Differential Privacy - An Overview #2

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

Outline

Outline

Recap from An Overview # 1

- Key idea: add noise to the output of queries made to databases
- Added noise is random; depends on a predetermined privacy budget and the type of queries
- Two important implications:
  1. the added noise is positively related to the sensitivity
  2. the added noise negatively related to the privacy budget

Recap from An Overview # 1

- Key idea: add noise to the output of queries made to databases
- Added noise is random; depends on a predetermined privacy budget and the type of queries
- Two important implications:
  1. the added noise is positively related to the sensitivity
  2. the added noise negatively related to the privacy budget
- How to add noise then?

Plan

- The Laplace Mechanism
  - a mechanism that satisfies $\epsilon-$differential privacy
  - adds noise from a Laplace distribution to the query output
  - the parameters of the corresponding Laplace distribution depend on the sensitivity (Δf) and the privacy budget (ϵ)
- Properties of differential privacy
  - composition theorem
  - sequential composition
  - parallel composition
  - post-processing

Outline

1. Introduction
2. The Laplace Mechanism
3. Properties of differential privacy

The Laplace distribution

- A random variable has a $\mathrm{Laplace}(\mu, s)$ distribution if its probability density function is

$$
\begin{aligned}
f(x \mid \mu, s) &= \frac{1}{2s} \exp\left(-\frac{|x-\mu|}{s}\right) && (1)\\
&= \frac{1}{2s} \begin{cases} \exp\left(-\frac{\mu - x}{s}\right) & \text{if } x < \mu; \\ \exp\left(-\frac{x-\mu}{s}\right) & \text{if } x \geq \mu, \end{cases} && (2)
\end{aligned}
$$

  - μ is a location parameter
  - $s > 0$ is a scale parameter
  - when $\mu = 0, b = 1$, the positive half-line is an exponential distribution scaled by $\frac{1}{2}$.

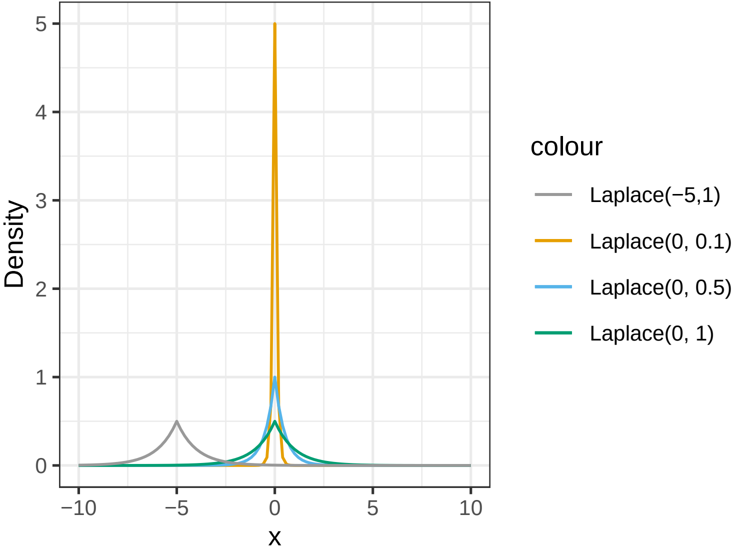
The Laplace distribution cont'd

- Like the normal distribution, the Laplace distribution is symmetric
- It is entered at its location parameter μ
- The scale parameter s controls its spread: larger s indicates bigger spread

The Laplace distribution cont'd

```r
require(rmutil)
cbPalette <- c("#999999",  "#E69F00", "#56B4E9",  "#009E73",
               "#CC79A7", "#D55E00",  "#F0E442", "#0072B2")
ggplot(data.frame(x = c(-10, 10)), aes(x)) +
  stat_function(fun = dlaplace, args = list(m = 0, s = 0.1),
                aes(color = "Laplace(0, 0.1)")) +
  stat_function(fun = dlaplace, args = list(m = 0, s = 0.5),
                aes(color = "Laplace(0, 0.5)")) +
  stat_function(fun = dlaplace, args = list(m = 0, s = 1),
                aes(color = "Laplace(0, 1)")) +
  stat_function(fun = dlaplace, args = list(m = -5, s = 1),
                aes(color = "Laplace(-5,1)")) +
  scale_colour_manual(values = cbPalette) + ylab("Density") +
  theme_bw(base_size = 10, base_family = "")
```

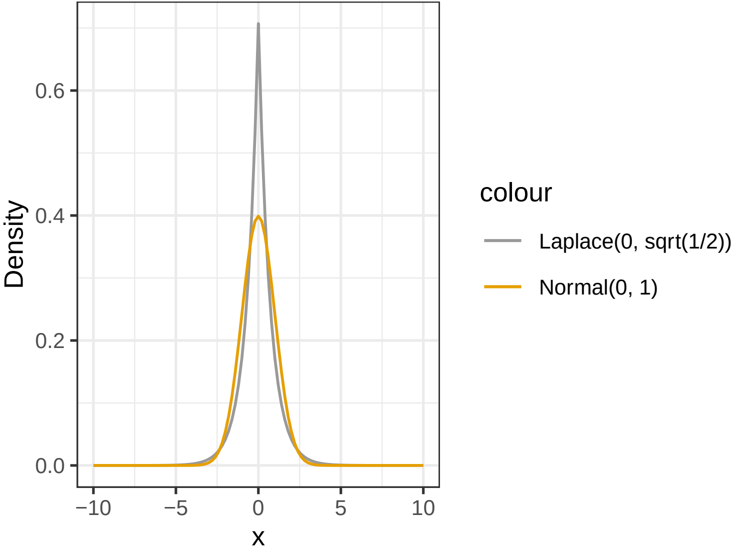
The Laplace distribution cont'd



The Laplace distribution vs the normal distribution

```
require(rmutil)
ggplot(data.frame(x = c(-10, 10)), aes(x)) +
  stat_function(fun = dlaplace, args = list(m = 0, s = sqrt(1/2)),
                aes(color = "Laplace(0, sqrt(1/2))")) +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1),
                aes(color = "Normal(0, 1)")) +
  scale_colour_manual(values = cbPalette) + ylab("Density") +
  theme_bw(base_size = 10, base_family = "")
```

The Laplace distribution vs the normal distribution cont'd



Laplace noise for privacy protection

- The Laplace Mechanism adds noise to the output with $\epsilon-$differential privacy guarantee
- The noise is drawn from a Laplace distribution
- Two important implications:
  1. the added noise is positively related to the sensitivity (Δf)
  2. the added noise negatively related to the privacy budget (ϵ)
- Also know that:
  1. the sensitivity (Δf) is dependent on the database and the query
  2. the privacy budget (ϵ) is independent of the database and the query

Laplace noise for privacy protection cont'd

- For given sensitivity Δf and privacy budget ϵ, the added noise to the output of a query sent to database $\mathbf{x}$, X^* is drawn from a Laplace distribution with mean 0, and scale $\frac{\Delta f}{\epsilon}$:

$$X^* \sim \mathrm{Laplace}\left(0, \frac{\Delta f}{\epsilon}\right) \tag{3}$$

Laplace noise for privacy protection cont'd

- For given sensitivity Δf and privacy budget ϵ, the added noise to the output of a query sent to database $\mathbf{x}$, X^* is drawn from a Laplace distribution with mean 0, and scale $\frac{\Delta f}{\epsilon}$:

$$X^* \sim \mathrm{Laplace}\left(0, \frac{\Delta f}{\epsilon}\right) \tag{3}$$

- The scale of a Laplace distribution controls its spread, and larger scale value indicates bigger spread
- If the added noise needs to be larger, we should draw it from a Laplace distribution with larger scale value, and vice versa

The scale $\frac{\Delta f}{\epsilon}$

$$X^* \sim \mathrm{Laplace}\left(0, \frac{\Delta f}{\epsilon}\right)$$

- The scale $\frac{\Delta f}{\epsilon}$ is the ratio of the ℓ_1-sensitivity and the privacy budget
- Connection to the two implications?
  1. the added noise is positively related to the sensitivity (Δf)
  2. the added noise negatively related to the privacy budget (ϵ)

The Laplace Mechanism

- Formally, given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as

$$\mathcal{M}_L(\mathbf{x}, f(\cdot), \epsilon) = f(\mathbf{x}) + (X_1^*, \cdots, X_k^*), \tag{4}$$

where X_i^* are <u>i.i.d.</u> random variables drawn from $\mathrm{Laplace}\left(0, \frac{\Delta f}{\epsilon}\right)$.

Examples: count query and average query

$$\mathcal{M}_L(\mathbf{x}, f(\cdot), \epsilon) = f(\mathbf{x}) + (X_1^*, \cdots, X_k^*),$$

- Query f: How many rural CUs are there in this sample?
  - $\Delta f = 1$
  - $k = 1$ (i.e. query f is 1−dimension)
  - what is the Laplace distribution the noise should be drawn from?

Examples: count query and average query

$$\mathcal{M}_L(\mathbf{x}, f(\cdot), \epsilon) = f(\mathbf{x}) + (X_1^*, \cdots, X_k^*),$$

- Query f: How many rural CUs are there in this sample?
  - $\Delta f = 1$
  - $k = 1$ (i.e. query f is $1-$dimension)
  - what is the Laplace distribution the noise should be drawn from?
- Another query f: What is the average income of this sample?
  - $\Delta f = \frac{b-a}{n}$
  - $k = 1$ (i.e. query f is $1-$dimension)
  - what is the Laplace distribution the noise should be drawn from?

The Laplace Mechanism preserves $\epsilon-$differential privacy

Proof: Let $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ and $\delta(\mathbf{x}, \mathbf{y}) = 1$, and let $f(\cdot)$ be some function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$. Let $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ denote the probability density functions of $\mathcal{M}_L(\mathbf{x}, f(\cdot), \epsilon)$ and $\mathcal{M}_L(\mathbf{y}, f(\cdot), \epsilon)$. We compare the two at some arbitrary output point $z \in \mathbb{R}^k$:

$$
\begin{aligned}
\frac{p_{\mathbf{x}}(z)}{p_{\mathbf{y}}(z)} &= \prod_{i=1}^{k} \left(\frac{\exp\left(-\frac{|f(\mathbf{x})_i - z_i|}{\Delta f/\epsilon}\right)}{\exp\left(-\frac{|f(\mathbf{y})_i - z_i|}{\Delta f/\epsilon}\right)} \right) && (5) \\
&= \prod_{i=1}^{k} \exp\left(\epsilon \frac{|f(\mathbf{y})_i - z_i| - |f(\mathbf{x})_i - z_i|}{\Delta f}\right) && (6) \\
&\leq \prod_{i=1}^{k} \exp\left(\epsilon \frac{|(f(\mathbf{x})_i - f(\mathbf{y})_i)|}{\Delta f}\right) && (7) \\
&= \exp\left(\epsilon \frac{||f(\mathbf{x}) - f(\mathbf{y})||_1}{\Delta f}\right) && (8) \\
&\leq \exp(\epsilon) && (9)
\end{aligned}
$$

CE example of a count query: step 1

- Calculate the true count of rural CUs

```
require(readr)
CEdata <- read_csv("CEdata.csv")
CEdata$s[CEdata$UrbanRural == 2] <- 1
```

```
## Warning: Unknown or uninitialised column: 's'.
```

```
CEdata$s[CEdata$UrbanRural == 1] <- 0
```

```
n_rural <- CEdata %>%
  summarize_at(vars(s), sum) %>%
  pull()
n_rural
```

```
## [1] 51
```

CE example of a count query: step 2

- Add Laplace noise to the true count

```
Delta_f_count <- 1
```

```
require(rmutil)
set.seed(123)
epsilon1 <- 0.1
rlaplace(1, n_rural, Delta_f_count/epsilon1) %>%
  round()
```

```
## [1] 45
```

CE example of a count query: step 2 cont'd

```
set.seed(123)
epsilon2 <- 1
rlaplace(1, n_rural, Delta_f_count/epsilon2) %>%
  round()
```

```
## [1] 50
```

- With the true count of 51 rural CUs, we can see that smaller privacy budget adds more noise, $51 - 45 = 6$ (when $\epsilon = 0.1$) versus $51 - 50 = 1$ (when $\epsilon = 1$)
- These outcomes are in line with our previously discussed implications, that when fixing the sensitivity value, the added noise is negatively related to the privacy budget

Outline

Composition theorem

- Idea: taking together the independent use of an ϵ_1-differentially private algorithm and an ϵ_2-differentially private algorithm, results in $(\epsilon_1 + \epsilon_2)-$differential privacy

Composition theorem

- Idea: taking together the independent use of an ϵ_1-differentially private algorithm and an ϵ_2-differentially private algorithm, results in $(\epsilon_1 + \epsilon_2)-$differential privacy
- Formally, let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1$ be an ϵ_1-differentially private algorithm, and let $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_2$ be an ϵ_2-differentially private algorithm. Then their combination, defined to be $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1 \times \mathcal{R}_2$ by the mapping: $\mathcal{M}_{1,2}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \mathcal{M}_2(\mathbf{x}))$ is $(\epsilon_1 + \epsilon_2)-$differentially private
- Proof omitted; check out handout

Composition theorem

- Idea: taking together the independent use of an ϵ_1-differentially private algorithm and an ϵ_2-differentially private algorithm, results in $(\epsilon_1 + \epsilon_2)-$differential privacy
- Formally, let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1$ be an ϵ_1-differentially private algorithm, and let $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_2$ be an ϵ_2-differentially private algorithm. Then their combination, defined to be $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1 \times \mathcal{R}_2$ by the mapping: $\mathcal{M}_{1,2}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \mathcal{M}_2(\mathbf{x}))$ is $(\epsilon_1 + \epsilon_2)-$differentially private
- Proof omitted; check out handout
- A generalization: let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_i$ be an ϵ_i-differentially private algorithm for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \rightarrow \prod_{i=1}^{k} \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \cdots, \mathcal{M}_k(\mathbf{x}))$, then $\mathcal{M}_{|k|}$ is $(\sum_{i=1}^{k} \epsilon_i)-$differentially private.

Sequential composition

- If m queries are sent to the same dataset, the privacy budget needs to be divided by m:

$$\epsilon_{new} = \frac{\epsilon}{m}. \tag{10}$$

Sequential composition

- If m queries are sent to the same dataset, the privacy budget needs to be divided by m:

$$\epsilon_{new} = \frac{\epsilon}{m}. \tag{10}$$

- Example: adding Laplace noise to four queries
  1. What is the average income of this sample?
  2. What is the average expneditures of this sample?
  3. What is the variance of income in this sample?
  4. What is the variance of expenditures in this sample?

```
epsilon <- 0.1
m <- 4
epsilon_new <- epsilon/m
rlaplace(1, income_average, Delta_f_average_income/epsilon_new)
rlaplace(1, expenditures_average, Delta_f_average_exp/epsilon_new)
rlaplace(1, income_variance, Delta_f_variance_income/epsilon_new)
rlaplace(1, expenditures_variance, Delta_f_variance_exp/epsilon_new)
```

Sequential composition cont'd

```
epsilon <- 0.1
m <- 4
epsilon_new <- epsilon/m
```

```
rlaplace(1, income_average, Delta_f_average_income/epsilon)
```

results in smaller noise (larger ϵ, same Δf)

```
rlaplace(1, income_average, Delta_f_average_income/epsilon_new)
```

results in larger noise (smaller ϵ, same Δf)

Parallel composition

- If m queries are sent to the same database but on non-overlapping subsets of the dataset, the privacy budget does not need to be divded by m

Parallel composition

- If m queries are sent to the same database but on non-overlapping subsets of the dataset, the privacy budget does not need to be divded by m
- Examples: adding Laplace noise to two queries
    1. What is the average income of rural CUs?
    2. What is the average income of urban CUs?

$$\epsilon_{new} = \epsilon$$

Post-processing

- Dealing with contingency tables (e.g. counts of observations in each category of a categorical variable)
- The post-processing property indicates that for a contingency table with c cells:

$$\epsilon_{new} = \frac{\epsilon}{c-1}. \tag{11}$$

- This is because knowing the noisy counts of $c - 1$ cells determines the count of the c-th cell

Post-processing cont'd

- Examples: adding Laplace noise to six queries (i.e. $c = 6$)
  1. How many CUs with reference person's race category as White are there in this sample?
  2. How many CUs with reference person's race category as Black are there in this sample?
  3. How many CUs with reference person's race category as Native American are there in this sample?
  4. How many CUs with reference person's race category as Asian are there in this sample?
  5. How many CUs with reference person's race category as Pacific Islander are there in this sample?
  6. How many CUs with reference person's race category as Multi-race are there in this sample?

```
epsilon <- 0.1
c <- 6
epsilon_new <- epsilon / (c-1)
Race_6 <- n - Race_1 - Race_2 - Race_3 - Race_4 - Race_5
```

References

- Dwork, C. and McSherry, F. and Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the Third Conference on Theory of Cryptography, 265-284.