KTH Computer Science and Communication

Computational Complexity: Problem Set 4

Due: Friday January 8, 2016, at 23:59 AoE. Submit your solutions as a PDF file by e-mail to `jakobn at kth dot se` with the subject line `Problem set 4: ⟨your full name⟩`. Name the PDF file `PS4_⟨YourFullName⟩.pdf` with your name written in CamelCase without blanks and in ASCII without national characters. State your name and e-mail address at the very top of the first page. Solutions should be written in LaTeX or some other math-aware typesetting system with reasonable margins on all sides (at least 2.5 cm). Please try to be precise and to the point in your solutions and refrain from vague statements. *Write so that a fellow student of yours can read, understand, and verify your solutions.* 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 people are allowed—and indeed, encouraged—but you should write down your own solutions individually and understand all aspects of them fully. You should also acknowledge any collaboration. State at the very top of the first page of your problem set solutions if you have been collaborating with someone and if so with whom. (Note that collaboration is on a per problem set basis, so you should not discuss different problems on the same problem set with different people.)

Reference material: Some of the problems are "classic" and hence 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. Anything said during the lectures or in the lecture notes, or which can be found in chapters of Arora-Barak covered in the course, should be fair game, though, unless you are specifically asked to show something that we claimed without proof in class. All definitions used should be as given in class or in Arora-Barak and cannot be substituted by definitions from other sources. It is hard to pin down 100% watertight formal rules on what all of this means—when in doubt, ask the lecturer.

About the problems: Some of the problems are meant to be quite challenging and you are not necessarily expected to solve all of them. A total score of around 85 points should be enough for grade E, 120 points for grade D, 155 points for grade C, 190 points for grade B, and 225 points for grade A on this problem set. Any corrections or clarifications will be given at `piazza.com/kth.se/fall2015/dd2445/` and any revised versions will be posted on the course webpage `www.csc.kth.se/DD2445/kplx15/`.

1 (10 p) The language QuadEq consists of systems of quadratic equations $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ that have 0/1-solutions, where by a quadratic equation E_ℓ we mean an equation on the form

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{\ell,i,j}x_i x_j = b_\ell$$

for $a_{\ell,i,j}, b_\ell \in \{0,1\}$, and where all arithmetic is over GF(2).

Show that there is a reduction R from CircuitSat to QuadEq with the property that if C is a circuit of size s with t inputs, then $R(C)$ is a set of quadratic equations with $O(s)$ equations over s variables such that in any solution to $R(C)$ the first t variables encode a satisfying input for C.

2 (10 p) Suppose that $\mathcal{P}$ is a strongly testable property, where $\mathcal{P} = \bigcup_{n=1}^{\infty} \mathcal{P}_n$ for $\mathcal{P}_n \subseteq \{f : \mathsf{D}_n \to \mathsf{R}\}$ (i.e., some subset of all functions from some family of domains $\mathsf{D_n}$ to some fixed range R). Prove that there is a non-adaptive one-sided tester for $\mathcal{P}$ with constant query complexity.

3 (20 p) Given a network $G = (V, E)$ of computational agents $V = [n]$ connected by communication links $(i, j) \in E$, where we assume that G is a connected (undirected) graph of diameter D, give a deterministic distributed algorithm that computes the total number of edges $|E|$ in G in $\mathrm{O}(D)$ rounds.

You can assume that all agents know n and that they know the index of their own vertex (as well as the indices of their neighbours), but the diameter D is *not* known in advance. At the end of the protocol, all agents should know $|E|$, and they should also know that the algorithm has terminated. In every round $\mathrm{O}(\log n)$ bits can be sent (in full duplex) over every link $(i, j) \in E$.

4 (20 p) In our lecture on property testing, we studied the 2^n-dimensional vector space of functions $f : \{\pm 1\}^n \to \mathbb{R}$ with inner product

$$\langle f, g \rangle = \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} f(x) g(x) \ .$$

In class, we claimed without too much of a proof that the set of functions $\{\chi_\alpha\}_{\alpha \subseteq [n]}$ defined by $\chi_\alpha(x) = \prod_{i \in \alpha} x_i$ form an orthonormal basis for this vector space, namely the *Fourier basis* that we then used to analyze the linearity test.

Fill in the details to establish this claim! That is, show that

$$\langle \chi_\alpha, \chi_\beta \rangle = \begin{cases} 1 & \text{if } \alpha = \beta; \\ 0 & \text{otherwise;} \end{cases}$$

and argue that this implies that $\{\chi_\alpha\}_{\alpha \subseteq [n]}$ is indeed an orthonormal basis.
Hint: Consider the symmetric difference $\gamma = \alpha \triangle \beta = (\alpha \cup \beta) \setminus (\alpha \cap \beta)$ and prove that it holds that $\sum_{x \in \{\pm 1\}^n} \chi_\gamma(x) = 0$ if $\gamma \neq \emptyset$.

5 (20 p) One of the two key technical lemmas in Irit Dinur's proof of the PCP theorem is a *gap amplification lemma* which can be stated as follows:

> *For every $q_0, \ell \in \mathbb{N}^+$ there exist $W \in \mathbb{N}^+$, $\kappa > 0$ and $\epsilon_0 > 0$ such that there is a polynomial-time reduction R from $\textsc{Max}q_0\textsc{CSP}_2$ to $\textsc{Max2CSP}_W$ satisfying the following properties:*
>
> - *If $val(\varphi) = 1$, then $val(R(\varphi)) = 1$.*
> - *If $val(\varphi) \leq 1 - \epsilon$ for $\epsilon < \epsilon_0$, then $val(R(\varphi)) \leq 1 - \ell\epsilon$.*
> - $|R(\varphi)| \leq \kappa \cdot |\varphi|$.

In this lemma, the blowup of the alphabet size is from q_0 to some unspecified W, but this W is independent of the size of the $\textsc{Max}q_0\textsc{CSP}_2$ instance φ.

Show that if we instead allow the alphabet blowup to be a function of the instance size (and drop the condition that the reduction should be polynomial-size), then for every $\epsilon' > 0$ there exist $W = W(|\varphi|)$ such that there is a reduction R from $\textsc{Max}q_0\textsc{CSP}_2$ to $\textsc{Max2CSP}_W$ as above except that if $val(\varphi) < 1$, then $val(R(\varphi)) \leq \epsilon'$.

6 (20 p) Let us define the MAXqCSP$_W$ problem for some fixed $W \in \mathbb{N}^+$ to consist of instances that are collections of m constraints $(C_1, I_1), (C_2, I_2), \ldots, (C_m, I_m)$ over a set of n variables, where each $C_j : [W]^q \to \{0,1\}$ is some q-ary predicate and each $I_j = \{i_{j,1}, i_{j,2}, \ldots, i_{j,q}\} \in [n]^q$ is a set of q variable indices. An assignment $\alpha \in [W]^n$ satisfies (C_j, I_j) if $C_j\big(\alpha_{i_{j,1}}, \alpha_{i_{j,2}}, \ldots, \alpha_{i_{j,q}}\big) = 1$, and the task is to compute the maximal number of constraints that can be satisfied by any assignment. Determine for which values of $q, W \in \mathbb{N}^+$ the MAXqCSP$_W$ problem is easy and for which values it is NP-hard.
Remark: Note that MAXqCSP$_W$ is not a decision problem, and so it does not quite make sense to ask whether it is NP-complete or not. However, we can still prove that it is NP-hard in the sense that there is a polynomial-time reduction from some NP-complete problem such that we could decide this problem efficiently if we had a polynomial-time algorithm for MAXqCSP$_W$.

7 (30 p) Prove that if $G = (V, E)$ is an (n, d, λ)-spectral expander, then for every subset of vertices $S \subseteq V$ with $|S| \leq n/2$ it holds that

$$\Pr_{(u,v)\in E}[u \in S \text{ and } v \in S] \leq \frac{|S|}{n}\left(\frac{1}{2} + \frac{\lambda}{2}\right) \ .$$

Remark: There are quite generous hints in Arora-Barak on this problem referring to material in Chapter 21. You are allowed (and encouraged) to make use of these hints, but you may not use any statements from Chapter 21 without proof. You have to provide a stand-alone solution (except that it can be based on what we did in class as specified in the handwritten lecture notes), and for any results you want to use from Chapter 21 you also have to provide complete, written proofs in your solution why these results hold.

8 (50 p) Prove the following more general version of the result shown in class during Danupon Nanongkai's guest lectures:

> *For any $b > 0$ and any function $f : \{0,1\}^b \times \{0,1\}^b \to \{0,1\}$ there is a distributed network $G(b)$ with $\Theta(b^2)$ nodes and diameter $\mathrm{O}(\log b)$ such that if there is a distributed algorithm A on $G(b)$ that computes f in $T < b/2$ rounds, then there is a 2-party deterministic communication protocol for $f : \{0,1\}^b \times \{0,1\}^b \to \{0,1\}$ using at most $\mathrm{O}\big(T\mathrm{polylog}(n)\big)$ bits of communication.*

Hint: Note that there is fairly detailed information in the slides posted on the course webpage how the two parties Alice and Bob in the communication protocol for $f : \{0,1\}^b \times \{0,1\}^b \to \{0,1\}$ should be embedded in a network $G(b)$. Your task is to go through the weaker result proven in class and then work out the additional details needed to establish the stronger theorem above.

9 (80 p) Let PCP$_{c,s}[r(n), q(n)]$ denote the class of languages which have PCP systems with completeness c, soundness error s, randomness $r(n)$, and query complexity $q(n)$. (Note that we follow the standard definition here in that we do *not* add big-ohs for the randomness and query complexity in the definition as in Arora-Barak. The verifier is still non-adaptive, though.) In this problem we want to discuss how the class of languages captured by PCP$_{c,s}[r(n), q(n)]$ changes as we vary the parameters.

9a (10 p) Show that PCP$_{1,0}[0, \mathrm{poly}(n)] = \mathsf{NP}$.

9b (10 p) Show that PCP$_{1,0}[\log n, 42] = \mathsf{P}$.

9c (30 p) Show that $\mathsf{PCP}_{1,2^{-100}}[\mathrm{poly}(n), 42] = \mathsf{coRP}$.

9d (30 p) For how large values of the soundness error s are you able to show that it holds that $\mathsf{PCP}_{1,s}[\mathrm{poly}(n), 42] = \mathsf{coRP}$? Can you give some argument why for large enough values $s' > 2^{-100}$ (but still as small as possible) it might hold that $\mathsf{PCP}_{1,s'}[\mathrm{poly}(n), 42] \neq \mathsf{coRP}$ (under some more or less believable complexity-theoretic assumption, say)? How large do you need s' to be for this argument?

Remark: Please motivate all your answers carefully. If something is "obvious," then make sure to indicate *why* this is so.

10 (110 p) The purpose of this problem is to study some of the details left out during the overview of the proof of Lemma 22.9 in Arora-Barak (the *Powering lemma*), and in the end to try to patch together a complete proof of the lemma.

It is important to note however, that all of the subproblems below can be solved in isolation, regardless of whether you want to try to produce a full proof of Lemma 22.9 or not. Also, results claimed in previous subproblems can be used freely in solutions of later subproblems even if you personally did not solve those previous problems.

10a (20 p) Let $G = (V, E)$ be an undirected d-regular graph, possibly with multiple copies of edges but *without* self-loops. Let $\ell \in \mathbb{N}^+$ be fixed. Consider the following experiment:

1. Uniformly at random pick a vertex $v_0 \in V$.
2. For $i = 1, 2, \ldots, \ell$, uniformly at random pick an edge (v_{i-1}, v_i) incident to v_{i-1} and walk to v_i.
3. Output the edge $(v_{\ell-1}, v_\ell)$.

Prove that $(v_{\ell-1}, v_\ell)$ is a uniformly random edge in E (where we have to distinguish distinct copies of edges if G is a multi-graph).

Now suppose that G might also have self-loops. Show that the experiment above results in a distribution over the edges in E where every self-loop has half the probability of being chosen compared to any proper edge between distinct vertices. Hence, the distribution is still uniform if we count all self-loops (v, v) once and all proper edges (u, v), $u \neq v$, twice (once for u and once for v).

Remarks: Note that each self-loop (v, v) only counts as *one* edge and hence only adds 1 to the degree of v (or, equivalently, only adds a term $1/d$ in the normalized adjacency matrix). Just to give a concrete example, the graph $G = (V, E)$ with $V = \{1, 2, 3\}$ and $E = \{(1, 1), (1, 2), (2, 3), (3, 3)\}$ is a 2-regular graph with normalized adjacency matrix

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

(where indeed self-loops are only "counted once" and other edges are "counted twice").

Also note that since it is in some sense "intuitively obvious" that in a regular graph without self-loops the edge $(v_{\ell-1}, v_\ell)$ is uniformly random, we really want a formal proof that the experiment above generates a uniform distribution over the edges.

10b (30 p) For two random variables X and Y ranging over a finite set Ω of outcomes, their *statistical distance* (also known as *total variation distance*) is defined as

$$\Delta(X,Y) = \max_{S\subseteq\Omega}\bigl\{\bigl|\Pr[X\in S]-\Pr[Y\in S]\bigr|\bigr\} = \frac{1}{2}\sum_{z\in\Omega}\bigl|\Pr[X=z]-\Pr[Y=z]\bigr| \ .$$

Let S_t be the binomial distribution over t balanced coins (i.e., $\Pr[S_t=k]=\binom{t}{k}2^{-t}$). Prove that for every $\delta<1$, it holds that $\Delta\bigl(S_t,S_{t+\delta\sqrt{t}}\bigr)\le 10\delta$.

Hint: Use Stirling's formula (and possibly other useful stuff that you can find in Appendix A of Arora-Barak).

10c (20 p) Suppose that V is a non-negative discrete random variable and that we want to prove a lower bound on $\Pr[V>0]$. First show that the expected value $\mathrm{E}[V]$ does not say anything about $\Pr[V>0]$ in the sense that $\mathrm{E}[V]$ can be arbitrarily large and $\Pr[V>0]$ arbitrarily small at the same time. Then prove the inequality

$$\Pr[V>0]\ge\frac{(\mathrm{E}[V])^2}{\mathrm{E}\bigl[V^2\bigr]} \ .$$

Hint: Consider the random variable V' distributed as V conditioned on $V>0$. Show that $\mathrm{E}\bigl[(V')^2\bigr]\ge\mathrm{E}[V']^2$ and use this to derive the inequality.

10d (40 p) Using the results in the subproblems above and in Problem 7 (regardless of whether you solved these problems or not), present a complete, self-contained proof of the Powering lemma. The goal of this exercise is (at least) twofold:

- To have you work out the proof in detail and make sure you understand it.
- To train your skills of mathematical writing.

When you write the proof, you can freely consult the lecture notes as well as the relevant material in Arora-Barak, but you need to fill in any missing details. Also, the resulting write-up should stand on its own without referring to the proof of the lemma as presented in Arora-Barak or in the lecture notes. Your write-up should be accessible to a fellow student who has studied and understood the material presented in this course *except* for the final two lectures when we discussed the Powering lemma.

For a full score you need to produce (a) a crisp, clear, easy-to-read exposition where (b) all the technical details skipped over in class and/or the textbook have been taken care of properly. You could think of the goal as producing nice, LaTeX:ed lecture notes for a hypothetical 24th lecture on the full proof of the Powering lemma.