Today

Streaming

Frequent Items.

Deteminstic Algorithm.

Alg:
(1) Set, \( S \), of \( k \) counters, initially 0.
(2) If \( x_i \in S \) increment \( x_i \)'s counter.
(3) If \( x_i \not\in S \)
   - If \( S \) has space, add \( x_i \) to \( S \) w/value 1.
   - Otherwise decrement all counters. Delete zero count elts.

Example:

State: \( k = 3 \)

Stream

\[ [(1, 2, 3), (2, 3, 2), (3, 3, 3)] \]

Previous State

\[ [(1, 2, 3), (2, 3, 2), (3, 3, 3)] \]

Streaming

Stream: \( x_1, x_2, x_3, \ldots, x_n \)

Resources: \( O(\log^2 n) \) storage.

Today's Goal: find frequent items.

Turnstile Model and Randomization

Alg:
(1) Set, \( S \), of \( k \) counters, initially 0.
(2) If \( x_i \in S \) increment \( x_i \)'s counter.
(3) If \( x_i \not\in S \)
   - If \( S \) has space, add \( x_i \) to \( S \) w/value 1.
   - Otherwise decrement all counters.

Estimate for item:
- if in \( S \), value of counter.
- otherwise 0.

Underestimate clearly.
- Increment once when see an item, might decrement.
- Total decrements, \( T \)? \( n \)? \( n/k \)? \( k \)?
- decrement \( k \) counters on each decrement.
- \( Tk \) total decrementing \( n \) items. \( n \) total incrementing.
- \( \leq \frac{\varepsilon}{k} \)
- Off by at most \( \frac{\varepsilon}{k} \)
- Space? \( O(k \log n) \)

Frequent Items: deterministic.

Additive \( \frac{\varepsilon}{k} \) error.

Accurate count for \( k + 1 \)th item?
- Yes?
- No?
- \( k + 1 \)st most frequent item occurs \( < \frac{n}{k} \)
- Off by 100%. 0 estimate is fine.
- No item more frequent than \( \frac{n}{k} \)?
- 0 estimate is fine.

Only reasonable for frequent items.
Count Min Sketch

Sketch – Summary of stream.
(1) t arrays, A[i], of k counters.
(2) Process elt (j, c),
A[i][h(j)]+ = c.
(3) Item j estimate: min, A[i][h(j)].

→ Additive \(|f|_1/k\) error on average for each of t arrays.
Why t buckets? To get high probability.

Analysis
(1) \(g_i : U \rightarrow [-1,+1]\), \(h_i : U \rightarrow [k]\)
(2) Elt (j, c)
A[i][h(j)] = A[i][h(j)] + gi(j)cj
(3) Item j estimate: median, A[i][h(j)].

Intuition:
- Items cancel each other out!
- Exercise: proof of Markov. (All above average?)

Total Space: \(O(\log \frac{1}{\delta} \log n)\)

Count min sketch: analysis
(1) t arrays, A[i], of k counters.
h1,...,ht from 2-wise ind. family.
(2) Process elt (j, c),
A[i][h(j)]+ = c.
(3) Item j estimate: min, A[i][h(j)].

Markov: \(Pr[X > 2\frac{t\delta}{k}] \leq \frac{1}{2}\)
Exercise: proof of Markov. (All above average?)
t independent trials, pick smallest.

Error \(\epsilon|f|_1\) if \(\epsilon = \frac{2}{k}\).
Space? \(O(k \log \frac{1}{\delta} \log n)\) \(O(\frac{1}{\epsilon} \log \frac{1}{\delta} \log n)\)

Sum up

Deterministic:
stream has items
Count within additive \(\frac{\epsilon}{k}\)
\(O(k \log n)\) space.
Within \(\epsilon n\) with \(O(\frac{1}{\epsilon} \log \frac{1}{\delta} \log n)\) space.

Count Min:
stream has \(\pm\) counts
Count within additive \(\epsilon|f|_1\)
with probability at least \(1 - \delta\)
\(O(\log \frac{1}{\delta} \log \frac{1}{\delta})\).

Count Sketch:
stream has \(\pm\) counts
Count within additive \(\epsilon|f|_2\)
with probability at least \(1 - \delta\)
\(O(\log \frac{1}{\delta} \log \frac{1}{\delta})\).

See you on Tuesday.