CSC2419:Algorithms and Complexity in Private Data Analysis Winter 2017

Lecture 8 — March 14, 2017

Aleksandar Nikolov *Scribe: Sepehr Abbasi Zadeh*

Projection Mechanism Analysis

Theorem 1. *The projection mechanism has an average error of at most α when $n \geq \frac{2\sigma_{\epsilon,\delta}\ell^*(K_Q)}{\sqrt{m}\alpha^2}$.*

Proof. Assuming that each coordinate of $\tilde{y}$ and y corresponds to a reported query answer and a true query answer, respectively, and $\hat{y}$ is the projection of the $\tilde{y}$, we want to show $\mathbb{E}(\frac{1}{m}\|\hat{y}-y\|_2^2) \leq \alpha^2$ when $n \geq \frac{2\sigma_{\epsilon,\delta}\ell^*(K_Q)}{\sqrt{m}\alpha^2}$. Here, the euclidean norm squared is the sum of squared errors and we are averaging it over m queries. Forgetting about the expectation, we can rewrite the distance inside the expectation as (see Figure 1):

$$\|\hat{y}-y\|_2^2 = \langle \hat{y}-y, \tilde{y}-y\rangle + \langle \hat{y}-y, \hat{y}-\tilde{y}\rangle \tag{1}$$

we claim that:

$$\langle \hat{y}-y, \hat{y}-\tilde{y}\rangle \leq 0 \leq \langle \hat{y}-y, \tilde{y}-y\rangle \tag{2}$$

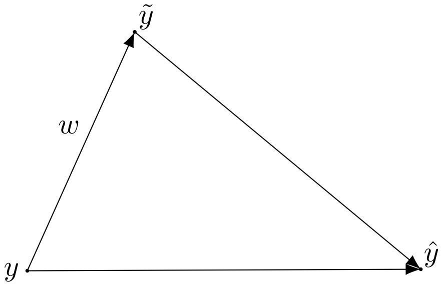

Figure 1: Vector view of query answers.

Now let us prove this claim. Considering the convex hull K_Q, we define line ℓ such that $\ell \perp (\hat{y}-\tilde{y})$ and $\hat{y} \in \ell$ (see Figure 2). It can be shown that ℓ separates K_Q and $\tilde{y}$ (ℓ does not intersect K_Q), because if it was not the case, then we could find a closer $\hat{y}$ to $\tilde{y}$ than the current one. Since $\hat{y}\tilde{y} \perp \ell$ and $y\hat{y}$ is inside K_Q we can infer that β should be an obtuse angle. Because if it was not, then the projection mechanism could find a closer point than the current $\hat{y}$. Thus:

$$\beta \geq \pi/2 \Rightarrow \langle \hat{y}-y, \hat{y}-\tilde{y}\rangle = \|\hat{y}-y\| \,.\, \|\hat{y}-\tilde{y}\| \,.\cos\beta \leq 0 \tag{3}$$

Since the distance in equation 1 is a positive value, we can infer claim 2 using equation 3. In addition, we can get:

$$\|\hat{y}-y\|_2^2 \leq \langle \hat{y}-y, \tilde{y}-y\rangle \tag{4}$$

Please note that the value $\tilde{y}-y$ can be interpreted as the value of the noise that was added to the true answer of the query (w in Figure 1). So we can replace it with a Gaussian noise of

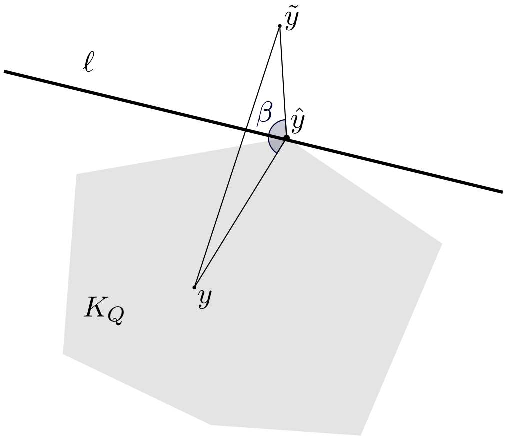

Figure 2: $\hat{y}$ is the projection of $\tilde{y}$ to the convex hull K_Q.

$w = (\sigma_{\epsilon,\delta}.\sqrt{m}/n)g$ where g is a normal Gaussian noise (i.e., $g \sim \mathcal{N}(0, I)$ and I is the identity matrix). Now we replace this noise in our last equation and take its expectation after dividing it by m:

$$\begin{aligned}
\mathbb{E}(\frac{1}{m}\left\|\hat{y}-y\right\|_2^2) = \mathbb{E}(\frac{1}{m}\langle \hat{y}-y, \tilde{y}-y\rangle) &= \mathbb{E}(\frac{1}{m}\langle \hat{y}-y, w\rangle) \\
&= 1/m\big(\,\mathbb{E}(\langle \hat{y}, w\rangle) + \mathbb{E}(\langle y, -w\rangle)\big) \\
&\leq 1/m\big(\,\mathbb{E}(h_{K_Q}(w)) + \mathbb{E}(h_{K_Q}(-w))\big) \\
&= \frac{1}{m}\frac{\sigma_{\epsilon,\delta}\sqrt{m}}{n}\big(\,\mathbb{E}(h_{K_Q}(g)) + E(h_{K_Q}(-g))\big) \\
&= \frac{2\sigma_{\epsilon,\delta}}{\sqrt{m}n}\ell^*(K_Q)
\end{aligned}$$

We used the facts that $h_{K_Q}(tw) = th_{K_Q}(w)$ for any nonnegative real t, and also that a centered Gaussian random variable g has the same distribution as $-g$. Finally, plugging in $n \geq \frac{2\sigma_{\epsilon,\delta}\ell^*(K_Q)}{\sqrt{m}\alpha^2}$ into the previous equation would give $\mathbb{E}(\frac{1}{m}\left\|\hat{y}-y\right\|_2^2) \leq \alpha^2$. □

Marginal Queries

Consider a database $\mathcal{X}^n$ where $\mathcal{X} \in \{0,1\}^d$ (i.e., a database of n rows of records with d binary attributes).

Definition 2. *A k-way marginal query is query over a conjunction of a subset of size k of attributes or their negations. It evaluates to one for a row, if this row satisfies this query and to zero otherwise. The answer to a marginal query on $\mathcal{X}^n$ is the fraction of rows which evaluate to one.*

For example, $c_1 \wedge \overline{c_2} \wedge c_3$ a 3-way conjunction query which evaluates to one for querying a row x if and only if $x_1 = 1, x_2 = 0$, and $x_3 = 1$ for this specific row. For example, c_1, c_2, and c_3 can respectively encode the gender, whether this person is smoking or not, and does he/she have lung cancer in each row. This specific query asks for the fraction of male people in this database who do not smoke and have lung cancer.

It can be seen that we have $\binom{d}{k}2^k$ different k-way marginal queries.

Like before, we can compute lower bounds for n when we want to have less than α error on running distinct mechanisms such as Private Multiplicative Weights (PMW) and Gaussian Noise with marginal queries:

$$n \geq \begin{cases} \frac{Cd^{k/2}\sqrt{\log 1/\delta}}{\alpha\epsilon} & Gaussian \\ \frac{C\sqrt{d}\log d\sqrt{\log 1/\delta}}{\alpha^2\epsilon} & PMW \end{cases}$$

In the above equations, C denotes a large enough *constant* which is independent from ϵ, δ and α.

We can show that the Gaussian noise mechanism's running time is $\text{poly}(n, d^k)$ and PMW's running time is $\text{poly}(n, 2^d)$.

When we are using the projection mechanism, we want to solve the following minimization problem:

$$\underset{z \in K_Q \subseteq \mathbb{R}^m}{\text{argmin}} \{\|\tilde{y} - z\|_2^2\}$$

Basically, it means that we are minimizing a convex function subject to a convex constraint. To do so, we need a **separation oracle**. A separation oracle for a convex set $K_Q \subseteq \mathbb{R}^m$ is an algorithm which takes a point $z \in \mathbb{R}^m$ as the input and returns *"inside"* if it is inside K_Q or otherwise it returns a hyperplane that separates K_Q and the queried point.

By having a polynomial separation oracle for the K_Q we can solve the minimization problem of the projection mechanism. However, the structure of the K_Q polytope is so complicated when we have k-way marginal queries for $k \geq 3$. Next time we will talk more specifically about 2-way marginals and their separation oracles.