Streaming Algorithms

Stream: $A = \langle a_1, a_2, \ldots, a_m \rangle$
\[ai \text{ in } [n] \text{ size } \log n\]
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?

\[ \frac{k}{m-1} \cdot \left( 1 - \frac{k}{m} \cdot \frac{1}{k} \right) = \frac{m-1}{m} \]

= $k/m$ -- good.

[kept] [not replaced by $a_m$]

Inductively, ignoring $a_{i+1} \ldots a_m$

what is probability $a_i$ should be kept to that point? $k/i$

Assume $a_{i+1} \ldots a_m$ 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 (kept) * (not replaced) = $(k/i) \cdot (i/m) = k/m$ -- good.

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

(eps,delta)-Approximate Counts:

Consider Interval $I$ subset $[n]$

\[ \text{count}(I) = \left| \left\{ a_i \in A \mid a_i \in I \right\} \right| \]

Goal: Data structure $S$ s.t. for query interval

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

++++++++++++++++++++++++++++

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

\[
\Pr[|M - \sum_i E[X_i]| > r * \alpha] < 2 \exp(-2 \alpha^2 / \sum_i (\Delta_i)^2)
\]

often: \( \Delta = \max_i \Delta_i \) and \( E[X_i] = 0 \) then:
\[
\Pr[|M| > r * \alpha] < 2 \exp(-2 \alpha^2 / r \Delta^2)
\]

Let \( S \) be a random sample of size \( k = O(1/\epsilon^2 \log (1/\delta)) \)
\( S(I) = |\{S \cap I\}| * (m/k) \)

Each \( s_i \) in \( I \) wp (count(I)/m)
\( \rightarrow \) RV \( Y_i = \{1 \text{ if } s_i \text{ in } I, 0 \text{ if } s_i \notin I\} \)
\( \quad E[Y_i] = \text{count}(I)/m \)
\( \rightarrow \) RV \( X_i = (Y_i - \text{count}(I)/m)/k \)
\( \quad E[X_i] = 0 \)
\( \Delta < 1/k \)

\( M = \sum_i X_i \) == error on count estimate by \( S \)

\[
\Pr[|M| > \epsilon] < 2 \exp(-2 \epsilon^2 / (k * (1/k^2))) < \delta
\]

Solve for \( k \) in \( \epsilon, \delta \):
\[
2 \exp(-2 \epsilon^2 k) < \delta
\]
exp(2 \ eps^2 \ k) > 2/\delta
2 \ eps^2 \ k > \ln(2/\delta)
k > (1/2) \ (1/\eps^2) \ ln \ (2/\delta)
= \ O((1/\eps^2) \ log \ (1/\delta))