# Private Data Analysis - Homework 2

Yanning Dai (212008)

October 9, 2024

1. In Lecture 5 and 6 we introduced the privacy loss as a random variable $L_{D,D'} = \ell_{D,D'}(Y)$ where $Y \sim A(D)$ and $\ell_{D,D'} = \ln\left(\frac{p_A(D)(y)}{p_A(D')(y)}\right)$. What is the privacy loss when $A$ is the Laplace mechanism in one dimension? To make things concrete: assume $f(D) = 0$ and $f(D') = 1$ and we add noise $Lap\left(\frac{1}{\epsilon}\right)$.

**Answer:**

The PDF of the Laplace distribution is:

$$p_{Lap(b)}(y) = \frac{1}{2b}\exp\left(-\frac{|y-\mu|}{b}\right)$$

where $b = \frac{1}{\epsilon}$ is the scale parameter, and $\mu$ is the true value of $f(D)$ or $f(D')$. Given that $f(D) = 0$ and $f(D') = 1$, we obtain:

$$p_A(D)(y) = \frac{\epsilon}{2}\exp\left(-\epsilon|y|\right)$$

$$p_A(D')(y) = \frac{\epsilon}{2}\exp\left(-\epsilon|y-1|\right)$$

Thus, the privacy loss is:

$$\begin{aligned} L_{D,D'} &= \ln\left(\frac{\frac{\epsilon}{2}\exp\left(-\epsilon|y|\right)}{\frac{\epsilon}{2}\exp\left(-\epsilon|y-1|\right)}\right) \\ &= \epsilon\left(|y-1| - |y|\right) \end{aligned}$$

The privacy loss depends on the values of $y$:

- For $y \geq 1$: $L_{D,D'} = \epsilon\left((y-1) - y\right) = -\epsilon$
- For $0 \leq y < 1$: $L_{D,D'} = \epsilon\left((1-y) - y\right) = \epsilon(1-2y)$
- For $y < 0$: $L_{D,D'} = \epsilon(1)$

2. (Adding Uniform Noise) Suppose we add uniform noise to a count query $f : \{0,1\}^n \to \mathbb{R}$ with $f(D) = \sum_{i=1}^{n} x_i$, that is, we release $A(D) = f(D) + Z$ where $Z \sim U_{[-\lambda,\lambda]}$, $U_{[-\lambda,\lambda]}$ is the uniform distribution on the interval $[-\lambda, \lambda]$. How large must $\lambda$ be to satisfy $(\epsilon, \delta)$-DP? Do both $\epsilon$ and $\delta$ matter in this setting? When $\delta < \frac{1}{n}$, will this mechanism produce useful information?

**(1) Answer:**

Support of $A(D) : [f(D) - \lambda, f(D) + \lambda]$

The sensitivity of this query is $\Delta f = 1$, as adding or removing a single element changes the count by at most 1. So in the worst-case scenario:

Support of $A(D') : [f(D) + 1 - \lambda, f(D) + 1 + \lambda]$

Thus, the overlap length between these two intervals is $2\lambda - 1$, and the non-overlapping length is 1.

Since the probability density function of a uniform distribution is $\frac{1}{2\lambda}$, the probability of the non-overlapping region is:

$$\delta = \frac{1}{2\lambda}$$

Therefore, to satisfy $(\epsilon, \delta)$-DP, $\lambda$ must be at least $\frac{1}{2\delta}$. Increasing $\lambda$ reduces the probability of the non-overlapping region, thus improving privacy protection. However, it may also introduce more noise, potentially reducing the utility of the output.

**(2) Answer:**

For all measurable sets $S$:

$$\Pr[A(D) \in S] \leq e^{\epsilon} \Pr[A(D') \in S] + \delta$$

In this case, when $y$ is in the overlapping region:

$$\Pr[A(D) = y] = \Pr[A(D') = y] = \frac{1}{2\lambda}$$

Thus,

$$\Pr[A(D) = y] \leq e^{0} \Pr[A(D') = y] + 0$$

When $y$ is in the non-overlapping region:

$$\Pr[A(D) = y] = \frac{1}{2\lambda}, \quad \Pr[A(D') = y] = 0$$

Thus,

$$\Pr[A(D) = y] \leq e^{0} \cdot 0 + \delta$$

Since $\Pr[A(D) = y] = \delta$, the inequality holds.

Therefore, the mechanism cannot limit the privacy loss to a finite $\epsilon$ across all outputs due to the infinite privacy loss in non-overlapping regions. The privacy guarantee relies solely on $\delta$, which bounds the probability of outputs leading to infinite privacy loss.

**(3) Answer:**

Set $\delta < \frac{1}{n}$, and use the relationship between $\lambda$ and $\delta$:

$$\lambda \geq \frac{1}{2\delta} > \frac{1}{2} \times n = \frac{n}{2}$$

Therefore, to ensure $\delta < \frac{1}{n}$, the noise parameter must satisfy $\lambda \geq \frac{n}{2}$. The noise $Z$ is uniformly distributed over:

$$[-\lambda, \lambda] = \left[-\frac{n}{2}, \frac{n}{2}\right]$$

The maximum noise magnitude $\frac{n}{2}$ is half of the maximum possible true count $n$, which leads to a very low signal-to-noise ratio. This makes it challenging to extract the true signal from the noisy data, severely degrading utility. The mechanism does not produce useful information.

3. (Implementation of Noisy-max Mechanism and Exponential Mechanism) You can find a selection problem (you have to say the output space, the score function and its sensitivity) and try to implement the noisy-max mechanism and the exponential mechanism. Write a report on your findings.

**Report:**

**1. Problem Description**

We consider a selection problem where the goal is to find the top 3 values from a set of 100 randomly generated numbers. Two differential privacy mechanisms, Noisy-Max and Exponential Mechanism, will be explored.

**(1) Output Space**

The output space consists of the set of 100 random numbers:

$$X = \{x_1, x_2, \ldots, x_{100}\}$$

We aim to select the top 3 values from this set.

**(2) Score Function**

The score function ranks each number based on its value. For each $x_i$, the score function is:

$$f(x_i) = x_i$$

**(3) Sensitivity**

The sensitivity is the maximum change in the score function when a single element is altered. In this case, the sensitivity is:

$$\Delta f = 1$$

since modifying one number can change its score by at most 1.

**2. Mechanism Implementation**

**(1) Noisy-Max Mechanism**

For each $x_i$, compute the score $f(x_i) = x_i$. In noisy-max mechanism, we add Laplace noise to each score. The noisy score for each candidate is computed as:

$$f(x_i) + \text{Lap}\left(\frac{1}{\epsilon}\right) = x_i + \text{Lap}\left(\frac{1}{\epsilon}\right)$$

Once all candidates have their scores perturbed by noise, we select the top 3 candidates with the highest noisy score:

$$\hat{X}_{top3} = \text{Top-3}_{x_i}\left(x_i + \text{Lap}\left(\frac{1}{\epsilon}\right)\right)$$

**(2) Exponential Mechanism**

For each candidate $x_i$, compute its score $f(x_i) = x_i$. In exponential mechanism, the probability of selecting a candidate $x_i$ is proportional to the exponential score:

$$P(x_i) \propto \exp\left(\frac{\epsilon \cdot f(x_i)}{2\Delta f}\right) = \exp\left(\frac{\epsilon \cdot x_i}{2}\right)$$

where $\Delta f$ is the sensitivity of the score function, which is 1 in this case. Then, select the top 3 candidates according to these probabilities. The candidates selected with the highest probabilities are:

$$\hat{X}_{top3} = \text{Top-3}_{x_i}\left(\exp\left(\frac{\epsilon \cdot x_i}{2}\right)\right)$$

## 3. Experiment Design

### (1) Dataset Construction

The experiment will test different privacy budgets $\epsilon$: [ 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 ]. For each $\epsilon$ value, randomly generate 20 datasets, each containing 100 continuous values in the range [0, 100]. For each dataset, apply the Noisy-Max and Exponential mechanisms, and perform 100 experiments for each mechanism to mitigate the effects of randomness.

### (2) Utility Analysis

For each experiment, record the overlap between the Top-3 elements selected by the two mechanisms and the actual Top-3 elements. Specifically, calculate how many of the selected Top-3 elements match the true Top-3. Average the overlap scores across multiple experiments to evaluate the average accuracy of both mechanisms.

Calculate the variance of the results to measure the fluctuation in performance for each mechanism. The variance will provide insights into the stability of the mechanisms under different privacy budgets.

## 4. Results and Discussion

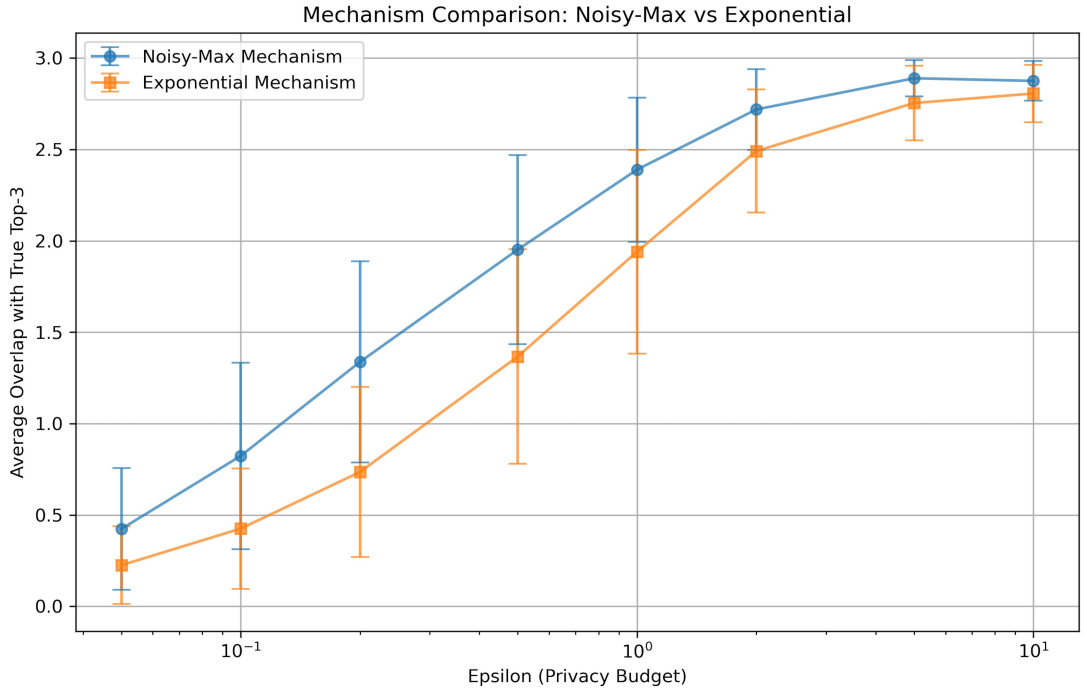


Figure 1: Experimental results of the two machanisms

The experimental results of the two mechanism are shown in Figure 1. This section summarizes the key findings.

### (1) Pravicy Protection Performance

Across all $\epsilon$ values, the Noisy-Max mechanism consistently showed higher accuracy in selecting the top-3 values compared to the Exponential mechanism. This advantage is particularly pronounced at lower $\epsilon$ values, where privacy protection is stronger. Noisy-Max retains better accuracy even under strict privacy constraints.

As $\epsilon$ increases (weaker privacy protection), both mechanisms improve in accuracy, but Noisy-Max remains slightly superior in most cases.

### (2) Variance and Stability

Noisy-Max shows higher variance, especially at smaller $\epsilon$ values, indicating less stability under strong privacy constraints.

The Exponential mechanism has lower variance, indicating more stable performance, though at the cost of lower accuracy across all $\epsilon$ values.

**(3) General Trends**

As $\epsilon$ increases, both mechanisms show improved accuracy, nearing optimal performance (accuracy approaching 3.0) with the highest $\epsilon$ values. This aligns with the expected behavior: reduced noise allows for more accurate selections.

**5. Conclusion**

- Noisy-Max delivers better accuracy overall but at the cost of greater variability, especially with stronger privacy protection (low $\epsilon$).
- The Exponential mechanism offers more stable results, but its accuracy is generally lower, making it a better option when stability is prioritized over peak accuracy.

In conclusion, Noisy-Max is preferable for applications requiring higher accuracy, while the Exponential mechanism is better suited for scenarios where stability and consistency are more important.

4. (Comparison with Gaussian and Laplace Mechanism) Generate a dataset $D = \{x_1, \ldots, x_n\}$ where each $x_i \in \{0,1\}^d$. Consider answering the average query $f(D) = \frac{1}{n}\sum_{i=1}^{n} x_i$ via Laplace and Gaussian mechanism. Implement these two mechanisms with varying $n, d, \epsilon$ (You cases must at least include $d = 1$ and $d \gg 1$). Write a report on your findings.

**Report:**

**1. Problem Description**

We are given a dataset $D = \{x_1, x_2, \ldots, x_n\}$ where $x_i \in \{0,1\}^d$, and we aim to answer the average query function $f(D) = \frac{1}{n}\sum_{i=1}^{n} x_i$ using two differential privacy mechanisms: the Laplace mechanism and the Gaussian mechanism.

**(1) Laplace Mechanism**

In the Laplace mechanism, noise is added to the query function result to ensure privacy. The perturbed function is given by:

$$\hat{f}(D) = f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$$

where $\Delta f$ is the sensitivity of the query and $\epsilon$ is the privacy budget. For the average query, the sensitivity $\Delta f = \frac{1}{n}$.

**(2) Gaussian Mechanism**

The Gaussian mechanism adds Gaussian noise to the query result:

$$\hat{f}(D) = f(D) + \mathcal{N}\left(0, \sigma^2\right)$$

where $\sigma = \frac{\Delta f \cdot \sqrt{2\ln(1.25/\delta)}}{\epsilon}$, and $\delta$ is an additional privacy parameter.

**2. Experiment Design**

**(1) Dataset Construction**

The experiment will test various combinations of the dataset size $n$, dimensionality $d$, and privacy budget $\epsilon$. The parameters are chosen as follows:

- $n$ is set to values: [ 10, 50, 100, 500, 1000 ];
- $d$ is set to values: [ 1, 5, 10, 50, 100];
- $\epsilon$ is set to values: [ 0.05, 0.1, 0.5, 1, 2, 5 ].

For each combination of $n$, $d$, and $\epsilon$, generate 10 datasets where each $x_i \in \{0,1\}^d$.

**(2) Implementing the Mechanisms**

For each dataset, calculate the query function $f(D) = \frac{1}{n}\sum_{i=1}^{n} x_i$, and then apply both the Laplace and Gaussian mechanisms to perturb the result. Perform 100 experiments for each mechanism to reduce the effect of randomness.

**(3) Utility Analysis**

For each combination of $n$, $d$, and $\epsilon$, evaluate the performance of both mechanisms using the following metrics:

- **Mean Squared Error (MSE)**:

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

◦ **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.

## 4. Experimental Results

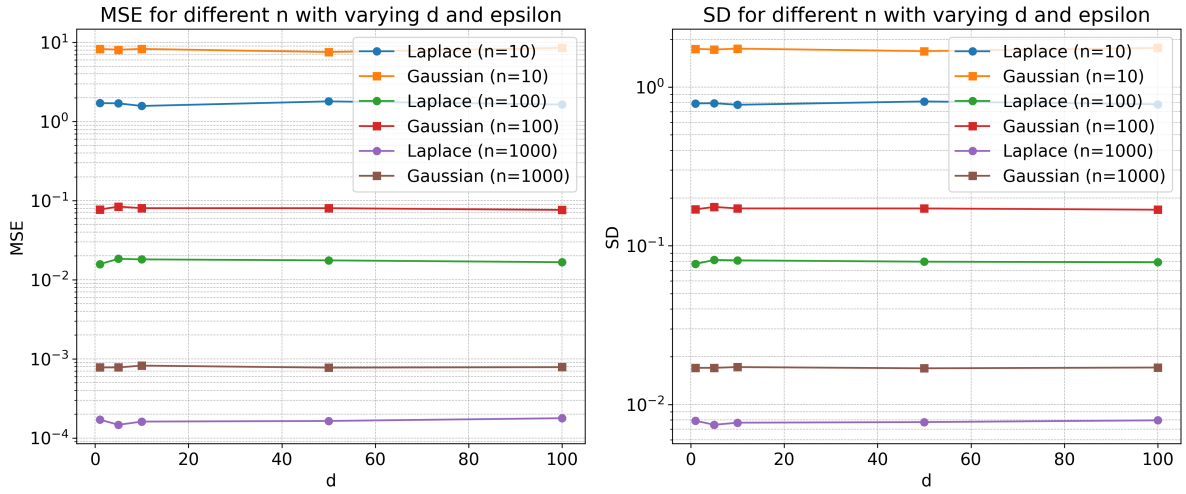


Figure 2: Plot of $d$ vs $\epsilon$ with fixed $n$

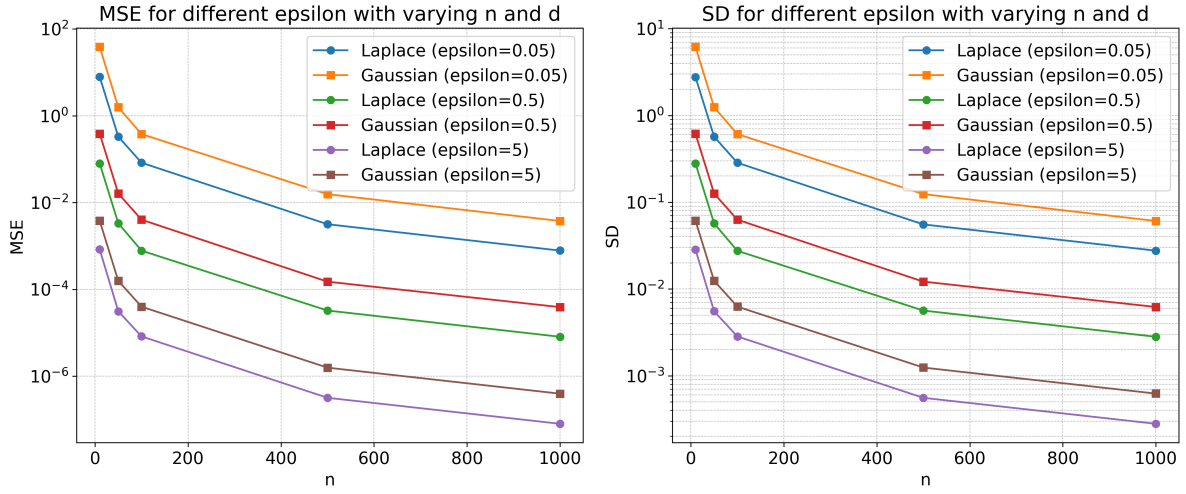


Figure 3: Plot of $n$ vs $d$ with fixed $\epsilon$

◦ MSE and SD curves for different $d$ with varying $n$ and $\epsilon$ (refer to Figure 2):

- Both Laplace and Gaussian mechanisms exhibit a consistent decline in MSE and SD as the dimensionality $d$ increases.
- The Laplace mechanism tends to perform slightly better in terms of accuracy (lower MSE) for lower $n$, particularly when $\epsilon$ is small.
- As $d$ increases, the difference between Laplace and Gaussian mechanisms diminishes, with both showing stable performance.

◦ MSE and SD curves for different $\epsilon$ with varying $n$ and $d$ (refer to Figure 3):

- The influence of the privacy budget $\epsilon$ is prominent, with higher $\epsilon$ values (weaker privacy) leading to better performance for both mechanisms.

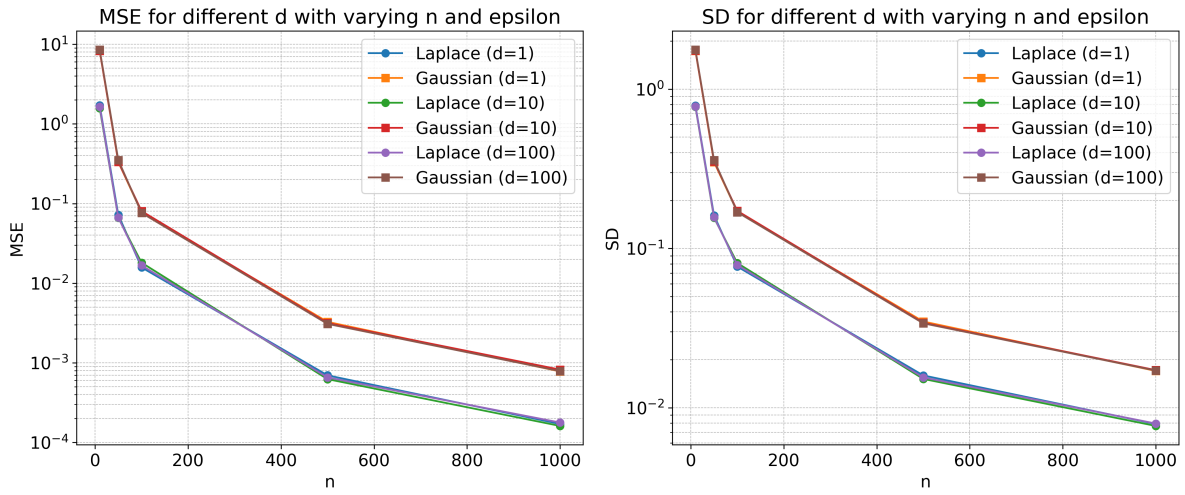


Figure 4: Plot of $n$ vs $\epsilon$ with fixed $d$

- The Laplace mechanism consistently outperforms the Gaussian mechanism, especially at smaller values of $\epsilon$, but the performance gap narrows as $\epsilon$ increases.
- MSE values show a significant improvement as $n$ increases, with the Gaussian mechanism catching up with Laplace for larger datasets ($n \geq 1000$).

◦ MSE and SD curves for different $n$ with varying $d$ and $\epsilon$ (refer to Figure 4):

- The results highlight that as the dataset size $n$ increases, both the MSE and SD values decrease substantially, indicating the importance of larger datasets for accurate results in both mechanisms.
- The Gaussian mechanism struggles with higher MSE values at smaller dataset sizes, while the Laplace mechanism handles small datasets more effectively, especially at lower $\epsilon$.

## 5. Discussion and Conclusion

### (a) Accuracy and Stability

◦ Accuracy (MSE): Across most configurations, the Laplace mechanism generally shows lower MSE than the Gaussian mechanism, particularly when privacy is strong ($\epsilon$ is low). Both mechanisms benefit significantly from increased dataset sizes. As $n$ grows, MSE decreases across the board, with the performance gap between the two mechanisms narrowing.

◦ Stability (SD): The Laplace mechanism, while more accurate, shows slightly higher variability (SD) than the Gaussian mechanism, particularly at lower dataset sizes and privacy budgets. As the dataset size increases, the variance in both mechanisms becomes stable, but the Gaussian mechanism tends to maintain slightly more consistent results.

### (b) Impact of Dimensionality and Dataset size

◦ Dimensionality ($d$) shows a smaller impact on MSE and SD compared to the dataset size. Both mechanisms seem robust to changes in $d$, though performance can vary slightly for extreme values of $d$. The impact of increasing $d$ seems to be more evident at lower privacy budgets ($\epsilon \leq 0.5$).

◦ Increasing the dataset size $n$ consistently improves both mechanisms, with MSE dropping as $n$ increases. This demonstrates the crucial role of data size in achieving high accuracy while maintaining privacy.

### (c) Privacy Budget

- Increasing $\epsilon$ (less privacy) improves both accuracy and stability, with MSE and SD dropping significantly. For high privacy settings ($\epsilon = 0.05$), the Laplace mechanism proves to be more reliable, but as $\epsilon$ increases, both mechanisms show comparable performance.

From these experiments, we can conclude that:

- Laplace Mechanism Superiority: The Laplace mechanism is generally more accurate (lower MSE), particularly for small privacy budgets ($\epsilon \leq 0.5$). This is true across various dataset sizes and dimensionalities.
- Gaussian Mechanism for Stability: The Gaussian mechanism, while less accurate in small data scenarios, offers slightly more stable (lower SD) performance, particularly in high-dimensional or large dataset scenarios.