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

Instructor: Jonathan Ullman

Problem 1 (Noisy Histograms).

In this problem you will see how to accurately answer *exponentially* many statistical queries on a dataset $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ 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.

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

Consider the following *noisy histogram algorithm*: output

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

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

(a) For what value of σ does this algorithm ensure (ε, δ)-differential privacy? Justify your answer using results we've seen (you don't need to rederive any results).

To achieve differential privacy using Gaussian noise, it suffices to add noise $N(0, \sigma^2)$ to each coordinate where

$$\sigma = O\left(\frac{GS_2(h) \cdot \sqrt{\log(1/\delta)}}{\varepsilon}\right)$$

where $GS_2(h)$ is the global ℓ_2-sensitivity of the function h. If x, x' differ on one row, then h can go down by $1/n$ in one coordinate and up in another coordinate, so

$$\max_{x \sim x'} \|h(x) - h(x')\|_2 \le \frac{2}{n}.$$

Thus it suffices to set $\sigma = O\left(\frac{\log(1/\delta)}{\varepsilon n}\right)$.

(b) Consider a statistical query $q(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$ for some $\phi : \mathcal{X} \to [0, 1]$. 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, $est(h(x), q) = q(x)$.

By definition we have

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = \sum_{j \in X} \phi(j) \cdot \left(\frac{1}{n}\sum_{i : x_i = j} 1\right) = \sum_{j \in X} \phi(j) \cdot h(x)_j.$$

Now if we define the vector $\vec{q} = (\phi(1), \dots, \phi(X))$, we can write $q(x) = \langle \vec{q}, h(x) \rangle$. Thus we define $est(h(x), q) = \langle h(x), \vec{q} \rangle$.

(c) Let $\mathcal{Q} = \{q_1, q_2, \dots\}$ 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 α as small as possible,

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

where α 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]

Consider any query q. By our definition from part (b), we have

$$est(\hat{h}, q) = \langle \hat{h}, \vec{q} \rangle = \sum_{j \in X} \hat{h}(x)_j \cdot \phi(j) = \sum_{j \in X} (h(x)_j + Z_j) \cdot \phi(j) = q(x) + \sum_{j \in X} Z_j \cdot \phi(j).$$

Since every Z_j is an independent sample from $N(0, \sigma^2)$, the distribution of $\sum_{j \in X} Z_j \cdot \phi(j)$ is precisely $N(0, \psi^2)$ for $\psi^2 = \sum_{j \in X} \sigma^2 \phi(j)^2 \leq \sigma^2 |X| = O\left(\frac{|X| \log(1/\delta)}{\varepsilon^2 n^2}\right)$.

The Gaussian distribution has the property that if $Y \sim N(0, \psi^2)$, then for some $c > 0$ (I think $c = 1$)

$$\mathbb{P}\left[|Y| \leq c\psi\sqrt{\ln(1/\beta)}\right] \leq 2\beta.$$

Thus, for any single query $q \in Q$ we have

$$\mathbb{P}\left[|est(\hat{h}, q) - q(x)| > O\left(\frac{\sqrt{|X| \ln(1/\delta) \ln(1/\beta)}}{\varepsilon n}\right)\right] \leq \mathbb{P}\left[|est(\hat{h}, q) - q(x)| > c\psi\sqrt{\ln(1/\beta)}\right] \leq 2\beta$$

By taking a union bound over all queries $q \in Q$, we have

$$\mathbb{P}\left[\exists q \in Q \quad |est(\hat{h}, q) - q(x)| > O\left(\frac{\sqrt{|X| \ln(1/\delta) \ln(1/\beta)}}{\varepsilon n}\right)\right] \leq 2\beta |Q|.$$

Now setting $\beta = \frac{1}{200|Q|}$ gives

$$\mathbb{P}\left[\exists q \in Q \quad |est(\hat{h}, q) - q(x)| > O\left(\frac{\sqrt{|X| \ln(1/\delta) \ln|Q|}}{\varepsilon n}\right)\right] \leq \frac{1}{100},$$

as desired.

(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?

Since the answer to a statistical query is in $[0, 1]$, to obtain non-trivial accuracy we need the error to be $\ll 1$. Comparing to the error bound from part (c), we see that non-trivial error is possible only when $|X| \ll \frac{\varepsilon^2 n^2}{\log(1/\delta)}$. For reasonable choices of the privacy parameters $\varepsilon = 1/10, \delta = 1/n^2$, we get non-trivial error when $|X| \ll \frac{n^2}{\log(n)}$.

[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.

If we compare to Gaussian noise, which requires error $\tilde{O}\left(\frac{\sqrt{|Q|\ln(1/\delta)}}{\varepsilon n}\right)$, we see that the noisy histogram does better when $|Q| \gg |X|$, and does much worse otherwise. So, specifically, this algorithm is an improvement of Gaussian noise roughly when the universe is small and the number of queries is large, i.e. $|X| \ll \min\{|Q|, n^2\}$.