Approximate Medians in the Vanilla Streaming model via Sampling

Let $\sigma = \{a_1, a_2, \ldots, a_m\}$ and define $\text{rank}(y) = |\{a_i : a_i \leq y\}|$

For simplicity assume all a_i's distinct.

Problem: Find an ϵ-approximate median of σ, i.e. y such that

$$\frac{m}{2} - \epsilon m \leq \text{rank}(y) \leq m/2 + \epsilon \cdot m$$

Algorithm: Sample t values from σ with replacement and return median of the sampled values. (easy if stream length m is known)

Lemma: If $t = \frac{7 \log(2\delta^{-1})}{\epsilon^2}$ then the algorithm returns an ϵ-approximate median with probability $1-\delta$.

Later in semester: ~~[illegible]~~ Quadratically better dependence on $\frac{1}{\epsilon}$.

Proof: Partition the a_i's into 3 groups:

$$S_L = \{a_i : \text{rank}(a_i) < m/2 - \epsilon m\}$$

$$S_M = \{a_i : \frac{m}{2} - \epsilon m \leq \text{rank}(a_i) \leq \frac{m}{2} + \epsilon m\}$$

$$S_U = \{a_i : \text{rank}(a_i) \geq \frac{m}{2} + \epsilon m\}$$

If fewer than $\frac{t}{2}$ elements of <u>both</u> S_L and S_U are in sample, then median of the sample is in S_M :) (Seeing this is more subtle than it appears at first glance.)

Assume ...

Let $X_j = 1$ if j^{th} sample is in S_L and $X_j = 0$ otherwise

Let $X = \sum_j X_j$, note $\mathbb{E}[X] = t \cdot \frac{|S_L|}{m} \cdot (\frac{1}{2} - \epsilon)$. By Chernoff bound, if $t > 7\epsilon^{-2}\log(2\delta^{-1})$

$$
\begin{aligned}
\Pr[X \geq \tfrac{t}{2}] &\leq \Pr[(X-\mu) > \epsilon t] \\
&\overset{t \geq 2\mu}{\leq} \Pr[(X-\mu) > 2\epsilon\mu] \\
&\overset{\text{Chernoff}}{\leq} 2\exp\left(-\frac{\mu \cdot (2\epsilon)^2}{3}\right) \\
&= 2\exp\left(-\frac{t(\frac{1}{2}-\epsilon)(2\epsilon)^2}{3}\right) \\
&= 2\exp\left(-\frac{7\epsilon^{-2}\log(2\delta^{-1})(\frac{1}{2}-\epsilon)(2\epsilon)^2}{3}\right) \\
&= 2\exp\left(\frac{-7\cdot(\frac{1}{2}-\epsilon)\cdot 4}{3}\log(2\delta^{-1})\right) \\
&\leq 2\exp\left(2\log(2\delta^{-1})\right) \\
&\leq \frac{\delta}{2}.
\end{aligned}
$$

Similarly, there are $\geq \frac{t}{2}$ elements from S_U with probability $\leq \frac{\delta}{2}$.

By a union bound, there are $< \frac{t}{2}$ elements from both of S_L and S_U with probability $1-\delta$ ■

How to compute a random sample (with replacement) of size t from a stream when you don't know the stream length n in advance?

Consider $t=1$ (for general t, run the "sample size 1" algorithm t times independently in parallel.

Algorithm: · Initially $s = x_1$.

· On seeing a_i, set $s \leftarrow a_i$ with probability $\frac{1}{i}$

Analysis: What is the probability that $s = a_i$ at some time $j \geq i$?

$$Pr[s = a_i] = \frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \left(1 - \frac{1}{i+2}\right) \times \left(1 - \frac{1}{j}\right)$$

$$= \frac{1}{j}.$$

prove this by induction.

Obvious when $j = i$. Assume it's true for j, let us show it is true for $j+1$. By induction,

$$\left(\frac{1}{i}\right) \times \left(1 - \frac{1}{i+1}\right) \times \cdots \times \left(1 - \frac{1}{j}\right) = \frac{1}{j}.$$

Hence, $\frac{1}{i} \times \left(1 - \frac{1}{i+1}\right) \times \cdots \times \left(1 - \frac{1}{j}\right)\left(1 - \frac{1}{j+1}\right)$

$$= \frac{1}{j} \cdot \left(1 - \frac{1}{j+1}\right)$$

$$= \frac{1}{j} - \frac{1}{j \cdot (j+1)}$$

$$= \frac{j+1-1}{j(j+1)} = \frac{1}{j+1}$$

There is a variant for sampling t items without replacement (on Wikipedia article). This also has the benefit of having $O(1)$ update time rather than $O(t)$

An overview of uniform sampling algorithms for various streaming problems.

- We just saw uniform random sampling gives an $O\left(\frac{\log(\frac{1}{\delta}) \cdot \log n}{\epsilon^2}\right)$- space streaming algorithm for outputting an ϵ-approximate median with probability $\geq 1-\delta$. (insertion-only streams)
    - $\boxed{\text{Suboptimal dependence on } \frac{1}{\epsilon}}$ we'll see a better approximate median algorithm later in the course.

- Uniform random sampling can also give an algorithm using space $O\left(\frac{\log n \cdot \log(\frac{1}{\delta})}{\epsilon^2}\right)$ for answering ϵ-approximate point queries i.e. outputting a summary of ~~size~~ the above size such that, with probability $\geq 1-\delta$, for any $i \in [n]$, an estimate $\hat{f}_i$ of f_i can be derived from the summary, satisfying $|f_i - \hat{f}_i| \leq \epsilon \cdot$ (again, works in insert-only model).

Algorithm: Sample $O\left(\frac{\log(\frac{1}{\delta})}{\epsilon^2}\right)$ stream updates, and output the estimate $\hat{f}_i := (\# \text{ of samples equal to } i) \cdot \frac{m}{t}$.

Analysis: Easy to see $\mathbb{E}_{\text{samples}}[\hat{f}_i] = f_i$. Bound the probability of $|\hat{f}_i - f_i| > \epsilon \cdot m$ using additive Chernoff bounds (exercise).

- suboptimal dependence on $\frac{1}{\epsilon}$. We saw a different algorithm in Lecture 2 using $O\left(\frac{\log n}{\epsilon}\right)$ space.

Are there any streaming problems for which random sampling is optimal?

Answer: Yes. Itemset frequency estimates.

Consider a database of grocery purchases. Each row is a reciept, each column is a product, $D_{ij} = 1$ if person i purchased item j, and $D_{ij} = 0$ otherwise.

	mops	gloves	bread	butter	...		hotdogs
Alice's receipt	1	0	1	0	0	1	0
Bob receipt	1	0	1	1	0	0	0
⋮	0	1	1	0	1	1	1
	0	0	0	0	1	1	0
	0	0	1	0	1	1	0
Joe's receipt	0	1	0	0	0	0	1

n rows

An itemset of size k is a set of k columns.

- The frequency of an itemset S is $f_S :=$ # of rows with a 1 in all columns in S

e.g. $f_{\{mops, gloves\}} =$ # of people who bought both mops and gloves.

- Identifying frequent itemsets is very well-studied in the data mining community.

- Goal: Output a summary of the database capable of returning for any k-itemset S, an estimate $\hat{f}_S$ satisfying $|\hat{f}_S - f_S| \le \epsilon m$.

- One simple summary: Sample t rows (they each take d bits to write down), for $t = O\left(\frac{\log\left(\binom{d}{k}/\delta\right)}{\epsilon^2}\right)$.

Output for each k-item set S the estimate

$$\frac{m}{t} \cdot \left(\begin{array}{l}\text{\# of sampled rows with all columns in } S\\ \text{containing 1s}\end{array}\right)$$

Additive Chernoff bounds imply this is a good summary with probability $\geq 1-\delta$.

- [LMTV16]: This space cost is <u>optimal</u> (even among summaries <u>not</u> computed by streaming algorithms).

- Intuitively, a key difference between point queries & itemset frequency queries is that for point queries there can be at most $\frac{1}{\epsilon}$ items i with frequency $f_i \geq \epsilon \cdot m$ (and for all other items it is okay to output the estimate 0). This is <u>not</u> the case for itemsets since a single row can contribute to the ~~frequency~~ frequency of <u>many</u> itemsets