Differential Privacy: 6 Key Equations Explained

Abhishek Tiwari

Citation: *A. Tiwari*, "Differential Privacy: 6 Key Equations Explained", Abhishek Tiwari, 2024. doi:10.59350/ntarj-tg210

Published on: December 01, 2024

Differential Privacy is a powerful framework for ensuring privacy in data analysis by adding controlled noise to computations. Its mathematical foundation guarantees that the presence or absence of any individual's data in a dataset does not significantly affect the outcome of an analysis. Here are six key equations that capture the essence of differential privacy and its mechanisms, along with references to their origins and explanations.

Definition of Differential Privacy

[1] formalises privacy through the concept of indistinguishability between neighbouring datasets. A mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S]$$

Where,

- Pr denotes the probability of an event.
- D, D': Neighboring datasets differing in one entry.
- $\mathcal{M}$: The randomized mechanism applied to the dataset.
- S: Subset of possible outputs.
- ϵ: Privacy budget controlling the privacy-accuracy tradeoff.

This definition ensures that the outputs of $\mathcal{M}$ are statistically indistinguishable for any two neighboring datasets.

Laplace Mechanism

The [1] is a common approach for achieving ϵ-differential privacy. It adds noise drawn from a Laplace distribution to the output of a function. The amount of noise is determined by the function's sensitivity and the privacy budget ϵ, balancing privacy and accuracy. The noise is defined as:

$$\eta \sim \text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$$

Here:

- $\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$: Sensitivity of the function f, measuring the maximum difference in output for neighboring datasets.
- ϵ is privacy budget.

The perturbed result is:

$$\mathcal{M}(D) = f(D) + \eta$$

Gaussian Mechanism

For scenarios where a small probability of failure is acceptable, (ϵ, δ)-differential privacy can be achieved using the [2]. The noise is sampled from a normal distribution:

$$\eta \sim \mathcal{N}(0, \sigma^2)$$

The standard deviation of the noise is calculated as:

$$\sigma = \frac{\Delta f \cdot \sqrt{2 \ln(1.25/\delta)}}{\epsilon}$$

Where:

- Δf: Sensitivity of the function f.
- δ: Probability of a privacy failure.
- ϵ is privacy budget.

The perturbed output is:

$$\mathcal{M}(D) = f(D) + \eta$$

When applied iteratively or in compositions, the Gaussian Mechanism aligns well with advanced composition theorems (described below), making it practical for repeated queries. It is particularly useful when the function's output has high sensitivity or when (ϵ, δ)-differential privacy is needed instead of strict ϵ-differential privacy.

Composition Theorems

When multiple differentially private mechanisms are applied, the privacy budget accumulates. Differential privacy provides a framework for measuring and bounding the [3] from multiple analyses of information about the same individuals. Two key composition approaches are:

Sequential Composition

If k mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$ are applied, each with privacy guarantees $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$, the total privacy guarantee is:

$$\epsilon_{\text{total}} = \sum_{i=1}^{k} \epsilon_i$$

Suppose a dataset is queried three times using mechanisms with $\epsilon_1 = 0.5$, $\epsilon_2 = 0.3$, and $\epsilon_3 = 0.2$. The total privacy budget consumed is:

$$\epsilon_{\text{total}} = 0.5 + 0.3 + 0.2 = 1.0$$

This means the combined analysis satisfies 1.0-differential privacy.

Advanced Composition

A tighter bound is given by advanced composition, which accounts for small failures:

$$\epsilon_{\text{total}} = \sqrt{2k \ln(1/\delta)} \cdot \epsilon + k \cdot \epsilon^2$$

Suppose a dataset is queried 100 times, each query satisfying $(\epsilon = 0.1, \delta = 10^{-5})$-differential privacy. Using advanced composition:

$$\epsilon_{\text{total}} = \sqrt{2 \cdot 100 \cdot \ln(1/10^{-5})} \cdot 0.1 + 100 \cdot (0.1)^2$$

1. Calculate the first term:

$$\sqrt{2 \cdot 100 \cdot \ln(10^5)} \cdot 0.1 = \sqrt{2 \cdot 100 \cdot 11.5129} \cdot 0.1 = \sqrt{2302.58} \cdot 0.1 = 4.8$$

2. Calculate the second term:

$$100 \cdot 0.01 = 1$$

3. Combine the terms:

$$\epsilon_{\text{total}} = 4.8 + 1 = 5.8$$

Thus, after 100 queries, the total privacy guarantee is approximately $(\epsilon = 5.8, \delta = 10^{-5})$.

Comparing Sequential and Advanced Composition

Aspect	Sequential Composition	Advanced Composition
Privacy Parameter (ϵ)	Sum of all individual ϵ values.	Tighter bound with sublinear growth.
Failure Probability (δ)	Assumes $\delta = 0$.	Allows for small $\delta > 0$.
Scaling with k	Linear scaling: $\epsilon_{\text{total}} = k \cdot \epsilon$.	Sublinear scaling with $\sqrt{k}$.
Accuracy	Less noise is required for individual mechanisms but results in higher cumulative noise.	Supports smaller cumulative noise.
Use Cases	Suitable for strict privacy guarantees.	Suitable for practical settings with relaxed privacy.

Sensitivity

The [1] of a function quantifies its robustness to changes in individual data points. Sensitivity determines the amount of noise required to ensure privacy. Currently, the global and local sensitivity are being mainly used in differential privacy.

Global Sensitivity

[1] is the maximum change in the output of a function Δf_G when applied to **any two neighboring datasets** D and D', differing by a single element. For function Δf_G, the ℓ_1 or global sensitivity is defined as:

$$\Delta f_G = \max_{D,D'} ||f(D) - f(D')||_1$$

Where:

- D, D': Neighboring datasets differing in one record.
- $|| \cdot ||_1$: The ℓ_1-norm, representing the absolute difference in outputs.

> Global sensitivity is the maximum differences in output with consideration of all possible datasets and is therefore only dependent on the query and not the dataset.

Local Sensitivity

[4] is a finer measure defined for a specific dataset D. It measures the maximum change in the output of function Δf_L for all neighboring datasets D' of D:

$$\Delta f_L(D) = \max_{D'} ||f(D) - f(D')||_1$$

> Local sensitivity attempts to calculate the sensitivity for a local data set, where the possible changes are bound by the local data set and not the universe of all data sets.

While local sensitivity may be smaller than global sensitivity for specific datasets, it is less robust for privacy guarantees since it depends on the specific dataset and not the worst-case scenario.

Sensitivity for Common Functions

1. **Counting Queries**: For functions that count individuals satisfying a condition, $\Delta f = 1$.
2. **Summation Queries**: If data values are bounded (e.g., income within $[0, 100]$), the sensitivity is the range of possible values. For summing incomes:

$$\Delta f = \text{max value} - \text{min value} = 100 - 0 = 100$$

3. **Average Queries**: Sensitivity for averages depends on both the range of values and the dataset size n:

$$\Delta f = \frac{\text{max value} - \text{min value}}{n}$$

4. **Maximum or Minimum Queries**: Sensitivity is the largest possible change in the maximum or minimum when a single data point is added or removed.

Comparing Global Sensitivity vs. Local Sensitivity

Aspect	Global Sensitivity	Local Sensitivity
Definition	Measures the maximum change in the output of a function f over **all possible** neighboring datasets.	Measures the maximum change in the output of f for a **specific dataset** and its neighbors.
Scope	Evaluates worst-case sensitivity across all possible datasets.	Evaluates sensitivity for a particular dataset D.
Robustness	Independent of the specific dataset; provides **universal guarantees**.	Dependent on the specific dataset; may not generalize to other datasets.
Noise Requirement	Requires **more noise** to account for the worst-case scenario.	May allow **less noise** for datasets with lower local sensitivity.
Use Case	Commonly used for general differential privacy guarantees.	Suitable for improving accuracy when local sensitivity is significantly smaller.
Accuracy	May result in less accurate outputs due to higher noise.	Can yield more accurate results for specific datasets with low local sensitivity.
Computation Complexity	Straightforward to compute for many functions but can be conservative.	Requires evaluating sensitivity for every neighboring dataset, which can be complex.
Examples	For a count query: $\Delta f = 1$, as adding/removing one individual changes the count by 1.	For a count query: Sensitivity may be less than 1 if the specific dataset structure limits changes.

Exponential Mechanism

The [5] is useful for selecting outputs in a way that prioritizes utility while maintaining privacy. It assigns probabilities to each potential output r based on a utility function $u(D, r)$:

$$\Pr[\mathcal{M}(D) = r] \propto \exp\left(\frac{\epsilon \cdot u(D, r)}{2\Delta u}\right)$$

Where:

- $u(D, r)$: Utility function representing the quality of r.
- Δu: Sensitivity of the utility function.

- ϵ is the privacy budget.

Unlike mechanisms that add noise to the output (e.g., Laplace Mechanism), the Exponential Mechanism modifies the selection probability, making it suitable when adding noise would render the output meaningless or when the output space is non-numeric. It extends the concept of differential privacy to non-numeric data types and provides a foundation for private data analysis in more complex settings.

Laplace Mechanism vs. Gaussian Mechanism vs. Exponential Mechanism

Aspect	**Laplace Mechanism**	**Gaussian Mechanism**	**Exponential Mechanism**
Purpose	Adds noise to the output of a numeric query to achieve ϵ-differential privacy.	Adds noise to the output of a numeric query to achieve (ϵ, δ)-differential privacy.	Selects an output from a discrete set based on a utility function, ensuring ϵ-differential privacy.
Type of Output	Numeric values (e.g., sums, counts, averages).	Numeric values (e.g., sums, counts, averages).	Categorical or discrete outputs (e.g., choosing the best item or category).
Noise Distribution	Noise is drawn from a Laplace distribution: Lap$(\frac{\Delta f}{\epsilon})$.	Noise is drawn from a Gaussian (normal) distribution: $\mathcal{N}(0, \sigma^2)$.	Selection probabilities are proportional to $\exp\left(\frac{\epsilon \cdot u(D,r)}{2\Delta u}\right)$.
Parameters	ϵ (privacy budget).	ϵ (privacy budget), δ (failure probability).	ϵ (privacy budget), Δu (sensitivity of the utility function).
Sensitivity Requirement	Requires global sensitivity Δf of the function.	Requires global sensitivity Δf of the function.	Requires sensitivity Δu of the utility function.
Privacy Guarantee	Strict ϵ-differential privacy.	Approximate (ϵ, δ)-differential privacy.	Strict ϵ-differential privacy.
Noise Scaling	Noise scales linearly with $\frac{\Delta f}{\epsilon}$.	Noise scales with $\frac{\Delta f \cdot \sqrt{2\ln(1.25/\delta)}}{\epsilon}$.	No direct noise; selection probabilities are adjusted based on utility and ϵ.

Aspect	Laplace Mechanism	Gaussian Mechanism	Exponential Mechanism
Failure Probability (δ)	No failure probability; guarantees hold strictly for all cases.	Allows a small failure probability ($\delta > 0$) for more flexibility in noise calibration.	No failure probability; guarantees hold strictly for all cases.
Key Strength	Simple to implement and provides strict privacy guarantees.	Handles high-dimensional queries and provides a trade-off with a small failure probability (δ).	Ensures privacy while prioritizing outputs with higher utility, suitable for categorical data.
Key Limitation	May add excessive noise for high-dimensional queries.	Requires additional parameter (δ) and more noise for strict privacy.	Limited to discrete or categorical output spaces; requires well-defined utility functions.
Applications	Numeric data analysis, such as counting, summation, and mean estimation.	Machine learning, high-dimensional statistics, and approximate differential privacy settings.	Parameter selection, private feature selection, or any situation with categorical output.
Accuracy vs. Privacy Tradeoff	Balances accuracy and privacy by adjusting ϵ; may degrade accuracy with large sensitivity.	Allows finer trade-offs by adjusting both ϵ and δ; suitable for approximate guarantees.	Balances utility and privacy; prioritizes high-utility outputs while maintaining privacy.

Conclusion

Differential privacy provides a rigorous [6] for protecting individual data in computations. These six equations form the core of differential privacy and its mechanisms. By carefully choosing the noise and parameters, analysts can ensure a balance between privacy and accuracy. Understanding these equations empowers practitioners to implement robust privacy-preserving techniques in real-world applications.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, 2006. doi: 10.1007/11681878_14.

[2] I. Mironov, "On Significance of the Least Significant Bits for Differential Privacy," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 2012. doi: 10.1145/2382196.2382264.

[3] P. Kairouz, S. Oh, and P. Viswanath, "The Composition Theorem for Differential Privacy," *arXiv*, 2015. doi: 10.48550/arXiv.1311.0776.

[4] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, 2007. doi: 10.1145/1250790.1250803.

[5] F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 2007. doi: 10.1109/FOCS.2007.66.

[6] A. Tiwari, "Mathematical Guarantee," 2024, *Abhishek Tiwari*. doi: 10.59350/ghs12-1vq60.