Problem Set 3: COSC-548 (Streaming Algorithms)

Due November 15th, 2016 by end of class

1. Recall that the Count Sketch outputs the median of $\log(1/\delta)$ copies of the following Basic Estimator.

   - Let $k = 6/\varepsilon^2$. Initialize k counters, denoted $C[1], C[2], \ldots, C[k]$, to 0.
   - Choose a hash function $h\colon [n] \to [k]$ at random from a pairwise-independent hash family.
   - Choose a hash function $g\colon [n] \to \{-1, 1\}$ at random from a pairwise-independent hash family.
   - When processing stream update (a_i, δ_i):
     $C[h(a_i)] \leftarrow C[h(a_i)] + g(a_i) \cdot \delta_i$.
   - On query j, output $\hat{f}_j = g(j) \cdot C[h(j)]$.

   We proved in class that for any fixed j, with probability at least 2/3, $|\hat{f}_j - f_j| \leq \varepsilon \cdot \|f\|_2$. This problem will walk through a proof of the following stronger bound.

   The Stronger Bound. Let $\ell = 1/\varepsilon^2$ (assume ℓ is an integer), and let $f^{res(\ell)}$ denote the frequency vector f of the stream with the top ℓ entries of f set to 0. For any fixed j, with probability at least 2/3, $|\hat{f}_j - f_j| \leq \varepsilon \cdot \|f^{res(\ell)}\|_2$.

   (1a) Assume without loss of generality that $f_1 \geq f_2 \geq \ldots f_n$. Let E denote the event that for all $i \leq \ell$ such that $i \neq j$, it holds that $h(i) \neq h(j)$. Show that $\Pr[E] \geq 5/6$.

   (1b) Show that $\mathbb{E}[\hat{f}_j | E] = f_j$.

   (1c) Show that $\mathrm{Var}[\hat{f}_j | E] \leq \|f^{res(\ell)}\|_2^2/k$.

   (1d) Use Chebyshev's inequality to conclude that $\Pr\left[\left|\hat{f}_j - f_j\right| > \varepsilon \cdot \|f^{res(\ell)}\|_2\right] \leq \Pr[E] + 1/(k\varepsilon^2) \leq 1/6 + 1/6 \leq 1/3$.

2. In class, we showed that for streams in the strict turnstile update model, the Count-Min sketch can, for each item $i \in [n]$, return an estimate $\hat{f}_i$ for f_i such that the following holds. For each fixed $i \in [n]$, with probability at least $1 - \delta$, $0 \leq \hat{f}_i - f_i \leq \varepsilon \cdot \|f\|_1$, where $\|f\|_1 = \sum_i f_i$. The space usage of the sketch is $O(\varepsilon^{-1} \cdot (\log m + \log n) \cdot \log(1/\delta))$ bits, and the estimate $\hat{f}_i$ can be computed from the sketch in $O(\log(1/\delta))$ time.

   Let $\phi, \varepsilon \in (0, 1)$. Recall that in Problem Set 2, we showed that for insert-only streams, the Misra-Gries algorithm can easily output a list of items items such that:

   - For every item i such that $f_i \geq \phi \cdot m$, i appears in the list.

- Every item j that appears in the list satisfies $f_j \geq (\phi - \varepsilon) \cdot m$.

This problem will focus on achieving the same goal for streams in the strict turnstile model. Specifically, suppose our goal is to develop a randomized streaming algorithm that, with probability at least $1 - \delta$, outputs a list satisfying the above two properties.

(2a) Give a randomized streaming algorithm that uses $O(\varepsilon^{-1} \cdot (\log m + \log n) \cdot \log(n/\delta))$ bits of space and achieves the above. How long does it take your algorithm to generate the list?

Sub-problems (2b) and (2c) below will walk you through the design of a randomized algorithm that uses slightly more space, but generates the list in $O(\phi^{-1} \cdot \log n)$ time. Throughout this problem, for $a, b \in \mathbb{N}$, $[a, b]$ denotes the set of all integers between a and b (inclusive).

(2b) A dyadic range R is a range of the form $[x \cdot 2^y + 1, (x+1) \cdot 2^y]$ for parameters $x, y \in \mathbb{N}$. Let us restrict our attention to dyadic ranges that are subsets of $[n]$. Show that each point in $[n]$ is a member of exactly $\log_2 n$ dyadic ranges, one for each $y \in \{0, \ldots, \log_2(n) - 1\}$. Using this fact, give a streaming algorithm that processes each stream update in $O(\log n)$ time, uses $O(\varepsilon^{-1} \cdot (\log m + \log n) \cdot \log n \cdot \log(1/\delta))$ bits of space, and achieves the following. Define the frequency of a dyadic range R to equal $f_R := \sum_{i \in R} f_i$. The algorithm should be able to return an estimate $\hat{f}_R$ for f_R such that the following holds. For each fixed R, with probability at least $1 - \delta$, $0 \leq \hat{f}_R - f_R \leq \varepsilon \cdot \|f\|_1$.

(2c) Using (2b), give an algorithm that, with probability at least $1 - \delta$, outputs a list satisfying the two bulleted properties above. The space usage of your algorithm should be $O(\varepsilon^{-1} \cdot (\log m + \log n) \cdot \log n \cdot \log(\phi^{-1} \cdot \log n \cdot \delta^{-1}))$ (argue this). Your algorithm should generate the list in $O(\phi^{-1} \cdot \log n)$ time.

Hint: Start by identifying all dyadic ranges of the form $[x \cdot 2^y + 1, (x+1) \cdot 2^y]$ with $y = \log_2(n) - 1$ satisfying $\hat{f}_R \geq \phi \cdot \|f\|_1$. For each such dyadic range, split it into two (sub)-dyadic ranges, and investigate recursively whether one or both subranges also have estimated frequency at least $\phi \cdot \|f\|_1$. Proceed in this manner until you arrive at dyadic ranges R consisting of a single item, and return all such identified items.

(2d) Show that each interval $[a, b]$ with $1 \leq a \leq b \leq n$ is covered by $O(\log n)$ disjoint dyadic ranges. Use this to give an algorithm that uses $O(\varepsilon^{-1} \cdot (\log m + \log n) \cdot \log^2 n \cdot \log(\log n \cdot \delta^{-1}))$ bits of space and achieves the following. For any interval $R = [a, b]$ (not just dyadic), the algorithm can return an estimate $\hat{f}_R$ of $f_R := \sum_{a \leq i \leq b} f_i$ in $O(\log n \cdot \log(1/\delta))$ time. Moreover, for each fixed R, with probability at least $1 - \delta$ it holds that $0 \leq f_R - \hat{f}_R \leq \varepsilon \cdot \|f\|_1$.

3. Call a stream in the general turnstile model *pure* if there exists at most one item i such that $f_i \neq 0$. Give a streaming algorithm that uses $O(\log(m \cdot n))$ space, and satisfies the following properties.

- If the stream is pure, then the algorithm always outputs "PURE", along with the unique item i such that $f_i \neq 0$.

- If the stream is not pure, then with probability at least $1 - 1/n$, the algorithm outputs "NOT PURE".

Hint: Let p be a sufficiently large prime (remember to say how large p must be when describing your solution). Pick a random $r \in [p]$, and compute the following three quantities.

- $\rho := \sum_{i \in [n]} f_i \cdot i$.
- $\phi := \sum_{i \in [n]} f_i$.
- $\tau := \sum_{i \in [n]} f_i \cdot r^i \pmod{p}$.

Consider what happens if your algorithm tests whether $\tau = \phi \cdot r^{\rho/\phi} \pmod{p}$.

4. Recall that a graph is bipartite if its vertices can be split into two sets A, B such that every edge contains exactly one vertex in A and exactly one vertex in V. Give an algorithm that works for graph streams in the insert-only update model, uses $O(n \cdot \log n)$ bits of space, and determines whether or not a graph is bipartite. Extra credit: Give an algorithm that works for graph streams in the strict turnstile update model and uses space $O(n \cdot \log^3 n)$.

5. Prove that any connected, cycle-free graph on n nodes has exactly $n - 1$ edges.