Invetigation of the Elliptical Gaussian Noise in the case of multivariate normal data

Tim Sehested Poulsen

November 24, 2022

1 Preliminaries

1.1 Definitions

The dataset is denoted by X, where $X \in \mathbb{R}^{n \times d}$. I have that n denotes the number of entries in the dataset and d is the number of dimensions of the dataset. I will throughout the report refer to a single entry of the dataset as x_i and a single dimension of the dataset as $X^{(j)}$, and therefore $x_i^{(j)}$ denotes the j'th dimension of the i'th entry.

Differential privacy is the heuristic of releasing a database statistic whilst limiting the impact of any one entry. Differential privacy has multiple slightly different formal definitions, one such is (ϵ, δ)-Differential Privacy refered to as (ϵ, δ)-DP. A prerequisite for almost all of the different differential privacy definitions relies on the concept of neighbouring dataset.

Definition 1.1 (Neighbouring dataset [1])**.** *Two dataset $X, X' \in \mathbb{R}^{n \times d}$ are said to be neigbouring if they differ in at most a single entry. Neighbouring dataset are denoted with the relation $X \sim X'$ and defined as followed*

$$X \sim X' \iff |\{i \in \mathbb{N} \mid i \leq n \wedge x_i \neq x_i'\}| \leq 1$$

Definition 1.2 (Sensitivity [2])**.** *Let $f(X) : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ given by $f(X) = \sum_{i=1}^{n} x_i$ be the sum over all vectors in a dataset. The sensitivity is then the maximal possible difference in the output of our summation from two neighbouring dataset denoted as Δ. We denote the sensitivity of the j'th dimension as*

$$\Delta_j = \max_{X \sim X'} \left| f(X)^{(j)} - f(X')^{(j)} \right|$$

and then the total l_2-sensitivity is then

$$\|\Delta\| = \max_{X \sim X'} \|f(X) - f(X')\|$$

Definition 1.3 ((ϵ, δ)-Differential Privacy [1])**.** *A randomized algorithm $\mathcal{M} : \mathbb{R}^{n \times d} \to \mathcal{R}$ is (ϵ, δ)-differentially private if for all possible subsets of outputs $S \subseteq \mathcal{R}$ and all pairs of neighbouring dataset $X \sim X'$ we have that*

$$\Pr\left[M(X) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[M(X') \in S\right] + \delta$$

1.2 Problem setup

The problem consists of realeasing the sum of vectors in a dataset under differential privacy. More formally we whish to release the value of $f : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ given by

$$f(X) = \sum_{i=1}^{n} x_i$$

under (ϵ, δ)-DP.
The problem that the Elliptical Gaussian Mechanism solves is in the setting where all dimensions $X^{(j)}$ are restricted by some bound Δ_j [2]. This means that all $x_i^{(j)} \in [-\Delta_j/2, \Delta_j/2]$.

References

[1] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality 7*, 3 (2016), 17–51.

[2] Pagh, R., and Lebeda, C. Private vector aggregation when coordinates have different sensitivity.