CS7880: Rigorous Approaches to Data Privacy, Spring 2017
POTW #5

Instructor: Jonathan Ullman

Due Fri, Mar 3rd, 11:59pm
(Email to jullman+PrivacyS17@gmail.com)

- **You may work on this homework in pairs if you like. If you do, you must write your own solution and state who you worked with.**
- Solutions must be typed in LaTeX.
- Aim for clarity and brevity over low-level details.

Problem 1 (Determining the Scale via Stability)**.**

In Section 3.3 of Vadhan's survey, there is an algorithm that releases a histogram over a possibly infinite domain with error $O\left(\frac{\log(1/\delta)}{\varepsilon n}\right)$. In this problem we will see how to use this algorithm to find the *scale* of data from an unknown distribution.

Suppose we have data drawn from some distribution D over $\mathbb{R}$. The distribution is uniform on some unknown interval $[\mu-\sigma, \mu+\sigma]$. We will assume that the dataset x consists of $2n$ iid samples from D, and we will design an (ε, δ)-differentially private algorithm to approximate the width of the interval, 2σ.

Hint: I recommend reading through the entire problem before you start. Depending on how you do part (c), you may find it preferable to prove a slightly different statement in part (b). Any pair of statements that leads to the right conclusion is fine.

(a) Suppose X_1, X_2 are independent samples from D. What is the distribution of the random variable $|X_1 - X_2|$? Write its probability density function and its mean.

(b) Suppose we pair up our dataset $x \in \mathbb{R}^{2n}$ into a new dataset $y \in \mathbb{R}^n$ consisting of the n numbers $x_1 - x_2, x_3 - x_4, \ldots, x_{2n-1} - x_{2n}$. Using a Chernoff bound, prove that for sufficiently large n, with probability at least 15/16, at least $2n/3$ out of the n numbers are contained in some interval of width $c\sigma$, for some $c < 2$.

(c) Define the infinite set of "buckets" $B_i = [2^i, 2^{i+1})$ for $i \in \mathbb{N}$. For a given dataset y, the histogram of y specifying how many of y's elements fall into each bucket B_i can be computed using the algorithm referenced above. Show how to use this algorithm to design an algorithm that outputs an estimate $\hat{\sigma}$ with the guarantee that when $n = O\left(\frac{\log(1/\delta)}{\varepsilon n}\right)$, with high probability[1], $\hat{\sigma} \in [\sigma/2, 2\sigma]$.

[1] You can deduce a more precise failure probability from the proof in Vadhan's survey, but for this problem you can just assume that the algorithm succeeds "with high probability."