KTH Computer Science
and Communication

Communication Complexity: Problem Set 2

Due: September 30, 2012. Submit as a PDF file by e-mail to `lauria at kth dot se` with the subject line `Problem set 2:` ⟨`your name`⟩. Name the PDF file `PS2_`⟨`YourName`⟩`.pdf` (with your name coded in ASCII without national characters). Solutions should be written in LaTeX or some other math-aware typesetting system. Please try to be precise and to the point in your solutions and refrain from vague statements. In addition to what is stated below, the general rules stated on the course webpage always apply.

Collaboration: Discussions of ideas in groups of two to three people are allowed, but you should write down your own solution individually and understand all aspects of it fully. For each problem, state at the beginning of your solution with whom you have been collaborating.

Reference material: For some of the problems, it might be easy to find solutions on the Internet, in textbooks or in research papers. It is not allowed to use such material in any way unless explicitly stated otherwise. You can refer without proof to anything said during the lectures on in the lecture notes, except in the obvious case when you are specifically asked to show something that we claimed without proof in class. It is hard to pin down 100% formal rules on what all this means—when in doubt, ask the lecturer.

About the problems: Some of the problems in the problem sets are meant to be quite challenging and you are not necessarily expected to solve all of them. A total score of around 50 points should be enough for grade E, 100 points for grade C, and 150 points for grade A on this problem set. Any corrections or clarifications will be posted on the course webpage `www.csc.kth.se/utbildning/kth/kurser/DD2441/semteo12/`.

1 (10 p) In the guest lectures on property testing, we discussed among other things the problem of testing bipartiteness of graphs. The graph is given as an adjacency matrix, we want to query as few positions in the adjacency matrix as possible, and then answer whether the graph is bipartite or not and be correct with high probability. In class we saw a property tester that picks a subset of vertices of constant size, checks if the subgraph consisting of these vertices and the edges between them is bipartite, and decides that the whole graph is probably bipartite if this subgraph is. We also learned (although we did not see the proof) that this turns out to be an excellent property tester.

The purpose of this problem is to analyse another interesting suggestion for a property tester that was given in class. The idea here was to choose a vertex in the graph, do a breadth-first search in a small neighbourhood around this vertex (consisting of a constant number of vertices, say), and accept if the neighbourhood graph is bipartite.

Is this second tester a good tester for graph bipartiteness, in the sense that we get similar query complexity and error probability as for the property tester presented in class? To get full credit on this problem, you need to get the answer right, but you do *not* need to give the full details. Just sketch why the property tester above seems to work just as fine (if you think it does), or point out where you think the problems are (if you think there are problems).

2 (10 p) We proved in our lectures on streaming algorithms that any randomized algorithm that exactly computes the kth frequency moment F_k for any nonnegative integer $k \neq 1$, gets the answer right with probability at least $1 - \epsilon$ for $\epsilon < 1/2$, and makes just a constant number of passes over the input must use $\Omega(n)$ space. In the proof, we reduced to the set disjointness problem where Alice and Bob get $A, B \subseteq [n]$ and want to know whether $A \cap B = \emptyset$ or not. If S is the sequence consisting of A and B concatenated (and where we assume that $A \cup B \neq \emptyset$), we claimed that for $k \geq 2$ it holds that $A \cap B = \emptyset$ if and only if the kth frequency moment of S is $F_k = |A| + |B|$ and that for $k = \infty$ it holds that $A \cap B = \emptyset$ if and only if $F_k = 1$. Prove that this is so.

3 (20 p) The streaming algorithm in [AMS99] for approximating frequency moments is constructed in several steps. In the first step, one defines a random variable X with the right expectation but with terrible variance. In the second step, one lets $Y = \frac{1}{m}\sum_{i=1}^{m} X_i$ for a suitably large number m of independent copies X_i of the experiment X to bring down the variance, and using Chebyshev's inequality one can show that Y is λ-close to F_k with high (but constant) probability. To get arbitrarily high probability $1-\epsilon$ of being λ-close to F_k, in the third step one then samples a number of independent copies of Y and takes the median, where an application of a Chernoff bound concludes the analysis.

Suppose that instead of taking the median of Y's we would have answered with a mean $Y = \frac{1}{m}\sum_{i=1}^{m} X_i$ but for m large enough so that this Y is λ-close to F_k with probability $1 - \epsilon$. How would that have affected the space requirements?[1] In particular, roughly how much more space would be needed for this second algorithm compared to the one in [AMS99] if we want a guaranteed error probability of at most 0.1%?
Hint: Apply Chebyshev's inequality and calculate how many independent samples would be needed.

4 (20 p) Recall that a function $f : \{0,1\}^n \mapsto \{0,1\}$ is k-linear if there is a set $S \subseteq [n]$, $|S| = k$, such that $f(x_1, x_2, \ldots, x_n) = \sum_{i \in S} x_i$ (where all operations are in the field $\mathbb{F}_2 = \mathrm{GF}(2)$ of two elements). Recall also that the distance between two functions f and g is the fraction of inputs in $\{0,1\}^n$ on which f and g disagree, and that we say that f and g are δ-far from each other if this distance is at least δ.

In the proof of the property testing lower bounds for k-linearity from [BBM12] that we saw in class, we needed the fact that if $k' \neq k$, then any k'-linear function is $1/2$-far from any k-linear function. Prove that this is true.
Hint: If $f(x) = \sum_{i \in S} x_i$ and $g(x) = \sum_{i \in T} x_i$ for $S \neq T$, focus on some coordinate i that is in S but not in T or the other way round.

5 (30 p) Recall that for $x, y \in \{0,1\}^n$, we let $\mathsf{GT}_n(x, y)$ be the function that evaluates to 1 if $x > y$ interpreted as n-bit numbers and to 0 otherwise. What is the best 2-player randomized public-coin protocol that you can give for this problem?

[1] And hence also the time requirements, although we will not focus on that.

6 (30 p) In our lectures on streaming algorithms, we also proved that any deterministic streaming algorithms that approximates F_k, $k \neq 1$, well must use $\Omega(n)$ space. For this proof we needed the existence of a family $\mathcal{G}$ of subsets $G_i \subseteq [n]$ satisfying the following properties:

1. $\mathcal{G}$ consists of $2^{\Omega(n)}$ subsets G_i.
2. Every $G_i \in \mathcal{G}$ has size $|G_i| = n/64$.
3. For every two $G_i, G_j \in \mathcal{G}$ such that $G_i \neq G_j$, it holds that $|G_i \cap G_j| \leq n/128$.

Prove that such a family of subsets $\mathcal{G}$ exists.
Hint: Count how many such subsets G_i there are of the right size and then how many of them intersect too much. Show that you have enough sets left to get a family $\mathcal{G}$ with as many subsets as stated. In your calculations, you can use the standard inequalities $\left(\frac{n}{s}\right)^s \leq \binom{n}{s} \leq \left(\frac{en}{s}\right)^s$ (these inequalities are far from tight for the range of parameters we are studying here, but will be good enough for our purposes anyway). You might also want to use that for $c < k < \ell < n/2$ it holds that $\binom{n-\ell}{k}\binom{\ell}{k} > \binom{n-\ell}{k-c}\binom{\ell}{k+c}$.

7 (40 p) We saw in class that any adaptive 2-sided error property tester for mononoticity of functions $f : \{0,1\}^n \mapsto R$ has query complexity $\Omega(n)$ when the size of the range of the function is $|R| = \Omega(n)$. The proof was by a reduction from set disjointness.

More precisely, for fixed $A, B \subseteq [n]$ and any $x \in \{0,1\}^n$ we defined $f(x) = (-1)^{\sum_{i \in A} x_i}$, $g(x) = (-1)^{\sum_{i \in B} x_i}$, and $h(x) = 2 \cdot \mathrm{wt}(x) + f(x) + g(x)$, where $\mathrm{wt}(x)$ is the *Hamming weight* of x, i.e., the number of 1s in x. Then the following holds:

1. If $A \cap B = \emptyset$, then h is monotone.
2. If $A \cap B \neq \emptyset$, then h is $\frac{1}{8}$-far from monotone.
3. The range of h has size $\Omega(n)$ (it is $[-2, 2n+2]$).

By the technique developed in [BBM12], the fact that disjointness has randomized communication complexity $\Omega(n)$ now immediately implies a property testing lower bound of $\Omega(n)$.

We then claimed, but did not prove, that a similar property testing lower bound can be proved for functions $f : \{0,1\}^n \mapsto R$ with range of size $|R| = \mathrm{O}(\sqrt{n})$, and used this to obtain a more general result. The purpose of the current problem is to establish this claim.

7a Prove that at least a fraction $\frac{15}{16}$ of all bitstrings $x \in \{0,1\}^n$ of length n have Hamming weight satisfying $\frac{n}{2} - \frac{4\sqrt{n}}{2} \leq \mathrm{wt}(x) \leq \frac{n}{2} + \frac{4\sqrt{n}}{2}$.
Hint: Use Chebyshev's inequality.

7b With notation as above, define a new function

$$h'(x) = \begin{cases} +\infty & \text{if } \mathrm{wt}(x) \geq \frac{n}{2} + \frac{4\sqrt{n}}{2} \\ h(x) & \text{if } \frac{n}{2} - \frac{4\sqrt{n}}{2} \leq \mathrm{wt}(x) \leq \frac{n}{2} + \frac{4\sqrt{n}}{2} \\ -\infty & \text{if } \mathrm{wt}(x) \leq \frac{n}{2} - \frac{4\sqrt{n}}{2} \end{cases}$$

and prove that h' is monotone if $A \cap B = \emptyset$.[2]

[2] If you are uncomfortable with $\pm\infty$, you can think of them as denoting suitable large and small integers.

7c Prove that if $A \cap B \neq \emptyset$, then h' is δ-far from monotone for some constant $\delta > 0$.

7d Prove that the size of the range of h' is $\mathrm{O}(\sqrt{n})$.

Putting together the pieces in 7a–7d, the claimed improvement now follows (which you do not need to prove).

Note that you can solve the subproblems independently of one another to get partial credit.

8 (40 p) Alice, Bob, and Carol get inputs $x, y, z \in \{0,1\}^n$ and want to determine whether $x = y = z$ or not. They can only see their own input. They are all deterministic and cannot broadcast but only pass private messages from one sender to one receiver at any one given time. What is the best lower bound you can prove on the communication complexity of this problem in this 3-player deterministic number-in-hand message-passing communication model?

An important note is that in this problem we care about the concrete constants, so do not hide constants in big-oh notation but instead compute exactly the best lower bound you can get.