ISYE 8813: Algorithmic Foundations of Ethical ML **Fall 2021**

Lectures 3 and 4: The Randomized Response and Laplace Mechanisms

Lecturer: Juba Ziani

In the previous lecture, we went through the formal definition of differential privacy, what it means and protects against, and a simple relaxation of said definition. We, however, have not yet seen how to design algorithms and mechanisms that satisfy ε-DP. This lecture is aimed at introducing the most basic algorithms and building blocks for differential privacy.

1 Privacy for Binary Attributes: the Randomized Response Mechanism

In this section, we consider that each individual's data belong to $\mathcal{X} = \{0,1\}$. So, each data point here is binary, and corresponds to whether an individual or not has a certain attribute, exhibits a certain behavior, etc.

The RR algorithm Now, imagine I am to report some sensitive and attribute behavior to a data analyst. For example, I want to report to my professor whether I cheated on the exam, but I do not want the professor to know with certainty that I have cheated. How can I do so? The idea behind Randomized Response (RR) is simple: I just flip my bit (0 if I did not cheat, 1 if I did) with some probability.

Example 1. *Two simple cases of RR:*

1. *Let $b \in \{0,1\}$ my private bit. Consider the following mechanism: with probability $1/2$, I report $\mathcal{M}(b) = b$, and with probability $1/2$, I report $\mathcal{M}(b) = 1 - b$.*
   Now, take the point of view of the data analyst. Imagine the analyst sees report $\mathcal{M}(b) = 1$; what is his posterior on what the original value of b was? By Bayes rule, we have

$$\begin{aligned}
Pr\left[b = 1 | \mathcal{M}(b) = 1\right] &= \frac{Pr\left[\mathcal{M}(b) = 1 | b = 1\right] Pr[b = 1]}{Pr[\mathcal{M}(b) = 1]} \\
&= \frac{Pr\left[\mathcal{M}(b) = 1 | b = 1\right] Pr[b = 1]}{Pr[\mathcal{M}(b) = 1 | b = 0] Pr\left[b = 0\right] + Pr[\mathcal{M}(b) = 1 | b = 1] Pr\left[b = 1\right]} \\
&= \frac{1/2 Pr[b = 1]}{1/2(Pr[b = 0] + Pr[b = 1])} \\
&= Pr[b = 1].
\end{aligned}$$

   You can interpret $Pr[b = 1]$ as your prior probability/knowledge of the distribution of b, before using your mechanism. Then, here, I see that my posterior probability of $b = 1$ after I see the output of the mechanism is the same as my prior probability of

$b = 1$ *before seeing the outcome of the mechanism. Hence, the mechanism* gives me no new information compared to what I initially knew.
In probabilistic term, b is probabilistic-ally independent *of the outcome of the mechanism. I.e., $\mathcal{M}(b)$ encodes no information about b. So, here, we have perfect privacy (but $\mathcal{M}$ is useless to the professor).*

2. *Now, suppose that I flip my private bit with probability 0, i.e. $\mathcal{M}(b) = b$ always. Then the analyst knows that the bit his see is exactly the private bit, and perfectly knows b. Then, no privacy at all is guaranteed, but we have perfect accuracy on the agent's private bit.*

In reality, we want something in the middle: i.e., we flip the private bit with some probability in $[0, 1/2]$, and we get some imperfect but non-trivial accuracy. We formalize the RR mechanism as follows: given a probability $p \in [0, 1/2]$, we let

$$RR_p(b) = \begin{cases} 1 - b & \text{w.p. } p \\ b & \text{w.p. } 1 - p, \end{cases}$$

where $p \in [0, 1/2]$. Note that restricting to $[0, 1/2]$ is wlog, as RR_p and RR_{1-p} are equivalent: as an example, flipping with probability 1 means that the analyst knows he always sees $1 - b$, in which case he can infer b perfectly.

Remark 1. *In [1], the proposed mechanism is slightly different. Their mechanism is the following:*

1. *Flip a first coin.*
2. *If heads (probability $1/2$), report b. If tails, flip a second, but now biased coin: with probability q, output $1 - b$, and with probability $1 - q$, output b.*

This is equivalent to $RR_p(b)$ with $q = 2p$; indeed, a simple calculation shows that $P[\mathcal{M}(b) = 1 - b] = q/2$.

Privacy guarantees of RR We now study the differential privacy guarantees of RR_p:

Theorem 2. *Let $p \in [0, 1/2]$. RR_p is $(\varepsilon, 0)$-differentially private with*

$$\varepsilon = \max\left(\ln \frac{p}{1-p}, \ln \frac{1-p}{p}\right).$$

When $p \leq 1/2$, this reduces to $\ln \frac{1-p}{p}$.

Proof. Remember the definition of ε-DP: for any neighboring databases x and y, for any possible outcome S, we want

$$\frac{Pr[\mathcal{M}(x) \in S]}{Pr[\mathcal{M}(y) \in S]} \leq \exp(\varepsilon).$$

Here, the possible databases are either $x = \{0\}$ or $x = \{1\}$ (the private bit of a single participant). The possible outcomes are also $\{0\}$ or $\{1\}$ (ignoring the cases $S = \emptyset$ and $S = \{0, 1\}$, those are trivial).

To look at all possible combinations of neighboring databases and outcome sets, we need to upper bound, for $b \in \{0, 1\}$, neighbouring $b' = 1 - b$, and for possible outcomes $\mathcal{M}(b) = b$ and $\mathcal{M}(b) = 1 - b$,

$$\frac{Pr[\mathcal{M}(b) = b]}{Pr[\mathcal{M}(b') = b]} = \frac{Pr[\mathcal{M}(b) = b]}{Pr[\mathcal{M}(1 - b) = b]},$$

and

$$\frac{Pr[\mathcal{M}(b) = 1 - b]}{Pr[\mathcal{M}(b') = 1 - b]} = \frac{Pr[\mathcal{M}(b) = 1 - b]}{Pr[\mathcal{M}(1 - b) = 1 - b]}.$$

Note that the first quantity is equal to $\frac{1-p}{p}$, while the second quantity is equal to $\frac{p}{1-p}$. Therefore, we have that for any neighboring databases x and y, for any possible outcome S,

$$\frac{Pr[\mathcal{M}(x) \in S]}{Pr[\mathcal{M}(y) \in S]} \leq \max\left(\frac{p}{1-p}, \frac{1-p}{p}\right) \triangleq \exp(\varepsilon).$$

Taking the log gives us the result. □

Special cases, and how they relate to Example 1:

- Take $p = 1/2$. Then $\varepsilon = \ln \frac{1-p}{p} = \ln \frac{1/2}{1/2} = \ln 1 = 0$. This means that we have perfect privacy, as seen in Example 1.
- Take $p \to 0$, we have $\varepsilon \to \ln \frac{1}{0} = +\infty$, and we have no privacy at all.

Using RR to compute population means: Now, we imagine that we have n individuals in our database, with individual i having a private bit b_i. We want to compute $\frac{1}{n}\sum_{i=1}^{n} b_i$ privately, but we only ever have access to $\mathcal{M}(b_i)$ for each agent i. How can we do so?

Claim 3. *Fix $b \in \{0, 1\}$. Then,*

$$\mathbb{E}\left[\mathcal{M}(b)\right] = 1 - p + (2p - 1)b.$$

In turn,

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}(b_i) - (1 - p)}{2p - 1}$$

is an unbiased estimator of $\frac{1}{n}\sum_{i=1}^{n} b_i$.

Proof.

$$\begin{aligned}\mathbb{E}\left[\mathcal{M}(b)\right] &= Pr[\mathcal{M}(b) = b] \cdot b + Pr[\mathcal{M}(b) = 1 - b] \cdot (1 - b) \\ &= pb + (1 - p)(1 - b) \\ &= 1 - p + (2p - 1)b.\end{aligned}$$

□

Accuracy of the above estimator In practice, the goal is to take p as close to $1/2$ as possible, say $p = 1/2 - \gamma$ for some small γ. Intuitively, the more individuals I have in my database/the larger n is, the smaller I can take γ and the stronger my privacy guarantee can be while preserving the same level of accuracy (intuitively the more points you take the average over, the more noise you can add). We formalize this intuition below. To do so, we first remind the reader of Chebyshev's inequality:

Lemma 4 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then,*

$$\Pr\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}.$$

Theorem 5 (Accuracy of estimator $\hat{b}$)**.** *With probability at least $1 - \delta$,*

$$\left|\hat{b} - \mathbb{E}[b]\right| \leq \frac{\sqrt{1/\delta}}{2(1-2p)\sqrt{n}}.$$

This means that if we want to take $p = \frac{1}{2} - \gamma$ (leading to $1 - 2p = 2\gamma$) for a very small value of γ to give strong privacy guarantees, we need roughly $n = \Omega\left(\frac{1}{\alpha^2\gamma^2}\right)$ samples to get small additive error α.

Proof. First, we bound the variance of the estimator. We have that

$$\operatorname{Var}\left[\hat{b}\right] = \frac{1}{(2p-1)^2 n^2}\sum_{i=1}^{n} \operatorname{Var}\left[\mathcal{M}(b_i)\right] \leq \sum_{i=1}^{n} \frac{1}{4(2p-1)^2 n^2} = \frac{1}{4(2p-1)^2 n},$$

where the first equality follows from the fact that $\operatorname{Var}[cX] = c^2 \operatorname{Var}[X]$ and $\operatorname{Var}[X + c] = \operatorname{Var}[X]$ for a constant c, and the inequality follows from the fact that $\mathcal{M}(b_i)$ is a Bernoulli random variable and has variance at most $1/4$. Using Chebyshev with $k = \frac{1}{\sqrt{\delta}}$, we have that

$$\Pr\left[\left|\hat{b} - \mathbb{E}[b]\right| \geq \frac{\sqrt{1/\delta}}{2(1-2p)\sqrt{n}}\right] \leq \delta.$$

□

2 Answering Numerical Queries: the Laplace Mechanism

We now see how to compute more general, *numerical*, (i.e. real-valued) queries privately. The main idea to do so is to add "central" noise: i.e., we first compute the exact value of the query non-privately, then add carefully chosen noise to the answer to that query, and only release said noisy answer.

Formally, our goal here is to answer a query $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$, which is simply a function that takes a database as an input, and outputs a real-valued answer. Some example of queries are:

- How many people in the database are male? Have cancer?
- What fraction of people in the database are female?
- What is the average age of individuals in the database?
- etc.

Sensitivity of a query Now, the amount of noise that we need to add for privacy depends on the "sensitivity" of a query, which is a measure of how much the value of the query changes when a single individual in the database changes his data. Intuitively, this is because the larger the effect a single individual can have the value of a query is, the more noise we need to add to "hide" this change. Formally:

Definition 6. *The sensitivity of a function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$ is given by*

$$\Delta f \triangleq \max_{x,y neighbors} |f(x) - f(y)|.$$

In words, the sensitivity of a query f is the maximum amount by which $f(x)$ changes when we change a single entry in x and go to a neighboring database y. This is how much of an effect a single individual can have on f, *in the worst case*. For example:

- How many people in the database are male? Have cancer? $\Delta f = 1$.
- What fraction of people in the database are female? $\Delta f = 1/n$.
- What is the average salary of individuals in the database? $\Delta f = +\infty$.
- etc.

The Laplace mechanism The most well-known and widely used mechanism to answer numerical queries is the Laplace mechanism. It works by adding Laplace noise to the answer of the query. First, let us study the Laplace distribution:

- We write $Z \sim Lap(b)$ to say that a random variable Z is drawn from the Laplace distribution with parameter b.
- Pdf: $f(z|b) = \frac{1}{2b} \exp(-|z|/b)$.
- It has expectation $\mathbb{E}[Z] = 0$ and variance $Var(Z) = 2b^2$.

Figure 1 shows the pdf of the Laplace distribution for different values of b. Note that the smaller b is, the more concentrated the distribution is around 0. So, when b is very small, you add very little noise to your query with high probability; when b is large, you have a much higher probability of being away from 0 and adding a significant amount of noise. Informally, the larger b is, the more noise you add (in absolute value).

We can now move to the definition of the Laplace mechanism:

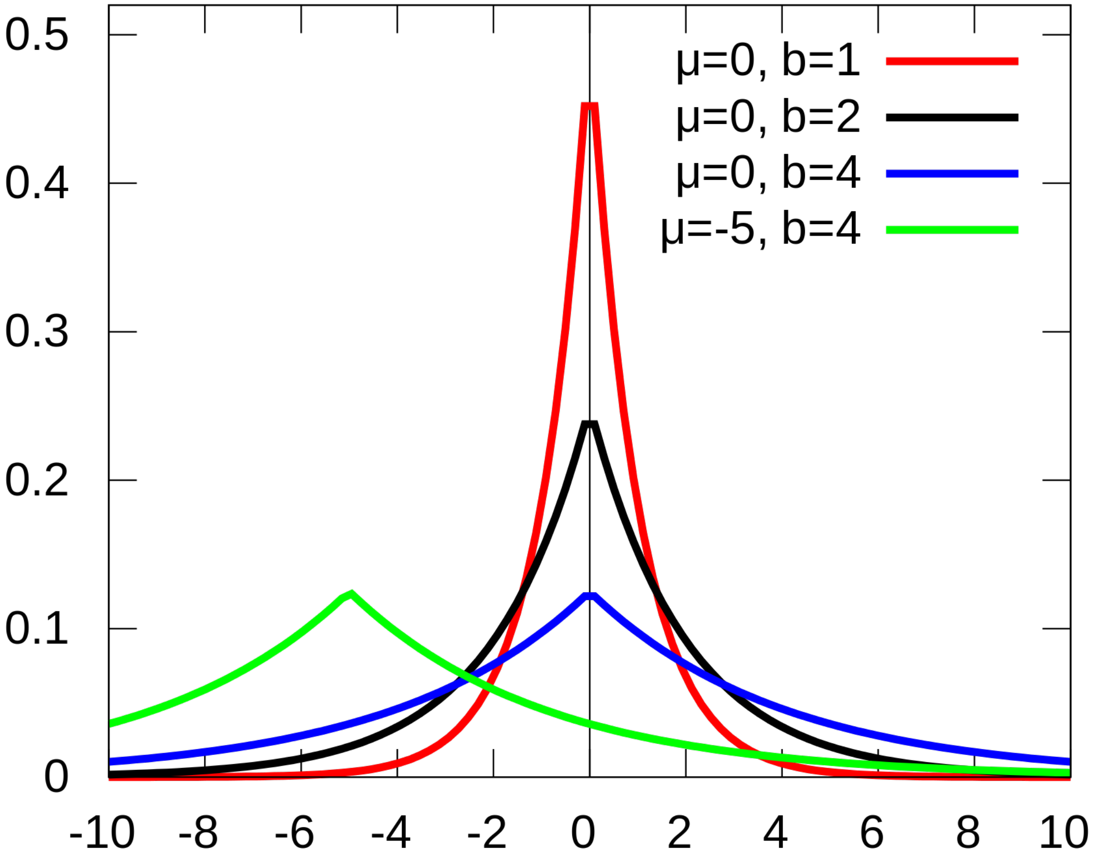

Figure 1: Laplace pdf for different values of b

Definition 7 (Laplace mechanism)**.** *Given a function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$, the Laplace Mechanism is given by*

$$\mathcal{M}_L(x, f, \varepsilon) = f(x) + Z,$$

where $Z \sim Lap\left(\frac{\Delta f}{\varepsilon}\right)$.

The mechanism works by computing the exact answer to the query f, then adding Laplace noise whose parameter b depends on the function's sensitivity and the desired level of noise. Since $b = \frac{\Delta f}{\varepsilon}$, we have that:

- The bigger Δf is/the more sensitive the query is, the bigger b is, and the more noise you add.
- The stronger the privacy guarantee you want to achieve is, the smaller ε is, hence the bigger b is, hence the more noise you add.

This follows our previous intuition: more privacy and more sensitive queries require more noise.

Privacy of the Laplace mechanism We now show the differential privacy guarantee that is obtained by the Laplace mechanism:

Theorem 8. *The Laplace Mechanism $\mathcal{M}_L(x, f, \varepsilon)$ is ε-differentially private.*

Proof. Consider two neighboring databases x and y. For any outcome s, for $X, Y \sim Lap\left(\frac{\Delta f}{\varepsilon}\right)$ (drawn independently from each other), we have that

$$\begin{aligned}
\frac{Pr\left[\mathcal{M}(x) = s\right]}{Pr\left[\mathcal{M}(y) = s\right]} &= \frac{Pr\left[f(x) + X = s\right]}{Pr\left[f(y) + Y = s\right]} \\
&= \frac{Pr\left[X = s - f(x)\right]}{Pr\left[Y = s - f(y)\right]} \\
&= \frac{\exp\left(-\frac{\varepsilon}{\Delta f}|s - f(x)|\right)}{\exp\left(-\frac{\varepsilon}{\Delta f}|s - f(y)|\right)} \\
&= \exp\left(\frac{\varepsilon}{\Delta f} \cdot (|s - f(y)| - |s - f(x)|)\right)
\end{aligned}$$

By the triangle inequality, we know that

$$|s - f(y)| - |y - f(x)| \leq |f(y) - f(x)|.$$

Further, by the definition of sensitivity, we have that

$$|f(y) - f(x)| \leq \Delta f.$$

Plugging this back in the equation above, we get that

$$\frac{Pr\left[\mathcal{M}(x) = s\right]}{Pr\left[\mathcal{M}(y) = s\right]} \leq \exp\left(\frac{\varepsilon}{\Delta f} \cdot \Delta f\right) = \exp(\varepsilon).$$

This concludes the proof. □

Accuracy of the Laplace mechanism As we have seen before, there is a trade-off between accuracy and privacy. The more privacy we want, the more noise we need to add to our query, hence the less accuracy we get. The following theorem quantifies this accuracy, by providing a high-probability bound on how far $f(x) + Z$ can be from $f(x)$, i.e. on how big Z is:

Theorem 9. *For any function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$,*

$$Pr\left[|f(x) - \mathcal{M}_L(x, f, \varepsilon)| \geq \ln\left(\frac{1}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] \leq \delta.$$

Proof. This immediately follows from the tails of the Laplace distribution. For $Z \sim Lap(b)$ and $z > 0$, we know that

$$Pr\left[|Z| \geq z\right] = \exp\left(-z/b\right).$$

Therefore, given that $f(x) - \mathcal{M}_L(x, f, \varepsilon)$ is Laplace with parameter $b = \frac{\Delta f}{\varepsilon}$, we have

$$Pr\left[|f(x) - \mathcal{M}_L(x, f, \varepsilon)| \geq z\right] = \exp\left(-z \cdot \frac{\varepsilon}{\Delta f}\right) \triangleq \delta.$$

Solving for z in $\exp\left(-z \cdot \frac{\varepsilon}{\Delta f}\right) = \delta$, we obtain

$$z \cdot \frac{\varepsilon}{\Delta f} = \ln\left(\frac{1}{\delta}\right),$$

hence

$$z = \ln\left(\frac{1}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right).$$

This concludes the proof. □

Note that this shows a trade-off between privacy and accuracy. As ε becomes smaller (i.e. we impose more stringent privacy guarantees), the accuracy of our mechanism (defined by how far the private answer to our query is from the true answer) worsens inversely proportionally to ε.

3 A Generalization: the Multi-Dimensional Laplace Mechanism

We now consider the case when $f(x) : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^d$, i.e. f is a *vector-valued function*. Think of $f(x)$ as a vector of d real values. For example, I may want to answer queries of the form:

- (”How many people in the data are female”, ”how many people in the database are male”). This is a 2-dimensional vector-valued query.
- etc.

In this case, we can still apply a vector-valued version to the Laplace mechanism. To do so, we need to first generalize our notion of sensitivity:

Definition 10. *The ℓ_1-sensitivity of a function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^d$ is given by*

$$\Delta f \triangleq \max_{x,y neighbors} \|f(x) - f(y)\|_1,$$

where $\|.\|_1$ denotes the ℓ_1-norm.

Remember that if z is a d-dimensional vector, we have that $\|z\|_1 = \sum_{i=1}^{d} |z_i|$. So here, for a vector valued query f, where we can write $f(x) = (f_1(x), \dots, f_d(x))$, we can rewrite

$$\Delta f = \max_{x,y \text{ neighbors}} \sum_{i=1}^{d} |f_i(x) - f_i(y)| \,.$$

In the example above, suppose we change one entry/swap the gender of a single individual in the database. Then, assuming binary genders (a strong assumption), the number of male students in the database changes by 1, and the number of female students in the database changes by 1. The sensitivity is therefore $1 + 1 = 2$.

However, suppose we use the addition-deletion definition of differential privacy here. Adding a male or female student only changes either the count of male students or the count of female students by 1, but not both simultaneously. In this case, the sensitivity of the function is $\Delta f = 1$.

We now generalize the Laplace mechanism below:

Definition 11 (Laplace mechanism)**.** *Given a function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^d$, the Laplace Mechanism is given by*

$$\mathcal{M}_L(x, f, \varepsilon) = f(x) + (Z_1, \dots, Z_d),$$

where $Z_i \sim Lap\left(\frac{\Delta f}{\varepsilon}\right)$, and $Z_1, \dots, Z_d$ are drawn independently from each other.

All previous results can then be extended:

Theorem 12. *The vector-valued Laplace mechanism $M_L(x, f, \varepsilon)$ is ε-differentially private. Further, for any function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$,*

$$Pr\left[|f(x) - \mathcal{M}_L(x, f, \varepsilon)| \geq \ln\left(\frac{d}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right) \right] \leq \delta.$$

Proof. Very similar to the 1-dimensional case. For more details, see pages 32 and 34 of [1]. □

Note the slight difference in the accuracy bound, which is now slightly worse and now depends on d, the dimension of the vector-valued function we consider; this comes from the fact that we have to add more noise, as we are answering more queries simultaneously. HoweverNicely, the dependency of d of the accuracy goes to the logarithm, meaning that as d increases, the accuracy worsens at a very slow rate. gracefully, at a very slow rate.

References

[1] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.