rangle \)
\( a_i \in [n] \) size \( \log n \)
Compute \( f(A) \) in \( \text{poly}(\log m, \log n) \) space

-------------------------

Goal: randomly sample \( k \) elements from stream
\( O(k \cdot \log n + \log m) \) space

-------------------------

Simpler question: randomly sample one element from stream
\( O(\log n + \log m) \) space

\( O(\log n) \) to store element \( S \)
\( O(\log m) \) to keep count of how many seen so far \( C \)

???
wp \( k/i \) keep \( a_i \) in register, replace old \( S \) w/ \( a_i \)
[Vitter '85]

Analysis:
What is probability \( a_m \) should be kept? \( k/m \) -- good.
What is probability \( a_{m-1} \) should be kept?
\[
(k/(m-1)) \cdot (1 - (k/m)(1/k) = (m-1)/m) = k/m \quad \text{-- good.}
\]
[kept] [not replaced by \( a_m \)]
Inductively, ignoring \( a_{i+1} \) ... \( a_m \)
what is probability \( a_i \) should be kept to that point? \( k/i \)
Assume \( a_{i+1} \) ... \( a_m \) kept with correct probability: total \( (m-i)/k \cdot k/m = (m-i)/m \)
\( a_i \) in \( S \) after processed wp \( k/i \)
not replaced afterwards wp \( 1-(m-i)/m = i/m \)
total \( (\text{kept}) \cdot (\text{not replaced}) = (k/i) \cdot (i/m) = k/m \quad \text{-- good.} \)

-------------------------

(\( \epsilon, \delta \))-Approximate Counts:

Consider Interval \( I \) subset \([n]\)
\( \text{count}(I) = |\{ a_i \in A \mid a_i \in I \}| \)
Goal: Data structure $S$ s.t. for query interval

$\Pr[|S(I) - \text{count}(I)| > \epsilon m] < \delta$

+++

Chernoff Inequality

Let $\{X_1, X_2, \ldots, X_r\}$ be independent RVs
Let $\Delta_i = \text{max}(X_i) - \text{min}(X_i)$
Let $M = \sum_i X_i$

$\Pr[|M - \sum_i E[X_i]| > \alpha] < 2 \exp(-2 \frac{\alpha^2}{\sum_i (\Delta_i)^2})$

often: $\Delta = \text{max}_i \Delta_i$ and $E[X_i] = 0$ then:
$\Pr[|M| > \alpha] < 2 \exp(-2 \frac{\alpha^2}{r \Delta^2})$

+++ 

Let $S$ be a random sample of size $k = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$

$S(I) = |\{S \cap I\}| * \left(\frac{m}{k}\right)$

Each $s_i$ in $I$ wp ($\text{count}(I)/m$)

$\rightarrow$ RV $Y_i = \{1$ if $s_i$ in $I$, $0$ if $s_i$ not in $I\}$

$E[Y_i] = \text{count}(I)/m$

$\rightarrow$ RV $X_i = \frac{(Y_i - \text{count}(I)/m)}{k}$

$E[X_i] = 0$

$\Delta < 1/k$

$M = \sum_i X_i \Rightarrow$ error on count estimate by $S$

$\Pr[|M| > \epsilon] < 2 \exp(-2 \frac{\epsilon^2}{k (1/k^2)}) < \delta$

Solve for $k$ in $\epsilon, \delta$:

$2 \exp(-2 \frac{\epsilon^2}{k}) < \delta$

$\exp(2 \frac{\epsilon^2}{k}) > 2/\delta$

$2 \epsilon^2 k > \ln(2/\delta)$

$k > \left(\frac{1}{2}\right) \left(\frac{1}{\epsilon^2}\right) \ln \left(\frac{2}{\delta}\right)$

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