# CS7880: Rigorous Approaches to Data Privacy, Spring 2017
# POTW #2

Instructor: Jonathan Ullman

**Due Fri, Jan 27th, 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** (Noisy Histograms)**.**

In this problem you will see how to accurately answer *exponentially* many statistical queries on a dataset $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ when $|\mathcal{X}|$ is reasonably small. The *histogram* representation of a dataset $x$ is a $|\mathcal{X}|$-dimensional vector where the $j$-th entry is the fraction of $x$'s rows that are equal to $j$.

$$h(x) := \left(h_1(x), \ldots, h_{|\mathcal{X}|}(x)\right) \qquad h_j(x) := \tfrac{1}{n}\left|\{i \in [n] \mid x_i = j\}\right|.$$

Consider the following *noisy histogram algorithm*: output

$$\hat{h}(x) := \left(h_1(x) + Z_1, \ldots, h_{|\mathcal{X}|}(x) + Z_{|\mathcal{X}|}\right)$$

where every $Z_j \sim N(0, \sigma^2)$ is an independent Gaussian.

**(a)** For what value of $\sigma$ does this algorithm ensure $(\varepsilon, \delta)$-differential privacy? Justify your answer using results we've seen (you don't need to rederive any results).

**(b)** Consider a statistical query $q(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$ for some $\phi : \mathcal{X}^n \to [0,1]$. Suppose you are given a (possibly noisy) histogram $h$. How would you estimate $q(x)$ using $h$? That is, design a function $est(h, q)$ such that for every statistical query $q$ and every dataset $x$, $est(h(x), q) = q(x)$.

**(c)** Let $\mathcal{Q} = \{q_1, q_2, \ldots\}$ be a set of statistical queries. Given a noisy histogram $\hat{h}(x)$, how accurately can you estimate the answers to every $q \in \mathcal{Q}$? Show that for some $\alpha$ as small as possible,

$$\forall x, \mathcal{Q} \qquad \mathbb{P}\left[\max_{q \in \mathcal{Q}} \left| Est(\hat{h}(x), q) - q(x) \right| \le \alpha \right] \ge .99,$$

where $\alpha$ is a function of $n, |\mathcal{X}|, |\mathcal{Q}|, \varepsilon, \delta$, and the probability is taken over the random Gaussian noise added to ensure privacy.[1]

**(d)** For what values of $|\mathcal{X}|$ does this algorithm provide a non-trivial accuracy guarantee? For what parameters does this algorithm improve on the approach of adding independent Gaussian or Laplacian noise to each query?

[1] Hint: A very useful fact about Gaussians is that if $Z_1 \sim N(\mu_1, \sigma_1^2)$ and $Z_2 \sim N(\mu_2, \sigma_2^2)$ are independent Gaussians, then their sum $Z_1 + Z_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ is also a Gaussian, and the means and variances add up.