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. It builds on the intuition that computing a statistics on a private dataset should not reveal any sensitive information about any one individual as long as that individual has little to no effect on the outcome. Differential privacy has multiple slightly different formal definitions, one such is (ε, δ)-Differential Privacy refered to as (ε, δ)-DP which will be introduced later on. 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 [3]). *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*

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

2 Algorithms

2.1 The Gaussian Mechanism

One of the most foundational algorithms for achieving (ε, δ)-DP is the Gaussian Mechanism [2]. It computes the real value of a statistic, where the l_2-sensitivity is known. That is it produces a (ε, δ)-DP estimate of a function $g : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ where $\|\Delta\|$ for the function g is known. It does so by computing the value of $g(X)$ and then adding noise to each dimension from the normal distribution $\mathcal{N}(0, \sigma^2_{\varepsilon,\delta})$. This can be seen as adding a noise vector η which is then distributed according to the multivariate normal distribution $\mathcal{N}(\overrightarrow{0}, \sigma^2_{\varepsilon,\delta} I)$. The algorithm can be seen in Algorithm 1.

Algorithm 1 The Gaussian Mechanism

```
  Input
      $\sigma_{\varepsilon,\delta}$            Standard deviation required to achieve $(\varepsilon, \delta)$-DP
      $X \in \mathbb{R}^{n \times d}$      Dataset
  Output
      $(\varepsilon, \delta)$-DP estimate of $g(X)$
  $\eta \leftarrow$ sample from $\mathcal{N}(\overrightarrow{0}, \sigma^2_{\varepsilon,\delta} I)$
  return $g(X) + \eta$
```

It is quite apparent that the main difficulty of the mechanism lies in determining a $\sigma_{\varepsilon,\delta}$ which achieves (ε, δ)-DP , and preferably the smallest such one.

The following theorem was initially proven

Theorem 1. [2] *Let $g : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ be an arbitrary d-dimensional function with l_2-sensitvity $\|\Delta\| = \max_{X \sim X'} \|g(X) - g(X')\|$, and let $\varepsilon \in (0, 1)$. The Gaussian Mechanism with $\sigma_{\varepsilon,\delta} = \|\Delta\| \sqrt{2\ln(1.25/\delta)}/\varepsilon$ is (ε, δ)-DP .*

The proof is rather long and is therefore ommitted here.

-Does introduce the same variance for all dimensions regardless of individual variance

3 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 [3]. 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] Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4 (2014), 211–407.

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