15-855: An Introduction to Computational Complexity Theory — V. Guruswami & M. Wootters

Problem Set #6 — Fall 2015

Due by 5 pm, Thursday, **December 10, 2015** (in person or by email to yuzhao1@cs.cmu.edu).

Instructions: Same as for problem set 1.

Solve any 4 out of the 5 problems. Note that just because a problem statement is long, that doesn't mean it is hard or has a long solution; in fact often it is the contrary. *If you turn in solutions to more than 4 problems, we will take your top scoring 4 problems.*

1. **(Why we didn't prove a lower bound for F_2)** For a data stream $a_1, a_2, \ldots, a_m$, with $a_i \in \{1, \ldots, n\}$ for all i, let f_j be the number of times that j appears in the stream. Define the k'th frequency moment by

$$F_k = \sum_{j=1}^{n} f_j^k,$$

with $F_\infty := \max_j f_j$ and $F_0 = |\{j : f_j > 0\}|$. In class, we saw that F_∞ required $\Omega(n)$ space to approximate using a randomized streaming algorithm. In this exercise, we'll show how to use $O(\log(n))$ space to approximate F_2 using a randomized streaming algorithm.[1]

For this exercise, you may assume the following fact:

> **Fact.** There is a family $\mathcal{H}$ of 4-wise independent hash functions $h : [n] \to \{\pm 1\}$ of size poly(n). Here, 4-wise independent means that for any distinct $x_1, x_2, x_3, x_4 \in [n]$ and any sign pattern $(s_1, s_2, s_3, s_4) \in \{\pm 1\}^4$,
>
> $$\mathbb{P}_{h \sim \mathcal{H}}[h(x_i) = s_i \forall i] = \frac{1}{16},$$
>
> where the probability is over h drawn uniformly at random from $\mathcal{H}$. (Notice that this is the natural extension of our definition of pairwise independent hash families).

Let $\mathcal{H}$ be as in the fact above, and let $b \in \mathbb{N}$ be a parameter. Consider the following randomized streaming algorithm, which runs on a stream $a_1, a_2, \ldots$

> Choose $h^{(1)}, \ldots, h^{(b)}$ independently, uniformly at random from $\mathcal{H}$.
> Initialize $Z^{(i)} = 0$ for $i = 1, \ldots, b$.
> For $j = 1, 2, \ldots$
> $Z^{(i)} = Z^{(i)} + h^{(i)}(a_j)$
> Return $Y = \frac{1}{b} \sum_{i=1}^{b} (Z^{(i)})^2$

(a) Argue that the algorithm above requires space $O\left(b \cdot (\log(m) + \log(n))\right)$.

(b) For each i, show that $\mathbb{E}_h[(Z^{(i)})^2] = F_2$, and hence $\mathbb{E}[Y] = F_2$.

[1] In fact, F_k can be computed with small space for $k \leq 2$, and for $k > 2$ it requires $\Omega(n)$; the algorithm for F_1 is easy—can you come up with an algorithm for F_0?

(c) For each i, show that $\mathrm{Var}[Y] \leq \frac{2F_2^2}{b}$. *(Hint: use 4-wise independence)*

(d) Show that taking $b = 2/(\epsilon^2 \delta)$,

$$\mathbb{P}_h\left[(1-\epsilon)F_2 \leq Y \leq (1+\epsilon)F_2\right] \geq 1 - \delta.$$

2. **(Public coins vs. private coins)** Recall that $R_\epsilon^{(private)}(f)$ denotes the randomized many-round communication complexity of $f : \{0,1\}^n \to \{0,1\}$, with error probability ϵ, where Alice and Bob have private randomness; $R_\epsilon^{(public)}(f)$ is the same with public randomness. Prove

$$R_{\epsilon+\delta}^{(private)}(f) \leq R_\epsilon^{(public)}(f) + O(\log(n/\delta)).$$

(Hint: Given a public coin protocol for f, suppose that there were strings $r_1, \ldots, r_t$, for $t = \mathrm{poly}(n/\delta)$ so that the protocol worked when it's random seed were drawn uniformly from $\{r_1, \ldots, r_t\}$; how could you turn this into a low-complexity private coin protocol? Now, use Chernoff bounds to show that there exist such strings $r_1, \ldots, r_t$.)

3. **(Low rank functions with a large monochromatic rectangle.)** Let $g : \mathbb{N} \to \mathbb{N}$ be a function so that the following holds: suppose that for any $f : X \times Y \to \{0,1\}$ with $\mathrm{rank}(f) = r$, there is a monochromatic rectangle in $M_f \subset \{-1,1\}^{X \times Y}$ of size $|R| \geq 2^{-g(r)}|X \times Y|$. As in class, $\mathrm{rank}(f)$ is the rank of M_f over $\mathbb{R}$, and M_f is the matrix with $(M_f)_{x,y} = (-1)^{f(x,y)}$.

   Show that any function $f : X \times Y \to \{0,1\}$ with $\mathrm{rank}(f) \leq r$ is computable by a deterministic many-round communication protocol with communication cost $O(\log^2(r) + \sum_{i=0}^{\log(r)} g(r/2^i))$.

   (Hint: Define a protocol that reduces the rank of f from r to $r/2$, and bound the number of leaves of this protocol. You may use the fact that any protocol with T leaves may be balanced to run with $O(\log(T))$ communication cost.)

   Notice that if the hypothesis holds with $g(r) = \mathrm{polylog}(r)$, this would imply the log rank conjecture.

4. **(To prove the log-rank conjecture, it suffices to show that low-rank functions have low randomized communication complexity)** In this exercise you will prove the following theorem.

   Theorem. Let $f : X \times Y \to \{0,1\}$. Suppose that $R^{(pub)}(f) = c$ is the randomized (public coin) communication cost of f. Then the deterministic communication cost of f is

   $$D(f) = O(c \log^2(\mathrm{rank}(f))),$$

   where as in class $\mathrm{rank}(f)$ is the rank (over $\mathbb{R}$) of the matrix $M_f \in \{-1,+1\}^{|X| \times |Y|}$ which has $(M_f)_{x,y} = (-1)^{f(x,y)}$.

   (a) Suppose that there is a randomized protocol for f with communication complexity c. For any $\epsilon > 0$, show that there's a deterministic protocol Π that partitions M_f into $N = 2^{O(\log(1/\epsilon)c)}$ rectangles $R_1, \ldots, R_N$, so that there is some rectangle R_i with

   $$|\{(x,y) \in R_i : f(x,y) = 1\}| \geq (1 - 2\epsilon)|R_i|,$$

and $|R| \geq \frac{|X \times Y|}{2N}$. That is, there is some reasonably large rectangle on which the value of f is nearly constant.

(Hint: Use a deterministic protocol which is correct with probability at least $1-\epsilon$ when x, y are drawn uniformly at random.)

(b) Prove the following claim:

Claim. Suppose that f has $\mathrm{rank}(f) = r$, and that there is some rectangle $R \subset X \times Y$ so that

$$|\{(x,y) \in R : f(x,y) = 1\}| \geq \left(1 - \frac{1}{4r}\right)|R|.$$

Then there is a sub-rectangle $R' \subseteq R$ with $|R'| \geq |R|/8$, so that $f(x,y) = 1$ for all $(x,y) \in R'$.

(c) Use parts (a) and (b) to prove the theorem. *(Hint: Use the previous problem)*

5. **(Log-rank conjecture for XOR function and parity decision tree.)** We call a function $F : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ an XOR function if for some Boolean function $f : \{0,1\}^n \to \{-1,+1\}$, we have $F(x,y) = f(x \oplus y)$ for all $x, y \in \{0,1\}^n$. (As in class, $\mathrm{rank}(F)$ is the rank of the matrix M_F over $\mathbb{R}$, and M_F is the matrix with $(M_F)_{x,y} = F(x,y)$). Recall that $\widehat{f}(\alpha)$ is a Fourier coefficient of f, and $\mathrm{sparsity}(f)$ be the number of nonzero $\widehat{f}(\alpha)$ among $\alpha \in \{0,1\}^n$ (See Homework 4).

(a) Show that $\mathrm{rank}(F) = \mathrm{sparsity}(f)$.

(b) A *parity decision tree* is a variant of a decision tree in which the nodes are allowed to query arbitrary parities of the input variables. We denote by $\mathrm{DT}^{\oplus}(f)$ the depth of the shortest parity decision tree that computes f. Show that

$$\frac{1}{2} \log \mathrm{sparsity}(f) \leq \mathrm{DT}^{\oplus}(f) \leq \mathrm{sparsity}(f).$$

(You only need to show the second inequality. The first inequality is similar to Problem 5 in Homework 4.)

(c) Show that the deterministic communication cost of F satisfies

$$D(F) \leq 2\,\mathrm{DT}^{\oplus}(f).$$

Conclude that if $\mathrm{DT}^{\oplus}(f) = O(\log^c(\mathrm{sparsity}(f)))$ (though this is still open), then the log-rank conjecture holds for the XOR function.