CmpE 587: Intro to Research in TCS **Boğaziçi University, Fall 2014**

Homework 1

Due: October 22, 11:59pm.

Email your PDF as an attachment to cmpe587homework@gmail.com

The attached file should be called `homework2-yourfirstname-yourlastname.pdf`

Homework policy: Please try to do the homework by yourself. If you get stuck, working in a group of two is okay, three at the most. Naturally, acknowledge any sources you worked with at the top of your solutions. LaTeX typesetting with pdf output is mandatory. Questions about the homework, LaTeX, or other course material can be asked on Piazza.

---

1. **A two-part problem on Chernoff bounds.**

   (a) Suppose $\boldsymbol{X}$ is a random variable which is 1 with probability p and 0 with probability $1-p$; however, *you don't know* p. You wish to estimate it from samples.[1] Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent random variables with the same distribution as $\boldsymbol{X}$ and let

   $$\widehat{p} = \frac{1}{n}(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n).$$

   Show that for any $0 < \delta, \epsilon < 1$, if

   $$n \geq \frac{3}{\epsilon^2} \ln(2/\delta),$$

   then

   $$\mathbf{Pr}[|\widehat{p} - p| \leq \epsilon] \geq 1 - \delta.$$

   In other words, it's very unlikely — at most a δ chance — that the estimate $\widehat{p}$ is more than ϵ-far from the true mean p. Remark: if you get a different constant than "3" in your bound for n, that's okay.

   (b) Suppose we form a "random 3CNF constraint satisfaction problem" as follows. Let $x_1, \dots, x_n$ denote Boolean variables. For each $1 \leq i \leq m$ independently, we choose a "random 3-clause constraint" $\boldsymbol{C}_i$ uniformly at random from the $8\binom{n}{3}$ possibilities. (A 3-clause is an expression like $x_7 \vee \overline{x}_{12} \vee \overline{x}_{85}$ — the OR of exactly three different variables/negated-variables.)

   Each truth assignment to the variables $x_1, \dots, x_n$ will satisfy some constraints from $\boldsymbol{C}_1, \dots, \boldsymbol{C}_m$ and falsify others. Let OPT denote the fraction of constraints satisfied by the best truth assignment. Show that there is an absolute constant K such that if $m \geq Kn$, then

   $$\mathbf{Pr}[\mathrm{OPT} \geq .875001] \leq 2^{-n}.$$

   (Hint: you will need to combine the mighty Chernoff bound with the lowly Union Bound.)

[1] For example, suppose you are a pollster and you are trying to figure out what is the fraction p of the population that supports the president. In this case, think of $\boldsymbol{X}$ as what you get if you pick a random citizen and ask them to report 1 if they support the president and 0 otherwise. Naturally, you will pick a bunch of random citizens and average their answers in an attempt to estimate p.

2. Let n be a positive odd integer. Give an exact expression $C(n)$ for the number of perfect matchings in K_{n+1}. Here K_{n+1} denotes the complete graph on $n+1$ vertices (in which all possible $\binom{n+1}{2}$ edges are present) and a "perfect matching" is collection of edges that touch each vertex exactly once. Further, "asymptotically determine" $C(n)$; that is, express $C(n) = D(n)(1 \pm o(1))$ as $n \to \infty$, where $D(n)$ is a "simple expression".

3. On Planet Negezeg there are m days in the year. On this planet, we choose a group of n people, where $n \leq m$. Assume each person's birthday is uniformly random one of the m days. Let $p_{n,m}$ denote the probability that all n people have distinct birthdays.

   (a) Derive an exact expression for $p_{n,m}$.

   (b) Show that $p_{n,m} \leq \exp(-\frac{(n-1)^2}{2m})$. (Here we are using the notation $\exp(t) = e^t$.) (Hint: you will want to recall that $\exp(x) \geq 1 + x$ for all real x.)

   (c) Deduce that if $n \geq \sqrt{2\ln 2}\sqrt{m} + 1$ then there is at least a 50% chance that two of the n people share a birthday.

   (d) Assuming $n \leq m/2$, prove that $p_{n,m} \geq \exp(-\frac{(n-1)^2}{2m})(1 - O(n^3)/m^2)$. (Hint: unfortunately, $\exp(x) \leq 1 + x$ is not true, but you should prove something like it.)

4. Consider the Word RAM model. Suppose we wish to "allocate" and use an "array" of n words. An underspecified aspect of the Word RAM model is whether we can assume that all memory words are initially set to 0^w (the w-bit word of all zeros), or whether — as in many real computers — memory words are initialized to unknown "junk" strings in $\{0,1\}^w$. The point of this problem is to show that it doesn't really matter. Precisely, you should show that even in the "junk" model, it is possible to "allocate an array of n words" and assume that they are each initialized to 0^w. Your data structure should support initialization, reading-a-word, and setting-a-word all in $O(1)$ time. You may use $O(n)$ additional space.

5. Let $\pi(n)$ denote the number of prime numbers less than or equal to n (so $\pi(1) = 0$, $\pi(2) = 1$, $\pi(3) = 2$, $\pi(4) = 2$, etc.). In this problem you will give a simple derivation of a good asymptotic lower bound on $\pi(n)$.

   (a) For $n \geq 1$, let $C_n = \binom{2n}{n} = \frac{(2n)!}{n!n!}$. Give a very simple proof that $C_n \geq 2^n$.

   (b) Let p be prime, and suppose that the largest power of p dividing C_n is p^k. Show that

$$k = \left(\left\lfloor \frac{2n}{p} \right\rfloor - 2\left\lfloor \frac{n}{p} \right\rfloor\right) + \left(\left\lfloor \frac{2n}{p^2} \right\rfloor - 2\left\lfloor \frac{n}{p^2} \right\rfloor\right) + \left(\left\lfloor \frac{2n}{p^3} \right\rfloor - 2\left\lfloor \frac{n}{p^3} \right\rfloor\right) + \cdots$$

   (c) Show that each term in parentheses above is either 0 or 1. Deduce that $p^k \leq 2n$.

   (d) Conclude that $\pi(2n) \geq n/\log_2(2n)$ and therefore $\pi(n) = \Omega(n/\log n)$.

   Remark: actually, the following stronger statements are true:

   - $\pi(n) = \Theta(\frac{n}{\ln n})$. (You did the lower bound. Proving $\pi(n) = O(\frac{n}{\ln n})$ is also quite easy.)
   - $\pi(n) \sim \frac{n}{\ln n}$. (This is called the *Prime Number Theorem.*)
   - $\pi(n) = \frac{n}{\ln n} \cdot \left(1 \pm O\left(\frac{1}{\log n}\right)\right)$.
   - $\pi(n) = \mathrm{Li}(n) \cdot (1 \pm e^{-\Omega(\sqrt{\log n})})$, where $\mathrm{Li}(n) := \int_2^n \frac{dt}{\ln t}$. (This is the bound derived in the first proofs of the Prime Number Theorem.)

- $\pi(n) = \mathrm{Li}(n) \cdot (1 \pm e^{-\Omega((\log n)^{3/5}/(\log\log n)^{1/5})})$. (This is the best known bound, due to Vinogradov and Korobov.)
- $\pi(n) = \mathrm{Li}(n) \pm O(\sqrt{n}\log n)$. (Actually this statement is not known to be true, but it's strongly believed. It's **equivalent** to the famous "Riemann Hypothesis".)