Streaming Algorithms

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

Goal: randomly sample \( k \) elements from stream
\( O(k \log n + \log m) \) 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?
\( (k/(m-1)) \times (1 - (k/m)(1/k)) = (k-m)/(m-1) \) -- 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 \( (m-i)/k \times k/m = (m-i)/m \)
\( a_i \) in \( S \) after processed wp \( k/i \)
not replaced afterwards wp \( 1-(m-i)/m = i/m \)
total (kept) * (not replaced) = \( k/i \times i/m = k/m \) -- good.

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

Consider Interval \( I \) subset \([n]\)
\( \text{count}(I) = |\{a_i \text{ in } A : a_i \text{ in } I\}| \)

Goal: Data structure \( S \) s.t. for query interval
\( \Pr[|S(I) - \text{count}(I)| > \varepsilon \times m] < \delta \)

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]| > \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| > alpha ] < 2 \exp(- 2 \alpha^2 / \Delta^2)

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

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

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

- RV \( Y_i = \begin{cases} 1 & \text{if } s_i \text{ in } I, 0 & \text{if } s_i \text{ !in } I \end{cases} \)
  
  \( E[Y_i] = \frac{\text{count}(I)}{m} \)

- RV \( X_i = \frac{Y_i - \text{count}(I)/m}{k} \)
  
  \( E[X_i] = 0 \)

\( \Delta < \frac{1}{k} \)

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

Pr[ |M| > \varepsilon ] < 2 \exp(- 2 \varepsilon^2 / (k \Delta^2) ) < \delta

Solve for \( k \) in \( \varepsilon, \delta \):

- \( 2 \exp(- 2 \varepsilon^2 k) < \delta \)
- \( \exp(2 \varepsilon^2 k) > \frac{2}{\delta} \)
- \( 2 \varepsilon^2 k > \ln(\frac{2}{\delta}) \)
- \( k > (\frac{1}{2}) \frac{1}{\varepsilon^2} \ln \frac{2}{\delta} \)
  
  \( = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta}) \)