Homeworks and Research Projects on Streaming Algorithms

Qin Zhang

May 2, 2011

1 Homeworks

1. Try to analyze the *Space-saving* algorithm (discussed in lecture 1). Can the running time be $O(1)$? Does it hold if every element has a weight?

2. Figure out how to use AMS sampling for estimating F_k (discussed in lecture 2).

3. (Stable distribution, discussed in lecture 2, use the Cauchy distribution) Input: Stream from two sources $\langle x_1, x_2, \ldots, x_m \rangle \in ([n] \cup [n])^m$. Goal: Estimate difference between distribution of red values and blue values, e.g., $\sum_{i \in [n]} |f_i - g_i|$ where $f_i = |k : x_k = i|$ and $g_i = |k : x_k = i|$.

4. (#Distinct items in the flat model) Suppose there are k players each holding a set of items $S_i \subseteq [n]$. Design an efficient communication protocol that estimates the number of distinct items in the union, i.e., $|S_1 \cup \ldots \cup S_k|$, with a relative $(1+\epsilon)$-error with probability at least 3/4. The total communication cost should be $O(k + 1/\epsilon^2)$.

5. (Triangle counting) Prove that $\Omega(n^2)$ space is required to determine if # Triangles $\neq 0$, even for randomized algorithms with error no more than 1/3.

   Notice: last time we do not have time to discuss lower bounds. We will discuss it again in lecture 4. To solve this question now you probably want to go over Lectures 13-15 in Amit Chakrabarti's lecture notes, which is available in the course website.

2 Research Projects

1. Edit Distance to Monotonicity (will be discussed in lecture 3)

(a) Can we get a $(1+\epsilon)$-approx algorithm for the Edit Distance to Monotonicity? Or prove a lower bound?

(b) Can we get a $(2+\epsilon)$-approx algorithm for the Edit Distance to Monotonicity over the time-based sliding window? (the current best is a $(4+\epsilon)$-approx algorithm).

(c) Close the gap between the upper bounds and the lower bounds for general edit distance.

2. Earth-Mover Distance (will be discussed in lecture 4)

(a) Can we embed the Earth-Mover distance to product norm spaces? Note that in the algorithm we discussed we have "binary decisions" in the reconstruction algorithm $\mathcal{R}$, which we do not know how to perform the embedding. Related paper: Overcoming the l_1. Non-Embeddability Barrier: Algorithms for Product Metrics by Andoni, Indyk and Krauthgamer.

(b) Can we prove any lower bounds on the size of *sketch*?