15-496/859X: Computer Science Theory for the Information Age — Spring 2012
Carnegie Mellon University — V. Guruswami & R. Kannan

PROBLEM SET 3
Due by 6 pm, Friday, March 9

INSTRUCTIONS

- You should think about *each* problem by yourself for *at least 30 minutes* before choosing to collaborate with others.
- You are allowed to collaborate with fellow students taking the class in solving the problems (in groups of at most 3 people for each problem). In fact this is encouraged so that you interact with and learn from each other. However, *you must write up your solutions on your own.* If you collaborate in solving problems, you should clearly acknowledge your collaborators for each problem.
- Reference to any external material besides the course text and material covered in lecture is not allowed. In particular, you are not allowed to search for answers or hints on the web. You are encouraged to contact the instructors or the TA for a possible hint if you feel stuck on a problem and require some assistance.
- Solutions typeset in LaTeX are preferred.
- Feel free to email the instructors or the TA if you have any questions or would like any clarifications about the problems.
- As always, you are strongly urged to start work on the problem set early. The problems require some thought, and are best approached without time pressure.

---

1. (20 points) In this problem, you will use a method similar to the one we used for max-cut to solve the following problem:

   Given an undirected graph $G(V, E)$, find within additive error εn^2 the minimum number of edges we need to remove from the graph to make it 3-colorable. [A graph is 3-colorable if we can assign three colors to is vertices so that no two adjacent vertices get the same color.]

   (a) Suppose A is the adjacency matrix of the graph. First suppose A was a rank-1 matrix and is the outer product of two vectors u, v, namely, $A = \sigma uv^T$, where σ is a positive real number and u, v are unit length vectors. For any subset S of vertices, denote $\sum_{i \in S} u_i$ by $u(S)$ and similarly for v. For a partition of the vertex set into 3 parts (the 3 colors) S, T, U, define a 6-vector $f(S, T, U)$ by $(u(S), v(S), u(T), v(T), u(U), v(U))$. Write down a formula for the number of edges which need to be removed if we color the vertices as per the partition S, T, U in terms of $f(S, T, U)$.

   (b) Next, still assuming that $A = \sigma uv^T$, suppose we approximate each coordinate of $f(S, T, U)$ to within $\pm\delta$ (where we will choose δ later). How many possible vectors $f(S, T, U)$ can we get from such approximations?

(c) Describe a dynamic programming algorithm to enumerate all such possible approximate vectors we can get.

(d) Now in general, approximate A by a rank k matrix B by SVD, where, k is not too large. Suppose $B = \sum_{t=1}^{k} \sigma_t u_t v_t^T$, where, u_t, v_t are now vectors.

For a partition S, T, U of the vertices, argue that the number of edges to be removed to color as in S, T, U is easy to get from the $6k-$vector:

$$f(S,T,U) = (u_1(S), u_1(T), u_1(U), v_1(S), v_1(T), v_1(U), u_2(S), u_2(T), \ldots) .$$

Approximate this vector, enumerate all possible such approximate vectors again by dynamic progamming.

Finally choose δ and argue error bounds so that we get an overall error of at most εn^2.

2. (10 points) In trying to estimate the number of distinct elements in a data stream, it was suggested in class that if we had $k-$wise independence for large enough k, then perhaps we can get an approximation to relative error ε (instead of a factor of 10 or 2). Alas, this does not work, as you will show in this exercise:

Suppose $x_1, x_2, \ldots, x_l$ are mutually independent Bernoulli (0-1) random variables such that the probability of x_1 being 1 is at most $(1-\varepsilon)/l$, so the expected number of x_i which are 1 is at most $1-\varepsilon$. Then what is the probability that at least one of them is 1?

Use this to prove what we want.

3. (20 points) Suppose we have a data stream consisting of symbols from $\{1, 2, \ldots, m\}$ with a $\pm$ sign in front of each symbol and again we define for $s \in \{1, 2, \ldots, m\}$, $f_s =$ the number of occurrences of s in the stream, where, we add or subtract 1 from f_s depending on the sign we see. [Note: f_s may be positive, negative or zero.] We are to find the first moment, namely:

$$\sum_s |f_s|.$$

(a) First argue that it won't suffice to just keep a running total (in 1 unit of space) as we could do in the case when there was no negative sign.

(b) Consider the real-valued continuous random variable Λ with probability density function $\frac{1}{\pi}\frac{1}{1+x^2}$. What is the median value of the random variable $|\Lambda|$? (That is, the value $\alpha \in \mathbb{R}$ for which $\Pr[|\Lambda| \leq \alpha] = 1/2$.)

(c) Let Z be any real-valued random variable with a continuous density function. Let $k = C\log(1/\delta)/\epsilon^2$ for a large constant C, and let $s_1, s_2, \ldots, s_k$ be independent copies of the random variable Z. Then prove that $M = \text{median}(s_1, s_2, \ldots, s_k)$ satisfies

$$1/2 - \epsilon \leq \Pr[Z \leq M] \leq 1/2 + \epsilon$$

with probability $1-\delta$ over the choice of the samples s_i. (You can assume k is odd so that the median $\text{median}(s_1, s_2, \ldots, s_k)$ is well-defined.)

Hint: Chernoff-Hoeffding bounds.

(d) *(Extra credit — this part alone; you can use this part for later ones even if you don't solve it.)*

Recall the random variable Λ from Part (3b), and let α be its median value. Suppose $t > 0$ is such that $1/2 - \epsilon \leq \Pr[|\Lambda| \leq t] \leq 1/2 + \epsilon$ for some small enough $\epsilon > 0$. Prove that $\alpha - 10\epsilon \leq t \leq \alpha + 10\epsilon$.

(e) The following is a rather remarkable fact about the above random variable Λ from part (3b) (we won't prove it here): Let $X_1, X_2, \ldots, X_s$ be i.i.d samples of the random variable Λ. Then for any integers $a_1, a_2, \ldots, a_s$, the linear combination $a_1 X_1 + a_2 X_2 + \cdots + a_s X_s$ has a distribution identical to $(\sum_{i=1}^{s} |a_i|)\Lambda$.

Using this fact and the previous parts, describe a sub-linear space streaming algorithm for estimating the first moment $\sum_s |f_s|$ of the data stream within a $(1 \pm \epsilon)$ factor with probability $1 - \delta$. Briefly justify why your algorithm works, showing how the previous pieces fit together in the analysis.

(For the sake of this problem, you can ignore issues of precision, and ignore the space needed to store samples from Λ.)

4. (10 points) In class we saw that there is a collection C of n^4 strings in $\{0,1\}^n$ which are 4-wise independent, i.e., for any 4 indices $1 \leq i_1 < i_2 < i_3 < i_4 \leq n$, for a random string $x \in C$, the distribution of $(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4})$ is uniform over $\{0,1\}^4$.

Turns out one can achieve 4-wise independence with an even smaller collection of $O(n^2)$ strings. We won't see this here, but in this exercise, you will show that quadratic is as good as it can get, i.e., every 4-wise independent collection of n-bit strings must have size at least $\Omega(n^2)$.

(a) Let $C \subseteq \{0,1\}^n$ be a 4-wise independent set. For each pair of indices $1 \leq i_1 < i_2 \leq n$, define a vector $v^{(i_1,i_2)} \in \mathbb{R}^C$ indexed by elements of C by $v^{(i_1,i_2)}(a) = (-1)^{a_i + a_j}$.

Prove that for two distinct pairs (i_1, i_2) and (j_1, j_2) (i.e., either $i_1 \neq j_1$ or $i_2 \neq j_2$ or both), the vectors $v^{(i_1,i_2)}$ and $v^{(i_1,i_2)}$ are orthogonal, i.e., $\langle v^{(i_1,i_2)}, v^{(j_1,j_2)} \rangle = 0$.

(b) Argue that the collection of vectors $v^{(i_1,i_2)}$ for pairs (i_1, i_2), $1 \leq i_1 < i_2 \leq n$, are linearly independent over reals.

(c) Deduce that $|C| \geq \binom{n}{2}$.