Computer Science 345
The Efficient Universe

Homework 5
Due Wednesday, March 29, 2006

You may collaborate with other students,
but you should write up the solutions entirely on your own.

1 $\mathcal{NP}$-Completeness

Definition 1 *A language $L \subset \{0,1\}^*$ is $\mathcal{NP}$-complete if*

- $L \in \mathcal{NP}$;
- *Every language L' in $\mathcal{NP}$ is Karp-reducible to L:*

$$L' \leq_K L.$$

Denote by $\mathcal{NPC} \subseteq \mathcal{NP}$ the set of all $\mathcal{NP}$-complete languages.

Problem 1 Let L be an $\mathcal{NP}$-complete language. Prove that

1. If $L \in P$, then $\mathcal{P} = \mathcal{NP} = \mathcal{NPC}$;
2. If $L \notin P$, then $\mathcal{P} \cap \mathcal{NPC} = \emptyset$.

2 Circuits

Problem 2 For every $f : \{0,1\}^n \to \{0,1\}$, let $s(f)$ be the smallest (in terms of the number of gates) Boolean circuit computing f. Show that for most function f, $s(f) > 2^{n/3}$.

Hint: Prove that the number of circuits of size s is at most 2^{s^2}. Count the number of functions and compare.

3 Probability

In this section we remind some basic theorems of probability theory.

Theorem 2 (Linearity of Expectation) *For every random variables (not necessarily independent) $\xi_1, \xi_2, \ldots, \xi_n$:*

$$\mathbb{E}\left[\xi_1 + \xi_2 + \ldots + \xi_n\right] = \mathbb{E}\left[\xi_1\right] + \mathbb{E}\left[\xi_2\right] + \ldots + \mathbb{E}\left[\xi_n\right].$$

Theorem 3 *For every independent random variables $\xi_1, \xi_2, \ldots, \xi_n$:*

1.
$$\mathbb{E}\left[\prod_{i=1}^{n} \xi_i\right] = \prod_{i=1}^{n} \mathbb{E}\left[\xi_i\right].$$

2.
$$\mathrm{Var}\left[\xi_1 + \xi_2 + \ldots + \xi_n\right] = \mathrm{Var}\left[\xi_1\right] + \mathrm{Var}\left[\xi_2\right] + \ldots + \mathrm{Var}\left[\xi_n\right].$$

Theorem 4 (Markov's Inequality) *Let ξ be a random variable taking nonnegative real values. Then for every positive t*

$$\Pr\left(\xi \geq t\right) \leq \frac{\mathbb{E}\left[\xi\right]}{t}.$$

Theorem 5 (Chebyshev's Inequality) *For every random variable ξ and every positive t*

$$\Pr\left(\left|\xi - \mathbb{E}\left[\xi\right]\right| \geq t\right) \leq \frac{\mathrm{Var}\left[\xi\right]}{t^2}.$$

Theorem 6 (Bernstein's Inequality or the Chernoff Bound) *Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent identically distributed Bernoulli random variables with mean p. In other words each ξ_i takes 1 with probability p and 0 with probability $q = 1 - p$. Then for any positive ε*

$$\Pr\left(\left|\frac{\xi_1 + \xi_2 + \cdots + \xi_n}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-\varepsilon^2 n/4}$$

Remark: *This inequality is usually known as Bernstein's Inequality in Mathematics and as the Chernoff Bound in Theoretical Computer Science.*

3.1 Problems

Problem 3 In this problem we shall see how the XOR of many flips of a biased coin reduces the bias quickly. Define the bias of a Bernoulli random variable ξ as follows:

$$\mathrm{Bias}(\xi) = \left|\Pr\left(\xi = 1\right) - \Pr\left(\xi = 0\right)\right|.$$

Now consider n independent identically distributed Bernoulli random variables $\xi_1, \xi_2, \ldots, \xi_n$ with mean p. Let ξ be the parity of the sum $\xi_1 + \xi_2 + \cdots + \xi_n$. In other words

$$\xi = \begin{cases} 0, & \text{the number of 1's among } \xi_1, \xi_2, \ldots, \xi_n \text{ is even;} \\ 1, & \text{the number of 1's among } \xi_1, \xi_2, \ldots, \xi_n \text{ is odd.} \end{cases}$$

Show that

$$\mathrm{Bias}(\xi) = (\mathrm{Bias}(\xi_1))^n.$$

What is the limit of Bias(ξ) as n tends to infinity?

Bonus: Compute the bias of a coin defined to be the *majority* vote on the outcome of 3 independent tosses. This improvement of bias can also be iterated. Compare using this method to using the parity when the total number n of independent coins is large.

In the next exercise we will see that randomness may drastically improve communication complexity.

Problem 4 Two friends Alice and Bob are connected via a slow Internet connection. Alice has a long binary string (or file) $A = a_1 \dots a_n$ of length n; and Bob has a string $B = b_1 \dots b_n$. They suspect that $A = B$ and want to check whether this is indeed the case.

If the Internet connection was fast, Alice could send here string A to Bob. Then Bob would compare A and B and tell the result to Alice. Unfortunately this protocol requires transmitting n bits (and we assume that n is large). Can we hope to reduce the communication complexity?

1. Show that any *deterministic, errorless* protocol that checks whether $A = B$ requires transmitting at least n bits.

Hint: Show that if less than n bits were exchanged, there must be a pair of inputs A, B on which Alice and Bob will make a mistake.

It turns out that if Alice and Bob can toss coins and are allowed to make errors (with small probability), then they can do much better.

First Alice picks a (uniform) random prime number between 1 and n^2. Then instead of transmitting the whole string A to Bob, she sends a fingerprint (or a hash value) of her string $m_A = M(A, p)$ (M defined below), which is much shorter than the string itself! She then transmits m_A, and the selected prime number p to Bob. Bob computes a fingerprint of his string $m_B = M(B, p)$ and compares m_A with m_B. If $m_A = m_B$, then he decides that $A = B$, otherwise $A \neq B$. The function $M(S, p)$ is simply the reduction of S modulo p, when S is considered an integer. Formally, M is defined as by:

$$M(S, p) = \sum_{i=1}^{n} 2^{i-1} \cdot s_i \mod p.$$

See page 5, for the step-by-step description of the algorithm.

3. Prove that if $A = B$, then Alice and Bob always decide that $A = B$.
4. If $A \neq B$, then the probability that Alice and Bob decide that $A = B$ is at most $O(1/n)$.

Hint: You may find the following theorem useful. First you may show that this probability is $O(\log n/n)$.

Theorem 7 (Prime number theorem [Chebyshev, Hadamard and Vallée Poussin]) *The number of prime numbers between 1 and n is $\Theta(n/\log n)$.*

5. How many bits do Alice and Bob send to each other, as a function of n?
6. Show that Alice and Bob can reduce the error probability by repeating the algorithm many times. What is the tradeoff between communication and error?

7. Assume that the size of their files is 100MB (roughly 800 million bits) and the connection speed is 50,000 bits per second. Estimate the running time of the first (deterministic) and second (probabilistic) protocol.

8. Recall that in class we saw a different method of "fingerprinting" a string S - simply computing its inner product (modulo 2) with a random Boolean vector v. Namely, $H(S, v) = \sum_i s_i v_i \mod 2$. The fingerprint in this case is just 1 bit, and we proved in class (please verify) that if $A \neq B$, then $\Pr[H(A, v) = H(B, v)] \leq 1/2$. What is the disadvantage of using this H here, as opposed to M?

Problem 5 A function $f : \{0,1\}^n \to \{0,1\}$ is linear if for every two vectors x and y, $f(x + y) = f(x) + f(y)$ (where $+$ is componentwise exclusive or). You are given a program P which computes some unknown linear function f, but makes an error on an unknown set of 1% of the inputs. Design a probabilistic program Q, which can use P as a subroutine, that for EVERY input z will satisfy $\Pr(Q(z) = f(z)) \geq 98\%$.

Algorithm 1 Algorithm M for computing a string fingerprint.

Input: a string $S = s_1 \dots s_n$, a prime number p.
Output: a number m_S.

- Compute

$$m_S = \sum_{i=1}^{n} 2^{i-1} \cdot s_i \mod p.$$

- Return m_S.

Algorithm 2 Protocol between Alice and Bob.

Alice's Input: a binary string A.
Bob's Input: a binary string B.
Output: Both programs return either "$A = B$" or "$A \neq B$".

Protocol:

- Alice picks a random (uniform) prime number p between 1 and n^2.
- Alice computes the fingerprint $m_A = M(A, p)$.
- Alice sends m_A and p to Bob.
- Bob computes $m_B = M(B, p)$.
- If $m_A = m_B$, then Bob returns "$A = B$"; otherwise "$A \neq B$".
- Alice returns the same message as Bob.
