CS 127 Spring 2020 - Homework Zero

Due on **Thursday January 30 2020 at midnight** but I recommend you do this homework even before the first lecture.

Some policies: (See the course syllabus for the full policies.)

- You can collaborate with other students that are currently enrolled in this course in brainstorming and thinking through approaches to solutions but you should write the solutions on your own and cannot share them with other students.
- Sharing questions or solutions with anyone outside this course, including posting on outside websites, is a violation of the honor code policy. Collaborating with anyone except students currently taking this course or using material from past years from this or other courses is a violation of the honor code policy.
- The submitted PDF should be typed and in the same format and pagination as ours. Please include the text of the problems and write **Solution X:** before your solution. (This is easy to do if you use our markdown template.) Please mark in gradescope the pages where the solution to each question appears. Points will be deducted if you submit in a different format or do not mark the pages.

By writing my name here I affirm that I am aware of all policies and abided by them while working on this problem set:

Your name: (Write name and HUID here)

Collaborators: (List here names of anyone you discussed problems or ideas for solutions with)

Number of late days used so far: (not including this pset; it is your responsibility to make sure you do not go over the late days budget)

Number of late days used for this pset:

Probability questions

As we'll see in the first lecture, much of cryptography relies on probability theory, and so basic knowledge of probability will be essential. The CS 121 probability review is one source for the probability theory we will need. During the course I will assume you are familiar with (or can pick up on your own) all notions presented there. However, if you find these questions unfamiliar or difficult you should not despair! There are plenty of sources on probability on the web, and in particular Harvard STAT 110 and its textbook are of course wonderful resources. If any of the notation is unfamiliar, looking at the lecture notes might help, and otherwise feel free to ask questions on Ed, even before the semester starts!

Notation: While often in probability theory people use the name "random variable" for a distribution over the set $\mathbb{R}$ of real numbers, it will be convenient for us to generalize this to arbitrary sets, and hence we will use the following notation. We define a *random variable* or *distribution* X over a finite set S to correspond to the probabilistic experiment where we draw an element x from S with some probability, which we denote by $\Pr[X = x]$. All we need from these probabilities is that they are non-negative and sum up to one. (One can also consider distributions over infinite sets, though almost always in this course we will restrict ourselves to the finite case.) We use $x \leftarrow_R X$ as shorthand for saying that x is drawn accoring to the distribution X. If $f : S \to T$ is a function, then the random variable $f(X)$ corresponds to the probabilistic experiment where we draw $x \leftarrow_R X$ and output $f(x)$.

Question 1 (30 points): In this question we will study the notion known as Total Variation or statistical distance. It is a basic notion of distance between probability distribution, and its computational analog is fundamental for cryptography.

Question 1.1 (15 points): If X and Y are two distributions over the same set S, we define the *statistical distance* of X and Y, denoted as $\Delta(X, Y)$ (also known as *total variation* distance of X and Y) to be $\sum_{x \in S} |\Pr[X = x] - \Pr[Y = x]|$. Prove that for every function $f : S \to [0, 1]$, $|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq \Delta(X, Y)$.

Solution 1.1:

Question 1.2 (15 points): Prove that the statistical distance satisfies the *triangle inequality*: For every three distributions X, Y, Z over the same set S, $\Delta(X, Z) \leq \Delta(X, Y) + \Delta(Y, Z)$.

Solution 1.2:

Question 2 (40 points + 10 points bonus): We will now use the notion of statistical distance to study one of the most basic questions in probability theory: if we are given a coin that is either completely unbiased, or has bias $\epsilon > 0$ towards "heads", how many tosses will it take for us to distinguish between the two cases.

Question 2.1 (15 points): Prove that we can distinguish between an unbiased coin and one that has ϵ bias towards "heads" using at most $O(1/\epsilon^2)$ coin tosses. Specifically prove that if $k > 100/\epsilon^2$, then there exists some function $f : \{0,1\}^k \to \{0,1\}$ such that if X is the uniform distribution over $\{0,1\}^k$ and Y is the distribution obtained by tossing k independent coins, each equaling 1 with probability $1/2 + \epsilon$ and equaling 0 with probability $1/2 - \epsilon$, then $\Pr[f(X) = 0] > 0.9$ and $\Pr[f(Y) = 1] > 0.9$, and hence f can distinguish between X and Y.[1]

Solution 2.1:

Question 2.2 (15 points): We now study the converse problem: showing a *lower bound* on the number of coin tosses needed. (This problem involves some notation, so take your time reading it carefully and parsing what it means.) We will derive this bound using the combination of Question 2.2 and Question 2.3.

For every $k \in \mathbb{N}$, $0 \leq i \leq k$, and $\epsilon > 0$, let $X_i^{k,\epsilon}$ be the following distribution over $\{0,1\}^k$: the first i bits of $X_i^{k,\epsilon}$ are chosen independently and uniformly at random, and the last $k - i$ bits are chosen independently at random but each is equal to 1 with probability $1/2 + \epsilon$ and equal to 0 with probability $1/2 - \epsilon$. Prove that for very k, $i < k$ and ϵ, $\Delta(X_i^{k,\epsilon}, X_{i+1}^{k,\epsilon}) \leq 10\epsilon$.

Solution 2.2:

Question 2.3 (10 points): Prove that, in the notation of Question 2.1, $\Delta(X,Y) \leq 10k\epsilon$. Show also that this implies that if $k < 1/(100\epsilon)$, there does not exists a function $f : \{0,1\}^k \to \{0,1\}$ such that $\Pr[f(X) = 0] > 0.9$ and $\Pr[f(Y) = 1] > 0.9$. Use your solution for Question 2.1, the triangle inequality proved in Question 1.2, and the result of Question 1.1 (this is a special case of the Hybrid Argument which is used time and again in cryptography).

Solution 2.3:

Question 2.4 (bonus problem: optional and more challenging - 10 points bonus): Prove that in fact $\Delta(X,Y) \leq O(\sqrt{k}\epsilon)$. See footnote for hint.[2] Conclude that the method of Question 2 is essentially *optimal* in the sense that there exist some absolute constant δ (independent of ϵ) such that for every ϵ and distributions X, Y as in Question 2, if $k < \delta/\epsilon^2$ then there does not exist a function $f : \{0,1\}^k \to \{0,1\}$ such that $\Pr[f(X) = 0] > 0.9$ and $\Pr[f(Y) = 1] > 0.9$.

Solution 2.4:

[1] **Hint:** Use the Chernoff bound.

[2] **Hint:** One way to prove this is to use the notion of KL divergence which is another notion of distance between the distributions that satisfies the triangle inequality. You can show that the KL divergence of $X_i^{k,\epsilon}$ and $X_{i+1}^{k,\epsilon}$ is $O(\epsilon^2)$ and then use the Pinsker Inequality that shows that the statistical distance between two distributions is at most the square root of their KL divergence. You can read about both KL divergence and the Pinsker Inequality in Wikipedia as well in several other sources.

And now for a little crypto

Question 3 (very important! 10 points): Read the lecture notes for lecture 1: introduction.

Solution 3: Write here "I affirm that I read all of the lecture notes". Unlike CS 121, my lectures in CS 127/227 will be under the assumption that all students had read seriously the lecture notes before each lecture. The notes are unpolished (to say the least) and so you may well run into typos/bugs and unclear points while reading them. When you do, feel free to ask questions on Ed and or to post issues or pull requests on the GitHub repository.

Question 4.1 (20 points): Prove that an encryption scheme (E, D) with messages of length ℓ and keys of length n is *perfectly secret* if and only if for every $x, x' \in \{0,1\}^\ell$, $\Delta(Y^x, Y^{x'}) = 0$, where Y^x is the distribution obtained by choosing a random $k \leftarrow_R \{0,1\}^n$ and outputting $E_k(x)$.

Solution 4.1:

Question 4.2 (20 points - bonus but please do attempt it): Let $\epsilon > 0$. Define an encryption scheme (E, D) with messages of length ℓ and keys of length n to be ϵ-secret if for every $x, x' \in \{0,1\}^\ell$, $\Delta(Y^x, Y^{x'}) < \epsilon$. Suppose that (E, D) is ϵ secret. Prove that for every adversary Eve, which we model as function $Eve : \{0,1\}^m \to \{0,1\}$ where m is the length of the ciphertexts, and a pair of messages $x_0, x_1 \in \{0,1\}^\ell$, the probability that Eve wins in the game described below is less than $1/2 + 10\epsilon$.

The game is defined as follows:

1. We pick k at random in $\{0,1\}^n$.
2. We pick b at random in $\{0,1\}$.
3. We let $y = E_k(x_b)$.
4. We let $b' = Eve(y)$.
5. Eve *wins* iff $b' = b$.

Solution 4.2: