Today

Streaming.

Frequent Items.

Streaming

Stream: $x_1, x_2, x_3, \ldots x_n$

Resources: $O(\log^c n)$ storage.

Today's Goal: find frequent items.

Frequent Items: deterministic.

Additive $\frac{n}{k}$ error.

Accurate count for $k+1$th item?

Yes?
No?
$k+1$st most frequent item occurs $< \frac{n}{k+1}$
Off by 100%. 0 estimate is fine.
No item more frequent than $\frac{n}{k}$?
0 estimate is fine.

Only reasonable for frequent items.

Deteministic Algorithm.

Alg:
(1) Set, S, of k counters, initially 0.
(2) If $x_i \in S$ increment x_i's counter.
(3) If $x_i \notin S$
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[(1,2)(2,2)(2,2)(3,0)[3,0)] stream7

1,2,311223343124 stream7

Previous State

[(1,2[(1,1)(2,2)(2,1)(3,0)]

Deterministic Algorithm.

Alg:
(1) Set, S, of k counters, initially 0.
(2) If $x_i \in S$ increment x_i's counter.
(3) If $x_i \notin S$
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 decremting
n items. n total incrementing.
$\leq \frac{n}{k}$.
Off by at most $\frac{n}{k}$

Space? $O(k \log n)$

Turnstile Model and Randomization

Stream: $\ldots, (i, c_i), \ldots$
item i, count c_i (possibly negative.)
Positive total for each item!

Estimate frequency of item: $f_j = \sum c_j$.

$|f|_1 = \sum_j f_j$ Smaller than $\sum_i c_i$.

Approximation:

Additive $\varepsilon |f|_1$ with probability $1 - \delta$

Space $O(\frac{1}{\varepsilon} \log \frac{1}{\delta} \log n)$.

Count Min Sketch

Sketch – Summary of stream.

(1) t arrays, $A[i]$, of k counters.
$h_1, \ldots, h_t$ from 2-wise ind. family.
(2) Process elt (j, c_j),
$A[i][h_i(j)] += c_j$.
(3) Item j estimate: $\min_i A[i][h_i(j)]$.

Intution: $|f_1|/k$ other "counts" in same bucket.

$\rightarrow$ Additive $|f_1|/k$ error on average for each of t arrays.

Why t buckets? To get high probability.

Analysis

(1) $\cdots$ $g_i : U \rightarrow [-1, +1], h_i : U \rightarrow [k]$
(2) Elt (j, c_j)
$A[i][h(j)] = A[i][h_i(j)] + g_i(j)c_j$
(3) Item j estimate: median of $g_i(j)A[i][h_i(j)]$.

Notice: $A[1][h_1(j)] = g_1(j)f_j + X$

$X = \sum_i Y_i$
$Y_i = \pm f_i$ if item $h_1(i) = h_1(j)$ $Y_i = 0$, otherwise
$E[Y_i] = 0$ $Var(Y_i) = \frac{f_i^2}{k}$.

$E[X] = 0$ — Expected drift is 0!

$Var[X] = \sum_{i \in [m]} Var(Y_i) = \sum_i \frac{f_i^2}{k} = \frac{|f|_2^2}{k}$

Cheybshev: $Pr[|X - \mu| > \Delta] \leq \frac{Var(X)^2}{\Delta^2}$

Choose $k = \frac{4}{\varepsilon^2}$: $Pr[|X| > \varepsilon|f|_2] \leq \frac{|f|_2^2/k}{\varepsilon^2|f|_2^2} \leq \frac{\varepsilon^2|f|_2^2/4}{\varepsilon^2|f|_2^2} \leq \frac{1}{4}$.
Each trial is close with probability $3/4$.
If > half tosses close, median is close!
Exists $t = \Theta(\log \frac{1}{\delta})$ where $\geq \frac{1}{2}$ are correct with probability $\geq 1 - \delta$

Total Space: $O(\frac{\log \frac{1}{\delta}}{\varepsilon^2} \log n)$

Count min sketch:analysis

(1) t arrays, $A[i]$, of k counters.
$h_1, \ldots, h_t$ from 2-wise ind. family.
(2) Process elt (j, c_j),
$A[i][h_i(j)] += c_j$.
(3) Item j estimate: $\min_i A[i][h_i(j)]$.

$A[1][h_j(j)] = f_j + X$, where X is a random variable.

Y_i - item $h_1(i) = h_1(j)$
$X = \sum_i Y_i f_i$

$E[X] = \sum_i E[Y_i] f_i = \sum_i \frac{1}{k} f_i = \frac{|f|_1}{k}$

Markov: $Pr[X > 2\frac{|f|_1}{k}] \leq \frac{1}{2}$
Exercise: proof of Markov. (All above average?)

t independent trials, pick smallest.

$\Pr[X > 2\frac{|f|_1}{k}$ in all t trials$] \leq (\frac{1}{2})^t$
$\leq \delta$ when $t = \log \frac{1}{\delta}$.

Error $\varepsilon|f|_1$ if $\varepsilon = \frac{2}{k}$.

Space? $O(k \log \frac{1}{\delta} \log n)$ $\quad O(\frac{1}{\varepsilon} \log \frac{1}{\delta} \log n)$

Sum up

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

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

Count Sketch:
stream has $\pm$ counts
Count within additive $\varepsilon|f|_2$
with probability at least $1 - \delta$
$O(\frac{\log n \log \frac{1}{\delta}}{\varepsilon^2})$.

Count sketch.

Error in terms of $|f|_2 = \sqrt{\sum_i f_2^2}$.

$\frac{|f|_1}{\sqrt{n}} \leq |f|_2 \leq |f|_1$.

Could be much better. E.g., uniform frequency $\frac{|f|_1}{\sqrt{n}} = |f|_2$

Alg:

(1) t arrays, $A[i]$:
t hash functions $h_i : U \rightarrow [k]$
t hash functions $g_i : U \rightarrow [-1, +1]$
(2) Elt (j, c_j)
$A[i][h(j)] = A[i][h_i(j)] + g_i(j)c_j$
(3) Item j estimate: median of $g_i(j)A[i][h_i(j)]$.

Buckets contains signed count (estimate cancels sign.)

Other items cancel each other out!
Tight! (Not an asymptotic statement.)

Do t times and average?

No! Median! Two ideas! One simple algorithm!

See you on Tuesday.