CSC2419:Algorithms and Complexity in Private Data Analysis Winter 2017

Lecture 7 — March, 7^{th}, 2017

Aleksandar Nikolov *Scribe: Tyrone Strangway*

1 Recap

As before we are interested in the query release problem.

Definition 1. *The query release problem is defined by a set of m linear queries, $\mathcal{Q} = q_1, \ldots, q_m$ and a database $x \in \mathcal{X}^n$. For a single row r, $q_i(r) \in \{0,1\}$ and $q_i(x) = \frac{1}{n}\sum_{j\in[n]} q_i(x_j)$, that is the fraction of rows in x with property q_i. Using a mechanism $\mathcal{M}$ to approximate the queries, we want to accurately estimate $\mathcal{Q}$ on x, that is for all i $|\mathcal{M}(x)_i - q_i(x)| \leq \alpha$. We also want to ensure the private information of the participants is protected, that is $\mathcal{M}$ should be (ϵ, δ)-Differentially Private. Our main goal is to find what the smallest size of data base, in terms of rows n, for which these tasks are possible.*

We have seen several mechanisms to solve this problem:

1. **Gaussian Mechanism:** For each $i \in [m]$ output $q_i(x) + w_i$, where $w_i \sim \mathcal{N}(0, \sigma^2_{\epsilon,\delta}\frac{m}{n^2})$. That is we add independent Gaussian noise to each query answer. We showed that for some $\sigma_{\epsilon,\delta} = O\Big(\frac{\sqrt{\log\frac{1}{\delta}}}{\epsilon}\Big)$, the Gaussian Mechanism is (ϵ, δ)-Differentially Private. It is easy to show that to get accuracy α for every query, we need to set $n \in \Omega(\frac{\sqrt{m\log m}\sqrt{\log\frac{1}{\delta}}}{\alpha\epsilon})$. The running time of the Gaussian Mechanism is running time is $O(mn)$.

2. **Laplace Mechanism:** For each $i \in [m]$ output $q_i(x) + w_i$, where $w_i \sim Lap(0, b)$. That is we add independent Laplacian noise to each query answer. Again with the correct parameters the mechanism is ϵ-Differentially Private. We also know it can obtain accuracy α when $n \in \Omega(\frac{m\log m}{\alpha\epsilon})$. Its running time is $O(mn)$.

3. **smallDB:** Using the Exponential Mechanism privately generate a synthetic database that approximates x with respect to $\mathcal{Q}$, use this database in place of x. This mechanism is ϵ-Differentially Private and has an error of α when $n \geq \frac{\log|\mathcal{X}|\cdot\log m}{\alpha^3\epsilon}$. Its running time is polynomial in $|\mathcal{X}|$, the number of potential database rows.

4. **Multiplicative Weights:** Starting with a database generated independently of x, iteratively refine it until it is similar to x with respect to $\mathcal{Q}$. The refinement is based on the Multiplicative Weights algorithm from machine learning, hence the name. This mechanism can be (ϵ, δ)-Differentially Private with an accuracy of α when $n \in \Omega(\frac{\sqrt{\log|\mathcal{X}|}\log m\sqrt{\log\frac{1}{\delta}}}{\alpha^2\epsilon})$. This mechanism requires running time $O(mnL|\mathcal{X}|)$, where $L = O(\frac{\log|\mathcal{X}|}{\alpha^2})$ is the number of refinements we execute.

The first two mechanisms are conceptually simple, but do not get the same accuracy performance as the more complex ones. Both smallDB and Multiplicative Weights, are able to reduce the m term to $\log m$ in the lower bound on n. But they also introduce a $\log |\mathcal{X}|$ term, which the Laplace Mechanism does not have. That being said we will later see a natural regime of parameters for which this is not an issue. On the flip side the first two are far faster, and do not run in time relative to $|\mathcal{X}|$ which can reasonably be exponential in the other terms.

2 The Projection Mechanism

Here we study the Projection Mechanism. For notation we define $\mathcal{Q}(x)$ to be the m dimensional column vector where entry i is $q_i(x)$.

Conceptually it is very simple, it is a slight modification of the Gaussian Mechanism, and has good accuracy. But there is a caveat, the error on the estimation will no longer be in the worst case but instead on average. This turns out not to be a huge burden as later on we will see we can take average errors and transform them into worst case errors by borrowing the technique of boosting from Machine Learning.

The Projection Mechanism is also interesting because it and its analysis make use of the geometry of the query space. Intuitively $\{\mathcal{Q}(x) : x \in \mathcal{X}^m\} \subseteq \mathbb{R}^m$ is just a set of points in m dimensional space, so the answer to some query on some database is just a point in this space.

2.1 Specifying the Mechanism

Before we introduce the Projection Mechanism we need the following items:

Definition 2. *The convex hull of a set S is the smallest convex set which contains S. That is* $\text{conv}(S) = min\{R : S \subseteq R \text{ and } R \text{ is convex}\}$. *A convex set is a set where the line segment connecting any two points in the set is itself entirely contained in the set. For a finite set of points S a convex hull can also be equivalently defined as* $\text{conv}(S) = \{\sum_{x \in S} \alpha_x x : (\forall x, \alpha_x \geq 0) \wedge \sum_{x \in S} \alpha_x = 1\}$. *The second definition will be useful in our proofs.*

Let $S_{\mathcal{Q}} = \{\mathcal{Q}(r) : r \in \mathcal{X}\}$, that is all possible query answers on single row databases. Because of the normalizing term in the definition of $\mathcal{Q}(x)$, we get $\forall x \in \mathcal{X}^n$ $\mathcal{Q}(x) \in \text{conv}(S_{\mathcal{Q}})$. Let $K_{\mathcal{Q}} = \text{conv}(S_{\mathcal{Q}})$.

We can now define the Projection Mechanism.

Algorithm 1 Projection Mechanism, with D.B. x and parameters ϵ, δ, α:

1: Let $\tilde{y} = \mathcal{Q}(x) + W$, where $W \sim \mathcal{N}(0, \sigma_{\epsilon,\delta}^2 \cdot \frac{m}{n^2} \cdot I)$ (I is the identity matrix and $\sigma_{\epsilon,\delta} \approx \frac{\sqrt{\log \frac{1}{\delta}}}{\epsilon}$)
2: Return $\hat{y} = \arg\min_z \{\|\tilde{y} - z\|_2 : z \in K_{\mathcal{Q}}\}$

The first line is the (ϵ, δ)-Differentially Private Gaussian Mechanism. Alone the mechanism may push the answer far away from $\mathcal{Q}(x)$, that is it may end up outside the area of feasible answers $K_{\mathcal{Q}}$. The second line simply projects the answer back onto the closest point in $K_{\mathcal{Q}}$. If the first line did not move the answer out of $K_{\mathcal{Q}}$ the second line does nothing. Using the fact that $K_{\mathcal{Q}}$ is convex and applying the triangle inequality we can get that the projection step can never make the accuracy worse, and in some cases will improve it.

2.2 Privacy of the Mechanism

Theorem 3. *The projection mechanism is (ϵ, δ)-Differentially Private.*

The proof of this is straight forward. As shown in a previous lecture, the Gaussian Noise Mechanism with the above parameters is (ϵ, δ)-Differentially Private. The actual projection step is simply post processing, it only relies on publicly available information, thus it is $(0, 0)$-Differentially Private. By composition the Projection Mechanism is (ϵ, δ)-Differentially Private. □

2.3 Accuracy of the Mechanism

We first formalize the notion of average error.

Definition 4. *A mechanism $\mathcal{M}$ has average error α if for all databases $x \in \mathcal{X}^m$:*

$\sqrt{\mathbb{E}\frac{\sum_{i=1}^{m}(\mathcal{M}(x)_i - q_i(x))^2}{m}} \leq \alpha$

This can be equivalently formulated as:

$\sqrt{\mathbb{E}\frac{1}{m}\|\mathcal{M}(x) - \mathcal{Q}(x)\|_2^2} \leq \alpha$

For both these terms the expectation is taken over the randomness of the mechanism.

That is, for some queries the error could be quite high, but on others it would be very low and on average it should be under α. In contrast, for worst case error error analysis we required each of the m queries be close to what the mechanism returned. Rather than measuring the error in expectation, we can ask for a high probability guarantee, and introduce another parameter β for the probability that the average error is larger than α. For the projection mechanism this would increase the lower bound on n by a $\log \frac{1}{\beta}$ term. We do not pursue this further, and instead stick with expectation to avoid dealing with yet another parameter.

We now state our first result regarding the accuracy of the Projection Mechanism.

Theorem 5. *The Projection Mechanism has average error at most α as long as $n \geq \frac{c\sqrt{\log|\mathcal{X}|}\cdot\sqrt{\log\frac{1}{\delta}}}{\alpha^2\cdot\epsilon}$, where c is a constant.*

To show this we will prove a more precise average error guarantee that relies on the geometry of $K_{\mathcal{Q}}$, and in particular on its size. The less precise result is useful for comparison with other mechanism's non geometric bounds. Interestingly unlike all of the previously seen bounds, this does not rely on the number of queries m. This is a byproduct of viewing the average error and not the worst case. Before we state the improved bound we need to define some measure of the "size" of $K_{\mathcal{Q}}$. To do so we first introduce the notion of a support function:

Definition 6. *The support function of a set $K \subseteq \mathbb{R}^m$ is $h_K(y) = \sup\{\langle x, y\rangle : x \in K\}$, where $y \in \mathbb{R}^m$.*

For some intuition, we mention that when y is of unit Euclidean norm, i.e. $\|y\|_2 = 1$, then $h_K(y) + h_K(-y)$ is its *width* in the direction of y. I.e. it is the smallest w so that we can sandwich K between two parallel hyperplanes, both orthogonal to y, and distance w apart.

The support function satisfies the following properties, $\forall x, y \in \mathbb{R}^m$:

1. $\forall t \geq 0$, $h_K(t \cdot y) = t \cdot h_K(y)$
2. $h_K(x + y) \leq h_K(x) + h_K(y)$
3. If K is bounded then $h_K(x) = 0$ only for the zero vector.

If K is bounded then by definition $h_K(y)$ is a norm. Since $K_{\mathcal{Q}}$ is compact we do get $h_{K_{\mathcal{Q}}}(y)$ is a norm (we will not end up using this fact). Also since $K_{\mathcal{Q}}$ is convex we can write $h_K(y)$ as $\max\{\langle x, y\rangle : x \in S_{\mathcal{Q}}\}$, where $y \in \mathbb{R}^m$. That is, when computing the support function on a convex set we only need to look at the extremal points (vertices) of the set, instead of all points in it.

Now we can introduce a way to measure the average width of $K_{\mathcal{Q}}$. We first define the mean width:

Definition 7. *The mean width of a convex set K is $M^*(K) = \mathbb{E}h_K(y)$, where y is chosen uniformly at random from the set of vectors where $\|y\|_2 = 1$. That is y is chosen according to the unique rotationa invariant probability measure on the unit sphere centred about the origin.*

We also define the Gaussian width as:

Definition 8. *The Gaussian width of a convex set K is $\ell^*(K) = \mathbb{E}h_K(g)$, where $g \sim \mathcal{N}(0, I)$ and I is the m dimensional identity matrix. That is g is chosen according to the standard m dimensional Gaussian centred about the origin.*

It turns out the Gaussian width and the mean width are closely related, actually we can show $\ell^*(K) = (\mathbb{E}\|g\|_2) \cdot M^*(K) = (c \cdot \sqrt{m}) \cdot M^*(K)$ where c is a constant.

We can now state our more precise accuracy measure which is in terms of the Gaussian width and thus also the mean width.

Theorem 9. *The Projection Mechanism, which is (ϵ, δ)-Differentially Private, has average error α if $n \geq \frac{c_1 \cdot \ell^*(K_{\mathcal{Q}}) \cdot \sqrt{\log \frac{1}{\delta}}}{\alpha^2 \cdot \epsilon \cdot \sqrt{m}} = \frac{c_2 \cdot M^*(K_{\mathcal{Q}}) \cdot \sqrt{\log \frac{1}{\delta}}}{\alpha^2 \cdot \epsilon}$, where c_1 and c_2 are constants.*

Now that we have the geometric bound, we can prove our original bound on the accuracy. To do so we will show that $\ell^*(K_{\mathcal{Q}}) \in O(\sqrt{m \log |\mathcal{X}|})$, that is the Gaussian width of $K_{\mathcal{Q}}$ is not too large. Plugging this into the geometric accuracy bound gets the original one.

To prove this we will use the following three facts:

1. For any single query $q_i : \mathcal{X} \to [0, 1]$, so for of all m queries on a single row, $S_{\mathcal{Q}} \subseteq [0, 1]^m$. Thus we get $\forall y \in S_{\mathcal{Q}}, \|y\|_2 \leq \sqrt{m}$.
2. The inner product between some vector and a Gaussian is itself Gaussian. So for the Gaussian g we used in the definition of Gaussian width, $g \sim \mathcal{N}(0, I)$, we get $\forall y, \langle y, g\rangle \sim \mathcal{N}(0, \|y\|_2^2)$
3. A standard Gaussian tail bound: if $y \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{P}(|y| \geq t\sigma) \leq e^{-\frac{t^2}{2}}$.

We can now prove our bound on $\ell^*(K_{\mathcal{Q}})$. First we will prove a bound on how likely large values of the support function are with input $g \sim \mathcal{N}(0, I)$:

$$\begin{aligned}\mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq \sqrt{t^2 + 2\log|\mathcal{X}|}\sqrt{m}) &= \mathbb{P}(\max_{x \in S_{\mathcal{Q}}} \langle g, x\rangle \geq \sqrt{t^2 + 2\log|\mathcal{X}|}\sqrt{m}) \\ &\leq \sum_{x \in S_{\mathcal{Q}}} \mathbb{P}(\langle g, x\rangle \geq \sqrt{t^2 + 2\log|\mathcal{X}|}\sqrt{m})\end{aligned}$$

Here the probability is taken over g. The equality is because $K_{\mathcal{Q}}$ is the convex hull of the points in $S_{\mathcal{Q}}$. The inequality is just the union bound. Now, using the fact that $g \sim \mathcal{N}(0, I)$ and combining with the above fact (2) we get $\langle g, x\rangle \sim \mathcal{N}(0, \|x\|_2^2)$. The standard deviation of this new Gaussian is $\|x\|_2$, so by fact (1), and the fact $x \in S_{\mathcal{Q}}$, we get the standard deviation of this Gaussian is at most $\sqrt{m}$. Now we can apply the concentration bound from fact (3) and get:

$$\begin{aligned}&\sum_{x \in S_{\mathcal{Q}}} \mathbb{P}(\langle g, x\rangle \geq \sqrt{t^2 + 2\log|\mathcal{X}|}\sqrt{m}) \\ &\leq \sum_{x \in S_{\mathcal{Q}}} e^{\frac{-(\sqrt{t^2+2\log|\mathcal{X}|})^2}{2}} \\ &= \sum_{x \in S_{\mathcal{Q}}} \frac{1}{|\mathcal{X}|} e^{\frac{-t^2}{2}} \leq e^{\frac{-t^2}{2}}\end{aligned}$$

The final equality is because $|S_{\mathcal{Q}}| = |\mathcal{X}|$. So overall we get that:

$$\mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq (t + \sqrt{2\log|\mathcal{X}|})\sqrt{m}) \leq \mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq \sqrt{t^2 + 2\log|\mathcal{X}|}\sqrt{m}) \leq e^{\frac{-t^2}{2}}$$

To actually show $\ell^*(K_{\mathcal{Q}}) \in O(\sqrt{m\log|\mathcal{X}|})$, we recall the Gaussian width is defined to be $\mathbb{E}h_{K_{\mathcal{Q}}}(g)$ where $g \sim \mathcal{N}(0, I)$. By integration by parts, we can write this expectation as:

$$\begin{aligned}\mathbb{E}h_{K_{\mathcal{Q}}}(g) &= \int_0^\infty \mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq t)dt \\ &= \int_0^{2\sqrt{m}\log|\mathcal{X}|} \mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq t)dt + \int_{2\sqrt{m}\log|\mathcal{X}|}^\infty \mathbb{P}(h_{K_{\mathcal{Q}}}(g) \geq t)dt \\ &\leq 2\sqrt{m}\log|\mathcal{X}| + \sqrt{m}\int_0^\infty e^{-t^2/2}dt \\ &= \sqrt{m}\left(\sqrt{\frac{\pi}{2}} + 2\log|\mathcal{X}|\right).\end{aligned}$$

In the inequality we used the bound on the probability that $h_{K_{\mathcal{Q}}}(g)$ is large, derived above, and change of variables.