E0234: Assignment 10

Due: Monday, April 11, 2016

It is highly recommended you do not google for the answers to the questions below. You can discuss with your friends, but then mention that in your submission. The writing should solely be your own.

1. In our class discussion of the Count-Min sketch, we focused on the *strict turnstile* model. Here, each update increments or decrements a coordinate of a vector $\mathbf{x} \in \mathbb{R}^n$ in such a way that $\mathbf{x}_i \geq 0$ for all $i \in [n]$ always.

   Now, consider the more general *turnstile* model where each update can increment or decrement a coordinate of $\mathbf{x}$ arbitrarily. In this problem, you will modify the Count-Min sketch so that it works in the turnstile model.

   Recall the notation used. The Count-Min sketch uses a table C of width w and depth d. It maintains d hash functions $h_1, \ldots, h_d : [n] \to [w]$ where each h_ℓ is pairwise independent. The Count-Min sketch maintains that for all $\ell \in [d]$ and $j \in [w]$:

   $$C[\ell, j] = \sum_{i : h_\ell(i) = j} \mathbf{x}_i$$

   (a) Argue that for any $i \in [n]$ and $\ell \in [d]$,

   $$\mathbf{E}[|C[\ell, h_\ell(i)] - \mathbf{x}_i|] \leq \frac{1}{w} \|\mathbf{x}\|_1$$

   (b) Let $\hat{\mathbf{x}}_i = \text{median}(\{C[\ell, h_\ell(i)] : \ell \in [d]\})$. Suppose $w = 2/\varepsilon$ and $d = O(\log \delta^{-1})$. Argue that:

   $$\Pr[|\hat{\mathbf{x}}_i - \mathbf{x}_i| > 3\varepsilon \|\mathbf{x}\|_1] < \delta$$

   The resulting algorithm is often called the *Count-Median* sketch in the literature.

2. In this problem, you will show how to use the Count-Min sketch to answer range queries for a stream. Assume that we are in the insertion-only model, so that each stream update only corresponds to incrementing the frequency of one item by one. So, given a stream $a_1, a_2, \ldots, a_m$ with each a_k from $[n]$, let $f(a)$ be defined as the number of times a occurs in the stream.

   A *range query* is the following problem: given $1 \leq i \leq j \leq n$, estimate $f([i, j]) = f(i) + f(i+1) + \cdots + f(j)$. We saw in class that the Count-Min sketch can be used to solve the range query problem when $i = j$.

(a) Assume n is a power of 2. Define the following partitions of $[n]$ into intervals:

$$\begin{aligned}
\mathcal{I}_0 &= \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\ldots,\{n\}\} \\
\mathcal{I}_1 &= \{\{1,2\},\{3,4\},\{5,6\},\{7,8\},\ldots,\{n-1,n\}\} \\
\mathcal{I}_2 &= \{\{1,2,3,4\},\{5,6,7,8\},\ldots,\{n-3,n-2,n-1,n\}\} \\
\mathcal{I}_3 &= \{\{1,2,3,4,5,6,7,8\},\ldots,\{n-7,n-6,n-5,n-4,n-3,n-2,n-1,n\}\} \\
&\vdots \\
\mathcal{I}_{\log_2 n-1} &= \{\{1,2,3,4,\ldots,n/2\},\{n/2+1,n/2+2,n/2+3,n/2+4,\ldots,n\}\}
\end{aligned}$$

Thus, for any $0 \le k < \log_2 n$, $\mathcal{I}_k$ is the partition of $[n]$ into $n/2^k$ many intervals of length 2^k. Each interval in any of these sets is called a *dyadic interval*.

Show that any interval $[i,j]$ can be written as the union of $\le 2\log_2 n$ dyadic intervals. For example, for $n = 256$,

$$[48,107] = [48,48] \cup [49,64] \cup [65,96] \cup [97,104] \cup [105,106] \cup [107,107]$$

(b) Construct $\log_2 n$ many CM-sketches, one for each $\mathcal{I}_k$, such that for any $I \in \mathcal{I}_k$, you can estimate $f(I)$. Use this idea and the above decomposition into dyadic intervals to give a streaming algorithm that, given any interval $I \subseteq [n]$ and parameters $\varepsilon, \delta > 0$, computes $\tilde{f}(I)$ such that: (1) $f(I) \le \tilde{f}(I) \le f(I) + \varepsilon m$ with probability at least $1-\delta$ and (2) the algorithm uses space $O((\log^3 n)/\varepsilon \cdot \log 1/\delta)$.