Foundations of Privacy

Lecture 3

Resume of previous lecture

Differential Privacy (continuous case) Let $\mathcal{K} : \mathcal{X} \to \mathcal{D}\mathcal{Z}$ be a randomized mechanism. We say that $\mathcal{K}$ is ε-differentially private if for every pair of databases $x_1, x_2 \in \mathcal{X}$ such that $x_1 \sim x_2$, and for every measurable $\mathcal{S} \subseteq Z$, we have:

$$p(Z \in \mathcal{S}|X = x_1) \leq e^{\varepsilon} p(Z \in \mathcal{S}|X = x_2)$$

where $p(Z \in \mathcal{S}|X = x)$ represents the probability that on the database x the mechanism reports an answer in $\mathcal{S}$

Properties

- Differential privacy is independent from the prior and the side knowledge of the adversary. In general by prior knowledge we mean the prior probabilistic knowledge about x, which represent the private values of the participants in the database (prior = before knowing the reported answer $z = \mathcal{K}(x)$). By side knowledge we mean every other knowledge of the adversary.

- Differential privacy is compositional, namely: given two mechanisms $\mathcal{K}_1$ and $\mathcal{K}_2$ on $\mathcal{X}$ that are respectively ε_1 and ε_2-differentially private, their composition $\mathcal{K}_1 \times \mathcal{K}_2$ is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

Resume of previous lecture

The meaning of differential privacy can be better understood in Bayesian terms. In the following, x_i represents the value of a participant i in the database, and x_{others} represents the value of all other participants.

Bayesian characterization

ε-differential-privacy is equivalent to the following property:

For all $(x_i, x_{others}) \in \mathcal{X}$, for all $z \in Z$,

$$e^{-\varepsilon} \leq \frac{p(X_i = x_i | X_{others} = x_{others}, Z = z)}{p(X_i = x_i | X_{others} = x_{others})} \leq e^{\varepsilon}$$

Namely: assuming that the adversary knows the value of all the other participants in the database, the reported answer does not increase significantly his probabilistic knowledge of the value of the participant i, with respect to his prior knowledge.

Note that the above property is **not** comparable with the following one:
For all $(x_i, x_{others}) \in \mathcal{X}$, for all $z \in Z$,

$$e^{-\varepsilon} \leq \frac{p(X_i = x_i | Z = z)}{p(X_i = x_i)} \leq e^{\varepsilon}$$

Namely, if we remove the conditioning on X_{others}, we obtain a different formula, which is neither stronger, nor weaker, than the previous one.

Resume of previous lecture

Laplace mechanism: Given a query $f: \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a subset of the real numbers, the Laplace mechanism $\mathcal{K}$ is obtained by adding Laplacian noise to the answer of f. Namely, if $f(x) = y$, then $\mathcal{K}(x)$ is a distribution on reals with a probability density function defined as:

$$dP_y(z) = c\, e^{-\frac{|z-y|}{\Delta_f}\varepsilon}$$

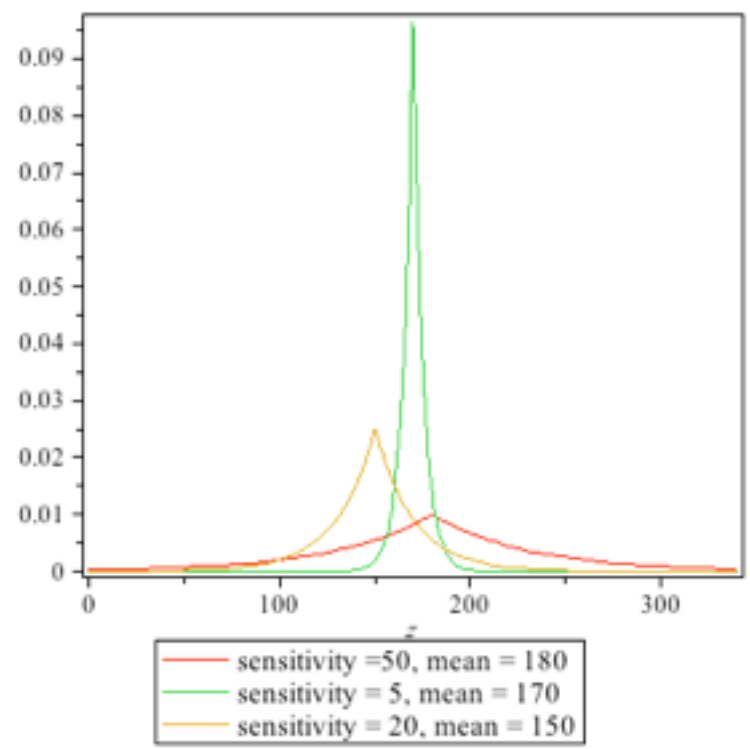
where Δ_f is the *sensitivity* of f:

$$\Delta_f = \max_{x \sim x' \in \mathcal{X}} |f(x) - f(x')|$$

and c is a normalization factor:

$$c = \frac{\varepsilon}{2\,\Delta_f}$$

Note that the Laplace mechanism is oblivious: the reported answer depends only on y, not on x



Resume of previous lecture

Geometric mechanism: This is an oblivious mechanism similar to the Laplace, but on integers rather than reals. Namely: Given a query $f: \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a subset of the integer numbers, the Geometric mechanism $\mathcal{K}$ is defined as follows: if $f(x) = y$, then $\mathcal{K}(x)$ is a distribution on integers with a probability distribution defined as:

$$p(Z = z | f(X) = y) = c\, e^{-\frac{|z-y|}{\Delta_f}\varepsilon}$$

where c is a normalization factor, given by

$$c = \frac{1-\alpha}{1+\alpha} \quad \text{with} \quad \alpha = e^{-\frac{\varepsilon}{\Delta_f}}$$

Exercise: show that the normalization factors of the Laplace and the geometric mechanisms are indeed those indicated

Intuition behind the Laplace distribution

Assume for example

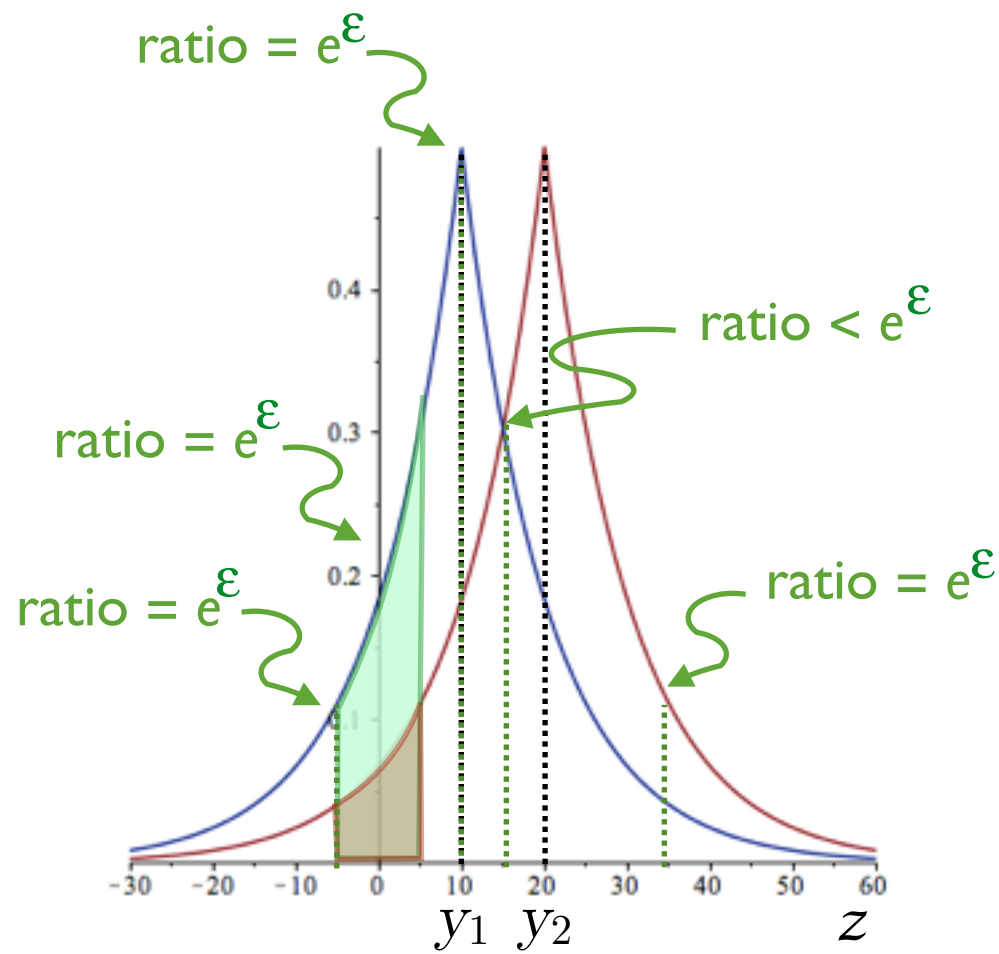
- $\Delta_f = |f(x_1) - f(x_2)| = 10$
- $y_1 = f(x_1) = 10$, $y_2 = f(x_2) = 20$

Then:

- $dP_{y_1}(z) = \frac{\varepsilon}{2 \cdot 10} e^{\frac{|z-10|}{10}\varepsilon}$
- $dP_{y_2}(z) = \frac{\varepsilon}{2 \cdot 10} e^{\frac{|z-20|}{10}\varepsilon}$

The ratio between these distribution is

- $= e^{\varepsilon}$ outside the interval $[y_1, y_2]$
- $\leq e^{\varepsilon}$ inside the interval $[y_1, y_2]$



Note that the distance between y_1 and y_2 is greatest when y_1 and y_2 correspond to the sensitivity of f. In this case the ratio between the respective Laplaces is e^{ε}. In all other cases, the distance between y_1 y_2 is smaller, and therefore also the ratio is smaller. Similar considerations hold for the geometric mechanism.

Example

Consider a query of the form

$$f(x) = \text{average age of the people in } x$$

Assume that:

- The DB contains at least 100 people
- The age of people ranges in $[0, 150]$

We want to define the Laplace mechanism for this query. For this purpose, we just need to compute the sensitivity of f.

$$\begin{aligned}
\Delta_f &= \max_{x_1 \sim x_2} |f(x_1) - f(x_2)| \\
&= \max_{a, v_1, v_2 \in [0,150]\ n \geq 100} \left|a - \tfrac{a\,n - v_1 + v_2}{n}\right| \\
&\quad \text{where } a = \text{average age in } x_1, \\
&\quad \text{and } v_1, v_2 = \text{ages of an individual in } x_1 \text{ and } x_2 \text{ resp.} \\
&= \max_{a, v_1, v_2 \in [0,150]\ n \geq 100} \left|\tfrac{v_1 - v_2}{n}\right| \\
&= \left|\tfrac{150 - 0}{100}\right| \\
&= 1.5
\end{aligned}$$

Hence: $dP_y(z) = \frac{\varepsilon}{3}\, e^{-\frac{|z-y|}{\Delta_f}\varepsilon}$

Example: Counting Queries

- A counting query for a certain property $\mathcal{P}$ is a query of the form:

  $f(x)$ = number of individuals in the database x who satisfy $\mathcal{P}$

- **Exercise:** determine the sensitivity of a counting query

Solution of the Exercises

Bob wants to find out whether Don is affected by a certain disease d. He knows Don's age and weight, and that Don is going to check in a hospital that maintains a database of all patients, and that can be queried with queries of the form:

- How many patients are affected by the disease d ?

- What is the average age and weight of the patients affected by the disease d?

Is it possible for Bob to determine, with high probability, whether Don has the disease ? If you answer yes, what is the strategy ? If you answer no, what other kind of queries or knowledge should Bob have at his disposal?

Solution of the Exercises

1. Show that the Laplace mechanism is ε-differentially-private
2. Prove that differential privacy is compositional (slide 2)
3. Prove that differential privacy is equivalent to its Bayesian characterization (slide 3)