Private Data Analysis - Homework 1

Yanning Dai (212008)

September 10, 2024

1. Properties of Laplace Distribution

Proof 1:

The probability density function of the Laplace distribution is:

$$f_Z(z) = \frac{1}{2\lambda} \exp\left(-\frac{|z|}{\lambda}\right), \quad z \in \mathbb{R}$$

Expectation of $\sqrt{\mathbb{E}(Z^2)}$ can be caluclated as:

$$\mathbb{E}(Z^2) = \int_{-\infty}^{\infty} z^2 f_Z(z)\, dz$$

Substitute $f_Z(z)$ into it:

$$\mathbb{E}(Z^2) = \int_{-\infty}^{\infty} z^2 \frac{1}{2\lambda} \exp\left(-\frac{|z|}{\lambda}\right)\, dz$$

Then, due to the symmetry of Laplacian distribution, we can get:

$$\mathbb{E}(Z^2) = 2 \cdot \int_{0}^{\infty} z^2 \cdot \frac{1}{2\lambda} \exp\left(-\frac{z}{\lambda}\right) dz$$

$$= \frac{1}{\lambda} \int_{0}^{\infty} z^2 \exp\left(-\frac{z}{\lambda}\right) dz$$

Set $u = \frac{z}{\lambda}$, gives:

$$\mathbb{E}(Z^2) = \lambda^3 \int_{0}^{\infty} u^2 \exp(-u) du$$

Using the standard gamma function: $\int_0^\infty u^2 \exp(-u) du = 2$, we can get:

$$\mathbb{E}(Z^2) = 2\lambda^2$$

Thus,

$$\sqrt{\mathbb{E}(Z^2)} = \sqrt{2\lambda^2} = \sqrt{2}\lambda$$

Proof 2:

The cumulative distribution function of the Laplace distribution for $Z > 0$ is:

$$P(Z > z) = \int_{z}^{\infty} \frac{1}{2\lambda} \exp\left(-\frac{x}{\lambda}\right) dx$$

Calculate the integral and we get:

$$P(Z > z) = \frac{1}{2\lambda} \left[-\lambda \exp\left(-\frac{x}{\lambda}\right)\right]_{z}^{\infty}$$

$$P(Z > z) = \frac{1}{2} \exp\left(-\frac{z}{\lambda}\right)$$

Then, substitute $z = \lambda t$:

$$P(Z > \lambda t) = \frac{1}{2} \exp(-t)$$

Thus,

$$P(Z > \lambda t) \leq \exp(-t)$$

2. Global Sensitivity

(a) Answer:

$$\Delta f = \frac{2}{n} \cdot \|x_i - x_i'\|_1 \leq \frac{2}{n}$$

Thus, the tight bound for global sensitivity is $\frac{2}{n}$.

(b) Answer:

$$\Delta f = \|x_i x_i^T - x_i' x_i'^T\|_1 = 2\|x_i x_i^T\|_1 \leq 2$$

Thus, the tight bound for global sensitivity is 2.

(c) Answer:

$$\Delta f = \begin{cases} 1, & \text{if median is changed} \\ 0, & \text{if median is not changed} \end{cases}$$

Thus, the tight bound for global sensitivity is 1.

(d) Answer:

When adding or removing an edge, the maximum change in the number of connected components is 1. Thus, the tight bound for global sensitivity is 1.

3. Name and Shame Mechanism

Proof: Let x_i be the value that differs between the datasets D and D'. The mechanism's output Y can take one of three possible values: "nothing," (i, x_i), or (i, x_i'). The corresponding probability differences between D and D' for these outputs are:

$$\Pr[A(D) = Y] - \Pr[A(D') = Y] = \begin{cases} 0, & \text{if } Y = \text{"nothing"} \\ \delta, & \text{if } Y = (i, x_i) \\ -\delta, & \text{if } Y = (i, x_i') \end{cases}$$

Thus, the absolute difference is bounded by δ:

$$|\Pr[A(D) = Y] - \Pr[A(D') = Y]| \leq \delta$$

Therefore, the mechanism A satisfies $(0, \delta)$-differential privacy.

4. Random Response and Laplacian Mechanism

Experiment Report

1. Problem Description

Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$ where $x_i \in \{0, 1\}$, and a function:

$$f(D) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

We use two mechanisms to protect its privacy:

(1) Random Response Mechanism:

Each x_i is perturbed as follows:

$$\hat{x_i} = \begin{cases} x_i & \text{with probability } \frac{e^\epsilon}{e^\epsilon+1} \\ 1 - x_i & \text{with probability } \frac{1}{e^\epsilon+1} \end{cases}$$

The perturbed mean function $\hat{f}(D)$ is then calculated as:

$$\hat{f}(D) = \frac{1}{n} \sum_{i=1}^{n} \hat{x_i}$$

(2) Laplacian Mechanism:

By adding noise to query results, the perturbed mean function $\hat{f}(D)$ becomes:

$$\begin{aligned} \hat{f}(D) &= f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) \\ &= f(D) + \text{Lap}\left(\frac{1}{n\epsilon}\right) \\ &= f(D) + \left(-\frac{1}{n\epsilon} \cdot \text{sgn}(U) \cdot \ln(1 - 2|U|)\right) \end{aligned}$$

here, the sensitivity of the query $\Delta f = \frac{1}{n}$, and U is a uniform random variable sampled from $[-0.5, 0.5]$.

2. Experiment Design

(1) Dataset Construction

The experiment will test different combinations of sample size n and privacy budget ϵ. The values of n and ϵ are chosen as follows:

- n is set to 6 values: [10, 50, 100, 500, 1000, 5000];
- ϵ is set to 8 values: [0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10].

For each combination of n and ϵ, randomly generate 20 binary datasets, where each dataset is represented as $D = \{x_1, \ldots, x_n\}$ and each $x_i \in \{0, 1\}$.

(2) Implementing Two Differential Privacy Mechanisms

For each dataset, calculate the query function $f(D) = \frac{1}{n} \sum_{i=1}^{n} x_i$. Then, apply the two DP mechanisms and perform 100 experiments for each mechanism to reduce the effect of randomness:

- **Randomized Response Mechanism** : Each x_i is randomly perturbed to produce $\hat{x_i}$, and the perturbed mean $\hat{f}(D)$ is computed.

◦ **Laplacian Mechanism**: Noise is directly added to the query result $f(D)$ by sampling from the Laplace distribution, yielding the perturbed result $\hat{f}(D)$.

(3) Utility Analysis

For each combination of n and ϵ, compute the utility of both mechanisms using the following evaluation metrics:

◦ Mean Squared Error (MSE):

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(D)_i - f(D))^2$$

◦ Mean Absolute Error (MAE):

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{f}(D)_i - f(D)|$$

◦ Standard Deviation (SD):

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{f}(D)_i - \overline{\hat{f}(D)})^2}$$

where $\overline{\hat{f}(D)}$ is the average perturbed result over multiple experiments.

3. Experimental Results

(a) Isolated Effect of n and ϵ on the two mechanisms

Figure 1 illustrates the impact of different sample sizes n and privacy budgets ϵ on both mechanisms (RR and Lap). The figures sequentially show the performance of both mechanisms in terms of error and stability (namely MSE, MAE, and STD). The results are plotted on a logarithmic scale for better clarity of the variations.

◦ **Randomized Response Mechanism (RR):**

Effect of n: With smaller sample sizes ($n = 10$), the error is moderate. For example, at $\epsilon = 0.05$, MSE is 0.045, and MAE is 0.169. As n increases, errors decrease significantly. For $n = 1000$ and $\epsilon = 0.05$, MSE drops to 0.0006, and MAE to 0.02, showing that larger sample sizes can effectively reduce noise. STD decreases as n increases, from 0.154 at $n = 10$ to 0.015 at $n = 1000$, indicating improved stability.

Effect of ϵ: With smaller ϵ, the errors are higher. At $\epsilon = 0.05$ and $n = 10$, MSE is 0.045 and MAE is 0.169. As ϵ increases, errors reduce significantly. At $\epsilon = 5.00$, MSE drops to 0.0007, and MAE to 0.007, showing that a higher privacy budget leads to less noise. STD decreases from 0.154 at $\epsilon = 0.05$ to 0.025 at $\epsilon = 5.00$.

◦ **Laplacian Mechanism (Lap):**

Effect of n: Due to sensitivity $\frac{1}{n}$, noise is substantial with small samples. For example, at $n = 10$ and $\epsilon = 0.05$, MSE is 8.03, and MAE is 1.993, showing that noise can overwhelm the true data. As n increases, noise reduces. For $n = 1000$ and $\epsilon = 0.05$, MSE drops to 0.0008, and MAE is 0.02. STD decreases from 2.796 at $n = 10$ to 0.028 at $n = 1000$, showing more stable results with larger samples.

Effect of ϵ: With small ϵ, the errors are high. For $\epsilon = 0.05$ and $n = 10$, MSE is 8.03, and MAE is 1.993. Increasing ϵ reduces noise significantly. At $\epsilon = 5.00$, MSE is 0.0008 and MAE is 0.02, indicating improved accuracy. STD decreases from 2.796 at $\epsilon = 0.05$ to 0.027 at $\epsilon = 5.00$, reflecting reduced data variability.

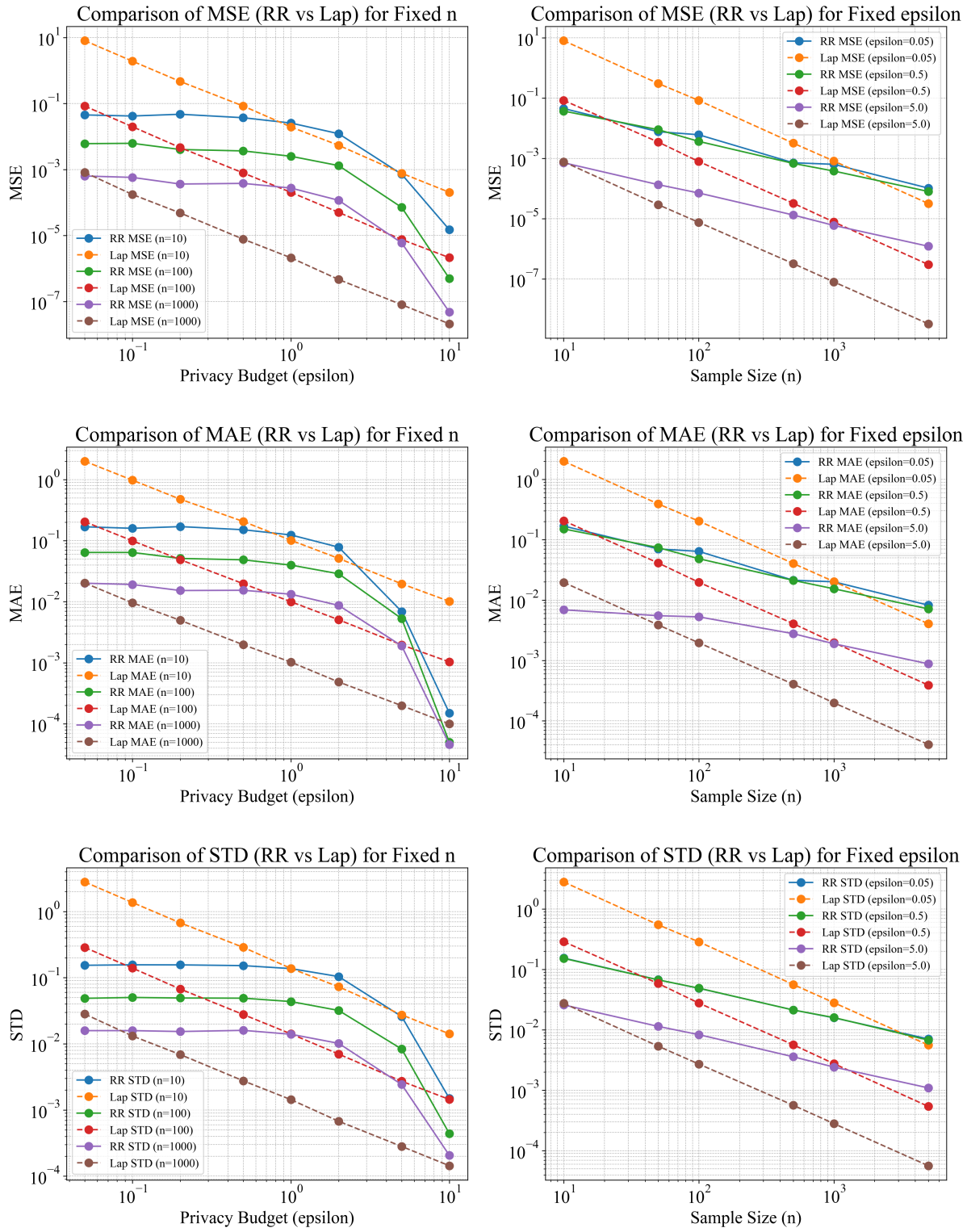

Figure 1: Comparison of MSE, MAE, and STD for Randomized Response and Laplacian Mechanisms across different n and ϵ.

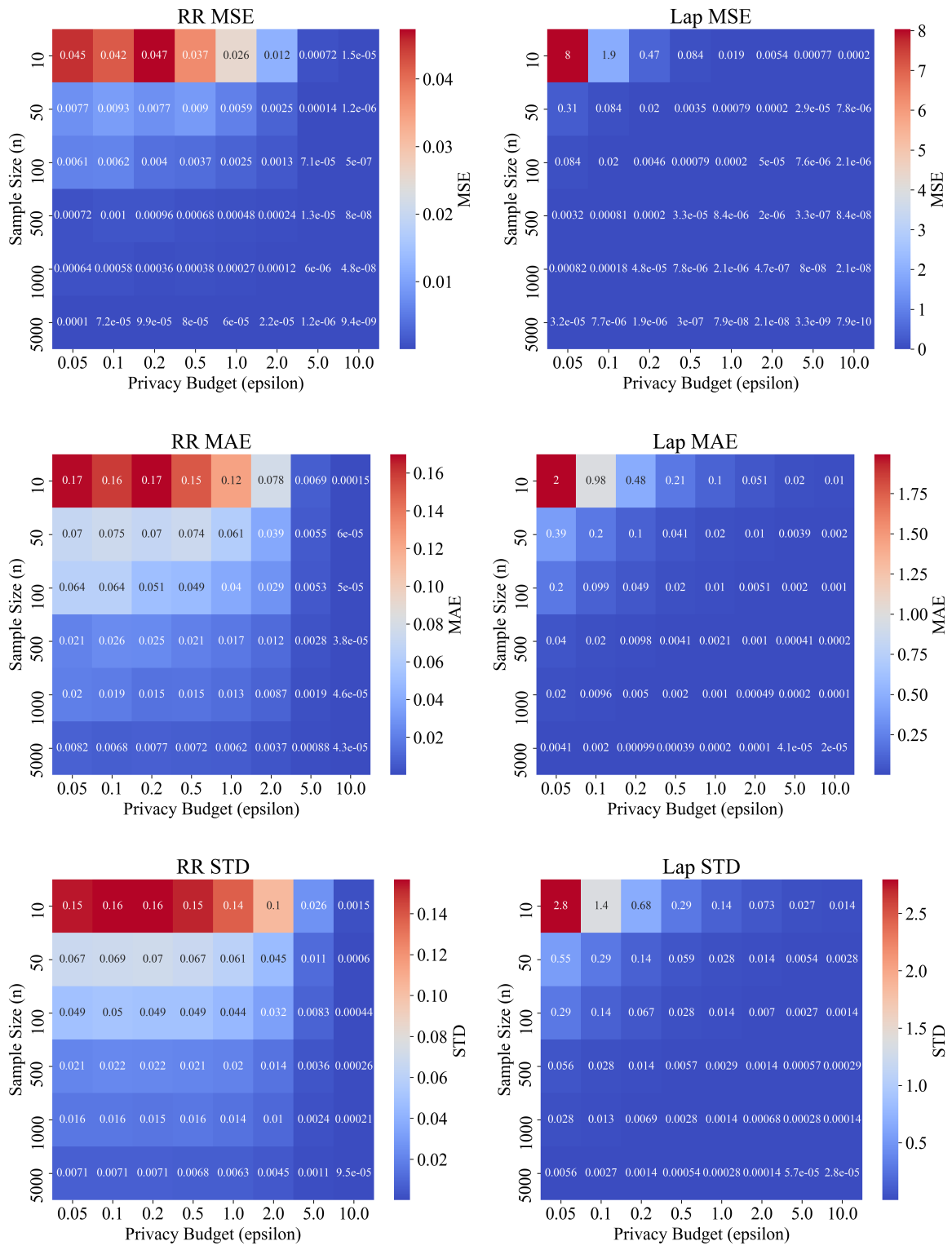

Figure 2: Heatmaps showing the combined effect of n and ϵ on the performance of both mechanisms in terms of MSE, MAE, and STD.

(b) Combined effect of n and ϵ on the two mechanisms

Heatmaps were generated to show how sample size n and privacy budget ϵ jointly affect the performance of both mechanisms. As shown in Figure 2, increasing n and ϵ consistently reduces errors (MSE, MAE, STD). Larger sample sizes and less stringent privacy constraints result in lower noise levels, improving performance across both mechanisms. However, the influence of n and ϵ are different:

- **Randomized Response Mechanism (RR)**: In RR, sample size n has a stronger impact than ϵ, especially when $n > 500$. Small n values lead to high errors regardless of ϵ, but larger n rapidly reduces MSE and MAE.
- **Laplacian Mechanism (Lap)**: In the Lap mechanism, both n and ϵ have an equal impact on error reduction, as noise is proportional to $\frac{1}{n\epsilon}$. Errors decrease steadily as either variable increases. Moreover, Lap mechanism remains more stable across a range of n and ϵ values.

4. Discussion and Conclusion

This experiment evaluated the effects of sample size n and privacy budget ϵ on the performance of two differential privacy mechanisms: Randomized Response and the Laplacian Mechanism. The results show that in the Randomized Response mechanism, sample size n plays a dominant role in reducing noise, particularly for larger n, while the influence of the privacy budget ϵ becomes less significant as n increases. In contrast, the Laplacian Mechanism exhibits a more balanced dependency on both n and ϵ, with noise being proportional to $\frac{1}{n\epsilon}$. This leads to more stable performance across different parameter combinations.

It is important to note that these results are specific to the average function used in this experiment. In practice, the influence of n and ϵ may differ depending on the sensitivity of the query. The Laplacian Mechanism, in particular, is more sensitive to the query's nature, introducing greater noise for highly sensitive queries.

In summary, Randomized Response is well-suited for individual-level privacy protection in settings like surveys, performing best with large datasets. Laplacian Mechanism, on the other hand, is better suited for aggregate statistics and offers more consistent privacy guarantees, but it is more sensitive to the nature and sensitivity of the query.