Notes on the Exponential Mechanism. (Differential privacy)

Boston University CS 558. Sharon Goldberg.

February 22, 2012

1 Description of the mechanism

Let $\mathcal{D}$ be the domain of input datasets.

Let $\mathcal{R}$ be the range of "noisy" outputs.

Let $\mathbb{R}$ be the real numbers.

We start by defining a scoring function score : $\mathcal{D} \times \mathcal{R} \to \mathbb{R}$ that takes in a dataset $A \in \mathcal{D}$ and output $r \in \mathcal{R}$ and returns a real-valued score; this score tells us how "good" this output r is for this dataset A, with the understanding that higher scores are better.

The exponential mechanism $\mathcal{E}$ takes in the scoring function score and a dataset A parameter ϵ and does the following:

$$\mathcal{E}(A, \text{score}, \epsilon) = \text{output } r \text{ with probability proportional to } \exp(\tfrac{\epsilon}{2}\text{score}(A, r))$$

1.1 Sensitivity

Next we need to define the sensitivity of a scoring function. The sensitivity Δ tells us the maximum change in the scoring function for any pair of datasets A, B such that $|A \oplus B| = 1$. More formally:

$$\Delta = \max_{r,A,B \text{ where } |A \oplus B|=1} |\text{score}(A, r) - \text{score}(B, r)|$$

1.2 Privacy of exponential mechanism

We can show that the privacy of the exponential mechanism depends on the sensititivy Δ.

Theorem 1. *If the score function has sensitivity Δ, the mechanism $\mathcal{E}(A, score, \epsilon)$ is $\epsilon\Delta$-differentially private.*

Proof. First, since we know the score has sensitivity Δ, we know that for any A, B where $|A \oplus B| = 1$, it follows that

$$-\Delta \leq \text{score}(A, r) - \text{score}(B, r) \leq \Delta \tag{1}$$

What if we have X, Z such that $|X \oplus Z| = a$? How can we bound the score function? Well, imagine we had $a + 1$ intermediate databases, where $Y_1 = X$ and $Y_{a+1} = Z$ and for each $i = 1, ..., a$ we have that $|Y_i \oplus Y_{i+1}| = 1$. Then, we can take the telescoping sum:

$$\text{score}(X, r) - \text{score}(Z, r) = (\text{score}(Y_1, r) - \text{score}(Y_2, r)) + (\text{score}(Y_2, r) - \text{score}(Y_3, r)) + ... + (\text{score}(Y_a, r) - \text{score}(Y_{a+1}, r)$$

Repeatedly applying (1) to the telescoping sum we get:

$$-a\Delta \leq \text{score}(X, r) - \text{score}(Y, r) \leq a\Delta$$

Now, let's take two databases X, Y where $|X \oplus Y| = a$. We can write:

$$\begin{aligned}
\frac{\Pr[M(X)=r]}{\Pr[M(Y)=r]} &= \frac{\frac{\exp(\frac{\epsilon}{2}\text{score}(X,r))}{\int_{\rho\in\mathcal{R}}\exp(\frac{\epsilon}{2}\text{score}(X,\rho))d\rho}}{\frac{\exp(\frac{\epsilon}{2}\text{score}(Y,r))}{\int_{\rho\in\mathcal{R}}\exp(\frac{\epsilon}{2}\text{score}(Y,\rho))d\rho}} \\
&= \frac{\exp(\frac{\epsilon}{2}(\text{score}(X,r)-\text{score}(Y,r)))}{\int_{\rho\in\mathcal{R}}\exp(\frac{\epsilon}{2}(\text{score}(X,\rho)-\text{score}(Y,\rho)))d\rho} \\
&\leq \frac{\exp(\frac{\epsilon}{2}a\Delta)}{\exp(-\frac{\epsilon}{2}a\Delta)\int_{\rho\in\mathcal{R}}d\rho} && \text{(Apply equation (1))} \\
&\leq \exp(\epsilon a\Delta) && (\text{If } \int_{\rho\in\mathcal{R}} d\rho \geq 1) \\
&= \exp(\epsilon\Delta|X\oplus Y|)
\end{aligned}$$

which gives us $\epsilon\Delta$-differential privacy. So we're done. □

1.3 So what's the point?

The point is, from now on we no longer have to design mechanisms – we need only design the scoring function! Then, working out the sensitivity of the scoring function, we can determine the differential privacy guarantee.

2 The Laplace mechanism

2.1 Laplace noise is an instance of the exponential mechanism

We can capture the laplace noise mechanism by setting the score function to be

$$\text{score}(A,r) = -2|\text{count}(A)-r|$$

To see how this works, notice first that with this scoring function, the exponential mechanism becomes

$$\mathcal{E}(A,\text{score},\epsilon) = \text{output } r \text{ with probability proportional to } \exp(-\epsilon|\text{count}(A)-r|)$$

we can equivalently write this as

$$\Pr[\mathcal{E}(A,\text{score},\epsilon)=r] \propto \exp(-\epsilon|\text{count}(A)-r|)$$

Meanwhile recall that for the laplace mechanism we have

$$\Pr[\text{count}(A)+\mathcal{L}(\tfrac{1}{\epsilon})=r] = \Pr[\mathcal{L}(\tfrac{1}{\epsilon})=r-\text{count}(A)] = \tfrac{\epsilon}{2}\exp(-\epsilon|\text{count}(A)-r|)$$

and so we can see the two are the same.

2.2 Sensitivity of this scoring function?

What is the sensitivity of this scoring function?

$$\begin{aligned}
\Delta &= \max_{r,A,B \text{ where } |A \oplus B|=1} |\text{score}(A,r) - \text{score}(B,r)| \\
&= \max_{r,A,B \text{ where } |A \oplus B|=1} |-2|\text{count}(A) - r| - (-2|\text{count}(B) - r|)| \\
&= 2 \max_{r,A,B \text{ where } |A \oplus B|=1} ||\text{count}(A) - r| - |\text{count}(B) - r|| \\
&\leq 2 \max_{r,A,B \text{ where } |A \oplus B|=1} |\text{count}(A) - \text{count}(B)| \qquad \text{(triangle inequality)} \\
&\leq 2
\end{aligned}$$

So, using this general purpose framework, it seems like the exponential mechanism provides 2ϵ-differential privacy. How could this be? We know the laplace mechanism gives us ϵ-DP! Where is the extra factor of 2 coming from?

(Think about it; it falls out of the proof of Theorem 1 – the idea is that we've used the general analysis of the exponential mechanism, so our result is less "tight".)

3 The median mechanism

We'll use the running of example of the median to explain how the exponential mechanism works. (Recall that the median of a list of numbers $A = \{1, 2, 4, 60, 71\}$ is 4; so we shall write $\text{med}(A) = 4$.)

For the median example, our scoring function will be:

$$\text{score}(A, x) = -\min |A \oplus B| \text{ such that } \text{med}(B) = x$$

What does this mean? The idea here is that we take in an A and a candidate median x, and we look for a B that is as similar as possible to A, such that $\text{med}(B) = x$.

What's the maximum possible score here? We'd expect the best possible score to be $\text{score}(A, a)$ where $\text{med}(A) = a$. Well, we can take $B = A$, and get that $\text{med}(B) = \text{med}(A) = a$; since for $A = B$ we have that $|A \oplus B| = 0$, we know that the score is 0. Similarly, we can show that for $x \neq \text{med}(A)$, it follows that $\text{score}(A, x) \leq 0$.

3.1 Sensitivity of this scoring function

Let's work out the insensitivity of this scoring function. I claim the sensitivity is 1.

To show this, I need to take two datasets X, Y such that $|X \oplus Y| = 1$, and show that *for any* such X, Y and *for any* "noisy output" r, it follows that

$$|\text{score}(X, r) - \text{score}(Y, r)| \leq 1$$

Let's do this now. For any X, r define s as $s = \text{score}(X, r)$. For the definition of the scoring function, we know that there must be some database B such that $|X \oplus B| = -s$ and $\text{med}(B) = r$.

Now, recall that Y must be such that $|Y \oplus X| = 1$. They key observation we can now make is that

$$|Y \oplus B| \leq |Y \oplus X| + |X \oplus B| \leq -s + 1$$

since $\text{med}(B) = r$, it follows that $\text{score}(Y, r) \geq s - 1$.

Now we can write

$$|\text{score}(X,r) - \text{score}(Y,r)| \le |s - (s-1)| = 1$$

Since this argument holds for every X, Y such that $|X \oplus Y| = 1$, we are done!

3.2 How accurate is this mechanism?

It depends! A careful look shows that the *accuracy of this mechanism depends on the input itself*! This is in contrast to every other mechanism we've seen so far in the course; for instance, the Laplace counting mechanism applies the same noise magnitude regardless of the input. To see how accurate this mechanism is, we'll consider two different datasets.

A "nice" dataset. Let's start with a "nice" dataset:

$$A = \{1, 100, 102, 104, 105, 200, 365\}$$

Notice that $\text{med}(A) = 104$. It follows that $\text{score}(A, 104) = 0$, since we can take the B in the scoring function as $B = A$, and we have that $\text{med}(B) = \text{med}(A) = 104$ while $|A \oplus B| = 0$.

Now, consider the following three datasets:

$$B_{102} = \{1, 100, 102, 102, 105, 200, 365\}; \qquad \text{med}(B_{102}) = 102$$

$$B_{103} = \{1, 100, 102, 103, 105, 200, 365\}; \qquad \text{med}(B_{103}) = 103$$

$$B_{105} = \{1, 100, 102, 105, 105, 200, 365\}; \qquad \text{med}(B_{105}) = 105$$

Using B_{102}, we can now see that

$$\text{score}(A, 102) = -2$$

. How can we see this? Well, first note that $|A \oplus B_{102}| = 2$ (not that this is 2, not 1, because A and B_{102} *differ* on the middle entry). Furthermore, we know that $\text{med}(B_{102}) = 102$, so applying the scoring function we can see that $\text{score}(A, 102) = -2$.

Using the similar argument and datasets B_{103} and B_{105} above, we can see that

$$\text{score}(A, 102) = \text{score}(A, 103) = \text{score}(A, 105) = -2$$

What about $\text{score}(A, 101)$? If we think about this a bit, we can come up with

$$B_{101} = \{1, 100, 100, 101, 105, 200, 365\}$$

and using this, show that $\text{score}(A, 101) = -4$.

So what does this all mean? Letting `NoisyMed` be our median mechanism, it means the following:

$$\begin{aligned}
\Pr[\texttt{NoisyMed}(A) = 104] &\propto 1 \\
\Pr[\texttt{NoisyMed}(A) = 102] = \Pr[\texttt{NoisyMed}(A) = 103] = \Pr[\texttt{NoisyMed}(A) = 105] &\propto e^{\frac{\epsilon}{2}(-2)} = e^{-\epsilon} \\
\Pr[\texttt{NoisyMed}(A) = 100] = \Pr[\texttt{NoisyMed}(A) = 101] &\propto e^{\frac{\epsilon}{2}(-4)} = e^{-2\epsilon}
\end{aligned}$$

In other words, `NoisyMed` is e^ϵ times more likely to output the correct answer 104 than the (not-too-noisy) answer 102,103, or 105. Similarly, the correct answer is $e^{2\epsilon}$ times more likely that the (slightly-more-noisy) answer 100 or 101. And so on.

So all this isn't too bad; our mechanism is very likely to output an answer that is very close to the correct answer, 104.

A "worst-case" dataset. But things aren't always so nice. Now let's consider the following "worst-case" database:

$$A = \{0, 0, 0, 0, 10^6, 10^6, 10^6\}$$

Obviously $\text{med}(A) = 0$, and usual a similar argument as above, we have that $\text{score}(A, 0) = 0$. Before we move on, let's consider the following few datasets:

$$B_1 = \{0, 0, 0, 1, 10^6, 10^6, 10^6\}; \qquad \text{med}(B_1) = 1$$

$$\cdots$$

$$B_i = \{0, 0, 0, i, 10^6, 10^6, 10^6\}; \qquad \text{med}(B_i) = i$$

$$\cdots$$

$$B_{10^6-1} = \{0, 0, 0, 10^6 - 1, 10^6, 10^6, 10^6\}; \qquad \text{med}(B_{10^6-1}) = 10^6 - 1$$

Now, looking at B_i we can apply the usual argument to find that $\text{score}(A, i) = -2$ (since $\text{med}(B_i) = i$ and $|A \oplus B_i| = 2$). This will hold for every $i = 1, ..., 10^6 - 1$. So what does this mean?

Well, now we know that

$$\begin{aligned} \Pr[\texttt{NoisyMed}(A) = 0] &\propto 1 \\ \Pr[\texttt{NoisyMed}(A) = i] &\propto e^{\frac{\epsilon}{2}(-2)} = e^{-\epsilon} \qquad \text{for } i = 1, ..., 10^{-6} - 1 \end{aligned}$$

so our mechanism is just a likely to output 1 as the median, as it is to output $10^6 - 1$ as the median! Notice how far off $10^6 - 1$ is from the true median 0; what this means is that when the input dataset is nasty, the mechanism can be highly inaccurate.

3.3 To wrap up

The bottom line is, while this mechanism gives us wonderfully clean privacy guarantees – it is ϵ-differentially private (for all datasets of course, since this is what the definition of DP requires from us) – the accuracy of the mechanism depends strongly on the dataset itself. In the first 'nice' dataset we saw, the mechanism's outputs were concentrated tightly around the correct answer of 104; meanwhile, for the worst case dataset, the mechanisms outputs were uniformly spread on $1, ..., 10^6 - 1$, which is pretty inaccurate.