CMPSCI 711 SPRING 2012: HOMEWORK 2
DUE 2:05 PM, MARCH 14TH

Rules: You may work in groups of at most four. Each group should submit (emailed or handed in at the start of class) one set of typed solutions (remember to include the names of everyone in the group). Cite all sources although work that uses published papers or material from related classes will receive no credit.

Question 1. *Consider a stream of $n+1$ numbers where each number is in the set $[n]$. Design a small space algorithm that returns an element that occurs at least twice in the stream.* **Hint:** *Use ℓ_1 sampling and consider the vector $g = (f_1 - 1, f_2 - 1, \ldots, f_n - 1)$ where f_i is the number of occurrences of i.*

Question 2. *Consider a stream that consists of the m (distinct) edges of a graph on n nodes. Let T be the number of triangles in the graph. Design a small space algorithm that approximate T up to additive error ϵmn.* **Hint:** *Use ℓ_0 sampling on some vector g of length $\binom{n}{3}$.*

Question 3. *In this question the goal is to modify the proof for sparse recovery in the ℓ_1 norm, to prove a similar result for the ℓ_2 norm. Some details will be similar but you should still include them.*

(1) *Using the Count-Sketch algorithm with the appropriate width and depth parameters, show that it is possible find $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_n)$ such that with probability at least $1 - 1/n^2$,*

$$\forall i \in [n], \quad |f_i - \tilde{f}_i|^2 \leq \frac{\epsilon^2 \mathrm{Err}_2^k(f)}{k} \tag{1}$$

where $\mathrm{Err}_2^k(f)$ is the sum of the $n-k$ smallest elements of $\{|f_1|^2, |f_2|^2, \ldots, |f_n|^2\}$.

(2) *Show that if $\tilde{f}$ satisfies Eq. 1 then*

$$\|f - \tilde{g}\|_2^2 \leq (1 + 9\epsilon) \min_{g : \|g\|_0 \leq k} \|f - g\|_2^2$$

where $\tilde{g}$ is the vector formed by zero-ing out all but the k largest elements of $\tilde{f}$.

Question 4. *Design an algorithm for estimating $F_2(f)$ based on Count-Sketch.* **Hint:** *Consider summing the squares of the entries of a row of the Count-Sketch table. What's the expectation and variance?*

Question 5. *We say $g = (g_1, g_2, \ldots, g_n)$ is a k-bucket histogram if there exists at most $k-1$ values of $i \in [n-1]$ such that $g_i \neq g_{i+1}$. In this question, we want to find a k-bucket histogram that approximates a vector $f = (f_1, f_2, \ldots, f_n)$. Suppose n is a power of two.*

(1) *Suppose that the entries of f are presented in order. Design a small-space algorithm that finds a k-bucket histogram $\tilde{g}$ such that*

$$\|f - \tilde{g}\|_\infty \leq (1 + \epsilon) \min_{g : g \text{ is a } k\text{-bucket histogram}} \|f - g\|_\infty \ ,$$

where $\|x\|_\infty = \max_{i \in [n]} |x_i|$. You may assume that all $f_i \in \{0, 1, \ldots, w\}$.

(2) *Suppose that the entries of f are presented in order. Design a small-space algorithm that finds a k-bucket histogram $\tilde{g}$ such that*

$$\|f - \tilde{g}\|_2 \leq (1 + \epsilon) \min_{g : g \text{ is a } k\text{-bucket histogram}} \|f - g\|_2 \ .$$

Hint: *First design a non-streaming dynamic programming algorithm for this problem based on computing* $C[b, t]$ *for* $b \in [k], t \in [n]$ *where* $C[b, t]$ *is the minimum error achievable when representing* $(f_1, f_2, \ldots, f_t)$ *by a* b*-bucket histogram. How could you save space if* $C[b, t] \approx C[b, t+1] \approx \ldots \approx C[b, t+r]$*, e.g., if* $C[b, t+r]/C[b, t] \leq (1+\gamma)$*?*

(3) *Haar wavelets are closely connected to histograms. Prove that a* k*-bucket histogram is* $2k \log_2 n$*-sparse in the Haar basis. Prove that a vector that is* k*-sparse in the Haar basis is a* $(3k+1)$*-bucket histogram.*